Universidad Nacional del Sur

Tesis de Doctorado en Ciencias de la Computación

Selección de Variables y Descubrimiento Causal a Partir de Textos de Artículos Periodísticos

Mariano Maisonnave

Bahía Blanca Argentina

2021

Prefacio

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico de Doctorado en Ciencias de la Computación, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito del Departamento de Ciencias e Ingeniería de la Computación durante el período comprendido entre el 1 de Abril de 2017 y el 27 de octubre de 2021, bajo la dirección de la Dra. Ana Gabriela Maguitman, y del Dr. Fernando Abel Tohmé.

.....

Mariano Maisonnave
mariano.maisonnave@cs.uns.edu.ar
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Bahía Blanca, 27 de octubre de 2021

UNIVERSIDAD NACIONAL DEL SUR
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el .../.../..., mereciendo la calificación de .....(.....)

Agradecimientos

No puedo evitar sentirme afortunado al ver que este objetivo de obtener un doctorado, que me propuse hace cinco años, se va convirtiendo en una realidad. Más aún al notar como, durante todo este tiempo, tanta gente increíble me acompañó en este viaje y me ayudó a cumplir este objetivo. Porque todo esto no hubiera sido posible sin tener el apoyo de tantos amigos, familiares y colegas, quiero dedicar los siguientes párrafos a todos los que hicieron este trabajo posible.

En primer lugar, a mis directores, Ana y Fernando. Quienes confiaron en mí desde el primer momento. Proponiéndome un trabajo y un área de investigación que me fascinaron, y acompañándome en cada uno de los pasos que daba en este nuevo mundo de la investigación. De igual manera, quisiera agradecer a Fernando D., por acompañarme desde el comienzo de este camino con la misma paciencia y dedicación que mis directores, transformándose en un tercer director. También quisiera dedicar un agradecimiento a Evangelos, quien también ofició de director desde el momento en el que empezamos a trabajar en conjunto en el año 2019. A todos ellos quiero extender un enorme agradecimiento. Sin su enorme experiencia, esfuerzo y dedicación esta tesis no hubiera sido posible. Gracias a su guía e invaluables consejos es que yo aprendí como desenvolverme en la tarea de la investigación.

También quiero expresar mi agradecimiento a mis padres, quienes siempre me dieron su amor y apoyo incondicional en cada etapa de mi vida, siempre motivándome a perseguir mis estudios con mucha dedicación y entusiasmo. Gracias a todas sus enseñanzas, especialmente a través del ejemplo, es que yo aprendí valores que me fueron fundamentales en muchas etapas de mi vida, en particular durante el desarrollo de esta tesis. También quiero hacer extensivo el agradecimiento a toda mi familia, a todos mis hermanos, a mis tíos y primos, que siempre estuvieron cerca, regalándome momentos felices y de distensión tan necesarios para lograr llevar una vida balanceada.

Quiero ofrecer un agradecimiento muy especial a una persona que me acompañó desde los primeros años de la universidad. Una persona con quien viví y vivo una amistad

increíble, llena de anécdotas, viajes, y momentos compartidos juntos. Quien siempre me empujó a ser una mejor persona en todos los aspectos posibles. Alguien que con su presencia y apoyo incondicional, sus incontables consejos y paciencia durante todos estos años me ayudaron a cumplir los objetivos que me propuse y me permitieron ser la persona que soy hoy, tanto desde lo profesional como desde lo personal. Por todo esto, y muchas cosas más, es que quiero brindar un enorme agradecimiento a mi novie y amor, Vir.

Porque durante todos estos años de desarrollo profesional, también estuvieron acompañados de momentos sociales y de distensión que fueron de gran importancia para mí, no quiero dejar de agradecer a todos los amigos que me acompañaron este tiempo. A todos los amigos que hice en la secundaria y que aún forman parte de mi vida, y especialmente a todos los que me acompañaron tan de cerca durante todo el desarrollo de la tesis desde las salitas de becarios, los pabellones 1 y 2. Así como también muchos colegas del Departamento de Ciencias e Ingeniería de la Computación, con los que compartí cátedra, momentos, juntadas y asados. De igual manera estoy agradecido al equipo de rugby Universidad Nacional del Sur. Volví a jugar al rugby en el momento en que arrancaba el doctorado, el mismo año que se creaba el equipo, y desde ese momento el rugby se transformó en pasatiempo, distracción y pasión que me terminaron acompañando hasta casi el final del doctorado. Muchas gracias a los entrenadores por su dedicación para ayudar a crear y mantener ese espacio para todos los alumnos de la universidad.

Por último, un agradecimiento para el Consejo Nacional de Investigaciones Científicas y Técnicas, al Departamento de Ciencias e Ingeniería de la Computación, y a la Universidad Nacional del Sur en su conjunto por brindarme un espacio de trabajo y recursos para hacer posible mi trabajo. Sumamente agradecido de la educación publica que hizo posible mi carrera. Un agradecimiento, también, a la Facultad de Ciencias de la Computación de la Universidad de Dalhousie, a Compute Canada, y a Google por los premios Latin American Research Awards (LARA), por facilitarme los recursos tan necesarios para la culminación de mi tesis. Así como también a mis cuatro directores que generosamente pusieron a mi disposición los recursos con los que ellos contaban para que yo pueda llevar a cabo mi trabajo de la mejor manera posible.

Eternamente agradecido a todos ustedes que hicieron este trabajo posible.

Maiso

27 de octubre de 2021

Resumen

La existencia de relaciones o dependencias estadísticas en los datos (correlaciones) se puede estudiar mediante herramientas estadísticas que se han desarrollado en los últimos dos siglos. Sin embargo, una pregunta tan simple de formular como: "¿Existe un vínculo causal entre estas dos variables correlacionadas?" presenta un desafío diferente que escapa a las respuestas que pueden brindar herramientas estadísticas clásicas, ya que, como se suele enseñar en todos los cursos de estadística: "correlación no es causalidad". La necesidad por parte de la comunidad científica de responder preguntas causales (¿El fumar causa cáncer? ¿Este medicamento es efectivo para tratar esta enfermedad?, etc.) generó un esfuerzo para la creación de herramientas formales que permitan descubrir y cuantificar efectos causales. Algunos ejemplos son la técnica basada en la Causalidad de Granger (GC por sus siglas en inglés) y la técnica de descubrimiento de estructuras causales PC (que recibe el nombre por las iniciales de sus autores).

Por otro lado, existe un gran interés por parte de la comunidad de procesamiento de lenguaje natural (NLP por sus siglas en inglés) en el descubrimiento de relaciones causales a partir de textos. Sin embargo, la mayoría de los esfuerzos están enfocados en recuperar información causal ya explícita en el texto. Por ejemplo, en la siguiente frase sobre la crisis argentina del 2001: "*Sucedió en el marco de una crisis mayor que se extendió entre 1998 y 2002, causada por una larga recesión que disparó una crisis humanitaria*" se tendría por objetivo extraer los dos vínculos causales que relacionan los tres eventos descritos (la recesión, una crisis económica y otra humanitaria). Estos trabajos, si bien tienen por objetivo el descubrimiento causal, utilizan herramientas más cercanas al área de NLP que a las herramientas usuales en la literatura sobre descubrimiento causal (tales como GC o PC).

Esta tesis propone un marco de trabajo (*framework*) en el que, a través de la utilización de herramientas como GC o PC, se plantea como objetivo el descubrimiento causal

entre variables extraídas de textos de artículos periodísticos cuya relación causal no necesariamente está explícita en el texto. De este modo se obtiene una red causal, donde cada nodo es una variable relevante y cada arco un vínculo causal. Para alcanzar este objetivo primero se proponen soluciones al desafío de extraer y filtrar variables relevantes a partir de textos. Este problema se resuelve mediante el uso de dos enfoques tomados de NLP: (1) una técnica de pesaje de términos y (2) un modelo de detección de menciones de eventos en curso a partir de textos de artículos periodísticos. Se crea un conjunto de datos utilizando las variables extraídas usando estas herramientas de NLP ((1) y (2)). Este conjunto de datos es usado en el paso posterior de extracción de relaciones causales. Se estudian nueve técnicas de descubrimiento causal, y se lleva a cabo un estudio comparativo de la aplicación de las técnicas en más de sesenta conjuntos de datos sintéticos y en un conjunto de datos real de demanda de energía eléctrica. Finalmente, un caso de uso es presentado donde se aplican las mejores técnicas de descubrimiento causal sobre los conjuntos de datos de variables extraídas de los textos de artículos periodísticos, dando lugar así a una demostración completa de la funcionalidad del *framework* (extracción de variables de textos y descubrimiento causal a partir de las mismas).

Los resultados obtenidos muestran la gran flexibilidad del *framework*, permitiendo la combinación de variables de diferentes tipos, con diferentes procesos de construcción, posibilitando la extracción causal posterior. Más aún, dando evidencia que información no textual podría ser incorporada al *framework* (por ejemplo, precios de materias primas, precios de acciones de la bolsa, indicadores socioeconómicos, entre otros). Este *framework* permitiría a un experto partir de un dominio, que puede ser un conjunto de textos periodísticos sobre algún episodio del mundo real, y obtener de manera automática un conjunto de variables relevantes a ese dominio (de las cuales puede elegir visualizar solo algunas, o todas). Posteriormente, se le mostraría al experto un conjunto de vínculos causales extraídos de manera automática, que vincularía a las diferentes variables relevantes al dominio. El grafo causal resultante (variables y vínculos relevantes a un dominio) puede representar una herramienta de gran interés para permitir a un experto tener una visión procesada y resumida de las interdependencias, permitiéndole un mejor entendimiento del dominio o posibilitando sacar conclusiones o explicaciones sobre eventos que se sucedieron o están sucediendo.

Las primeras dos contribuciones de esta tesis están enfocadas en la propuesta de técnicas novedosas de NLP para la etapa de extracción de variables. En esta etapa se propone,

primero, una herramienta nueva para pesaje de términos y estimación de puntajes de relevancia de términos asignados por usuarios. Segundo, se propone una tarea de NLP, de detección de eventos en curso (OED por sus siglas en inglés) para ser usados como variables en el *framework*. Se muestran los resultados de diferentes modelos para la tarea de OED, alcanzando un modelo superador con respecto a modelos existentes para tareas similares. Estas dos contribuciones permitieron la extracción de variables relevantes para ser usadas como nodos del grafo. Finalmente, la tercera contribución principal es la presentación de un análisis comparativo de nueve técnicas de extracción de causalidad y la posterior aplicación de las mejores para un ejemplo de un caso de uso del *framework* completo.

Abstract

The existence of statistical relationships or dependencies in the data (correlations) can be studied using well-known statistical tools that have been developed over the last two centuries. However, a question as simple to pose as "Is there a causal link between these two correlated variables?" entails a whole set of different challenges that escape from the answer that classical statistical tools can provide, since, as is usually taught in statistical courses: "correlation is not causation". The need by the scientific community to answers to causal questions (such as: "does smoking cause cancer?" or "is this drug effective in treating this disease?") generated an effort to create formal tools for detecting and quantifying causal effects. Some examples are the methods based on the Granger Causality (GC) test and the PC causal structure learning algorithm.

On the other hand, there is great interest from the natural language processing (NLP) community in discovering causal relationships from texts. However, most efforts are focused on recovering causal information already explicit in the text. For example, in the following sentence about the Argentine crisis of 2001: "*It happened in the context of a bigger crisis that lasted between 1998 and 2002, caused by a long recession that triggered a humanitarian crisis*" the goal would be to extract the two causal links that relate the three events described (the recession, an economic crisis, and a humanitarian crisis). In that literature, although the goal is also to detect causal relations, tools closer to the NLP field are used, instead of the usual tools in the literature of causal discovery (such as GC-based techniques or PC).

This thesis proposes a framework that aims at performing causal discovery between variables extracted from texts of newspaper articles using tools like GC and PC. In contrast to other approaches, the causal relationships do not need to be explicit in the texts. Using this framework, a causal network is obtained, where each node is a relevant variable and each edge is a causal link. To achieve this goal, the first challenge addressed

is to extract and select relevant variables from texts. This is achieved by the use of two NLP approaches: (1) a term weighting technique and (2) a model for detecting ongoing event mentions in news articles. A data set is built using these two types of variables extracted from texts using these two NLP approaches ((1) and (2)). This data set is used in the following stage of causal discovery. Nine causal discovery techniques are analyzed, and a comparative study of the application of these techniques is carried out in sixty-four synthetic data sets and in one real-world electricity demand data set. Finally, a use case is presented where the best causal discovery techniques are applied to the data sets of variables extracted from the texts of newspaper articles, thus giving rise to a complete demonstration of the functionality of the framework (extraction of text variables and causal discovery from them).

The results obtained show the great flexibility of the framework, which allows the combination of variables of different types (potentially with different generative processes), enabling the subsequent causal extraction. Furthermore, they provide evidence that non-textual information could be incorporated into the framework (for example, commodity prices, stock prices, and socioeconomic indicators, among others). This framework would allow an expert to start from a domain, which can be defined as a set of newspaper texts about some real-world episode, and automatically obtain a set of variables relevant to that domain (from which the expert could choose to visualize either a subset or the entire set). Subsequently, the expert would be shown a set of causal links extracted automatically, linking the relevant variables of the domain. The resulting causal graph (variables and edges relevant to a domain) can become a tool of great interest for an expert to process and summarize the variables and interdependencies in a domain, allowing a better understanding and making it possible to draw conclusions or find explanations for events that happened or are happening in the domain.

The first two contributions of this thesis are focused on the proposal of novel NLP techniques to be applied at the variable extraction stage. First, a new tool for weighing terms and estimating relevance scores of terms assigned by users is proposed. Secondly, an NLP task consisting of the detection of ongoing events (OED) from texts is proposed to use those events as variables in the framework. The results for different instances of the OED task are shown, indicating that the model outperforms state-of-the-art models for similar tasks. These two contributions allow the extraction of relevant variables to be used as nodes of the graph. Finally, the third main contribution is the presentation

of a comparative analysis of nine causality extraction techniques and the subsequent application of the best ones on a use case of the complete framework.

Índice general

Capítulo 1

Introducción

La crisis financiera global del 2008 generó un gran interés por investigar y entender la estructura del sistema financiero y los canales por los que propaga sus riesgos [Ahe16].También generó interés en encontrar herramientas analíticas que permitan identificar, monitorear y abordar el riesgo sistémico. El análisis de redes ha probado ser una herramienta crucial para modelar el complejo sistema de interconexiones de los sistemas financieros. En [Ahe16] se propone el uso de redes bayesianas para la construcción de estas herramientas para modelar interconexiones en los sistemas financieros.

Las redes bayesianas [Pea85, KF09] (BN por sus siglas en inglés) constituyen una clase de modelos gráficos probabilísticos en los que grafos dirigidos acíclicos codifican variables aleatorias y sus dependencias condicionales. Aunque las orientaciones de los arcos en una BN pueden no ser significativas, es una "buena" práctica que se correspondan con direcciones causales (ya sea causalidad directa o indirecta) [KF09]. Sin embargo, siempre que una BN represente correctamente una distribución probabilística subyacente la respuesta a preguntas probabilísticas va a ser la misma, independientemente de si la estructura de la BN es causal o no. Aunque las respuestas a preguntas probabilísticas sean las mismas, existe un tipo de razonamiento para el cual es crucial que la red represente una semántica causal: el razonamiento intervencional. Esto es donde se introduce un cambio en el curso natural de los eventos. Para este tipo de razonamiento los modelos $X \to Y$ y $Y \to X$, que son equivalentes como modelos probabilísticos, resultan modelos con semánticas causales totalmente distintas y con diferente respuesta a preguntas intervencionales. Por ejemplo, la cantidad de policías de un barrio y la tasa de crímenes pueden estar correlacionadas (porque más policías son asignados a barrios peligrosos y menos a menos peligrosos). Por

ende, ambas variables son buenos predictores de la otra variable. En este caso, sin embargo, la pregunta observacional es distinta que la intervencional: si observo muchos policías probablemente también observe una tasa alta de crímenes, pero si yo *asigno* muchos policías esto no implica que vaya a observar una tasa alta de criminalidad. Poder responder adecuadamente a este tipo de preguntas es fundamental para la toma de decisiones (porque ésta involucra intervenir en el sistema, no solo observar sus datos).

Mientras que en [Ahe16] proponen el uso de redes bayesianas para la construcción de herramientas para modelar interconexiones en los sistemas financieros, en [Var14] se ofrece una discusión sobre la importancia de la colaboración entre expertos de las áreas de econometría y aprendizaje automático para contribuir al desarrollo de herramientas de inferencia causal. Se argumenta, en dicho trabajo, que existen muchas técnicas de causalidad propuestas en la literatura de econometría que pueden ser utilizadas para el diseño de dichas herramientas[1]. También se observa que investigadores del área de computación, como Judea Pearl, han realizado contribuciones significativas al modelado causal útiles para el análisis de fenómenos económicos[Pea09, PM18]. Sin embargo, Varian opina que estos avances teóricos aún no se han incorporado en la práctica del aprendizaje automático en un grado significativo y que mucho de los esfuerzos se concentran en la "predicción pura".

Por otro lado, grandes volúmenes de datos acerca de la actividad económica son recolectados de forma rutinaria por organizaciones y programas nacionales e internacionales, encontrándose disponibles a través de diversas fuentes. Análogamente, gran cantidad de datos útiles pueden ser recolectados a través de otros medios tales como periódicos digitales o plataformas de *microblogging* o redes sociales. Esto abre nuevas oportunidades para la explotación de la sinergia entre aprendizaje automático y econometría, en particular, orientado a generar y poner a prueba conjeturas acerca de posibles relaciones causales en los datos.

La presente tesis responde a la necesidad de desarrollar herramientas que permitan modelar variables y sus interconexiones en modelos complejos, como, por ejemplo, los sistemas financieros [Ahe16]. En un contexto favorable de creciente disponibilidad de datos para la tarea propuesta, y motivado por las posibilidades que brinda la sinergia entre las

[1] Presenta como ejemplos la noción de causalidad de Granger [Gra69], el uso de variables instrumentales [AIR96], el concepto de discontinuidad en las regresiones [TC60] o la detección de diferencias en diferencias [CK93, Car90]. Un resumen de estas herramientas se puede encontrar en [Cun21].

áreas de econometría y aprendizaje automático para la inferencia causal, **se propone como objetivo detectar eventos del mundo real y extraer otras variables relevantes a partir de textos de noticias con el objetivo de aprender modelos causales**. El objetivo final es el estudio de la viabilidad de una herramienta que ayude a expertos a entender y explicar cómo se desarrollan eventos en escenarios complejos y sus interconexiones.

La propuesta de un marco de trabajo (*framework*) para la construcción de modelos causales a partir de textos de noticias *relevantes* puede ser dividida en dos etapas. La primera consiste en la construcción y selección de variables relevantes a partir de los textos. Cada una de estas variables es representada como una serie de tiempo que captura el momento y la frecuencia con la que dicha variable es mencionada en los textos. Dichas variables constituyen los nodos del modelo causal. Por otra parte, la segunda etapa consiste en la aplicación de una técnica de descubrimiento causal a las variables obtenidas en la etapa previa. El modelo causal es representado como un grafo dirigido $\mathcal{G}$ constituido por nodos y arcos, donde los nodos son las variables de interés y los arcos capturan las relaciones causales y sus direcciones.

La primera etapa de este trabajo (1) (construcción y selección de variables), es abordada de dos maneras distintas. (1.a) Primero a través de la recuperación de términos de interés (unigramas, bigramas y trigramas) de los textos *relevantes* de artículos periodísticos y su posterior filtrado para seleccionar los más relevantes. (1.b) La segunda forma en que se trata el problema es a través de un modelo de detección de eventos en curso, con el que se detectan todas las menciones de eventos en cada fragmento de texto *relevante* para su posterior procesamiento. Cada mención de evento individual es agrupada con menciones semánticamente similares y cada grupo constituye una variable. De esta manera, por ejemplo, diferentes menciones de subas del precio del dólar en un dado país X, pueden constituir un grupo que se identifica con la variable "aumento del dólar en X", capturando los diferentes momentos en los que ese evento es reportado en las noticias.

El primer enfoque para construir y seleccionar variables (1.a) consiste de construir un vocabulario de unigramas, bigramas y trigramas usando todos los textos *relevantes* para el usuario. Como el objetivo final del trabajo es la construcción de modelos causales para un cierto dominio de interés, primero hay que definir el dominio por medio de los fragmentos de textos considerados *relevantes*. Por ejemplo, si un usuario está interesado

en recuperar variables y eventos de interés de la crisis argentina del 2001, deberá disponer de artículos periodísticos que mencionen a Argentina durante el periodo de la crisis (y tal vez unos meses previos y posteriores). Usando los textos *relevantes* del dominio de interés y un conjunto de textos no relevantes, seleccionados al azar, se plantea la utilización de una técnica de pesaje de términos supervisada para ordenar los términos del vocabulario en función de su relevancia para el tópico. El usuario de la herramienta propuesta podría elegir entre los K términos más relevantes aquellos que se constituyen en variables para el modelo causal $\mathcal{G}$ y aquellos que no. Para llevar a cabo esta tarea se propone una nueva técnica de pesaje de términos supervisada denominada FDD_β.

El Capítulo 2 presenta, analiza y evalúa la técnica FDD_β, una técnica supervisada de pesaje de términos (*term-weighting*). La técnica FDD_β asigna un puntaje a los términos basándose en dos factores que representan el poder descriptivo y discriminativo de los términos con respecto a un cierto tópico. Dicho puntaje combina estos dos factores a través de un parámetro ajustable que permite favorecer distintos aspectos de la recuperación de información. Durante el desarrollo de los experimentos para dicho capítulo se alcanzaron las siguientes contribuciones: (1) La presentación y análisis de una nueva técnica de pesaje de términos con resultados prometedores para diferentes tareas relacionadas a la recuperación de información y la extracción de variables. La técnica es evaluada en las tareas de: (i) estimación de puntajes de relevancia de términos para tópicos asignados por usuarios, (ii) construcción de consultas para la recuperación de información temática, (iii) cobertura total (*total recall*), (iv) construcción de consultas de múltiples términos usando operadores disyuntivos y, finalmente, (V) en términos de su comportamiento en función del parámetro ajustable. La técnica demostró ser una parte fundamental del *framework* de recuperación de estructuras causales a partir de textos por su capacidad como estimador de la relevancia de términos para el tópico que le asigne el usuario a cada término (tarea (i)).

El segundo enfoque para construir y seleccionar variables (1.b) fue llevado a cabo implementando un modelo predictivo basado en redes neuronales para la detección de eventos en los textos de noticias relevantes al dominio de interés del usuario. Para poder construir las variables de tipo serie de tiempo necesarias para la etapa de recuperación causal (2), es necesario no solo recuperar la mención de los eventos relevantes, sino también saber el momento del tiempo en el que sucedieron. Para obtener el momento en el que sucedieron se puede usar la fecha de publicación de los textos **solo** si

los eventos encontrados se corresponden con **eventos en curso** al momento de que son reportados/mencionados en los textos. Por este motivo en el Capítulo 3 se presenta la definición de la tarea de detección de eventos en curso (*Ongoing Event Detection* (OED)), que es una tarea específica dentro de la tarea más general de Detección de Eventos (*Event Detection* (ED)). El objetivo de la tarea de OED es la detección de menciones de eventos en curso, en contraposición a eventos históricos, futuros, hipotéticos u otras formas de eventos que no están en curso ni son actuales. En el mismo capítulo se implementa y analiza el desempeño de un modelo de redes neuronales para la tarea de OED. El modelo demostró ser una parte fundamental del *framework* de descubrimiento de estructuras causales a partir de textos por su capacidad para detectar los eventos de interés para la posterior construcción de las variables de tipo evento.

Las contribuciones más importantes del capítulo de detección de eventos en curso (Capítulo 3) son: (1) la definición de la tarea de OED junto con un conjunto de datos manualmente etiquetado para dicha tarea, (2) el diseño y desarrollo de un modelo predictivo basado en Redes Neuronales Recurrentes (*Recurrent Neural Networks* (RNN)) para la tarea de OED, (3) la presentación de una extensa evaluación empírica que incluye exploración de arquitecturas, hiperparámetros y atributos, así como también la presentación de dos modelos de referencia (*baseline*) para la tarea.

Finalmente, **la segunda etapa de este trabajo (2)** que se apoya en las tareas de la primera etapa ((1.a) y (1.b)) consiste de la construcción de las variables utilizando las herramientas propuestas en los Capítulos 2 y 3 para la posterior aplicación de una técnica de descubrimiento causal sobre las mismas. Los dos tipos de variables obtenidas a partir de (1.a) y (1.b) son combinadas en el mismo conjunto de datos para ser usadas en la etapa posterior. Esta segunda etapa se describe en el Capítulo 4. En este capítulo se ofrece un extenso análisis comparativo de diferentes técnicas de aprendizaje de estructuras causales a partir de datos de tipo serie de tiempo para su posible aplicación a la etapa (2). Posteriormente, una descripción completa del *framework* de descubrimiento de estructuras causales a partir de textos es presentada. Se describe, a modo de demostración de la capacidad de alcanzar los objetivos propuestos, un caso de uso de dicho *framework* a partir del corpus de noticias del *New York Times*[2] [San08]. Utilizando un subconjunto de los textos del corpus a modo de dominio de interés se muestra la aplicación de cada

[2]https://catalog.ldc.upenn.edu/LDC2008T19

uno de los pasos ((1) y (2)) para ejemplificar cómo el *framework* cumple con las metas planteadas.

Durante el desarrollo del caso de uso, en el paso (1) se utilizan las herramientas de los Capítulos 2 y 3 para la construcción del conjunto de datos de variables relevantes. Para el caso de las variables tipo términos, la técnica FDD_β es utilizada para elegir los mejores K términos para presentar al usuario. En el caso de las variables de tipo evento, el modelo de detección de eventos en curso es utilizado para detectar todas las menciones de eventos presentes en fragmentos de textos relevantes al dominio de interés. Luego, a partir de todas las menciones individuales de eventos se crean grupos de menciones de eventos semánticamente similares para constituir cada variable (como por ejemplo la variable "aumento del dólar"). Filtrando los eventos por cohesión y cantidad de menciones se construye un conjunto de datos con variables de tipo evento, que es combinado con el conjunto de datos del término para crear el conjunto de datos final para ser usado en la etapa de descubrimiento causal (2). Finalmente el resultado final del caso de uso del *framework* es presentado y analizado en el mismo capítulo.

Las contribuciones principales del Capítulo 4 son: (a) la formulación completa del marco de trabajo (*framework*) para obtener estructuras causales para expertos a partir de artículos periodísticos (pasos (1) y (2)); (b) la presentación de un extenso análisis comparativo de nueve técnicas de aprendizaje de estructura causal en series de tiempo, analizando 64 conjuntos de datos sintéticos y un conjunto de datos reales sobre demanda eléctrica en el Gran Buenos Aires (GBA); y (c) la presentación de un caso de estudio de la aplicación del *framework* completo a texto.

Para concluir, en el Capítulo 5 se presentan las conclusiones y discusiones que surgen de los experimentos realizados en el contexto de esta tesis. Se analizan las contribuciones individuales de cada uno de los tres capítulos principales que en su conjunto constituyen el *framework* para obtener estructuras causales para expertos a partir de artículos periodísticos (los Capítulos 2, 3 y 4). La viabilidad de una herramienta interactiva que implemente el *framework* es puesta a discusión en ese capítulo, y se delinean posibles trabajos futuros que se apoyan en la flexibilidad del *framework* propuesto. Entre las posibles mejoras se incluye la incorporación de variables adicionales que pueden ser de naturaleza distinta (no solo obtenidas a partir de textos), así como también la incorporación de técnicas de recuperación de causalidad adicionales que también pueden ser de naturaleza distintas (basadas en procesamiento de lenguaje natural (NLP por sus siglas en inglés)).

Capítulo 2

Selección de Variables

Resumen

La recuperación de información temática (topic-based retrieval) es la tarea de buscar y recuperar material relacionado a un tópico de interés. Esta tarea involucra dos subtareas: (1) construir la consulta eligiendo términos adecuados y (2) ordenar por relevancia los resultados obtenidos. Los enfoques supervisados para evaluar la importancia de un término en un tópico (o clase) han demostrado ser efectivos en guiar la subtarea de selección de términos para la consulta (subtarea 1). El presente capítulo presenta, analiza y evalúa la técnica FDD_β, una técnica supervisada de pesaje de términos (term-weighting) que puede ser usada en el proceso de selección de términos para consultas (subtarea 1) en la recuperación de información temática. La técnica FDD_β asigna puntaje a los términos basándose en dos factores que representan el poder descriptivo y discriminativo de los términos con respecto a un cierto tópico. Dicho puntaje combina estos dos factores a través de un parámetro ajustable que permite favorecer distintos aspectos de la recuperación de información, como pueden ser la precisión (*precision*), la cobertura (*recall*) o un balance de los dos. Durante el desarrollo de los experimentos para el presente capítulo se alcanzaron las siguientes contribuciones: (1) La presentación y análisis de una nueva técnica de pesaje de términos con resultados prometedores para diferentes tareas relacionadas a la recuperación de información y la extracción de variables. La técnica es evaluada en las tareas de: (i) estimación de puntajes de relevancia de términos para tópicos asignados por usuarios, (ii) construcción de consultas para la recuperación de información temática, (iii)

cobertura total (*total recall*), (iv) construcción de consultas de múltiples términos usando operadores disyuntivos y (v) análisis de la técnica en función del parámetro ajustable.

(2) La técnica es empíricamente evaluada utilizando tres conjuntos de datos diferentes y comparada con 18 técnicas del estado del arte y tradicionales. Dos de los tres conjuntos de datos son conjuntos clásicos del área de NLP (*20 Newsgroups* y *Reuters-21578*), mientras que el tercero es un conjunto de datos generados durante el desarrollo experimental de este capítulo y es publicado para permitir la reproducibilidad del trabajo.

Es posible concluir que, a pesar de su simplicidad, FDD_β es competitivo con el estado del arte y ofrece una gran flexibilidad que otras técnicas no tienen. Esta flexibilidad le permite adaptarse a diferentes objetivos y resulta en que la técnica ofrezca un mecanismo conveniente para explorar diferentes enfoques para construir consultas complejas.

2.1. Introducción

La recuperación de información temática es el problema de buscar material almacenado en forma de texto (documentos, artículos periodísticos, tweets, etc.) relevantes a un cierto tópico y devolver este material como respuesta al usuario interesado en dicho tópico. Para obtener estos resultados, se deben generar consultas sobre tópicos mediante la selección estratégica de términos que puedan ayudar a obtener buen rendimiento en la recuperación de material relevante. Los términos pueden ser elegidos a partir de diferentes fuentes, dependiendo de cómo se representa el tópico de interés. Por ejemplo, un tópico puede ser representado usando un fragmento de un texto, o un conjunto de palabras, o un conjunto de documentos etiquetados como relevantes o irrelevantes para el tópico. Si dicha colección de documentos está disponible, se vuelve posible aplicar técnicas supervisadas para construir consultas para ese tópico.

En este trabajo se utilizan tópicos amplios (como economía, religión, electrónica, etc.) representados usando una colección de documentos (noticias, publicaciones en grupos de difusión, etc.) etiquetados como relevantes o irrelevantes para el tópico considerado. El objetivo es el de extraer términos (unigramas, bigramas o trigramas) relevantes a un tópico a partir de los textos y ordenarlos de acuerdo a su importancia para el tópico analizado. Por ejemplo, un término puede ser un buen descriptor de un tópico y otro puede ser un gran discriminador del tópico, lo cual los transforma en términos de consulta útiles

para recuperación de textos (aunque por sus distintas cualidades puede que sirvan para distintos objetivos). El objetivo final del trabajo propuesto es el de proveer al usuario con un conjunto de variables relevantes que sirvan para resumir un dado tópico y explorar estrategias para construir consultas para recuperación de textos basada en tópicos.

Los modelos tradicionales de recuperación de información típicamente aplican técnicas no supervisadas para determinar la importancia de un término. Estos enfoques asignan un valor numérico a cada término en un documento basado en el número de ocurrencias del término en el documento, factores basados en ocurrencia a nivel de documento se suelen llamar factores locales. Típicamente, cuanto mayor es la frecuencia del término en el documento mejor es su poder descriptivo. Adicionalmente, la mayoría de los modelos adoptan algún tipo de noción de especificidad del término, que usualmente está asociada al número de documentos en el que el término aparece. A este tipo de factores que mide ocurrencia a nivel de colección se los suele llamar factor global. En este sentido, cuanto más baja es la frecuencia de un término en una colección de documentos, su especificidad es más alta, por ende, más alto su poder discriminativo. Por el contrario, términos que aparecen demasiado frecuentemente en la colección de documentos tienen poco poder discriminativo, llegando al caso extremo de las palabras vacías (*stopwords*) que aparecen en casi todos los documentos y no aportan ningún tipo de información útil para la recuperación de información (ni otras tareas como clasificación de documentos). Estos dos factores son combinados de distintas formas en diversas técnicas de pesaje de términos que miden el nivel informativo de un término en un documento basándose en una combinación de estos factores. El valor informativo trata de medir qué tan bien describe o discrimina un término a un dado documento, pero es independiente del tópico de interés. Solo se basa en qué tan frecuentemente aparece un término en un documento y qué tan frecuentemente aparece ese término en la colección, sin hacer distinción si son documentos relevantes para el tópico o no. Por esto los puntajes obtenidos a partir de técnicas de pesajes de términos no supervisadas representan el nivel de información de un dado término para un documento dado un corpus, sin tener en cuenta si los documentos del corpus pertenecen o no al tópico. Este es el caso para la técnica no supervisada más utilizada, TF-IDF (term-frequency, inverse document frequency). Esta técnica se basa en un factor local (TF) y un factor global (IDF). De manera similar a TF-IDF, otras técnicas combinan de otras formas estos factores descriptivos y discriminativos([RJ76, TM94, Rob04, DGB$^+$10]). Sin embargo, por ser no supervisadas no se utiliza información del tópico del documento, y por ende, si se tiene esta información se estaría dejando intencionalmente de lado información

importante para la recuperación de información temática.

Otras técnicas de pesaje de términos adoptan un enfoque supervisado, en estos enfoques se combinan otros factores para medir la importancia de un término para un dado tópico o clase. En la mayoría de los casos, las técnicas de pesajes supervisada fueron formuladas en el contexto de tareas de clasificación, donde la importancia de un término para una clase es un valor fijo ([DS04, LSLT05, WZ13, DLY14, CZLZ16, VSHK16, FS16, WGG17, FLSZ18]). Mientras que usar un valor fijo puede ser apropiado para técnicas de clasificación, representa una limitación para recuperación de información temática, ya que un término puede ser más o menos efectivo dependiendo de si la tarea requiere alta cobertura, alta precisión o un balance de ambas.

En este capítulo se analiza la técnica de pasaje de términos supervisada FDD_β, propuesta en [MDTM21, MDTM19b, MDTM18] que utiliza un parámetro ajustable β para favorecer distintos aspectos de la recuperación de información. La técnica FDD_β ofrece una ventaja por encima de otras técnicas del estado del arte, ya que permite favorecer a la precisión, cobertura u obtener un balance de ambas a través de su parámetro ajustable. Adicionalmente, la técnica FDD_β tiene la ventaja de estar basada en una formulación muy sencilla derivada de nociones tradicionales usadas como métricas en recuperación de información. Esto contrasta con otras técnicas que se basan en nociones de teoría de información o estadística mucho más elaboradas, como entropía, información mutua, o distribuciones de probabilidad. Los resultados experimentales sugieren que estas nociones más complejas adoptadas por muchas técnicas de pesaje son útiles para técnicas no supervisadas, pero no necesariamente más efectivas que las nociones simples adoptadas para la técnica FDD_β en un escenario supervisado.

Este capítulo está estructurado como se detalla a continuación. En la Sección 2.2 se presentan los objetivos y contribuciones de los experimentos del presente capítulo. En la Sección 2.3 se presentan los conceptos base y se revisa la literatura relacionada. En la Sección 2.4 se describe la técnica FDD_β y se discute cómo puede ser aplicada a la recuperación temática de información a partir de un conjunto etiquetado de documentos, mostrando cómo puede favorecer la precisión, la cobertura, o un balance entre ambos.

En la Sección 2.5 se presentan los tres conjuntos de datos utilizados y en la sección 2.6 se examina el efecto del parámetro ajustable β en la técnica FDD_β. Luego, en la Sección 2.7 se presentan la aplicación de la técnica propuesta a la primera tarea, estimación de puntajes de relevancia para tópicos asignados por usuarios. En la sección 2.8 se pre-

senta una evaluación de la técnica propuesta para la tarea de recuperación de información temática, comparándola con otras técnicas del estado del arte. Finalmente, en la Sección 2.9 se describen algunas aplicaciones que podrían beneficiarse de la técnica de pesaje de términos propuesta y en particular se presenta un análisis de la técnica para la tarea de cobertura total). Una discusión de los resultados y sus implicancias es presentada en la Sección 2.10 y, por último, en la sección 2.11 se presentan las conclusiones y posibles trabajos futuros.

2.2. Objetivos y Contribuciones

La motivación de toda la experimentación, resultados y conclusiones presentados en este capítulo están centrados en evaluar el comportamiento y rendimiento de la técnica propuesta, FDD_β, para las tareas de estimación de puntajes de relevancia para un tópico asignados por usuarios y para la selección de términos para la tarea de recuperación temática de información. En particular, la primera tarea (estimación de puntajes de relevancia) es la tarea central a cumplir por la técnica FDD_β como parte del *framework* de recuperación de estructuras causales a partir de textos (los puntajes permiten determinar qué términos son elegidos para formar parte de las estructuras causales). Los distintos experimentos pretenden analizar diferentes aspectos de la técnica propuesta frente a diferentes problemas a resolver. Se proponen los siguientes objetivos a resolver a partir de las experimentaciones del presente capítulo.

1. Lograr un entendimiento del comportamiento de la técnica FDD_β como función de su parámetro ajustable β.

2. Determinar la capacidad de dicha técnica como un estimador de puntajes de relevancia de tópicos asignados por usuarios.

3. Evaluar la eficacia de dicha técnica para selección de términos para la tarea de recuperación de información temática, comparándola con otras técnicas del estado del arte y tradicionales.

4. Explorar la utilidad de la técnica FDD_β para generar consultas de múltiples términos para la tarea de recuperación temática de información.

5. Indagar en el desempeño de dicha técnica para la tarea de recuperación de información con el objetivo de alcanzar cobertura total.

Este capítulo presenta las siguientes contribuciones:

1. La técnica supervisada de pesaje de términos FDD_β es introducida.
2. Un amplio análisis del comportamiento de la técnica FDD_β como función de su parámetro ajustable β es presentado.
3. Un estudio es presentado donde se mide la eficacia de la técnica para estimar puntajes de relevancia para términos en tópicos en comparación con los puntajes asignados manualmente por ocho voluntarios.
4. Una comparación de la técnica FDD_β contra dieciocho técnicas tradicionales y del estado del arte es presentada. Los métodos seleccionados para llevar a cabo la comparación incluyen técnicas supervisadas y no supervisadas.
5. Una evaluación y comparación del desempeño de consultas disyuntivas construidas a partir de técnicas seleccionadas con FDD_β y otras técnicas del estado del arte es presentada.
6. Se presentan varias posibles aplicaciones y se muestran resultados para una de ellas en particular: recuperación de información con el objetivo de alcanzar cobertura total.
7. Se publican recursos para permitir la replicación de varios de los resultados obtenidos a lo largo de este capítulo, promoviendo trabajos que buscan encontrar más evidencia para las mismas hipótesis que aquí se plantean y para trabajos que buscan probar nuevas hipótesis a partir de los resultados y recursos obtenidos de este trabajo. Como parte de las contribuciones de este trabajo se deja disponible un conjunto de datos que contiene 1.789 artículos periodísticos del portal de noticias *The Guardian* etiquetado por expertos en el dominio como relevantes o irrelevantes para el dominio económico. También dejamos disponible el código fuente donde se define la técnica FDD_β y se implementan las dieciocho técnicas con las cuales se compara la propuesta. En dicho código fuente se muestra el resultado de varios experimentos mencionados en este capítulo. Para mayor detalle del código disponible ver la Sección 2.12.

Todo el análisis y las evaluaciones son realizadas sobre tres conjuntos de datos: dos conjuntos de datos populares de la literatura, *20 Newsgroups* y *Reuters-21578*, y el conjunto de datos etiquetado y publicado como parte de los experimentos de esta tesis, al cual se denomina *Economic Relevant News from The Guardian*.

Los resultados obtenidos de los análisis y evaluaciones demuestran que la técnica FDD_β ofrece un mecanismo útil para explorar diferentes enfoques de construcción de consultas complejas. Además, el análisis comparativo del desempeño muestra que, a pesar de su simplicidad, FDD_β alcanza resultados competitivos con el estado del arte sin basarse en nociones complejas de teoría de la información o estadística.

2.3. Conceptos Base y Trabajos Relacionados

El proceso de recuperación de información temática es el proceso de buscar y recuperar material relacionado a algún tópico de interés ([LML+16]). Una colección de documentos etiquetados puede ser usados para representar un tópico o tema de interés y ofrece una guía para el proceso de recuperación de información. La recuperación de información temática es diferente de la más conocida recuperación de información tradicional. Notar que la recuperación de información ad hoc es el proceso de obtener documentos relevantes como respuesta a una pregunta que realizó un usuario ([VH99]). En contraste, la recuperación de información temática tiene como objetivo obtener documentos relevantes a partir de un tópico, que típicamente es más amplio que la consulta del usuario en la recuperación de información tradicional, y está definido por un contexto temático. El contexto temático podría a la vez estar definido por una muestra de documentos relevantes al tópico o tema de interés. El contexto puede ser aumentado con información negativa, esto es, tener información de documentos que se sabe son irrelevantes al tópico etiquetados como tal, lo que facilita el proceso de entrenamiento. Esta colección puede ser explícitamente provista por el usuario o puede ser inferida a partir de los intereses o interacciones de un usuario con un sistema. Por ejemplo, un sistema puede estar monitoreando un historial de consultas, tiempo de permanencia, clics, etc.

La recuperación de información temática puede ser usada para generar alertas basadas en tópicos ([ES19]), o asistir a expertos mientras organizan el conocimiento de dominio de un tópico ([LML+16], construir portales verticales ([PFM18]), entre otras posibles aplicaciones. Dependiendo de la tarea que se esté abordando, el énfasis de la recuperación

temática puede ponerse en obtener alta precisión, alta cobertura o una combinación de ambas. La recuperación temática usualmente se basa en la formulación automática de consultas. Por ende, uno de los problemas más desafiantes a resolver es la generación automática de consultas poniendo énfasis en los diferentes objetivos que tenga la tarea que se esté resolviendo.

El problema de la generación automática de consultas para la recuperación temática puede ser formulado como un problema de optimización donde el objetivo es maximizar la eficacia de la consulta en términos de cierta métrica de desempeño de recuperación de información temática. Uno de los desafíos de este problema de optimización es que el problema no tiene una subestructura óptima, lo cual significa que la solución óptima no puede ser construida eficientemente a partir de soluciones óptimas a subproblemas. Como consecuencia, combinar buenos términos de consulta no necesariamente resulta en una consulta más larga con buen desempeño. Sin embargo, debido a la alta dimensionalidad del espacio de búsqueda y el carácter combinatorio del problema de búsqueda de una consulta óptima, los abordajes más comunes involucran selección de términos adoptando una estrategia naïve. Esto es, eligiendo cada término de forma independiente a los otros términos que son elegidos. En consecuencia, el algoritmo de selección típicamente computa, para cada término candidato, una medida de informatividad basada en algún esquema de pesaje de términos, y ordena los términos de manera decreciente usando esa medida, para luego usar cierta cantidad de los mejores términos candidatos para construir la consulta. El problema de generación de consultas para recuperación de información temática se reduce a definir un esquema de pesaje de términos y decidir cuántos de estos términos se van a usar y cómo se van combinar en la consulta (usando operadores disyuntivos, conjuntivos, o una combinación de ambos).

2.3.1. Técnicas No Supervisadas de Pesaje de Términos

Muchas medidas sobre la informatividad de un término han sido usadas para clasificación de textos y recuperación de información. Los esquemas de pesaje tradicional adoptan un enfoque no supervisado y se originan de la comunidad de recuperación de información. En [SB88] se menciona la existencia de dos factores fundamentales para la construcción de un esquema de pesaje de términos: un factor global y un factor local. Un tercer factor mencionado es el factor de normalización que se suele usar para corregir el rango de los pesos obtenidos. Los factores locales típicamente son usados para representar la frecuencia

de aparición de un término en un documento dado. Sirven para modelar la intuición de que los términos que aparecen varias veces en un documento están estrechamente relacionados con el contenido del mismo. Están diseñados para mejorar la cobertura de la técnica de recuperación. El más simple de los factores locales puede ser considerado un factor binario que toma el valor 1 para representar la presencia de ese término en un documento dado y el valor 0 para representar la ausencia del término en el documento. Notar que por tratarse de un factor local un mismo término tiene diferentes puntajes en diferentes documentos. Esto significa que el valor del puntaje depende tanto del término elegido como del documento considerado. Otro de los factores sencillos, y más conocidos, es el factor Frecuencia de Término (TF por sus siglas en inglés), que asigna de puntaje el número de apariciones del término elegido en el documento considerado. Algunas variantes del factor TF, son por ejemplo, el *término de frecuencia inverso* (ITF por sus siglas en inglés) ([LK02]), que normaliza el valor de TF al intervalo [0, 1] basándose en la ley de Zipf. Otra transformación al factor TF es presentada en [DS04], donde el factor TF para términos extremadamente frecuentes en el documento no aumenta de valor tan rápido como el clásico TF.

A diferencia de los factores locales, los factores globales representan qué tan frecuente o infrecuente es un término en una colección de documentos. Por ejemplo, una variación del factor local TF para transformarlo en un factor global mide la frecuencia del término a nivel de toda la colección de documentos. A esta técnica, en este trabajo, se la denomina Frecuencia Global del Término (TGF por sus siglas en inglés). Aunque TGF tiene como objetivo mejorar la cobertura, en general los factores globales son usados para penalizar términos demasiado frecuentes a nivel de la colección, mejorando así la precisión. La intuición detrás de estos factores es considerar que un término muy común es un mal discriminador. El ejemplo más común dentro de esta categoría es el factor frecuencia inversa en el documento, o IDF por sus siglas en inglés ([SB88]). Este método de pesaje penaliza en base a la cantidad de documentos de la colección que contienen el término, suponiendo que términos que aparecen en la mayoría de los documentos no aportan información ni permiten discriminar ya que son probablemente palabras frecuentes del idioma como palabras vacías (stopwords). Otro factor global es la frecuencia inversa pesada en el documento o WIDF por sus siglas en inglés ([TM94]). Este método penaliza términos muy frecuentes en la colección teniendo en cuenta el número de veces que aparece en cada documento. Un enfoque no supervisado al poder discriminativo y descriptivo de un término en un tópico es presentado en [MLRM04].

2.3.2. Técnicas Supervisadas de Pesaje de Términos

Más relacionados al problema de recuperación de información temática son los métodos que utilizan la información de clase, lo que da lugar a los métodos supervisados. Un método simple que utiliza información de clase puede ser computado contando la cantidad de documentos de la clase en los que aparece el término. Por ejemplo, TGF* es una variación supervisada de TGF que cuenta la cantidad de documentos en una clase que contienen al término. Una técnica tradicional de pesaje de términos supervisada está dada por la probabilidad condicional de que un término ocurra en una dada clase, lo que da origen al método *fracción de probabilidad* (o odds ratio (OR) por sus siglas en inglés) ([vHP81]). Otra técnica basada en probabilidad (*Prob*) presentada en [LLS09] se computa utilizando dos proporciones directamente relacionados con la capacidad del término de representar un tópico o categoría. La primera proporción aumenta con el número de documentos de la clase que contienen al término, mientras que la otra aumenta cuando la mayoría de las ocurrencias del término son dentro de la clase. Para representar el poder de discriminación de un término, otra técnica supervisada denominada frecuencia inversa de clase (ICF por sus siglas en inglés) y el método factor de relevancia de categoría (CRF por sus siglas en inglés) ([DTY$^+$02]) penalizan a un término proporcionalmente al número de diferentes clases en las que el término aparece. Las técnicas supervisadas de pesaje de términos derivadas de la teoría de información incluyen Información Mutua (MI por sus siglas en inglés), *chi-squared* (χ^2), *Ganancia de Información* (IG por sus siglas en inglés) y *proporción de ganancia* (GR por sus siglas en inglés).

En [DU19b], los autores demuestran el impacto medible de incorporar factores locales en técnicas supervisadas de pesaje de términos. En su análisis incluyen el uso de tres factores locales distintos, TF, la raíz cuadrada de TF (SQRT_TF), y el logaritmo de TF (LOG_TF). Estos factores locales son combinados con siete técnicas supervisadas de pesaje de términos. Sus experimentos indican que SQRT_TF ofrece los resultados más prometedores como factor local. Los coeficientes *Galavotti-Sebastiani-Simi* presentados en [GSS00] y el factor de diferencia de cobertura de categoría basado en entropía (ECCD por sus siglas en inglés) presentado en [LMG11] son esquemas de pesaje de términos adaptados de técnicas de selección de atributos (*feature selection*). Otro esquema usa factor de frecuencia de relevancia (RF por sus siglas en inglés), presentado en [LTSL09]. Está técnica busca favorecer términos cuya frecuencia en la clase positiva es más alta que en la clase negativa.

Varias otras definiciones de factores globales son modificaciones del factor global IDF. Por ejemplo, en [TLL20], los autores proponen una nueva técnica de pesaje de términos llamada frecuencia inversa en documentos exponencial (TF-IEF por sus siglas en inglés), para abordar algunas limitaciones encontradas en el factor clásico IDF. Otra variante de IDF es llamada frecuencia inversa en documentos excluyendo categoría (IDFEC por sus siglas en inglés) ([DMPS15]). Esta técnica penaliza términos frecuentes, pero no penaliza aquellos términos que ocurren varias veces en documentos pertenecientes a la clase relevante. La combinación de la técnica IDFEC y RF resulta en la técnica IDFEC_B scheme ([DMPS15]). En [SBMM19] los autores proponen las técnicas ifn-tp-icf, RFR y modOR, que están propuestas basándose en tres técnicas del estado del arte, iqf-qf-icf ([QWQ11]), RF ([LTSL09]) y OR ([vHP81]), respectivamente. Además de proponer estas tres técnicas nuevas, también proponen una técnica llamada ifn-modRF y analizan el desempeño de los métodos para la tarea de clasificar textos cortos (short-text classification). Otras técnicas de pesaje de términos fueron propuestas específicamente para el contexto de textos cortos ([SBMM19, AH19]). En [AX20], los autores proponen una transformación estadística, una estandarización, que es usada para definir una nueva técnica de pesaje de términos. Otra técnica de pesaje de términos para computar el poder discriminativo de una consulta fue presentada en [kSM12], y se basa en los rangos de los valores de similitud de los documentos relevantes y no relevantes recuperados por la consulta.

Inercia de Gravedad Inversa (IGM por sus siglas en inglés) [CZLZ16] es una técnica de pesaje de términos que incorpora una medida estadística de la concentración ínter-clase de un término. Para definir el valor de IGM de un término t_k es necesario computar la frecuencia de ese término en todas las clases. Por ende, cada término t_k tiene asociado un único valor de IGM, no uno por tópico o categoría. La distribución de la frecuencia del término es usada para evaluar el poder de discriminación de clase del término. En otras palabras, las frecuencias de un término en cada clase definen una lista ordenada $f_{k1} \geq f_{k2} \geq \ldots \geq f_{km}$, donde f_{kr} ($r = 1, 2, \ldots,$ m) es la frecuencia del término t_k en la r-ésima clase luego de ser ordenada. Si la distribución ínter-clase de un término es uniforme, entonces el centro de gravedad va a estar ubicado en el centro de la lista ordenada, y su poder de discriminación de clase va a ser mínimo. Por otro lado, si el término ocurre en una sola clase, el centro de gravedad va a estar al comienzo de la lista ordenada ($r = 1$). Entonces, la posición del centro de gravedad es usada para definir la métrica IGM como

sigue:

$$\text{IGM}(t_k) = \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \times r},$$

donde r es la posición en la lista ordenada, f_{kr} es la frecuencia del término t_k en la r-ésima clase luego de ser ordenada, y m es el número de clases. Notar que f_{k1} es la frecuencia del término t_k en la clase en la que tiene la frecuencia más alta, o sea la primera posición de la lista ordenada ($r = 1$).

Las métricas TGF* y IGM se pueden combinar para formar el esquema TGF*-IGM como sigue:

$$\text{TGF*-IGM}(t_k) = \text{TGF*}(t_k) \times (1 + \lambda \times \text{IGM}(t_k)) .$$

Se utiliza $\lambda = 7$ para los experimentos realizados con esta técnica ya que es el valor por defecto utilizado en [CZLZ16].

En [DU19a] los autores proponen una modificación a la técnica IGM con el objetivo de mejorar sus capacidades para definir los pesos de los términos, en específico para algunos casos que los autores definen como casos extremos. Los autores plantean varios escenarios donde se asigna el mismo valor de IGM para diferentes términos que aparecen en diferente cantidad de documentos pero que aparecen en la misma cantidad de clases. Por ejemplo, dos términos que aparecen en una sola clase tendrán el mismo IGM (i.e., $f_{k1}/(f_{k1} \times 1) = 1$) independientemente del valor de f_{k1}. Sin embargo, uno de esos términos puede aparecer 100 veces en la clase, mientras que el otro puede aparecer solo una vez. A la luz de estos escenarios, los autores incorporan el siguiente componente al cociente de IGM: $D_{total}(t_{k\_max})$. Este valor representa la cantidad de documentos de la clase A, donde la clase A es la clase donde el término t_k aparece más veces. Usando este componente adicional, los autores definen la fórmula de IGM mejorada como sigue:

$$\text{IGM}_{imp}(t_k) = \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \times r + \log_{10}\left(\frac{D_{total}(t_{k\_max})}{f_{k1}}\right)}.$$

A partir de esta métrica los autores combinan IGM_{imp} con la frecuencia del término y la raíz cuadrada de la frecuencia del término, dando lugar a las siguientes formulaciones:

$$\text{TGF*-IGM}_{imp}(t_k) = \text{TGF} * (t_k) \times (1 + \lambda \times \text{IGM}_{imp}(t_k)) ,$$

$$\text{SQRT-TGF*-IGM}_{imp}(t_k) = \sqrt{\text{TGF*}(t_k)} \times (1 + \lambda \times \text{IGM}_{imp}(t_k)) .$$

En este trabajo replicamos estas dos técnicas usando $\lambda = 7$ como se sugiere en [DU19a].

Técnica	Formulación
TGF	$A + C$
IDF	$\log\left(N/(A + C)\right)$
TGF*	A
TGF*-IDFEC	$A \times (\log((C + D)/\max(C, 1)))$
χ^2	$N((AD - BC)^2/((A + C)(B + D)(A + B)(C + D)))$
OR	$\log((\max(A, 1) \times D)/\max(B \times C, 1))$
IG	$(A/N)\log(\max(A, 1)/(A + C))-$ $((A + B)/N)\log((A + B)/N) + (B/N)\log(B/(B + D))$
GR	$IG/(-((A + B)/N)\log((A + B)/N) - ((C + D)/N)\log((C + D)/N))$
GSS	$\log(2 + ((A + C + D)/(\max(C, 1))))$
Prob	$\log(1 + (A/B)(A/C))$
RF	$\log(2 + (A/\max(C, 1))$
IDFEC	$\log((C + D)/\max(C, 1))$
TGF-IDFEC	$(A + C)(\log((C + D)/\max(C, 1)))$
MI	$\log((N \times \max(A, 1))/((A + B)(A + C)))$
IDFEC_B	$\log(2 + (A + C + D)/(\max(C, 1)))$
TGF*-IGM	$\text{TGF*}(t_k) \times (1 + \lambda \times \text{IGM}(t_k))$
TGF*-IGM$_{imp}$	$\text{TGF*}(t_k) \times (1 + \lambda \times \text{IGM}_{imp}(t_k))$
SQRT-TGF*-IGM$_{imp}$	$\sqrt{\text{TGF*}(t_k)} \times (1 + \lambda \times \text{IGM}_{imp}(t_k))$

Tabla 2.1: Definición de las técnicas de pesaje de términos.

En la Tabla 2.1 se muestran las definiciones de las técnicas del estado del arte utilizadas en este capítulo. La siguiente notación, adaptada de [LSLT05, DMPS15] es usada siempre que es posible:

- A denota el número de documentos que pertenecen a la clase c_k y contienen el término t_i.
- B denota el número de documentos que pertenecen a la clase c_k pero no contienen el término t_i.
- C denota el número de documentos que no pertenecen a la clase c_k pero contienen el término t_i.
- D denota el número de documentos que no pertenecen a la clase c_k y no contienen el término t_i.
- N denota el número total de documentos de la colección($N = A + B + C + D$).

Notar que algunas formulaciones incluyen la expresión $\max(X, 1)$ para evitar el problema de valores indefinidos, como cuando se divide por cero o se trata de calcular el valor $\log(0)$.

La técnica FDD$_\beta$ analizada en este capítulo es similar a algunas de las técnicas del estado del arte mencionadas, en el sentido que adopta un enfoque supervisado y considera

tanto el poder descriptivo como el poder discriminativo de un término para una clase o tópico. Sin embargo, se diferencia de otros esquemas en la incorporación de un parámetro ajustable que permite favorecer a la precisión, a la cobertura u obtener un balance entre ambas. También se basa en una formulación más sencilla e intuitiva, derivada de los conceptos básicos de precisión y cobertura del área de recuperación de información.

2.4. La Técnica de Pesaje FDD$_\beta$

La técnica FDD$_\beta$, presentada y evaluada en este capítulo, es una técnica de pesaje de términos que se basa en dos principios: (1) La etiqueta de clase o tópico aporta información útil para obtener el peso del término, y (2) la importancia de un término depende del objetivo específico que se tenga para la tarea (por ejemplo, obtener alta cobertura o alta precisión). El resultado es un método supervisado basado en un parámetro que distingue dos factores de relevancia. El primer factor, el cual mide la *relevancia descriptiva* (DESCR) es local a la clase y representa la importancia del término para describir la clase. Dado un término t_i y una clase c_k el factor DESCR se define como sigue:

$$\text{DESCR}(t_i, c_k) = \frac{|d_j : t_i \in d_j \wedge d_j \in c_k|}{|d_j : d_j \in c_k|},$$

el cual es equivalente a $A/(A+B)$, usando la notación presentada previamente, y representa el hecho de que aquellos términos que ocurren en muchos documentos de una dada clase son buenos descriptores de esa clase.

El segundo factor, el cual mide la *relevancia discriminativa* (DISCR), es global a la colección y es computado para un término t_i y una clase c_k como sigue:

$$\text{DISCR}(t_i, c_k) = \frac{|d_j : t_i \in d_j \wedge d_j \in c_k|}{|d_j : t_i \in d_j|},$$

lo cual es equivalente a $A/(A+C)$ y representa el hecho de que términos que tienden a ocurrir solo en documentos de esa clase son buenos discriminadores de dicha clase.

La técnica FDD$_\beta$ combina los factores DESCR y DISCR como sigue

$$\text{FDD}_\beta(t_i, c_k) = (1+\beta^2)\frac{\text{DISCR}(t_i, c_k) \times \text{DESCR}(t_i, c_k)}{(\beta^2 \times \text{DISCR}(t_i, c_k)) + \text{DESCR}(t_i, c_k)},$$

resultando en la siguiente formulación, de acuerdo a la notación definida previamente:

$$\text{FDD}_\beta(t_i, c_k) = (1+\beta^2)\frac{A/(A+C) \times A/(A+B)}{(\beta^2 \times A/(A+C)) + A/(A+B)}.$$

El parámetro ajustable β permite favorecer diferentes objetivos de la tarea de recuperación de información. Usando $\beta > 1$ podemos ponderar el poder descriptivo por encima del poder discriminativo, mientras que $beta < 1$ pondera el poder discriminativo por encima del descriptivo. La técnica FDD$_\beta$ es derivada de la fórmula tradicional F$_\beta$, usada en recuperación de información, que se define como sigue:

$$F_\beta = (1 + \beta^2)\frac{\text{precisión} \times \text{cobertura}}{(\beta^2\text{precisión}) + \text{cobertura}}$$

Esta fórmula mide el desempeño para la tarea de recuperación de información con respecto a un usuario que le asigna más o menos importancia a la precisión que a la cobertura basándose en el valor de β [Rij79]. Con valores de β cercanos a cero, el F_β tiende a priorizar precisión, y con valores de β cercanos a infinito, F_β tiende a priorizar cobertura.

La búsqueda temática de información puede ser formulada como un problema de aprendizaje supervisado donde el conjunto de entrenamiento es definido como un conjunto de documentos etiquetados como relevantes o irrelevantes para el tópico de interés. La colección de entrenamiento se puede usar para computar el valor de FDD$_\beta$ para cada término de la colección de acuerdo a la definición de la técnica dada previamente. La aplicación de la técnica FDD$_\beta$ no necesita estar limitada a palabras aisladas, puede ser también aplicada a n-gramas, conceptos o términos más complejos.

La técnica FDD$_\beta$ ofrece la posibilidad de identificar términos útiles partiendo de una colección de entrenamiento y ordenar esos términos de acuerdo a la capacidad de los mismos de reflejar las necesidades del usuario. Los términos aprendidos a partir de la colección de entrenamiento pueden ser usados posteriormente para la tarea de recuperación de información temática, esto es, para construir consultas con el objetivo de recuperar más material de colecciones no etiquetadas, por ejemplo, de la web, de Twitter u otras fuentes. Intuitivamente, si un usuario está buscando recursos de un tópico específico entonces se puede favorecer la precisión a través de la construcción de consultas con términos que posean FDD$_\beta$ alto usando valores de β pequeños. Este enfoque priorizará el poder discriminativo de los términos por encima de su valor descriptivo. Alternativamente, si el usuario busca tantos recursos relevantes como sea posible, entonces se puede favorecer la cobertura eligiendo términos que tengan FDD$_\beta$ alto utilizando valores de β altos. En este último caso, buenos descriptores van a ser priorizados por encima de buenos discriminadores. El análisis y la evaluación presentados en la siguiente sección proveen evidencia empírica mostrando que esta intuición detrás de esta idea es correcta.

2.5. Conjuntos de Datos

Durante el desarrollo del presente capítulo se realizan diferentes experimentos usando tres conjuntos de datos. El primer conjunto de datos usado es un conjunto de datos manualmente etiquetado específicamente para las tareas desarrolladas durante este capítulo, denominado *Noticias Relevantes para la Economía de The Guardian* (*ERNTG* por sus siglas en inglés). Este conjunto de datos fue etiquetado por expertos en economía y consiste de los textos completos de artículos periodísticos recolectados de las Secciones *Política*, *Noticias del Mundo*, *Negocios* y *Sociedad* del portal de noticias *The Guardian* (`https://www.theguardian.com/`). Se recolectaron el **total** de noticias correspondientes a esas cuatro secciones para el mes de enero del 2013 utilizando la API provista por el portal de noticias. Un total de 1.689 noticias fueron recolectadas para esas cuatro secciones durante ese periodo. Aunque las noticias tienen múltiples metadatos, solo el texto completo de la noticia y el título fue usado para este trabajo. Los textos fueron preprocesados usando la librería *spaCy* para procesamiento de lenguaje natural (NLP por sus siglas en inglés) versión 2.1.4. Cada texto fue separado en palabras (*tokens*) usando la herramienta de *spaCy* para dicha tarea y fueron posteriormente lematizados con la misma librería. Términos con guiones fueron considerados como dos términos separados. Finalmente, el vocabulario fue construido utilizando los *tokens* antes mencionadas, excluyendo términos irrelevantes como *stopwords*, términos que son muy infrecuentes en la colección (aquellos términos que ocurren en menos de 15 artículos) y términos con caracteres no alfabéticos. Para construir el conjunto de entrenamiento, dos expertos en economía leyeron la colección de 1.689 noticias y llegaron a un consenso en si la noticia era relevante para el dominio económico. Debido a que el proceso fue desarrollado siguiendo un enfoque de consenso, no se reportan puntajes de ínter-concordancia entre los anotadores. Como resultado del proceso de anotación, 537 noticias fueron etiquetadas como relevantes y 1.152 como irrelevantes. Vale la pena mencionar que el proceso de etiquetado fue necesario porque ninguna sección del portal de noticias coincidía con la definición de relevante para la economía que se necesitaba para este trabajo. Por ejemplo, la sección de *Negocios* tenía 418 noticias relevantes para el tópico considerado (de 512). Mientras que la sección *Política* tenía 39 relevantes (de 290), la sección *Noticias del Mundo* tenía 43 (de 650) y la sección *Sociedad* 37 relevantes (de 237). Otros 100 artículos periodísticos fueron etiquetados de la misma forma por los expertos para construir el conjunto de datos de prueba (test). Estos 100 artículos periodísticos fueron tomados aleatoriamente del periodo desde febrero

de 2013 hasta diciembre del 2015. El conjunto de datos etiquetado (los 1.789 artículos y sus etiquetas) fueron publicados como parte de los trabajos de esta tesis para facilitar la reproducibilidad del trabajo y permitir trabajos futuros [MDTM19a].[1]

El segundo conjunto de datos es uno ampliamente usado por la comunidad de NLP para tareas de clasificación y agrupación (*clustering*), el conjunto *20 Newsgroups*, también conocido como *NG20* (`http://qwone.com/~jason/20Newsgroups/`). Este conjunto de datos consiste de aproximadamente 20.000 documentos pertenecientes a 20 grupos de difusión distintos (*newsgroups*), cada uno de ellos con su propio tópico de discusión. Algunos de los grupos de difusión tiene tópicos similares entre sí (por ejemplo, comp.sys.ibm.pc.hardware y comp.sys.mac.hardware), mientras que otros son muy disimilares (misc.forsale y soc.religion.christian). De la misma forma que para el conjunto de datos *ERNTG*, el vocabulario fue preprocesado usando la librería de NLP de *spaCy*. El conjunto de datos fue separado en 80 %-20 % para entrenamiento y test, respectivamente.

El tercer conjunto de datos usado es otro conjunto de datos clásico de la literatura de clasificación de textos, la colección de textos *Reuters-21578*, también conocido como *Reuters* (`https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection`). Esta colección consiste de 21.578 reportes periodísticos recolectados de *Reuters newswire* en 1987. Los documentos fueron recolectados e indexados en categorías por personal de *Reuters Ltd.* Cada documento puede tener asociado cero o más categorías (tópicos), de un total de 120 tópicos posibles. No todos los reportes de la colección tienen texto asociados. Por ende, la colección final utilizada para este trabajo consistió en los textos de los 19.043 de las observaciones que sí tenían texto asociado, junto con sus correspondientes tópicos (cero o más de los 120 posibles). Algunos ejemplos de los tópicos de la colección son: *housing*, *livestock* y *jobs*. De la misma manera que para los dos conjuntos de datos anteriores, se realizó el preprocesado usando la librería de NLP de *spaCy*. El conjunto de datos fue separado en 80 %-20 % para entrenamiento y testeo, respectivamente.

2.6. El Rol del Parámetro β en la Técnica FDD$_\beta$

El comportamiento de la técnica FDD$_\beta$ en función de su parámetro β y su efectividad como técnica para elegir términos para generación de consultas para recuperación de

[1] http://dx.doi.org/10.17632/yt8j2f3hpp.

información fueron analizados sobre dos conjuntos de datos, *ERNTG* y *20NG*.

2.6.1. Análisis en el Conjunto de Datos ERNTG

Para analizar el comportamiento del parámetro β se analizó el desempeño para la recuperación de información de consultas generadas usando el término con FDD_β más alto para distintos valores de β dentro de un rango. Las consultas evaluadas fueron generadas usando los términos con el valor de FDD_β más alto en el conjunto de entrenamiento de *ERNTG* para distintos valores de beta, los valores de β se acotaron al rango 0 a 10. Se usaron todos los valores de β en el rango [0, 2] avanzando de a pasos de $+0,01$, y para el rango de valores de β de (2, 10] se usó un paso de $+0,1$ entre los sucesivos valores de β. Para este trabajo, se utilizó un vocabulario que contiene unigramas, bigramas y trigramas, y, por ende, los términos elegidos para las consultas pueden ser cualquiera de estos tres. Usando las consultas generadas, se computaron las métricas clásicas de precisión, cobertura y F_1 tanto en el conjunto de entrenamiento como en el de test. Esto es, utilizando la misma consulta (generada en el conjunto de entrenamiento), se realizan (y evalúan) dos búsquedas de recuperación de información, una en cada partición de datos (entrenamiento y test). En la Tabla 2.2 se ilustra el rol del parámetro β mostrando el desempeño de la técnica para distintos rangos de valores de β. Cada rango de valores de β representa un intervalo durante el cual el término con el mayor FDD_β siempre fue el mismo. Por ejemplo, el bigrama (‘`Royal`’, ‘`Bank`’) obtuvo el mejor valor de FDD_β en el conjunto de entrenamiento para $\beta \in [0,07;\ 0,14)$[2]. Esta consulta alcanzó una cobertura de $0,09$ en el conjunto de entrenamiento y $0,07$ en el de test. Obtuvo un valor de precisión de $0,97$ en el conjunto de entrenamiento y de $1,0$ en el de test. Finalmente, este mismo bigrama, obtuvo un F_1 de $0,16$ en el conjunto de entrenamiento y $0,13$ en el de test.

Para visualizar la dinámica de la precisión, cobertura y F_1 para diferentes valores de β, en la Figura 2.1 se muestran los valores de desempeño obtenidos con las consultas que maximizan el FDD_β aprendidas del conjunto de entrenamiento para valores de β variando de 0 a 10. El análisis de estos resultados muestra que el parámetro β tiene un efecto importante en el desempeño de la consulta para la tarea de recuperación de información. Como se esperaba, se puede observar que valores bajos de β favorecen la precisión y valores altos de β favorecen la cobertura. El valor óptimo de β para obtener un determinado desempeño

[2]Para la presente tesis se utiliza siempre que es posible la coma como separador de elementos, si los elementos tienen coma decimal entonces se utiliza el punto y coma.

Término	Rango de β	Cobertura	Precisión	F_1
(‘Capital’ ’Economics’)	$[0,00;\ 0,01)$	$0,04/0,03$	$1,00/1,00$	$0,08/0,06$
(‘bank’, ‘say’)	$[0,01,\ 0,04)$	$0,05/0,04$	$1,00/1,00$	$0,10/0,07$
(‘Federal’, ‘Reserve’)	$[0,04;\ 0,07)$	$0,05/0,07$	$1,00/1,00$	$0,10/0,13$
(‘Royal’, ‘Bank’)	$[0,07;\ 0,14)$	$0,09/0,07$	$0,97/1,00$	$0,16/0,13$
(‘economist’)	$[0,14;\ 0,24)$	$0,17/0,19$	$0,88/0,83$	$0,29/0,31$
(‘investor’)	$[0,24;\ 0,36)$	$0,26/0,25$	$0,81/0,74$	$0,39/0,38$
(‘growth’)	$[0,36;\ 0,49)$	$0,39/0,41$	$0,71/0,70$	$0,51/0,52$
(‘market’)	$[0,49;\ 1,25)$	$0,50/0,43$	$0,65/0,59$	$0,57/0,50$
(‘year’)	$[1,25;\ 1,34)$	$0,84/0,89$	$0,36/0,36$	$0,51/0,51$
(‘have’)	$[1,34;\ 0,0]$	$0,99/1,00$	$0,32/0,31$	$0,49/0,47$

Tabla 2.2: Términos que maximizan el valor de FDD_β en el conjunto de entrenamiento para distintos rangos de valores de β y su desempeño como consultas en el conjunto de entrenamiento y test (conjunto de datos *ERNTG*).

se puede aprender del conjunto de datos de entrenamiento. El análisis presentado en la Figura 2.1 muestra el desempeño de la consulta que maximiza el FDD_β en el conjunto de entrenamiento usada como consulta en el conjunto de test. Es posible verificar que el comportamiento observado en el entrenamiento y el test son similares, indicando que la selección de términos obtenidos con FDD_β no sufre de sobreajuste (*overfitting*) para este escenario analizado.

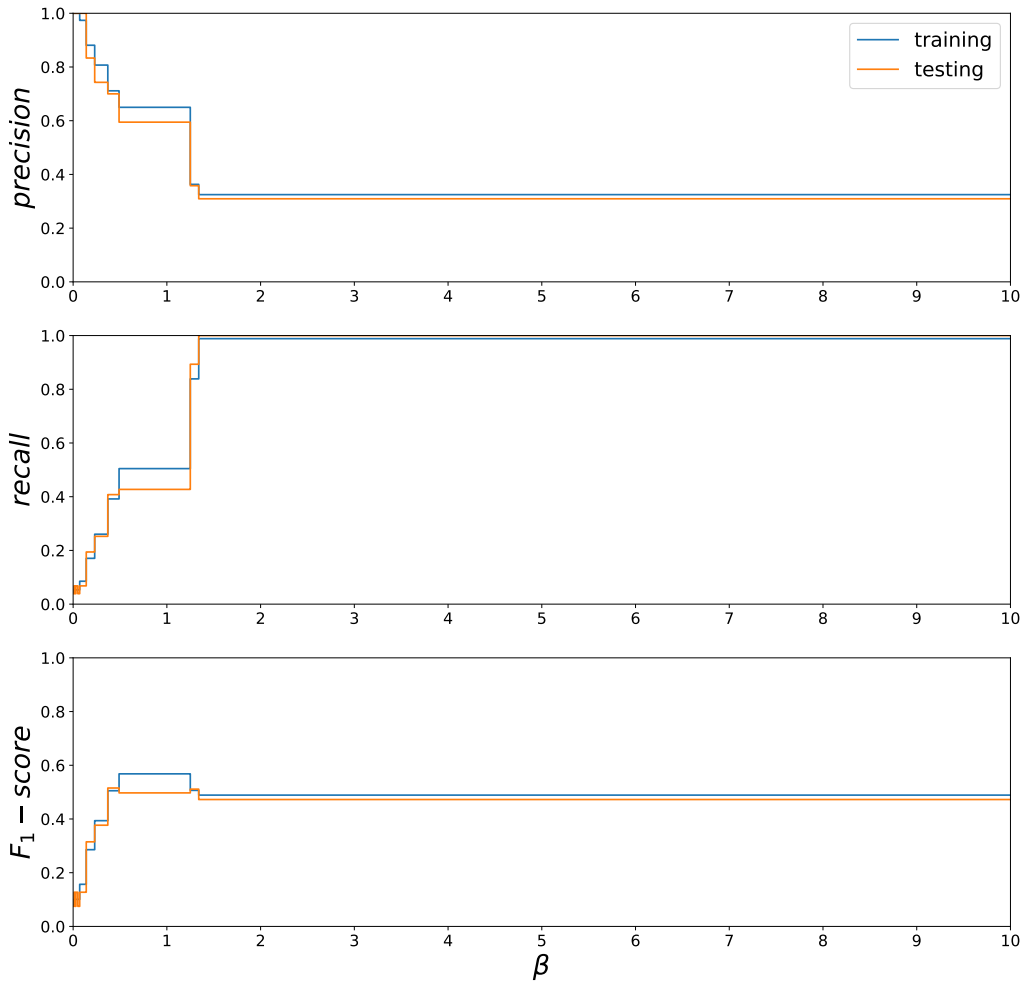


Figura 2.1: Desempeño de las consultas seleccionadas usando los términos que maximizan el valor de FDD$_\beta$ en el conjunto de entrenamiento usando diferentes valores de β. Las consultas son utilizadas para la tarea de recuperación de información tanto en el conjunto de entrenamiento como en el conjunto de test (conjunto de datos *ERNTG*).

2.6.2. Análisis en el Conjunto de Datos 20NG

Un análisis similar pero más extenso fue realizado para el conjunto de datos *20NG*. Por cada una de las 20 categorías de la colección, se seleccionaron los términos (unigramas, bigramas o trigramas) que maximizan el valor obtenido por la técnica FDD$_\beta$ para diferentes

valores de β. Se evaluó el desempeño de usar los términos seleccionados como consultas sobre el conjunto de entrenamiento y el de test. Una visualización de la dinámica de la precisión, cobertura y F_1 para diferentes valores de β en las particiones de entrenamiento y test es presentada en las Figuras 2.2 y 2.3, respectivamente. En estas Figuras, cada sombra representa el desempeño en una categoría. Los desempeños promediados a lo largo de las 20 categorías se muestran en la Figura 2.4. Como se había observado previamente, se puede ver que valores bajos de β favorecen precisión mientras que valores altos favorecen la cobertura. Finalmente, en la Figura 2.5 se muestran los valores de FDD_β en los conjuntos de entrenamiento y test para los mejores términos elegidos del entrenamiento usando la técnica FDD_β para diferentes valores de β. Los desempeños similares obtenidos entre el conjunto de entrenamiento y test indican, al igual que para los experimentos anteriores, que la estrategia de selección de términos basada en FDD_β no tiene problemas de sobreajuste (overfitting).

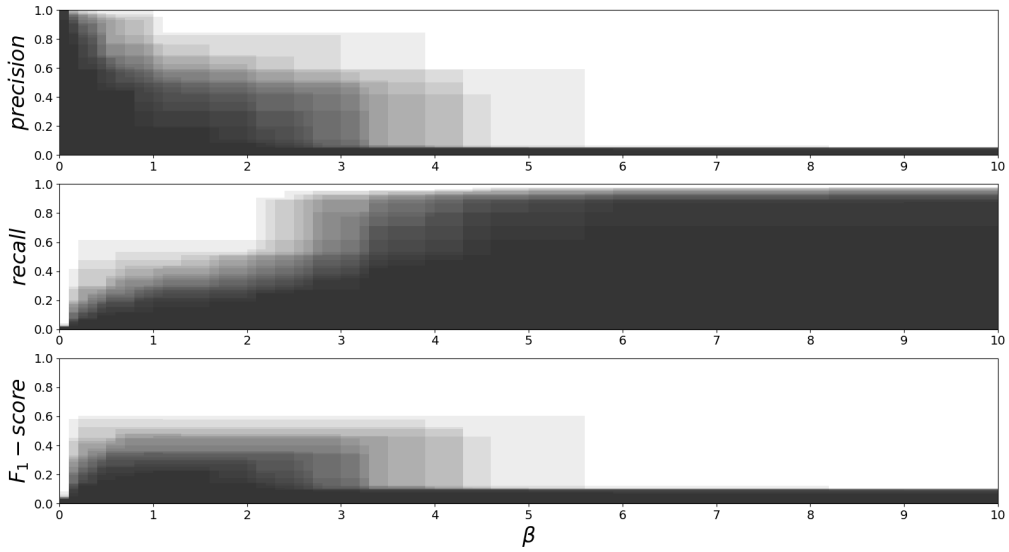


Figura 2.2: Desempeño en el conjunto de entrenamiento del mejor término de consulta seleccionado a partir del conjunto de entrenamiento usando la técnica FDD_β con distintos valores de β. Cada sombra representa una categoría diferente (conjunto de datos *20NG*).

Para profundizar en el análisis del parámetro β, se analizó qué términos (unigramas, bigramas o trigramas) son los que obtienen el mayor F_1 en el conjunto de entrenamiento para cada una de las 20 categorías del conjunto de datos *20NG*. En la Tabla 2.3 se reportan los resultados de estos experimentos, presentando para cada categoría el rango

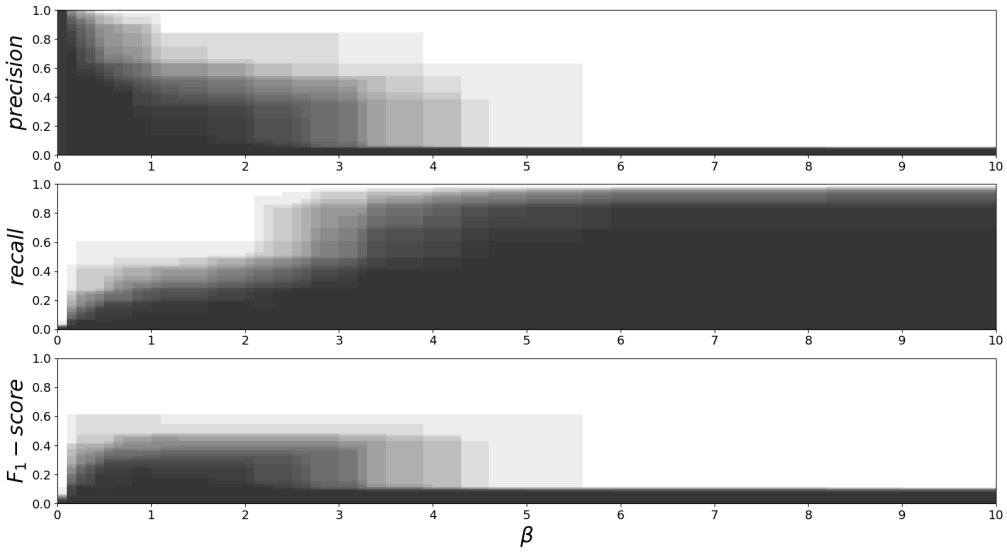


Figura 2.3: Desempeño en el conjunto de test del mejor término de consulta seleccionado a partir del conjunto de entrenamiento usando la técnica FDD$_\beta$ con distintos valores de β. Cada sombra representa una categoría diferente (conjunto de datos *20NG*).

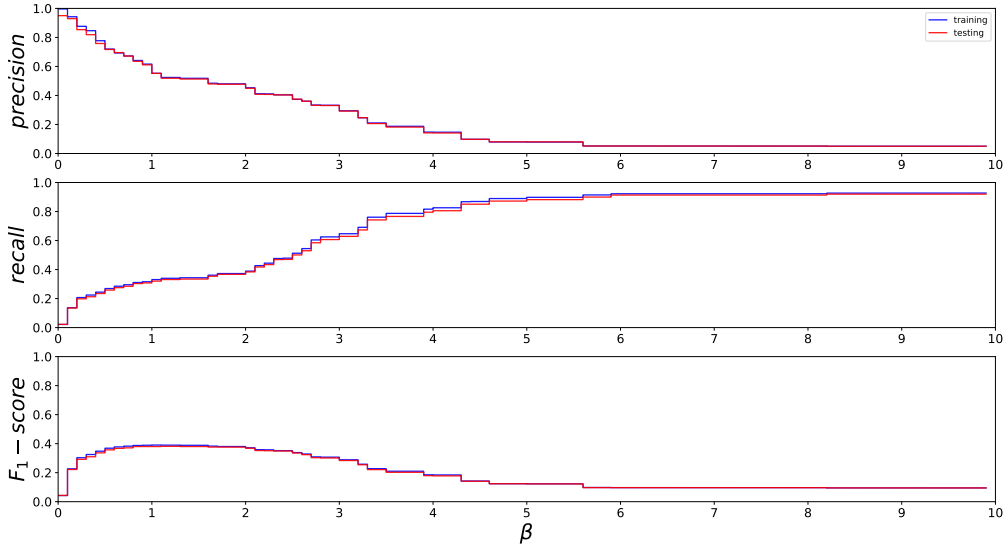


Figura 2.4: Desempeño en el conjunto de entrenamiento y test del mejor término de consulta seleccionado a partir del conjunto de entrenamiento usando la técnica FDD$_\beta$ con distintos valores de β promediado a lo largo de las 20 categorías (Conjunto de datos *20NG*).

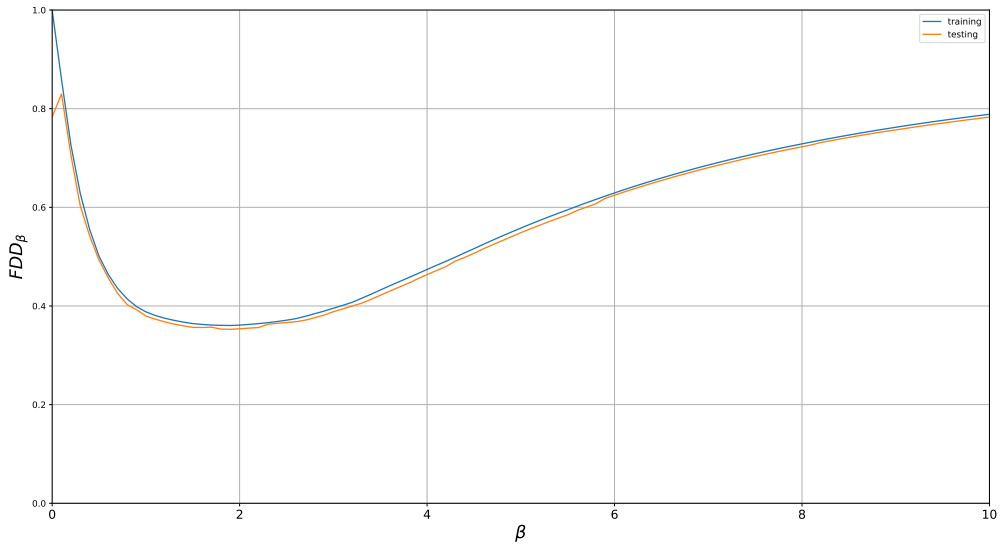


Figura 2.5: El valor de FDD_β obtenido en el conjunto de entrenamiento y test usando como consulta el término que maximiza el FDD_β en el conjunto de entrenamiento para distintos valore de β (conjunto de datos *20NG*).

de β que maximiza el F_1, el término que maximiza dicha métrica y el F_1 obtenido por ese término en los conjuntos de entrenamiento y test. En el Apéndice A se dejan disponibles tablas adicionales con los términos seleccionados usando el mejor FDD_β para diferentes rangos de β y su desempeño en términos de precisión, cobertura y F_1 en el conjunto de entrenamiento y test.

2.7. Validación con Estudio de Usuarios

Ocho voluntarios participaron de un experimento realizado de manera *online* a modo de validación. El grupo incluía cuatro voluntarios con PhD. en economía y cuatro voluntarios con formación universitaria pero sin formación, de grado ni de posgrado, en economía. Nos referimos al primer grupo como *expertos* y al segundo grupo como *no-expertos*. La motivación para examinar y comparar los desempeños de ambos grupos, expertos y no-expertos, es determinar si ambos grupos exhiben comportamientos distintos y evaluar la discrepancia entre la técnica y ambos grupos. Un conjunto de 50 términos (10 listas de 5 términos cada una) y otro conjunto de 100 términos (20 listas de 5 térmi-

Categoría	Rango de β	Término consulta	F_1 (entrenamiento/test)
alt.atheism	$[0,3;\ 1,9)$	(‘atheist’)	$0,32$ / $0,29$
comp.graphics	$[0,8;\ 2,0)$	(‘image’)	$0,27$ / $0,25$
comp.os.ms-windows.misc	$[0,2;\ 4,2)$	(‘Windows’)	$0,53$ / $0,47$
comp.sys.ibm.pc.hardware	$[0,8;\ 2,5)$	(‘card’)	$0,30$ / $0,33$
comp.sys.mac.hardware	$[0,3;\ 2,4)$	(‘Apple’)	$0,36$ / $0,37$
comp.windows.x	$[0,5;\ 3,2)$	(‘X’)	$0,40$ / $0,43$
misc.forsale	$[0,4;\ 3,1)$	(‘sale’)	$0,35$ / $0,37$
rec.autos	$[0,2;\ 5,5)$	(‘car’)	$0,60$ / $0,62$
rec.motorcycles	$[0,1;\ 1,0)$	(‘bike’)	$0,58$ / $0,61$
rec.sport.baseball	$[0,5;\ 1,0)$	(‘pitch’)	$0,35$ / $0,31$
rec.sport.hockey	$[0,7;\ 1,2)$	(‘team’)	$0,48$ / $0,46$
sci.crypt	$[1,0;\ 3,4)$	(‘key’)	$0,46$ / $0,48$
sci.electronics	$[0,1;\ 1,5)$	(‘circuit’)	$0,25$ / $0,20$
sci.med	$[0,4;\ 2,0)$	(‘doctor’)	$0,30$ / $0,31$
sci.space	$[0,9;\ 2,6)$	(‘space’)	$0,36$ / $0,34$
soc.religion.christian	$[0,6;\ 4,2)$	(‘God’)	$0,51$ / $0,46$
talk.politics.guns	$[0,5;\ 3,1)$	(‘gun’)	$0,45$ / $0,38$
talk.politics.mideast	$[0,6;\ 2,9)$	(‘Israel’)	$0,48$ / $0,48$
talk.politics.misc	$[1,0;\ 1,6)$	(‘government’)	$0,23$ / $0,17$
talk.religion.misc	$[0,8;\ 2,6)$	(‘Christian’)	$0,23$ / $0,27$

Tabla 2.3: Mejor intervalo de valores de β que resultan en mayor F_1 en el conjunto de entrenamiento para cada categoría, el término resultante es usado como consulta para medir el F_1 tanto en el conjunto de entrenamiento como en el de test (conjunto de datos *ERNTG*).

nos cada una) fueron elegidos estratégicamente del total de términos del conjunto de datos *ERNTG* (+10.000 términos distintos). La selección fue basada en la distribución de frecuencias basándose en la Ley de Zipf. Esta selección estratégica tenía por objetivo evitar que términos poco frecuentes (que son la mayoría) tengan más posibilidades de ser seleccionados que términos con alta frecuencia (que son unos pocos). Para completar la etapa inicial de ajuste del parámetro β, dos expertos anotaron por consenso la relevancia económica de cada término del conjunto de datos de 50 palabras. Los expertos asignaron puntajes a cada uno de estos términos usando un rango del 0 (irrelevante para el dominio económico) a 5 (muy relevante para el dominio económico). Se utilizaron estos puntajes para aprender el valor del parámetro de β que maximice la correlación de Pearson entre la estimación asignada por la técnica FDD_β utilizando ese β y los valores de relevancia de los usuarios. Como se puede observar en la Figura 2.6, el valor más alto de correlación de Pearson entre los valores de relevancia de los usuarios y la técnica FDD_β fue de $0,798$, el cual fue obtenido usando $\beta = 0,477$.

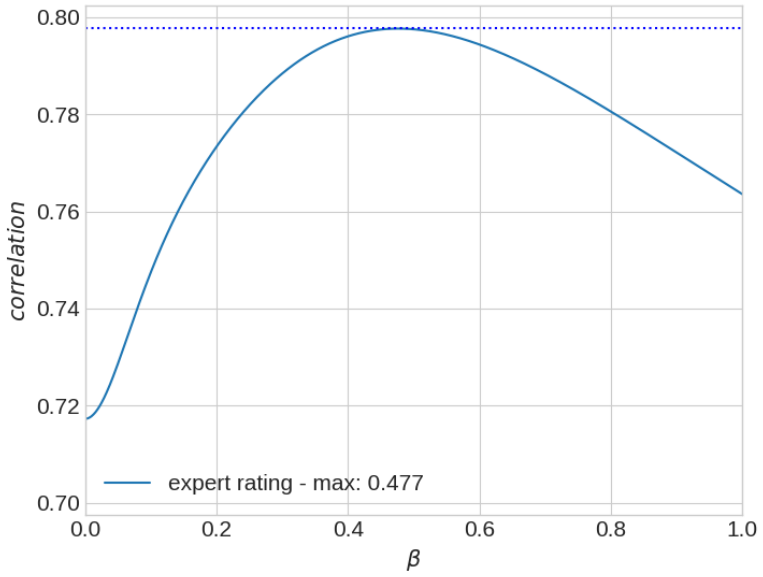


Figura 2.6: Aprendiendo el β óptimo a partir del conjunto de 50 términos etiquetado por dos usuarios expertos. Visualización de la correlación de Pearson entre los puntajes asignados por los usuarios y el puntaje calculado por la técnica. Máxima correlación de Pearson igual a 0,798 usando valor de $\beta = 0,477$.

Para completar la validación se utilizó el conjunto de datos etiquetado por los ocho voluntarios que consiste de 100 términos. A cada uno de estos 100 términos los ocho voluntarios les asignaron puntajes de relevancia para el dominio económico usando la escala

de 0 a 5 previamente mencionada. Adicionalmente, para cada término se computaron las métricas de DESCR, DISCR, $FDD_{0,477}$ y las primeras quince de las técnicas definidas en la Tabla 2.1. Primero se evaluó el nivel de acuerdo entre todos los pares de usuarios pertenecientes al grupo *no-expertos* y entre todos los pares de usuarios pertenecientes al grupo *expertos*, este nivel de acuerdo se midió utilizando el promedio de la correlación de Pearson entre cada par. En la Tabla 2.4 se muestran las medias y desvíos estándares que surgen de dicho análisis. Es posible observar que hay un alto nivel de acuerdo en cada grupo y un acuerdo aún mayor en el grupo de expertos.

no-expertos	**expertos**
$\mu = 0,839$; $\sigma = 0,039$	$\mu = 0,876$; $\sigma = 0,009$

Tabla 2.4: Medias (μ) y desvíos estándares (σ) de la correlación de Pearson usada para medir el nivel de acuerdo entre usuarios del grupo *no-expertos* y usuarios del grupo *expertos*.

En la Tabla 2.5 se presenta el resultado de medir el nivel de acuerdo entre los dos grupos (*expertos* y *no-expertos*), y el nivel de acuerdo entre los pesos obtenidos con la técnica $FDD_{0,477}$ y cada uno de los dos grupos (*expertos* y *no-expertos*). El nivel de acuerdo es medido con la correlación de Pearson media (promedio).

no-expertos y expertos	**no-expertos y $FDD_{0,477}$**	**expertos y $FDD_{0,477}$**
$\mu = 0,804$; $\sigma = 0,053$	$\mu = 0,686$; $\sigma = 0,055$	$\mu = 0,752$; $\sigma = 0,019$

Tabla 2.5: Medias (μ) y desvíos estándares(σ) de las correlaciones computadas entre los dos grupos (*expertos* y *no-expertos*), y entre los pesos obtenidos con la técnica $FDD_{0,477}$ y cada uno de los dos grupos (*expertos* y *no-expertos*).

Finalmente, para comparar la efectividad de las técnicas de pesaje de términos como predictores de los puntajes de relevancia asignados por los voluntarios se computó la correlación de Pearson entre los puntajes promedios asignados por los voluntarios y los puntajes asignados por cada una de las quince técnicas de ponderación de términos usadas en este estudio. En la Tabla 2.6 se muestra el resumen de todas estas correlaciones. Los valores reportados se corresponden con la correlación entre cada método y los diferentes grupos de voluntarios. En todos los casos se observa que la técnica $FDD_{0,477}$ supera a las demás, siendo TGF*-IDFEC la segunda técnica más efectiva en estimar la relevancia económica asignada por el usuario.

Técnica	**no-experto (promediado)**	**experto (promediado)**	**no-experto y experto (promediado)**
TGF	$0,284$	$0,365$	$0,332$
IDF	$-0,489$	$-0,564$	$-0,539$
TGF*	$0,574$	$0,643$	$0,623$
MI	$0,697$	$0,660$	$0,695$
χ^2	$-0,165$	$-0,088$	$-0,129$
OR	$0,433$	$0,307$	$0,378$
IG	$0,663$	$0,706$	$0,701$
GR	$0,663$	$0,706$	$0,701$
GSS	$0,723$	$0,757$	$0,758$
Prob	$0,654$	$0,697$	$0,692$
RF	$0,473$	$0,407$	$0,451$
IDFEC	$-0,226$	$-0,326$	$-0,283$
TGF-IDFEC	$0,604$	$0,677$	$0,656$
TGF*-IDFEC	$0,722$	$0,774$	$0,766$
IDFEC_B	$-0,221$	$-0,320$	$-0,277$
DESCR	$0,574$	$0,643$	$0,623$
DISCR	$0,662$	$0,611$	$0,652$
$\text{FDD}_{0,477}$	**0,735**	**0,792**	**0,782**

Tabla 2.6: Correlaciones entre los métodos analizados y los puntajes de relevancia obtenidos promediando los puntajes del grupo *no-expertos*, del grupo *expertos* y de todos los puntajes de los usuarios (*expertos* y *no-expertos*).

2.8. Evaluación del Desempeño para Recuperación de Información

En esta sección se analiza el desempeño de la técnica FDD_β al ser usada como una técnica para seleccionar términos de consulta para la tarea de recuperación temática de información. Se compara el desempeño de la técnica propuesta contra quince técnicas del estado del arte para el conjunto de datos *ERNTG* y un total de dieciocho técnicas del

estado del arte en los conjuntos de datos *NG20* y *Reuters*. Las dieciocho técnicas están descritas en la Tabla 2.1. Para el conjunto de datos *ERNTG* se usaron las primeras quince (omitiendo las técnicas basadas en IGM).

Para los experimentos realizados sobre el conjunto de datos *ERNTG* se obtuvo el término que maximiza el puntaje asignado por cada una de las quince técnicas utilizadas en el conjunto de entrenamiento y se la utilizó como consulta tanto sobre el conjunto de entrenamiento como en el conjunto de test. Se muestra el desempeño para diferentes valores de β para la técnica FDD_β en contraste con los valores obtenidos por las otras quince técnicas del estado del arte usadas (que tienen un valor constante para diferentes valores de β).

Para la evaluación comparativa sobre los conjuntos de datos *20NG* y *Reuters* se utilizó, también, una porción de los datos como entrenamiento y otra como test. El conjunto de entrenamiento se usó para seleccionar los términos (unigramas, bigramas y trigramas) con mejor ponderación para cada categoría y para cada método (los métodos usados son los reportados en la Tabla 2.1). Como estos conjuntos de datos tienen más de una categoría se reporta el desempeño promedio de cada método a lo largo de todas las categorías. Como no es posible reportar el desempeño de la técnica FDD_β para diferentes valores de β (debido a la mayor cantidad de categorías de estos dos conjuntos de datos), para poder ilustrar cómo la técnica FDD_β puede favorecer distintos objetivos de la recuperación de información, se reportan los resultados para FDD_β con (1) $\beta = 0,5$ para favorecer la precisión por encima de la cobertura, (2) con $\beta = 10$ para favorecer en gran medida la cobertura por encima de la precisión, y (3) con $\beta = 1$ para ponderar equitativamente la precisión y la cobertura. Consultas simples consistentes de un solo término (unigrama para *ERNTG* y unigrama, bigrama o trigrama para *NG20* y *Reuters*) fueron generadas usando los términos con mayor ponderación de acuerdo a cada esquema de puntaje de términos. Cada una de esas consultas simples fueron evaluada usando las métricas clásicas de recuperación de información: precisión, cobertura y F_1 sobre el conjunto de entrenamiento y de test.

2.8.1. Evaluación en el Conjunto de Datos *ERNTG*

En esta sección, se analiza el desempeño de la técnica FDD_β como un mecanismo de selección de términos para generar consultas de un único término utilizando el conjunto de datos *ERNTG*. Se compara el desempeño de la técnica propuesta en dicho conjunto

de datos contra otras quince técnicas del estado del arte para ponderación de términos. En particular se la compara con las primeras quince técnicas de la Tabla 2.1.

En primer lugar, se utiliza el conjunto de entrenamientos de la colección *ERNTG* para seleccionar el término con mayor puntaje asignado por la técnica FDD_β utilizando diferentes valores del parámetro β. Las consultas simples, de un único término, son generadas utilizando los términos seleccionados y luego evaluadas en términos de las métricas clásicas: precisión, cobertura y F_1. Los resultados de este análisis se muestran en la Figura 2.7. Como se esperaba, el valor más alto de cobertura obtenido utilizando la técnica FDD_β se obtiene con los valores más altos de β, mientras que los valores más altos de precisión obtenidos con esta técnica se obtienen con los valores de β más pequeños. Se puede notar que términos como (‘`uk`’) ocurren muy seguido en las noticias relevantes probablemente debido a que son noticias recolectadas de un portal británico de noticias. Como resultado de esto, el término (‘`uk`’) resulta en una consulta con alta cobertura. Sin embargo, dicho término, no es un buen discriminador de los artículos relevantes a la economía, por lo que resulta en una baja precisión. Por otro lado, términos como (‘`adp`’), (‘`jp`’), (‘`ubs`’), (‘`forecasts`’) y (‘`ftse`’) no son buenos descriptores pero tienden a aparecer únicamente en noticias relevantes a la economía, esto quiere decir que son buenos discriminadores. Estos términos, al ser usados como términos de consulta, pueden favorecer la precisión, aunque usualmente obteniendo una baja cobertura. Otros términos como (‘`sales`’), (‘`growth`’) y (‘`business`’) tienen un balance de poder descriptivo y discriminativo lo cual resulta en un buen desempeño en términos de F_1.

El término que obtiene mayor F_1 es el término (‘`growth`’), el cual es el término que maximiza el FDD_β para el rango de β entre $0,4$ y $1,2$. Notar que este rango incluye al valor $0,477$ el cual es el valor que obtuvo el mayor valor de correlación entre la técnica FDD_β y los valores de relevancia asignados por los expertos. Basado en este análisis preliminar, se puede observar que la técnica FDD_β obtiene un desempeño igual de bueno en términos de F_1 que las dos mejores técnicas del estado del arte (TGF-IDFEC and TGF*-IDFEC). El término con mejor puntuación de acuerdo a tres de las técnicas de ponderación de términos es (‘`growth`’). Es interesante notar que para valores pequeños de β la técnica FDD_β supera a estos dos métodos en términos de precisión y para valores más grandes de β los supera en términos de cobertura.

El conjunto de test de la colección *ERNTG* fue usado para determinar si la mejor consulta identificada en el conjunto de entrenamiento es efectiva en un conjunto distinto

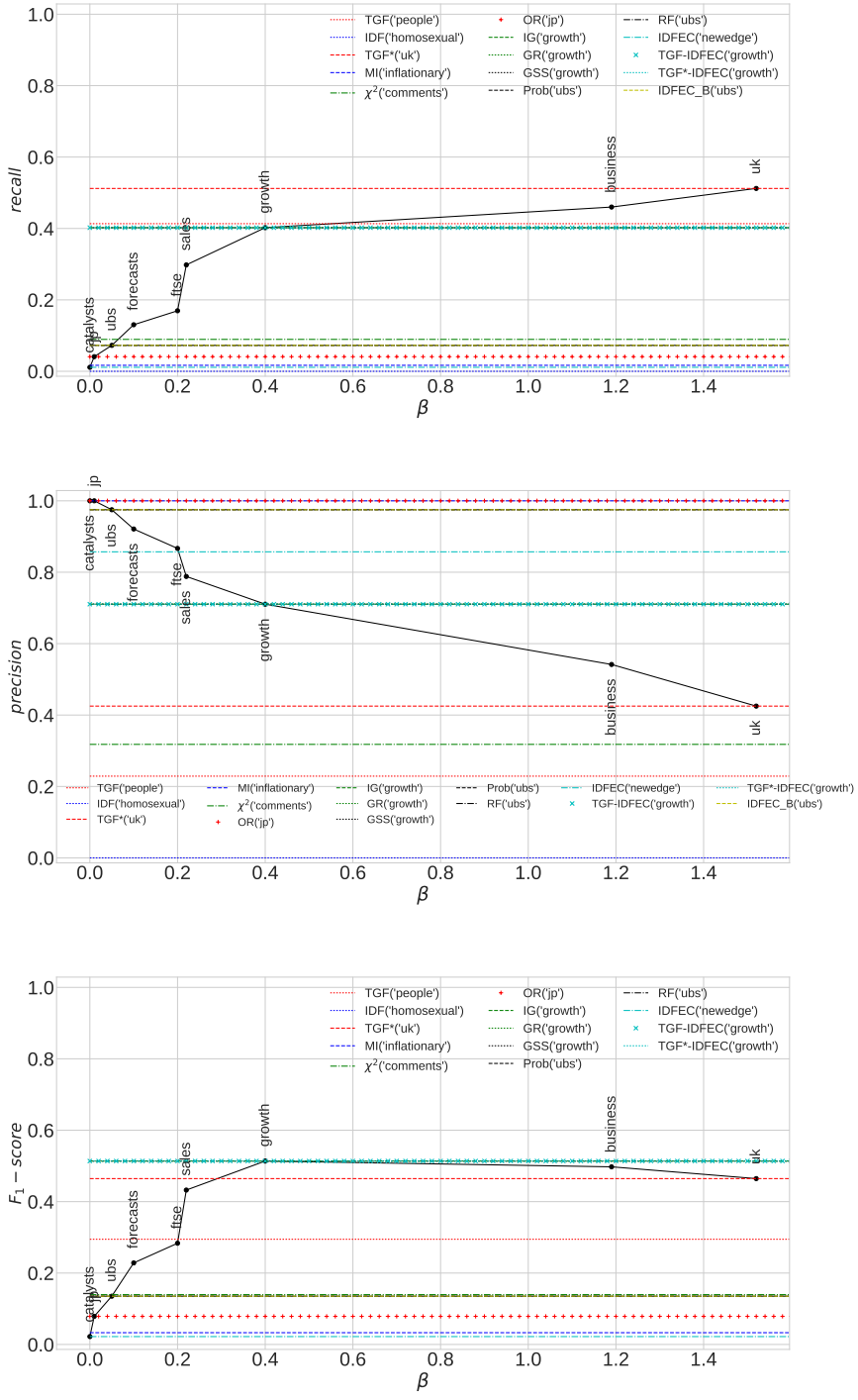


Figura 2.7: Eficacia en la partición de entrenamiento del conjunto *ERNTG* de las consultas generadas basadas en dieciséis técnicas de ponderación de términos distintas (la propuesta y quince del estado del arte). La curva sólida se corresponde con la eficacia de las consultas seleccionadas usando la técnica FDD_β usando diferentes valores de β.

de noticias. El resultado de las métricas de precisión, cobertura y F_1 sobre el conjunto de test se muestra en la Figura 2.8. Debido a que el conjunto de test es pequeño, muchos de los términos con alto poder discriminativo detectados durante el entrenamiento (como (‘`adp`’) y (‘`ubs`’)) no están presentes en el conjunto de test, resultando en una consulta con cero artículos recuperados. Sin embargo, aquellos términos con buen balance entre descripción y discriminación (como (‘`sales`’), (‘`growth`’) y (‘`business`’)) obtuvieron los valores más altos de F_1 cuando fueron usados sobre el conjunto de test. Este análisis preliminar indica que la técnica no sufre de sobreajuste al conjunto de entrenamiento, y que por ende puede ser utilizada para aprender buenos términos de consulta sobre colecciones etiquetadas para luego ser usados sobre colecciones no etiquetadas.

Para indagar más en la eficacia de la técnica FDD_β, se usaron los términos con mayor puntaje asignado por FDD_β usando diferentes valores de β para formular diferentes consultas. Las consultas evaluadas incluyen: consultas simples de un solo término (FDD_β), consultas disyuntivas con dos términos (FDD_β (OR(2))), consultas disyuntivas de tres términos (FDD_β (OR(3))), consultas conjuntivas de dos términos (FDD_β (AND(2))) y consultas conjuntivas de tres términos(FDD_β (AND(3))). Con motivos de comparación, se usó la técnica TGF*-IDFEC, la cual obtuvo uno de los mejores desempeños en el análisis previo y se procedió a formular diferentes consultas construidas a partir de los términos con mejor puntaje para esta técnica. Las consultas evaluadas son: consultas simples de un único término (TGF*-IDFEC), consultas disyuntivas de dos términos (TGF*-IDFEC (OR(2))), consultas disyuntivas de tres términos (TGF*-IDFEC (OR(3))), consultas conjuntivas de dos términos (TGF*-IDFEC (AND(2))) y consultas conjuntivas de tres términos (TGF*-IDFEC (AND(3))).

En la Figura 2.9 se muestran los resultados de eficacia de estas consultas sobre el conjunto de entrenamiento en términos de las métricas precisión, cobertura y F_1. Estos resultados indican que las consultas disyuntivas con tres términos seleccionados con FDD_β (FDD_β (OR(3))) obtienen el mismo F_1 que la mejor consulta basada en TGF*-IDFEC cuando se usan valores de β de un intervalo que incluye al valor $0,477$ (el cual es el valor de β que obtuvo el valor de correlación más alto entre la técnica FDD_β y los valores de relevancia asignados por expertos). Adicionalmente, es interesante notar que para valores altos de β, la consulta “FDD_β (OR(3))” supera a todas las consultas basadas en TGF*-IDFEC en términos de cobertura, mientras que para valores pequeños de β varias combinaciones de las consultas basadas en FDD_β superan a las consultas basadas en

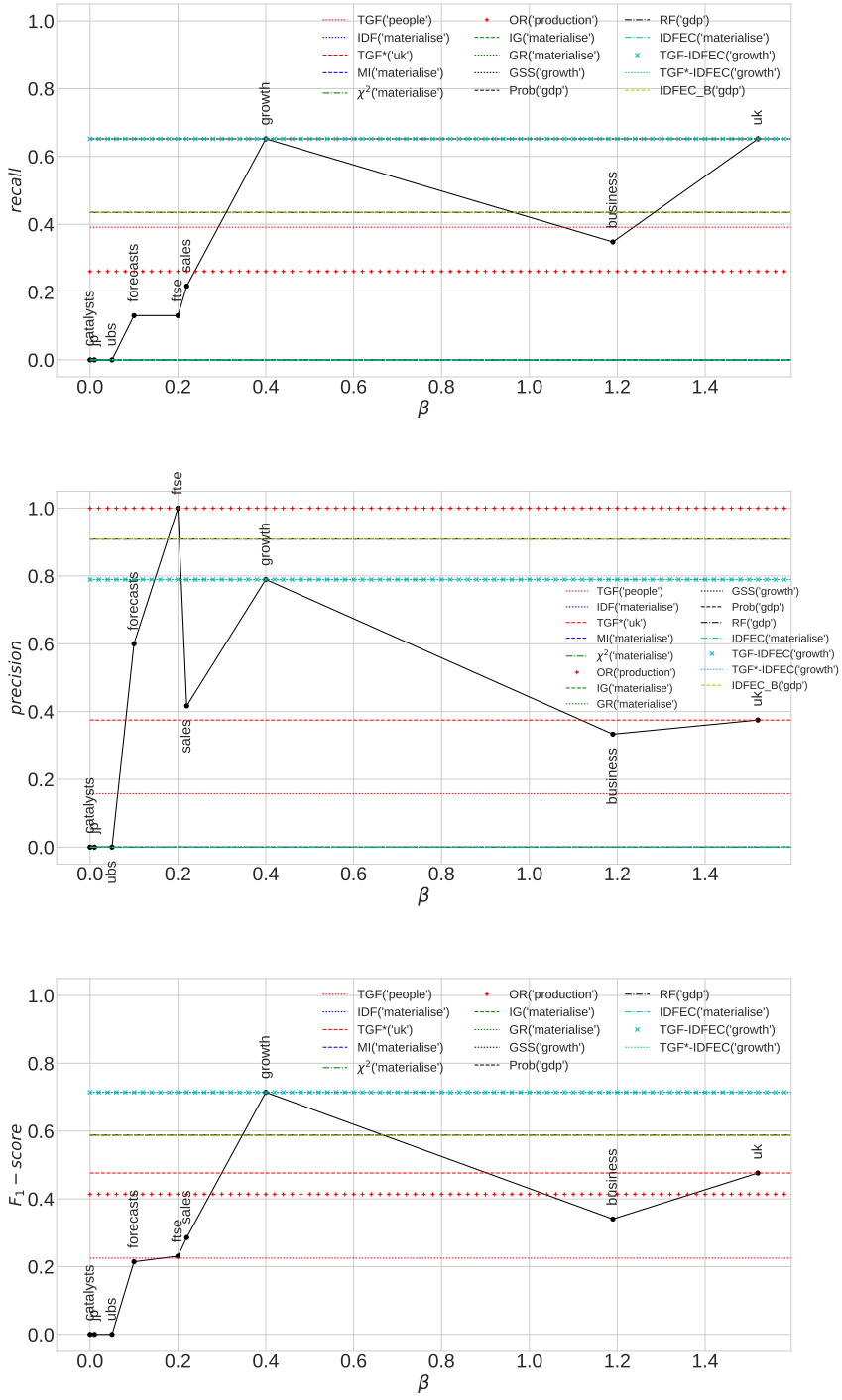


Figura 2.8: Eficacia en la partición de test del conjunto *ERNTG* de las consultas generadas basadas en dieciséis técnicas de ponderación de términos distintas (la propuesta y quince del estado del arte). La curva solida se corresponde con la eficacia de las consultas seleccionadas usando la técnica FDD_β con diferentes valores de β.

TGF*-IDFEC en términos de precisión.

En la Figura 2.10 se muestra el desempeño de las consultas previamente descritas sobre el conjunto de test. Estos resultados muestran que para algunos rangos de β, las consultas de un solo término (FDD_β) y algunas consultas disyuntivas con 3 términos (FDD_β (OR(3))) obtienen un F_1 igual al alcanzado por la mejor consulta basada en TGF*-IDFEC. Una vez más, algunas de las consultas generadas usando el esquema "FDD_β (OR(3))" superan a todas las demás en términos de cobertura. Por otro lado, varias consultas basadas en FDD_β y TGF*-IDFEC obtienen el máximo valor posible de precisión $(1, 0)$. Una vez más se puede observar que los esquemas de generación de consultas no sufren de sobreajuste al conjunto entrenamiento.

2.8.2. Evaluación en el Conjunto de Datos *20NG*

En esta sección se presentan los resultados y discusiones de los experimentos de recuperación de información realizados sobre el conjunto de datos *20NG*. Los resultados del análisis comparativo para las particiones de entrenamiento y test se muestran en las Figuras 2.11 y 2.12, respectivamente.

En base a estas evaluaciones, se puede observar que la técnica FDD_β con $\beta = 1$ obtiene mejor desempeño en términos de F_1 con respecto a todas las demás técnicas. En particular, el desempeño (en términos de F_1) es superior al obtenido con las técnicas de pesaje de términos más efectivas (TGF*-IGM, IG, GR, TGF*-IGM_{imp}, GSS y SQRT-TGF*-IGM_{imp}) tanto para el conjunto de entrenamiento como para el de test. Más aún, IG, GR y GSS son superadas en términos de precisión por la técnica FDD_β con $\beta = 0,5$ y superadas en cobertura por FDD_β con $\beta = 10$. Asimismo, mientras que algunos métodos, como OR, Prob, RF, IDFEC, MI y IDFEC_B son muy efectivos cuando son evaluados en términos de la precisión, tienen un bajo desempeño en términos de cobertura y F_1.

Por el contrario, otros métodos como TGF y TGF* obtienen valores de cobertura de los más altos, pero tienen bajo desempeño en términos de precisión y F_1. Finalmente, vale la pena mencionar que las técnicas no supervisadas que ponen énfasis en favorecer términos discriminativos, tales como IDF y χ^2 tienen bajo desempeño tanto en el conjunto de entrenamiento como en el de test. Esto se debe a que estos métodos tienden a favorecer términos raros del conjunto de entrenamiento independientemente de que sean relevantes o no para el tópico (porque son no supervisados).

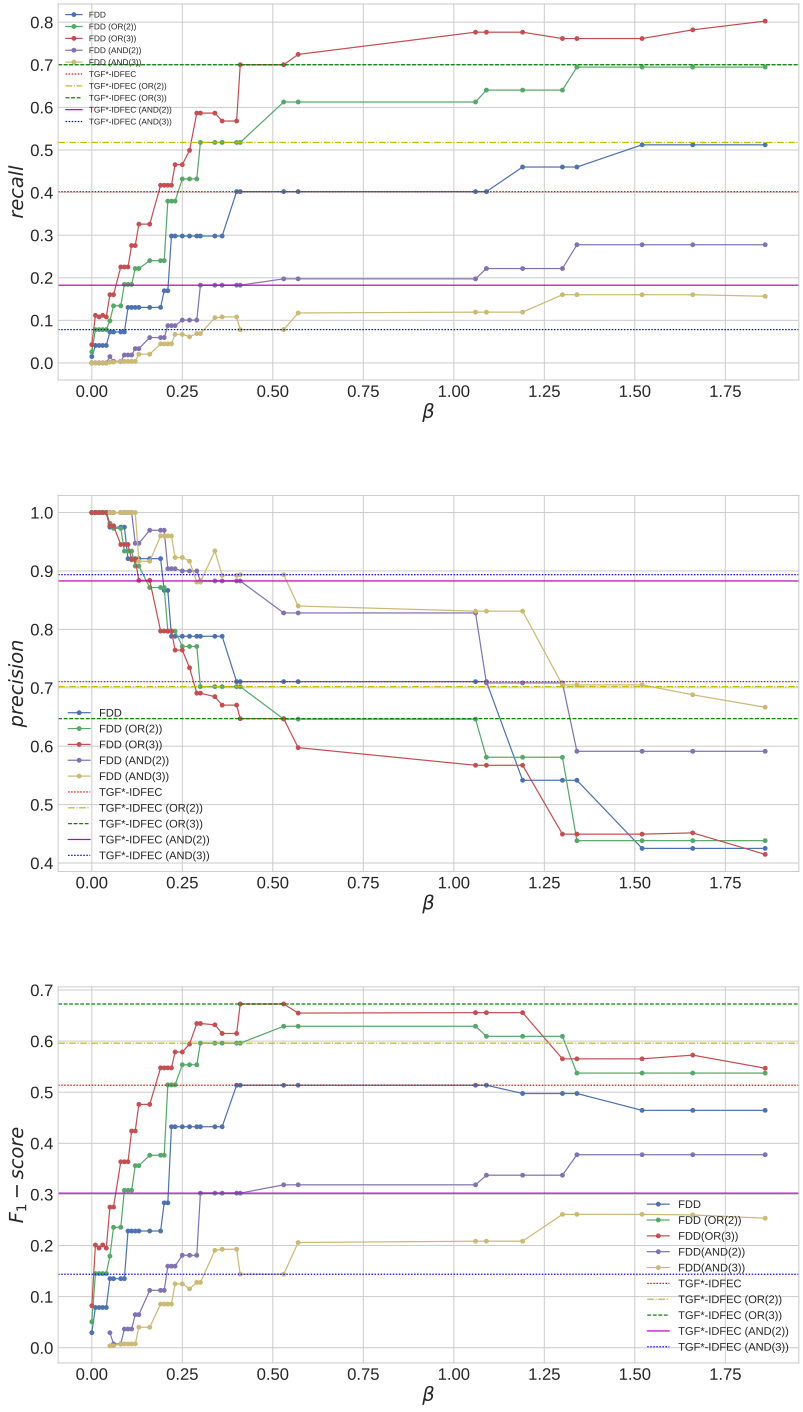


Figura 2.9: Eficacia en la partición de entrenamiento del conjunto *ERNTG* de las consultas de uno, dos y tres términos, tanto conjuntivas como disyuntivas, generadas usando las técnicas FDD_β y TGF*-IDFEC. Las líneas sólidas se corresponden con la eficacia de las consultas seleccionadas usando FDD_β con diferentes valores de β.

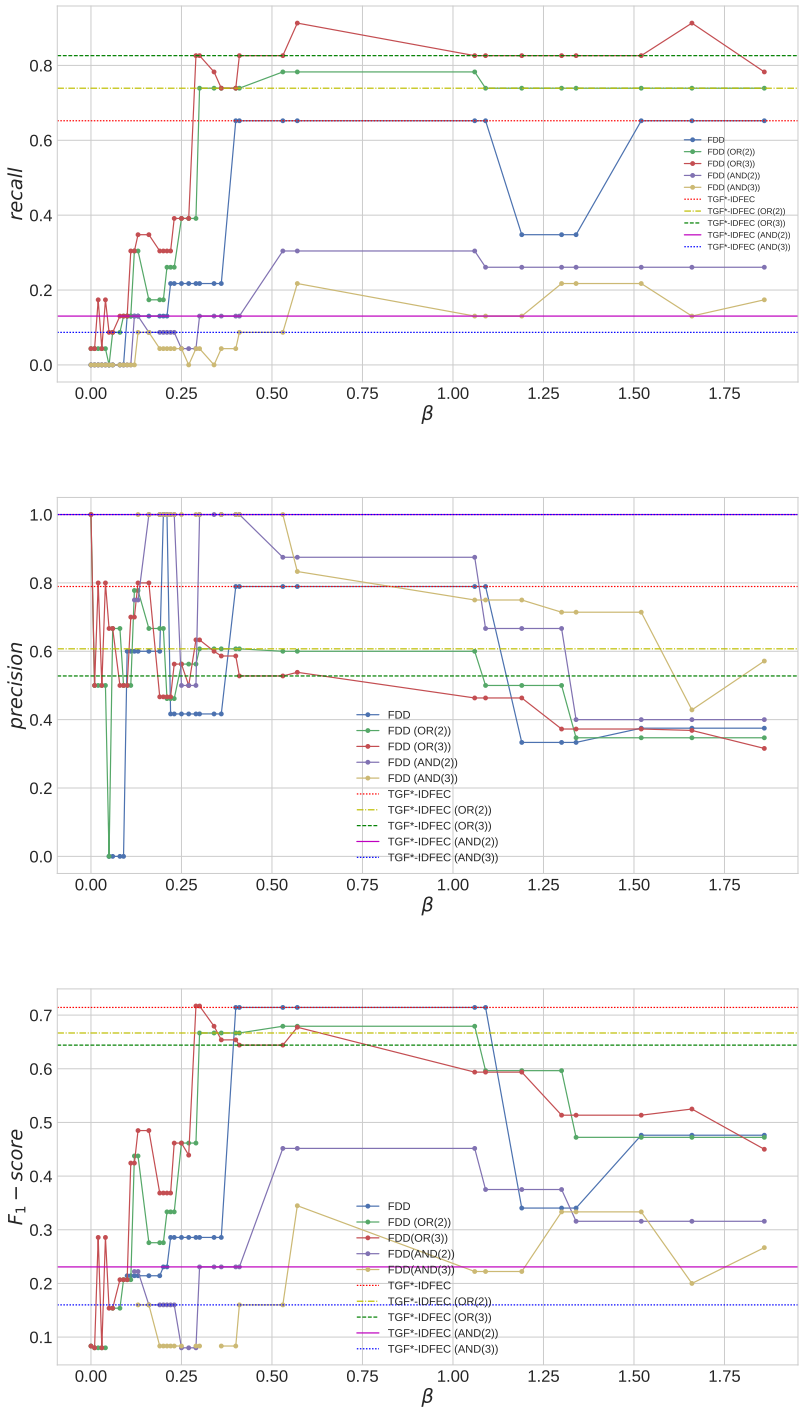


Figura 2.10: Eficacia en la partición de test del conjunto *ERNTG* de las consultas de uno, dos y tres términos, tanto conjuntivas como disyuntivas, generadas usando las técnicas FDD_β y TGF*-IDFEC. Las líneas sólidas se corresponden con la eficacia de las consultas seleccionadas usando FDD_β con diferentes valores de β.

Estos resultados sugieren la importancia de aprovechar la ventaja de las etiquetas de clase para el proceso de aprender buenos términos de consulta, así como también muestran la importancia de utilizar la técnica correcta dado los objetivos que se quieren alcanzar. Para el caso de la técnica FDD_β, la misma ya tiene un mecanismo incluido en la definición que permite ajustar a diferentes objetivos, esto es, el parámetro β. Vale destacar que todas las técnicas basadas en IGM tienen buen desempeño. Estos resultados sugieren que combinar el poder de discriminación de un término con un factor de frecuencia de término (que aporta cobertura) es una dirección prometedora para proponer una técnica de pesaje de términos. Las mejoras observadas al combinar discriminación con un factor que aporte cobertura (poder descriptivo), brinda evidencia para avalar la intuición en la cual está basada la técnica FDD_β.

El análisis previo se basa en el uso de consultas simples, esto es, consultas que consisten de un solo término. Consultas más complejas pueden ser formadas introduciendo más términos separados por operadores booleanos como "and" y "or". Como se mencionó previamente, un problema con las consultas más complejas es optimizar su construcción a partir de consultas simples. Esto quiere decir que combinar términos que sean buenos como consultas individuales no necesariamente resulta en una consulta más larga con buen desempeño. Aunque el análisis de consultas complejas puede involucrar muchos aspectos y dimensiones, por simplicidad, este trabajo se limita al análisis de consultas disyuntivas de tamaño dos y tres. Esta decisión está respaldada por trabajos previos que indican que las consultas disyuntivas tienden a tener mejor desempeño que consultas conjuntivas [CG09]. En la Figura 2.13 se presentan los resultados obtenidos con consultas disyuntivas obtenidas combinando términos obtenidos usando los métodos analizados. Por ejemplo, "$\text{FDD}_1$1 or $\text{FDD}_1$2" representa la consulta disyuntiva construida usando los dos términos mejor puntuados usando la técnica FDD_β con $\beta = 1$. De manera similar, "DISCR1 or DESCR1" representa la consulta obtenida de la disyunción del mejor discriminador (DISCR1) y del mejor descriptor (DESCR1). Notar que se seleccionaron un subconjunto del total de los métodos analizados para ser comparados. Esto se debe a que muchas veces dos técnicas tenían desempeños afines y estaban basadas en ideas similares, por ejemplo, IG y GR. Por simplicidad, en estos casos se consideró solo uno de los métodos como un representante del grupo.

Como se esperaba, los resultados muestran las ventajas de usar términos muy discriminativos ("DISCR1 or DISCR2" y "DISCR1 or DISCR2 or DISCR3") para obtener alta

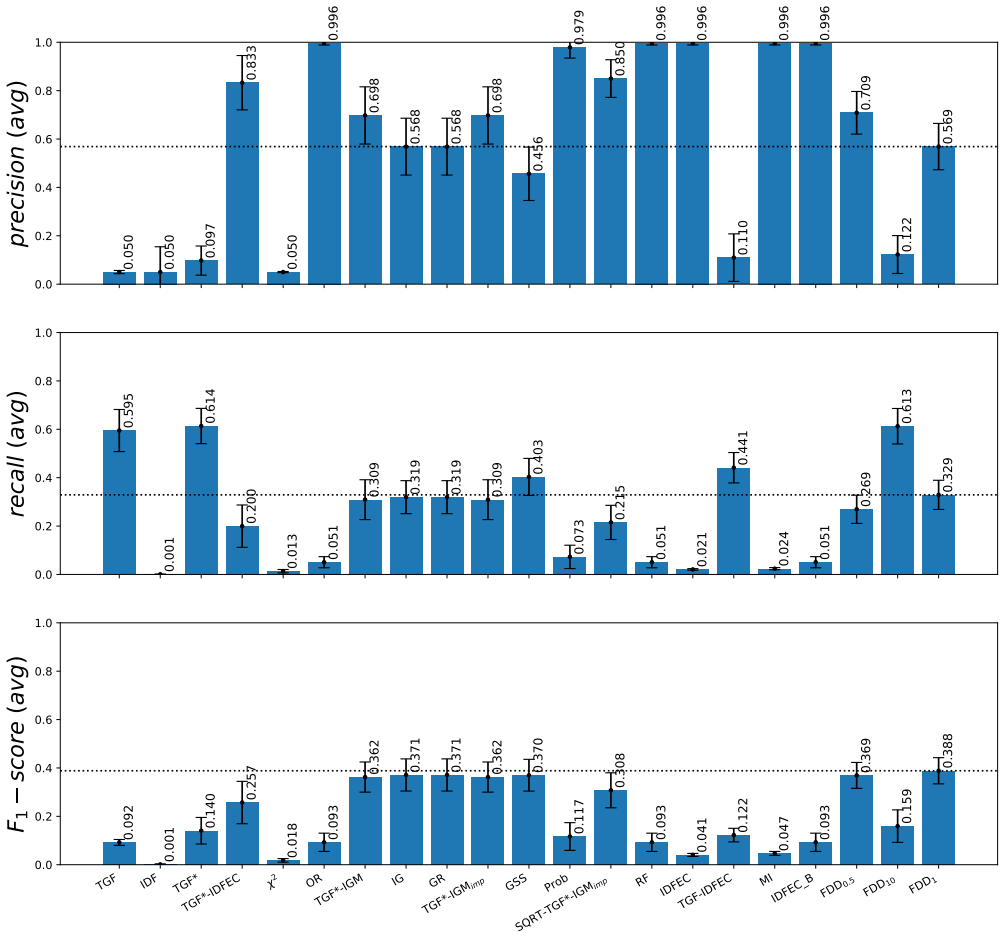


Figura 2.11: Comparación de los desempeños obtenidos con los métodos analizados sobre el conjunto de entrenamiento (conjunto de datos *20NG* - promediado a lo largo de las 20 categorías).

precisión, y usar buenos descriptores ("DESCR1 or DESCR2" y "DESCR1 or DESCR2 or DESCR3") para obtener alta cobertura. Finalmente, se puede observar que el mejor F_1 se obtiene utilizando la técnica FDD_1 ("$FDD_1$1 or $FDD_1$2" y "$FDD_1$1 or $FDD_1$2 or $FDD_1$3").

Notar que para la técnica FDD_1, usando consultas de dos términos ($FDD_1$1 or $FDD_1$2) se tiene un desempeño levemente mejor que usando una consulta de tres términos ($FDD_1$1 or $FDD_1$2 or $FDD_1$3). Esto da evidencia que indica que incorporar múltiples términos usando la técnica FDD_β con el mismo valor de β no necesariamente es beneficioso siem-

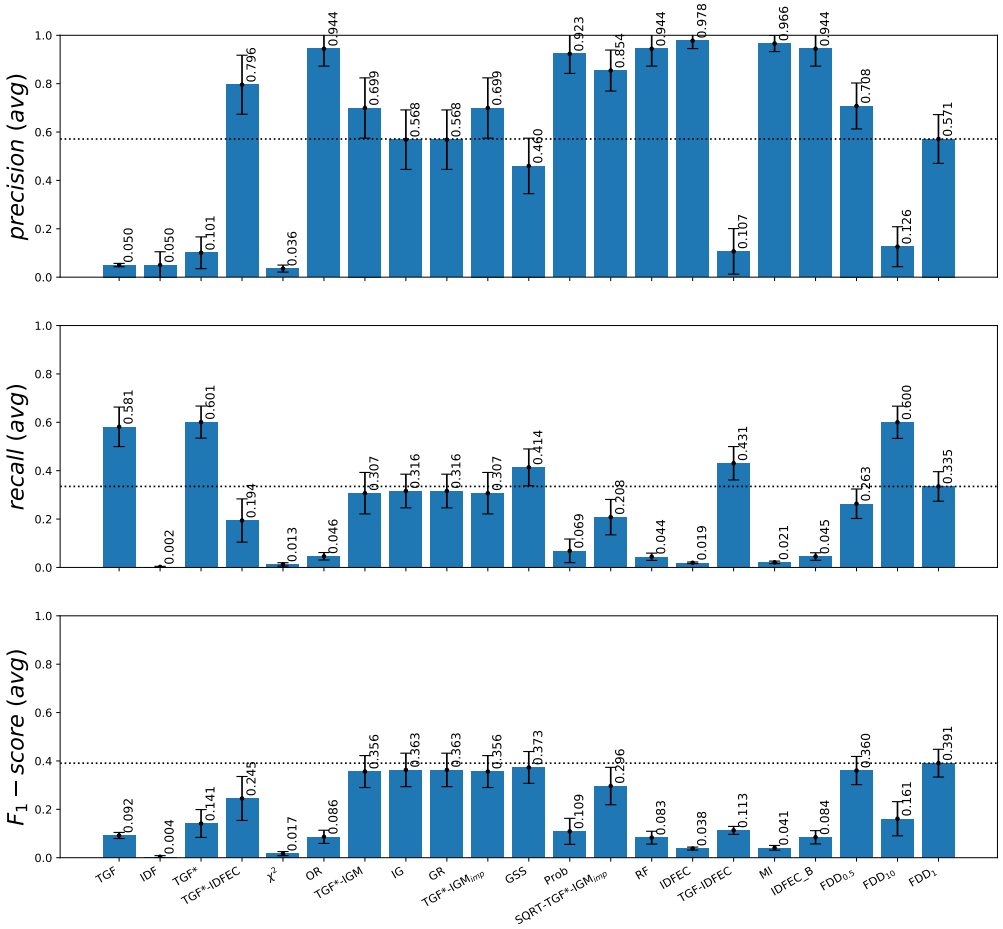


Figura 2.12: Comparación de los desempeños obtenidos con los métodos analizados sobre el conjunto de test (conjunto de datos *20NG* - promediado a lo largo de las 20 categorías).

pre. Esto indica que para obtener mejores resultados se debe complejizar el sistema de búsqueda (buscar términos obtenidos con distintos valores de β para combinar objetivos). La técnica propuesta, FDD$_\beta$, es una buena candidata para ser usada en un esquema más complejo de construcción de consultas por su capacidad de ser ajustada para diferentes objetivos (a través de su parámetro β).

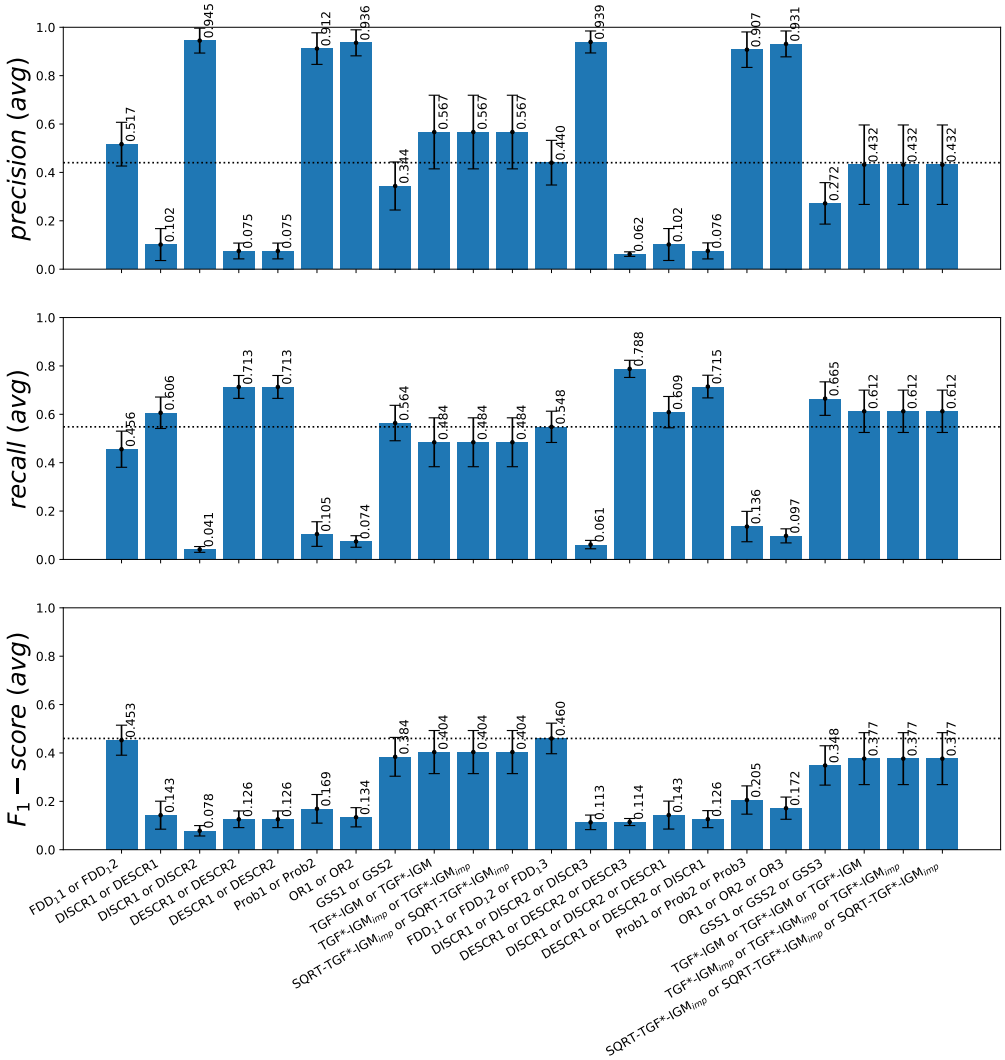


Figura 2.13: Comparación de los desempeños en el conjunto de test de las consultas complejas (de múltiples términos) obtenidas a partir de diferentes métodos de pesaje de términos (conjunto de datos *20NG* - promediado a lo largo de las 20 categorías).

2.8.3. Evaluación en el Conjunto de Datos *Reuters*

En esta sección se presentan y discuten los resultados obtenidos de comparar la técnica FDD_β con las otras dieciocho técnicas presentadas en 2.1 en el contexto de recuperación de información usando el conjunto de datos *Reuters*. Para realizar estas evaluaciones se usó el conjunto de entrenamiento para seleccionar los términos mejor ponderados de acuerdo

a cada técnica de ponderación de términos analizada. Luego, los términos seleccionados fueron usados para construir consultas de un único término, del mismo modo que se hizo para los conjuntos *ERNTG* y *20NG*. El desempeño de cada método fue evaluado usando las métricas clásicas de recuperación de información: precisión, cobertura y F_1 en ambas particiones de datos, la partición de entrenamiento y la de test. El desempeño sobre las particiones de entrenamiento y test es reportado en las Figuras 2.14 y 2.15, respectivamente.

Notar que el conjunto de datos *Reuters* define un escenario distinto a los anteriores dada las características de las categorías. En este conjunto de datos hay 120 tópicos (o categorías), y cada documento puede estar asociado a cero o más categorías (categorías no excluyentes). Mientras que en el conjunto de datos *20NG*, cada documento tiene estrictamente una categoría asociada. Debido a las características del conjunto de datos *20NG*, se pudo construir una distribución de clases de veinte dimensiones para computar todos los métodos basados en IGM. En el caso del conjunto de datos *Reuters*, para computar la distribución de clases el problema se tuvo que simplificar a categorías binarias (pertenece o no pertenece al tópico bajo análisis), esto se debió a la característica no excluyente de las categorías del conjunto de datos.

Nuevamente se puede ver que los resultados obtenidos por FDD_β obtienen el mejor F_1. Vale la pena mencionar que algunos métodos que obtenían buen desempeño en el conjunto *20NG*, como IG, GR y GSS, son consistentemente efectivos en el conjunto *Reuters*. Por otro lado, los métodos basados en IGM, los cuales estaban entre los más efectivos en el conjunto de datos *20NG*, tuvieron un bajo desempeño en *Reuters*. Este bajo desempeño es una consecuencia de la previamente mencionada simplificación de la distribución de clase (de un problema de 20 categorías se pasó a un escenario binario). Esto sugiere que los métodos basados en IGM no son adecuados para escenarios binarios.

También se observa que $FDD_{0,5}$ tiene un desempeño inferior a MI y RF en términos de precisión pero mejor en términos de cobertura y F_1. Finalmente, se observa que FDD_{10} obtiene una cobertura comparable, aunque levemente inferior, a la obtenida por TFG*, que es el método que más cobertura obtiene de todos los métodos del estado del arte. Sin embargo, todas las variantes de FDD_β fueron superiores a TFG* en términos de precisión y F_1. En general, se puede observar que el desempeño a nivel de F_1, precisión y cobertura en el conjunto de datos *Reuters* fue sistemáticamente inferior en todos los métodos analizados cuando se compara con el conjunto de datos *20NG*. Esto habla de la

dificultad intrínseca del conjunto de datos y de la diferencia que resulta del cambio de escenario (de 20 categorías excluyentes a categorías binarias no excluyentes).

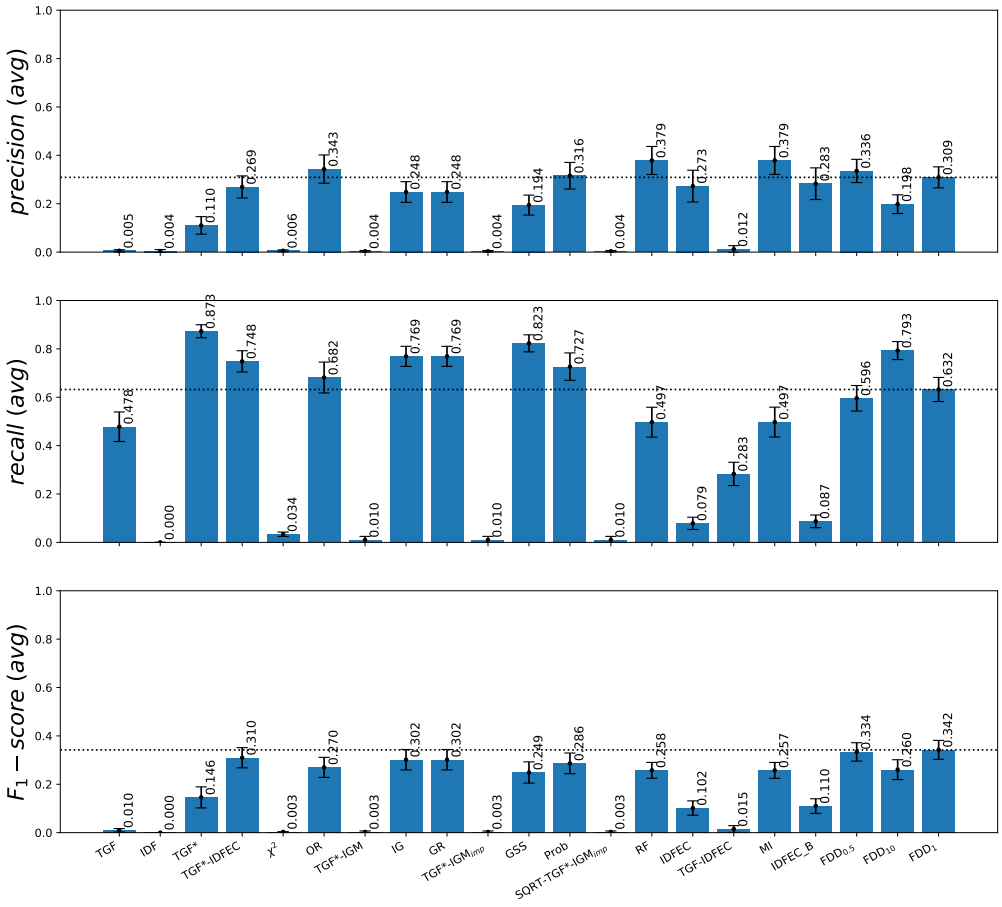


Figura 2.14: Comparación de desempeño de los métodos analizados sobre el conjunto de entrenamiento (conjunto de datos *Reuters-21578* - promediado a lo largo de los 120 tópicos).

2.9. Aplicación a Extracción de Variables, Modelado y Recuperación de Información

Como se puede observar de los resultados reportados en este capítulo, la técnica FDD$_\beta$ tiene un desempeño consistentemente bueno, no solo como un estimador de los puntajes de

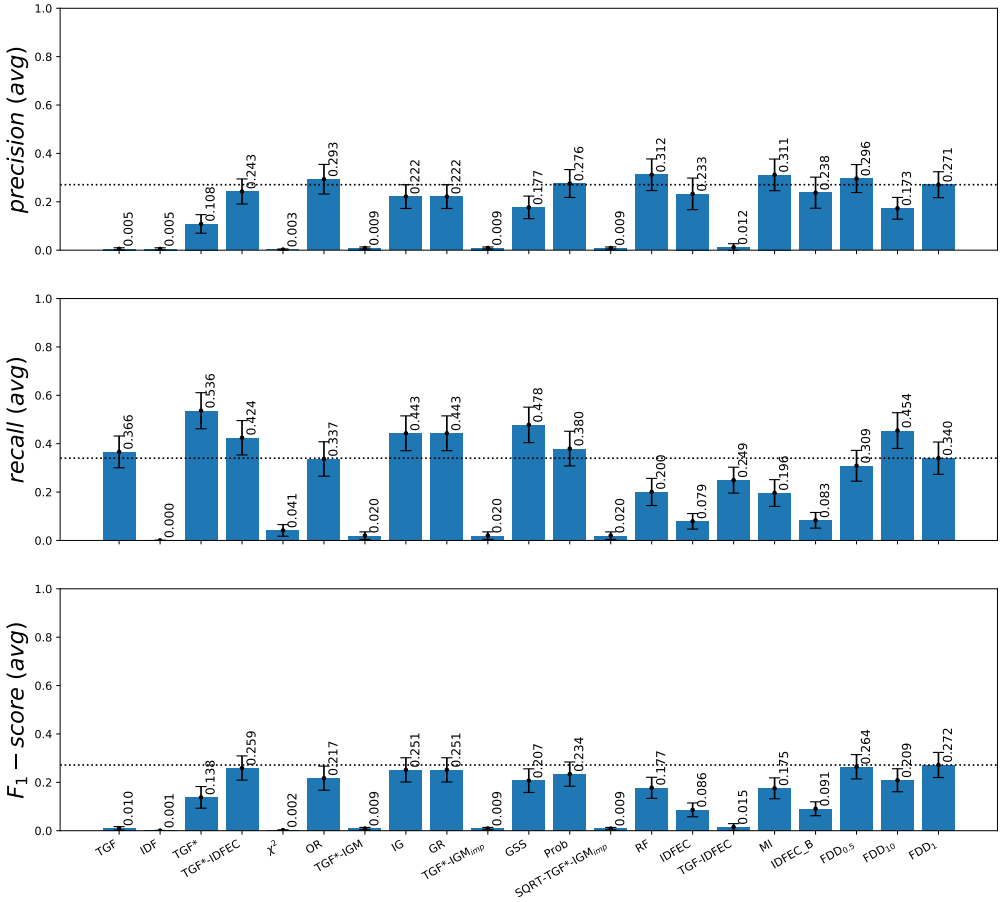


Figura 2.15: Comparación de desempeño de los métodos analizados sobre el conjunto de test (conjunto de datos *Reuters-21578* - promediado a lo largo de los 120 tópicos).

relevancia de los usuarios sino también como un método para guiar la selección de buenos términos de consulta para recuperación de información. Esto abre numerosas oportunidades para aplicar el esquema propuesto a diferentes escenarios. Esta sección describe algunas posibles aplicaciones que podrían beneficiarse de la técnica de pesaje de términos propuesta.

Figura 2.16: Nube de palabras con los términos mejor puntuados de acuerdo a la técnica $\text{FDD}_{0,477}$ en el conjunto de datos *ERNTG* (solo se muestran aquellos términos que cumplen $\text{FDD}_{0,477} > 0,6$, para evitar sobrecargar la imagen).

2.9.1. Creación de Modelos Predictivos

La técnica propuesta puede ser usada para extraer variables de medios periodísticos digitales con el objetivo final de construir modelos de predicción, explicación o descripción. En la Figura 2.16 se muestra una visualización de nube de palabras con los términos (unigramas) mejor puntuados de acuerdo a la técnica FDD_β en el conjunto de entrenamiento de la colección *ERNTG*, usando $\beta = 0,477$. Para no sobrecargar la imagen solo se muestran términos que cumplen $\text{FDD}_{0,477} > 0,6$.

Vale la pena mencionar que la técnica propuesta FDD_β puede ser computada no solo sobre palabras de una colección (como ha sido ilustrado hasta ahora) sino también sobre otro tipo de unidades léxicas como por ejemplo entidades, n-gramas o conceptos. En la Figura 2.17 se presenta una nube de palabras con las entidades mejor puntuadas de acuerdo a la técnica $\text{FDD}_{0,477}$ identificadas en el conjunto de entrenamiento de la colección *ERNTG*. La herramienta *Stanford Named Entity Recognition tagger* (Stanford NER) [FGM05] fue usada para identificar esas entidades.

Un próximo paso puede ser la identificación de dependencia entre estas variables extraídas. Por ejemplo, algunas *relaciones causales* que se pueden encon-

Figura 2.17: Nube de palabras construida con las entidades mejor puntuadas de acuerdo a la técnica $\text{FDD}_{0,477}$ en el conjunto de datos *ERNTG* (solo entidades con $\text{FDD}_{0,477} > 0,15$ son mostradas).

trar entre estas variables son `investment-growth-gdp`, `spending-market-recovery` y `sales-companies-investment-gdp`. Otro tipo de relaciones que se pueden identificar son *asociaciones estrechas* como por ejemplo: `credit-debt-banks`, `recession-decline`, `trading-stock-ftse` y `debt-bank-investors-trading-recession`. Se pueden encontrar *relaciones de simultaneidad* como por ejemplo *market-prices*. También es interesante notar que la variable ('`christmas`') puede estar capturando *estacionalidad en series causales*. Automáticamente identificar este tipo de relaciones es un problema desafiante que se resuelve parcialmente en este trabajo (solo para relaciones causales) y se plantea como trabajo futuro la extracción de otros tipos de relaciones. En particular en este trabajo se plantea el problema de construir grafos causales dirigidos a partir de variables extraídas de artículos periodísticos.

2.9.2. Asistir al Modelado de Conocimiento

Construir modelos de conocimientos (knowledge models) es una tarea costosa y difícil. Existen varias iniciativas que tienen por objetivo proveer asistencia inteligente para apoyar a los expertos en la tarea de construcción de modelos de conocimiento, como es el caso de la familia de los sistemas de sugerencias para mapas conceptuales descritos en [LMR14, LML+16].

Los mapas conceptuales [Nov90] son una forma de modelar conocimiento que fue propuesta para educación, para permitir a estudiantes externalizar sus conocimientos a través de la representación gráfica de los conceptos y sus relaciones. Diversos sistemas para el mapeo de conceptos han sido usados tanto por usuarios en la escuela primaria como así también por científicos para asistir en la generación, almacenamiento y acceso a mapas conceptuales electrónicos. Además de proveer operaciones básicas para dibujar y manipular mapas conceptuales, estos sistemas pueden ser equipados con herramientas que facilitan la extensión de conocimiento. En particular, la identificación automática de términos relevantes al dominio modelado permite extender el modelo de conocimiento más allá de la información ya capturada en él. La visualización presentada en las Figuras 2.16 y 2.17 puede asistir en la construcción de mapas conceptuales permitiendo la selección de términos relevantes a un dominio particular, lo que típicamente es el paso inicial en la etapa de construcción de modelos de conocimientos.

Los mapas conceptuales son usualmente organizados en jerarquías, donde los términos más generales suelen aparecer en la parte superior del mapa (cercanos a la raíz del mismo) mientras que los términos más específicos tienden a ocurrir en la parte inferior (cercanos a las hojas). La técnica de ponderación de términos propuesta puede ofrecer una nueva solución al problema de identificar diferentes términos y entidades que pueden ser usados como sugerencias para ir agregando en diferentes niveles del mapa conceptual. Como los descriptores tienden a representar términos que describen un tópico general mientras que los discriminadores tienden a ser términos específicos, se puede usar el parámetro ajustable β para ayudar en la selección de términos favoreciendo generalidad o especificidad de acuerdo a las necesidades del usuario.

2.9.3. Alcanzar Cobertura Total

La tarea de recuperación de información con cobertura total (*total recall*) [RCGC15, GCR16, AGZ+18] es el problema de encontrar todos (o casi todos) los documentos relevantes a un dado tópico de búsqueda. A diferencia de los escenarios donde el objetivo es encontrar unos pocos documentos altamente relevantes (por ejemplo, para responder una pregunta específica se puede necesitar alta precisión), los escenarios donde se requiere alta cobertura son aquellos que se concentran en recuperar todos los documentos relevantes sin tener una pérdida significativa de precisión (por ejemplo, recolectar todos los documentos relevantes a un tópico para construir un sitio web de un determinado tópico).

Una interrogante que surge de atacar el problema de cobertura total es la de saber cuántos términos son necesarios en una consulta disyuntiva para alcanzar el 100 % de cobertura. Para analizar esta interrogante se utilizó el conjunto de entrenamiento del conjunto de datos *ERNTG* para incrementalmente construir consultas disyuntivas mediante la inclusión de los términos mejor puntuados usando la técnica FDD_β con diferentes valores para el parámetro β. Como se discutió previamente, los términos con alto poder descriptivo para un tópico son aquellos que tienden a ocurrir frecuentemente en documentos relevantes al tópico. Por ende, se espera que consultas generadas usando valores de β altos (poder descriptivo alto) obtengan cobertura total con menos cantidad de términos en contraste con consultas con valores de β bajo (poder discriminativo alto). Esta intuición se puede verificar con los resultados reportados en la Figura 2.18, donde es posible ver que a medida que el valor de β aumenta, el número de términos necesarios para alcanzar cobertura total disminuye significativamente. En la Figura 2.19 se presenta un análisis

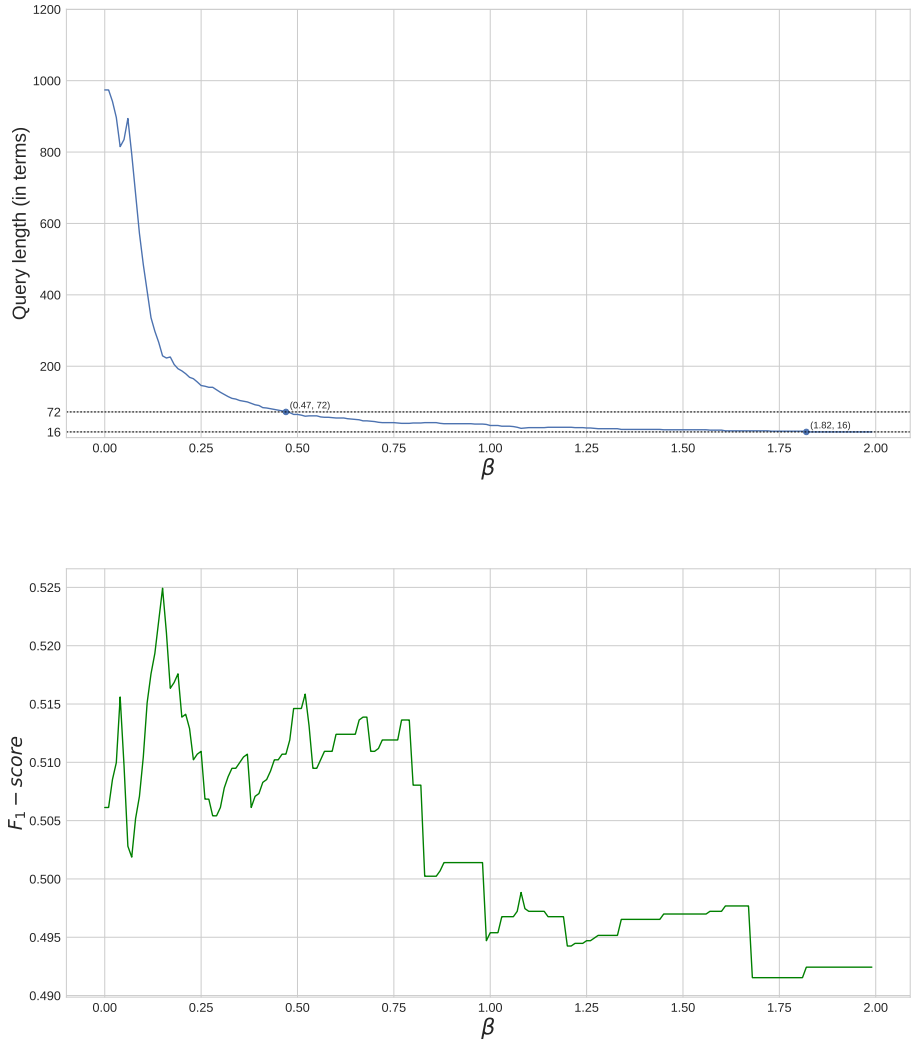


Figura 2.18: Cantidad de términos unidos por consultas disyuntivas necesarios para obtener cobertura total (cobertura=1,0) en la partición de entrenamiento del conjunto *ERNTG* (arriba). Los términos se ordenan por FDD_β para cada valor de β y se seleccionan los de mayor puntaje hasta obtener cobertura total. Valores de F_1 obtenidos por las consultas de cobertura total (abajo).

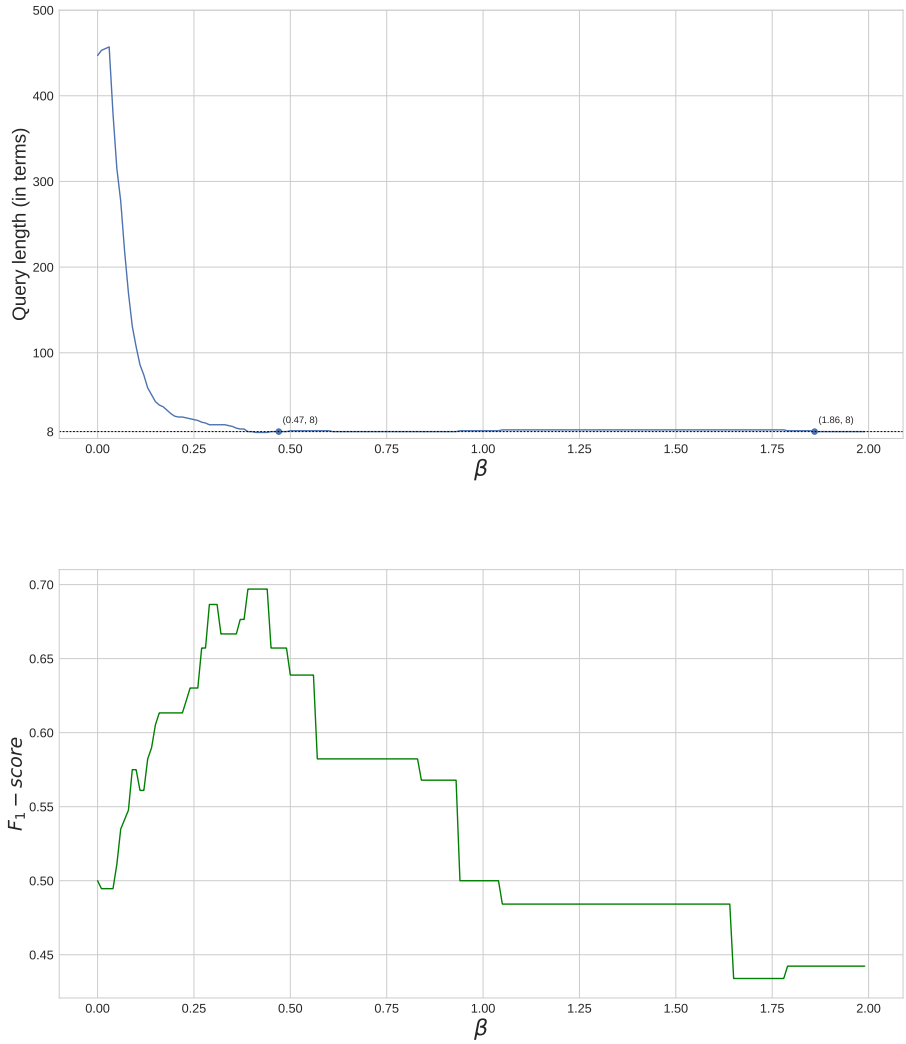


Figura 2.19: Cantidad de términos unidos por consultas disyuntivas necesarios para obtener cobertura total (cobertura=1,0) en la partición de test del conjunto *ERNTG* (arriba). Los términos se seleccionan del conjunto de entrenamiento, se ordenan por FDD_β para cada valor de β y se seleccionan los de mayor puntaje hasta obtener cobertura total. Valores de F_1 obtenidos por las consultas de cobertura total (abajo).

similar usando el conjunto de entrenamiento para generar las consultas y el conjunto de test para calcular las métricas. Como se vio previamente, los resultados indican que un valor mayor de β resulta en la construcción de consultas con menos términos para obtener cobertura total. Notar que el largo de la consulta para alcanzar cobertura total es muy dependiente del número de documentos relevantes en la colección. Por ende, las consultas para obtener cobertura total para el conjunto de test son significativamente más cortas que las consultas necesarias para el conjunto de entrenamiento.

Para analizar cómo la cobertura total impacta en el F_1 también reportamos dicha métrica para las consultas evaluadas en ambos conjuntos, el de entrenamiento (Figura 2.18 (abajo)) y el de test (Figura 2.19 (abajo)). Para ambas particiones es posible ver que para ciertos valores de β la cobertura total es posible sin una caída significativa del F_1. Estos resultados dan indicios del potencial de la técnica propuesta como un mecanismo para atacar el problema de la cobertura total.

2.10. Discusión

Los resultados presentados en la Sección 2.7 muestran el potencial de la técnica como un estimador de los puntajes de relevancia asignados por los usuarios, mostrando resultados prometedores de la técnica para la tarea de extracción de variables que posteriormente pueden ser usadas en modelos predictivos o para asistir al modelado del conocimiento de un dominio (a través de mapas conceptuales o grafos causales).

El análisis presentado en la Sección 2.6 permitió analizar en detalle el comportamiento del parámetro β en la técnica de pesaje de términos FDD_β. Se observó que el parámetro β tiene un rol crucial en el desempeño de la recuperación de información y que su valor óptimo para un dado escenario puede ser aprendido utilizando un conjunto de entrenamiento utilizando una métrica objetivo (precisión alta, cobertura alta, o un balance de ambas). Se observó que el desempeño de FDD_β fue similar en los conjuntos de entrenamiento y test, indicando que la técnica no sufrió de sobreajuste en los escenarios planteados.

Las evaluaciones reportadas en la Sección 2.8 muestran que la técnica FDD_β con β=1 consistentemente obtiene F_1 mayor que todas las demás técnicas evaluadas, tanto en el conjunto de entrenamiento como en el de test. Adicionalmente se vio que el parámetro β puede ser ajustado para obtener mejor desempeño que las otras técnicas en términos

de alguna de las métricas analizadas: precisión o cobertura. Esto se vuelve particularmente relevante cuando se necesitan resolver tareas con un objetivo específico como la tarea de cobertura total analizada en la Sección 2.9.3. Las evaluaciones también mostraron las ventajas de las técnicas supervisadas por encima de las no supervisadas. Es importante resaltar que la técnica FDD_β con β=1 también resultó en el mejor F_1 cuando se la utilizó para la construcción de consultas de múltiples términos usando operadores disyuntivos. Finalmente, se demostró la utilidad de generar consultas con varios términos usando buenos descriptores o discriminadores para obtener alta cobertura o precisión, respectivamente.

Un hallazgo importante que se desprende de los resultados de los experimentos de este capítulo es que a pesar de que ciertas nociones de las áreas de teoría de la información o estadística (como entropía, información mutua o ganancia de información) han mostrado ser muy útiles para construir técnicas no supervisadas de pesaje de términos, este nivel de complejidad no parece ser necesario para obtener buenos resultados en la tarea de recuperación temática de información. Como se puede ver de los resultados de este capítulo, la técnica FDD_β obtiene desempeños altamente competitivos con el estado del arte sin necesitar apoyarse en las nociones complejas antes mencionadas. Los resultados derivados de estos experimentos muestran que la técnica FDD_β ofrece un mecanismo útil para explorar diferentes enfoques de construcción de consultas complejas para la recuperación de información temática.

2.11. Conclusiones y Trabajo Futuro

Este capítulo define la técnica de ponderación de términos FDD_β, analiza su comportamiento y evalúa su desempeño para diferentes tareas: (i) identificación de términos específicos a un dominio (estimación de puntajes de relevancia elegidos por usuarios) (ii) selección de términos de consulta para la tarea de recuperación de información temática, (iii) generación de consultas para obtener cobertura total. La primera contribución de este capítulo es la definición de la técnica FDD_β y la presentación de un extenso estudio sobre el análisis y el impacto del parámetro β en el comportamiento de la técnica FDD_β. En este capítulo se presentan análisis para apoyar la idea de que la flexibilidad ofrecida por el parámetro β representa una importante ventaja por encima de otras técnicas existentes de ponderación de términos.

La segunda contribución es la presentación de un extenso estudio comparativo donde se evalúa la técnica presentada, FDD_β, contra otras dieciocho técnicas del estado del arte y tradicionales, incluyendo tanto técnicas supervisadas como no supervisadas. El código fuente donde se define la técnica FDD_β y las dieciocho técnicas replicadas se deja disponible para permitir la reproducibilidad del estudio. El estudio comparativo de desempeños muestra que, a pesar de su simplicidad, la técnica FDD_β obtiene resultados que son competitivos con el estado del arte, incluso cuando se los compara con técnicas que se basan en conceptos más complejos de teoría de la información o estadística.

La tercera contribución es una nueva perspectiva hacia la pregunta de cómo construir consultas de múltiples términos basándose en técnicas supervisadas de pesaje de términos. El análisis presentado en este capítulo demuestra que la técnica FDD_β ofrece un mecanismo que permite explorar diferentes objetivos de la construcción de consultas complejas.

Finalmente, la cuarta contribución es la creación y publicación de un conjunto de datos etiquetado por expertos en economía que puede ser usado para computar pesaje de términos basados en tópicos o aplicado para otras tareas supervisadas en el dominio económico. En este trabajo el conjunto de datos fue usado para hacer un análisis exhaustivo del comportamiento de la técnica FDD_β en función de su parámetro ajustable β. Dicho conjunto de datos también fue usado para evaluar el desempeño de las diferentes técnicas para la tarea de recuperación de información temática, para la tarea de extraer términos para el estudio de puntajes de relevancia asignados por usuarios a términos y para la tarea de recuperación de información con cobertura total.

A partir de este trabajo surgen diversas propuestas de trabajos futuros. En primer lugar, sería interesante evaluar estrategias de recuperación de información usando FDD_β en el contexto de construcción de consultas de múltiples términos con sintaxis más complejas (no solo disyuntivas o solo conjuntivas). Estas consultas pueden, potencialmente, considerar distintos aspectos de la recuperación de información (alta cobertura, alta precisión, cobertura total). En segundo lugar, se espera evaluar la técnica FDD_β para tareas de clasificación. En particular se propone estudiar el problema de encontrar el β óptimo para la tarea de clasificación durante el proceso de entrenamiento. Por último, también se propone la adaptación de la técnica FDD_β (que depende de un conjunto de entrenamiento) para obtener pesos para términos que aparecen en los conjuntos de validación o test, donde no hay información de clases.

2.12. Disponibilidad del Código Fuente

Los experimentos realizados para evaluar el rol del parámetro β en los conjuntos de datos *ERNTG* (Sección 2.6.1) y 20NG (Sección 2.6.2) se encuentran disponibles para permitir su reproducibilidad[3]. En el mismo enlace se encuentran disponibles también los experimentos realizados para evaluar el desempeño de la técnica FDD_β como herramienta de recuperación de información sobre los conjuntos de datos 20NG (Sección 2.8.2) y *Reuters* (Sección 2.8.3). En estos últimos experimentos la técnica es comparada contra las 18 técnicas del estado del arte presentadas en la Tabla 2.1.

[3] `http://cs.uns.edu.ar/~mmaisonnave/resources/FDD_code`

Capítulo 3

Detección de Eventos en Curso

Resumen

Este capítulo presenta la definición de la tarea de detección de eventos en curso (*Ongoing Event Detection* (OED)), que es una tarea específica dentro de la tarea más general de Detección de Eventos (*Event Detection* (ED)). El objetivo de la tarea de OED es la detección de menciones de eventos en curso, en contraposición a eventos históricos, futuros, hipotéticos u otras formas de eventos que no están en curso ni son actuales. Cualquier aplicación que necesite extraer información estructurada sobre eventos en curso a partir de textos no estructurados puede beneficiarse de un sistema de OED. Las contribuciones más importantes de este capítulo son:

1. Se introduce la tarea de OED junto con un conjunto de datos manualmente etiquetado para dicha tarea.

2. Se presenta el diseño y desarrollo de un modelo predictivo basado en Redes Neuronales Recurrentes (*Recurrent Neural Networks* (RNN)) para la tarea de OED. Dicho modelo utiliza *embeddings BERT* para construir representaciones contextuales de palabras y oraciones para ser usadas como atributos de entrada del modelo.

3. La presentación de una extensa evaluación empírica que incluye: (i) la exploración de diferentes arquitecturas e hiperparámetros, (ii) un estudio de ablación para medir el impacto de cada atributo y (iii) una comparación con un modelo del estado del arte replicado para la tarea de OED.

Los resultados permiten un entendimiento de la importancia de los *embeddings* contextuales e indican que el enfoque propuesto es efectivo para la tarea de OED, superando en desempeño los modelos base.

3.1. Introducción

La tarea de extracción de información (IE por sus siglas en inglés) consiste en la extracción de información estructurada a partir de textos de lenguaje natural (no estructurados). La tarea de Extracción de Eventos (EE) es una subtarea de IE, cuyo objetivo es detectar y recuperar eventos del mundo real de textos en lenguaje natural. La tarea de EE aborda el problema a nivel de procesamiento de lenguaje natural (NLP por sus siglas en inglés) recuperando menciones individuales de eventos a partir de textos, pudiendo haber múltiples menciones de eventos en cada fragmento de texto. No debe confundirse con la tarea de agrupar una colección de documentos de acuerdo a qué evento central es tratado en cada artículo [Mai19, WL21, APL98]. Un sistema de EE usualmente lleva a cabo dos pasos diferentes para completar la extracción de los eventos.

El primer paso es la detección del *event trigger*, que es la palabra que más claramente indica la ocurrencia del evento y la clasifica dentro de una de las categorías predefinidas de eventos. Por ejemplo, la frase "Jeff Bezos renuncia a su puesto de CEO de Amazon." contiene un solo evento, siendo el *event trigger* de dicho evento la palabra *renuncia*. Asumiendo que en la taxonomía de eventos predefinidos existe el tipo de evento "renuncia" (Abandonar voluntariamente un puesto de trabajo), dicho *event trigger* pertenecería a este tipo de eventos. A este primer paso de detectar el *event trigger* y clasificarlo se lo denomina Detección de Eventos (ED por sus siglas en inglés).

El segundo paso es la extracción de los argumentos (participantes o atributos). Por ejemplo, el tipo de evento "renuncia" puede tener tres argumentos definidos en la taxonomía: la persona que renuncia, el puesto al que renuncia y la empresa en la cual era el puesto (en el ejemplo anterior serían "Jeff Bezos", "CEO" y "Amazon", respectivamente).

La tarea de ED se ha abordado en la literatura tanto como una tarea independiente [NG18, LCLZ18, DHZ17, NG16, NG15, NFCG16, WBL+14, JY16], como una tarea que forma parte de un sistema de EE [ZJS19, LLH18, SQCS18, HJCV17, NCG16, CXL+15, LJH13, Ahn06]. Existe un gran incentivo para estudiar los sistemas de ED no solo porque

tienen un impacto directo en múltiples aplicaciones que hacen uso de los eventos detectados (como sistemas de alerta) sino también porque cualquier mejora en los sistemas de detección de eventos tendría un impacto directo en el desempeño de los sistemas de extracción de eventos que dependen de este primer paso para cumplir con la totalidad de la tarea. Los sistemas de EE y ED son esenciales en múltiples aplicaciones y dominios que necesitan obtener información estructurada a partir de grandes colecciones de datos no estructuradas. Algunos ejemplos de estos sistemas son: sistemas de pregunta-respuesta [SQCS18] y sistemas para generar resúmenes de textos [LCJ03]. Estos sistemas son útiles también para generar reportes de información disponible para un dominio dado [LSLL09, AOGD$^+$16, CZSL18]. Estos reportes pueden ayudar a un experto a tomar decisiones o generar políticas para abordar un problema.

En este capítulo, se define y aborda la tarea de OED como parte de un proyecto más amplio que trata de detectar eventos del mundo real en curso y otras variables relevantes de artículos periodísticos con el objetivo final de construir modelos causales [MDT$^+$20b]. En este proyecto más amplio, presentado en detalle en el capítulo 4, se utilizan técnicas de series de tiempo del área de econometría para aprender esos modelos causales [Gra69, Sch00]. Para construir esas series de tiempo, se requieren (i) todas las menciones de eventos de las noticias y (ii) el momento en el que esos eventos ocurrieron. Estos dos requerimientos definieron las necesidades del sistema de detección de eventos. Por (i), solo se requiere la detección de los eventos (*trigger detection*), no siendo necesaria la extracción de los argumentos de los eventos. Por (ii), se requiere que los eventos estén en curso cuando son reportados (descartando eventos pasados, futuros o hipotéticos). Si el evento no estaba en curso cuando fue reportado entonces la fecha del artículo periodístico no puede ser usada como el momento de ocurrencia del evento. Estos dos requerimientos resultaron en la definición de la nueva tarea propuesta, a la cual se la definió como Detección de Eventos en Curso (OED por sus siglas en inglés).

El proyecto de construcción de modelos causales se dividió, en líneas generales, en dos pasos, (1) la tarea de OED y (2) la construcción de series de tiempos y modelos causales a partir de esas series de tiempo. Estos dos pasos involucrados en la construcción del modelo causal están representados en la Figura 3.1. Es importante notar que dicha figura es una representación parcial del *framework* completo de descubrimiento de estructuras causales a partir de texto. Se muestra primero la detección de eventos en dos fragmentos de textos y los vínculos causales extraídos directamente del texto. En el *framework*

completo se utilizan múltiples menciones de múltiples eventos para la construcción de series de tiempo. El posterior descubrimiento causal puede encontrar relaciones causales que no necesariamente están explícitas en el texto. El trabajo reportado en este capítulo se concentra en el primer paso únicamente, esto es, la tarea de OED. Los detalles de la construcción del conjunto de datos de tipo serie de tiempo y el posterior descubrimiento causal se explican en el Capítulo 4.

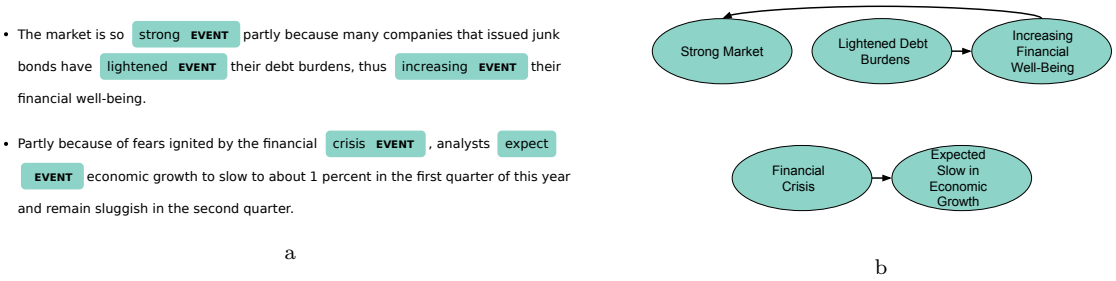


Figura 3.1: Dos casos de uso para la herramienta de OED, aplicada a dos oraciones (a). Un modelo causal extraído manualmente de las dos oraciones ejemplo (b).

Motivados por las limitaciones de propuestas previas (que no se ajustaban a las necesidades del proyecto propuesto en este trabajo) se delineó una definición de evento (descrita en la Sección 3.3) que, por un lado, no está limitado a una taxonomía fija de tipos de eventos, y, por otro lado, está orientada únicamente a eventos en curso. Usando esta definición de evento, se anotó manualmente un conjunto de datos de entrenamiento y test para la tarea [MDT+20a] (descrito en la Sección 3.4.1). Usando este conjunto de datos etiquetados se desarrolló un modelo basado en RNN para predicción de eventos (descrito en la Sección 3.4.2). Debido a que no hay estudios previos con el conjunto de datos usado en este capítulo, también se implementaron dos modelos de referencia (*baselines*) para fines de comparación. Primero, se implementó un modelo simple para OED basado en una herramienta clásica (un modelo SVM). Luego, se replicó, como un segundo *baseline*, un modelo del estado del arte (un modelo CNN) de ED aplicado a la tarea definida en este trabajo [NG15]. Ambos *baselines* son descritos en la Sección 3.4.3. Tanto el código fuente para los modelos *baseline*, como el código fuente para el modelo propuesto y el conjunto de datos usados están disponibles para permitir la reproducibilidad de los resultados y la reutilización de los datos en otros trabajos[1]. Los resultados, discusiones y conclusiones son presentados al final de este capítulo en las Secciones 3.5 3.5.2 y 3.6.

[1]El código se encuentra en https://cs.uns.edu.ar/~mmaisonnave/resources/ED_code/ y el conjunto de datos en https://cs.uns.edu.ar/~mmaisonnave/resources/ED_data/

Las contribuciones de este capítulo se pueden resumir de la siguiente manera:

1. Primero, se define la tarea de OED por primera vez y se presenta un conjunto de datos manualmente etiquetado para la tarea. La tarea definida para este trabajo ha mostrado ser una dirección prometedora para la implementación del primer paso del framework de aprendizaje de estructuras causales.

2. Segundo, se diseña e implementa un modelo predictivo basado en RNNs para la tarea de OED que incluye *embeddings* basados en *BERT* como atributos. El uso de *embeddings* basados en *BERT* para la tarea ha sido extensamente estudiado en este trabajo. Los resultados indican la gran utilidad de los *embeddings* sensibles al contexto para la tarea de OED. También se elaboran *baselines* para la tarea replicando un modelo del estado del arte [NG15], así como también un *baseline* basado en SVM. Ambos modelos son entrenados y evaluados con el conjunto de datos diseñado y anotado para la tarea de OED. El modelo propuesto, basado en RNN, supera los *baselines* evaluados.

3. Finalmente, se presenta un extenso estudio empírico con diferentes arquitecturas e hiperparámetros tanto para el *baseline* como para el modelo propuesto. También se realiza un estudio de ablación para medir el impacto de cada atributo en el desempeño. El código completo para todos los experimentos está disponible para permitir la reproducibilidad de este estudio.

3.2. Conceptos Base y Trabajos Relacionados

En el presente capítulo se define la tarea de detección de eventos en curso (OED) y se propone una metodología para abordar dicha tarea. Se diseña y publica un conjunto de datos para la tarea y se propone la utilización de *embeddings* contextuales para obtener un desempeño superador con respecto a métodos *baseline* existentes (replicados para la tarea). En esta sección se contrasta el conjunto de datos presentado con las colecciones de datos para la detección de eventos existentes, se menciona en qué se parece y en qué se diferencia la tarea de OED con tareas similares existentes y se discuten trabajos similares que incorporan la noción de eventos en curso o que mencionan la idea de información contextual.

El conjunto de datos más utilizado para la tarea de EE es el conjunto de datos *ACE 2005 Multilingual Training Corpus* [WSMM06], el cual contiene el conjunto completo de entrenamiento para la competencia *2005 Automatic Content Extraction (ACE) technology evaluation* [DMP+04] para los idiomas inglés, árabe y chino. Existen otros conjuntos de datos para EE y ED como, por ejemplo, *TimeML* [PCI+03] y *SentiFM* [JLH18]. Sin embargo, todos estos conjuntos de datos poseen una taxonomía fija de eventos y no cumplen con los requisitos necesarios para el trabajo aquí planteado.

Existen tres desventajas principales que aparecen al tratar de utilizar conjuntos de datos existentes de la literatura para la tarea de OED. Primero, como se mencionó previamente, estos conjuntos de datos limitan la anotación de eventos a un conjunto de posibles tipos de eventos fijados en una taxonomía, lo cual reduce el número de eventos presentes en el conjunto de datos. Este problema aparece porque en la tarea de EE se requieren detalles específicos sobre argumentos de eventos para cada nuevo tipo de evento que se agrega. Por ejemplo, si se quiere agregar el evento "casamiento" a una taxonomía se tiene que detallar que existen (además de un lugar y fecha) dos participantes. Mientras que si se quiere agregar el evento "bancarrota" se tiene que agregar solo el argumento de la empresa (además del lugar y fecha). Cada tipo de evento puede tener distintos tipos de argumento, por lo que los conjuntos de datos existentes deben predefinir todos los tipos de eventos posibles y sus argumentos. Sin embargo, para la tarea de ED en general, y en particular también para la tarea de OED, no es necesario considerar los argumentos ya que el objetivo solo es detectar los eventos, es decir, detectar los *event triggers*.

La segunda desventaja surge al utilizar un conjunto de datos existentes para un nuevo dominio. Al cambiar el dominio, la taxonomía y el origen de los artículos periodísticos usualmente no se preserva, y por ende el conjunto de datos no siempre se adapta exactamente al nuevo problema. La tercera desventaja aparece al tratar de utilizar conjuntos de datos existentes para ED o EE para la tarea de OED, dado que los enfoques existentes no consideran la diferencia entre eventos en curso de la misma manera que en la tarea de OED, y por lo tanto no se adaptan exactamente a las necesidades de este trabajo. Por estas tres desventajas se optó por crear un conjunto de datos específico para el desarrollo experimental de este capítulo.

El estado del arte para ED consiste en la generación de un modelo de clasificación (siendo los más recientes basado en redes neuronales) con una clase o categoría por cada posible tipo de evento y una categoría adicional para *no-eventos*. Tradicionalmente,

estos clasificadores usaban una gran variedad de atributos para representar información léxica, sintáctica o información sobre las entidades presentes en el texto. Típicamente estos atributos son el resultado de aplicar técnicas de NLP (por ejemplo, en [LJH13]). Con la llegada y popularización de las modelos de lenguaje (*language models*) basados en redes neuronales y la disponibilidad de representaciones continuas de palabras y oraciones aprendidas de manera no supervisada a partir de grandes colecciones de textos (*embeddings* de palabra y de oración), los atributos usados en los modelos de ED cambiaron radicalmente. Los atributos utilizados en la mayoría de los modelos actuales de ED son representaciones automáticas aprendidas de grandes colecciones de datos (como por ejemplo *Word2vec* [MSC+13] y *Fasttext* [BGJM17]), y en muchos de estos modelos estas representaciones siguen siendo actualizadas específicamente para la tarea, al ser entrenadas junto con el modelo mediante el método de descenso por el gradiente (*gradient descent*).

De acuerdo a [Bor18], los enfoques de EE caen en una de tres categorías: basados en patrones [KJRI91, HAT+92, Ril96a, Ril96b, YGTH00], basados en atributos [Fre98, CNL03, STA06, JG08, PR09, LG10, HR11, HZM+11, LJH13, BDL+15] y basados en redes neuronales [CXL+15, NG15, NCG16, NFCG16, FQL18]. Los primeros están basados en reglas o patrones creados manualmente, mientras que los segundos están basados en atributos léxicos, usualmente extraídos con herramientas de NLP. La tercera categoría utiliza redes neuronales tanto para la representación de los atributos como para la predicción. En este capítulo, nos concentramos en esta última categoría, que soluciona muchos de los problemas presentados por las primeras dos categorías, alcanzando resultados competitivos o superadores con respecto a las propuestas anteriores.

Los sistemas de EE típicamente siguen una de dos posibles arquitecturas: línea de montaje (*pipelined*) [CXL+15, JG08, LG10, HZM+11] o arquitectura paralela o conjunta (*joint*) [ZJS19, LLH18, SQCS18]. En la arquitectura de línea de montaje, la tarea de ED es el primer paso, que consiste en detectar y clasificar el *event triggers*. Posteriormente, el sistema lleva a cabo el resto de la tarea de EE que consiste en extraer los argumentos para esos *triggers*. En la arquitectura conjunta, la extracción de argumentos y la detección de *event triggers* se realizan en conjunto.

La noción de información contextual no es nueva para la tarea de ED. En [FQL18], los autores proponen un modelo RNN para capturar una noción simple de contexto para cada palabra, con el objetivo de mejorar el desempeño del modelo para la tarea. Este modelo

no utiliza como atributos *embeddings* contextuales, sino que utiliza *embeddings* clásicos y la noción de contexto viene dada por el estado interno de la RNN. En [FFSG20], los autores muestran cómo los *embeddings* basados en *BERT* pueden mejorar el desempeño de un modelo para la tarea de detección de eventos adversos a drogas (*Adverse Drug Event* (ADE)) en el dominio médico. En [TWC+20] se utilizan atributos basados en *embeddings* contextuales *BERT* junto con atributos de imagen para resolver la tarea de detección de eventos (ED) en noticias. Aunque estos enfoques no son aplicados al mismo dominio que el presentado en este trabajo o incorporan información no relacionada a textos, en ellos se puede encontrar evidencia que apoya la intuición presentada en este trabajo de que la información contextual es de vital importancia para la tarea de ED y OED.

Por otro lado, la noción de evento en curso como es definida en este capítulo — hasta donde se pudo investigar para este trabajo— ha sido introducida por primera vez aquí. En [HCJ+16], los autores usan clasificación de documentos para clasificar artículos periodísticos que contienen eventos en una de las posibles categorías: pasado, en curso, futuro, planificado, alerta futura, y futuro posible. Aunque ellos señalan la importancia de distinguir eventos en curso de pasados o futuros, su abordaje es diferente al aquí presentado. En dicho trabajo el enfoque es a nivel de documento, clasificando artículos periodísticos completos, mientras que en el presente trabajo se clasifican las palabras a nivel individual como *event-triggers* y *non-event-trigger*. Adicionalmente, el conjunto de datos introducido en dicho trabajo, *EventStatus*, está definido para un contexto diferente al presentado en este trabajo. *EventStatus* consiste de eventos de disturbios civiles, como por ejemplo protestas, demostraciones y huelgas.

3.3. Definición de la Tarea de Detección de Eventos en Curso (OED)

La tarea de OED es una tarea específica de ED cuyo propósito es detectar únicamente menciones de eventos en curso, en contraste con eventos históricos, futuros, hipotéticos u otras formas de eventos que no son están en curso ni son actuales. Descripciones de estados o situaciones actuales también son considerados eventos en curso, por ejemplo, una crisis o una recesión que está en curso es considerada un evento en curso, por más que no sea un evento que sucede en un lugar y tiempo específico (sino que es una combinación de características de un país o entidad). Tanto eventos puntuales que ocurren en un

lugar (por ejemplo incendios) como descripciones de estado (por ejemplo recesiones) son considerados eventos en curso en este trabajo.

En esta sección, se presentan las definiciones necesarias para la tarea de OED y posteriormente se define explícitamente la tarea de OED. A continuación, se presentan varios ejemplos ilustrativos para favorecer el entendimiento de la tarea. Todas estas definiciones, ejemplos y las pautas usadas por los anotadores para la construcción del conjunto de datos se pueden encontrar en la página oficial del conjunto de datos.[2]

Definición 3.1 (Evento en Curso) *Un evento en curso se identifica con cualquier fragmento de texto en una noticia que esté reportando un evento del mundo real que cumpla una de las siguientes condiciones: (i) se trata de un evento actual que acaba de comenzar, (ii) es un evento que comenzó tiempo atrás, pero sigue en curso, (iii) está describiendo el estado actual o situación de una entidad dada.*

Un ejemplo de (i) es un fragmento de un artículo periodístico que cubre un terremoto que acaba de suceder. Un ejemplo de (ii) es cuando un disturbio comenzó en una ciudad hace unos días y un artículo periodístico lo vuelve a mencionar mientras el disturbio aún está sucediendo (una nueva mención de un evento que ya está ocurriendo puede darse siempre que aparezca información nueva o se haga una recapitulación de lo que está sucediendo). Por último, un ejemplo de (iii) es cuando un artículo reporta una crisis o una recesión que está sucediendo en un país o región. Es importante notar que un fragmento de un artículo periodístico puede contener más de un evento, y en muchos casos estos eventos están relacionados o uno puede haber desencadenado al otro. Por ejemplo, una crisis en curso (un ejemplo de (iii)) puede desencadenar un disturbio (un ejemplo de (ii)).

El *event trigger* de un evento en curso y la tarea de detección de eventos en curso (OED) se definen análogamente a las definiciones de *event trigger* y la tarea de *Event Detection* (ED) [WSMM06], respectivamente, como sigue:

Definición 3.2 (*event trigger* de evento en curso) *El* event trigger *de un evento en curso es la* **palabra** *que más claramente expresa la ocurrencia del evento en curso (Definición 1).*

[2] `https://cs.uns.edu.ar/~mmaisonnave/resources/ED_data/`.

Definición 3.3 (Tarea de Detección de Eventos en Curso (OED)) *La tarea de detección de eventos en curso (OED) es la de la detección de* event triggers *de eventos en curso (Definición 2).*

Para la tarea de OED, el contexto de una palabra es crucial para determinar si se refiere a un evento en curso o no. Por ejemplo, consideremos la palabra "crisis" en las siguientes oraciones:

1. La *crisis* actual va a acelerar el desarrollo de la tecnología digital.
2. No habrá una *crisis* en el futuro previsible.
3. La misma tendencia se pudo observar durante la *Crisis* Financiera Global de hace más de una década.
4. Cualquier *crisis* financiera es catastrófica, y debemos reducir el riesgo de cualquier posible *crisis* futura.

Solo la referencia a "crisis" en la oración 1 es considerada un *event trigger* de un evento en curso, mientras que las otras menciones de la misma palabra no lo son. Esto surge de la necesidad de distinguir eventos en curso de eventos que no lo son, lo que constituye un requerimiento del objetivo final del trabajo de detectar relaciones causales entre eventos extraídos de noticias. La mayoría de los trabajos existentes no adoptan esta noción tan sensible al contexto del evento. Más aún, la definición aquí presentada considera como evento en curso al reporte del estado actual de un país o institución (por ejemplo, cuando se reporta que un estado o país está en recesión), lo que representa un evento no considerado en las propuestas y conjuntos de datos disponibles en la literatura.

Como otro ejemplo, consideremos el siguiente extracto de un artículo periodístico: "La devaluación no es una opción realista considerando el déficit actual ya que solo contribuiría a debilitar la credibilidad en las políticas económicas como sucedió en la última crisis." La única palabra en esta oración que puede ser un *event trigger* de evento en curso es la palabra "déficit" porque es la única que se refiere a un evento en curso. La palabra "devaluación" no es un *event trigger* de un evento en curso porque puede que nunca suceda. Similarmente, la palabra "debilitar" tampoco lo es porque se refiere a un evento hipotético. Finalmente, la palabra "crisis" no es considerada un *event trigger* de un evento actual porque se refiere a una crisis del pasado. Notar que las palabras "devaluación", "debilitar"

y "crisis" podrían haber sido *event triggers* de un evento en curso en otro fragmento de texto, pero solo si el contexto alrededor de esas palabras hubiera sido distinto, indicando que se trataba de eventos en curso. Esto señala la importancia del contexto para la detección de eventos en curso.

Detalles adicionales sobre la tarea de OED, incluyendo una descripción del sistema usado para asistir al proceso de etiquetado, pueden ser encontrados en las pautas utilizadas por los anotadores durante el proceso de etiquetado, disponibles en el sitio web oficial del conjunto de datos[3].

3.4. Configuración Experimental

3.4.1. El Conjunto de Datos

Para construir el conjunto de datos se procesó el conjunto de datos del *New York Times* (NYT) (1987-2007) [San08] usando la librería de procesamiento de lenguaje natural (NLP por sus siglas en inglés) de *spaCy*. La totalidad de los artículos fue dividida en oraciones usando la herramienta para esa tarea de la librería de *spaCy*. De la totalidad de las oraciones extraídas del corpus (∼64 millones), se seleccionó un subconjunto para ser etiquetadas manualmente.

Se seleccionaron tres episodios de crisis del mundo real: la crisis del peso mexicano de 1994, la crisis financiera de Rusia de 1998 y la crisis financiera de Asia de 1997. Se configuró el motor de búsqueda Lucene[4] (usando la configuración predeterminada) para buscar oraciones relacionadas a esos tres episodios. Se realizó la búsqueda utilizando palabras claves seleccionadas manualmente por expertos. Algunos ejemplos de estas palabras son: "Mexico", "crisis", "debt", "capital flights", y "devaluation". De los resultados obtenidos, se seleccionaron aleatoriamente 2.000 oraciones para formar parte del conjunto de entrenamiento. Además, se seleccionaron aleatoriamente un conjunto de 200 oraciones para ser usadas de test. La selección de episodios de crisis del mundo real tenía como fin elegir fragmentos de artículos periodísticos con una significativa cantidad de ejemplos de eventos y sucesos.

[3] https://cs.uns.edu.ar/∼mmaisonnave/resources/ED_data/

[4] https://lucene.apache.org/

Se prefirió utilizar oraciones en lugar de los textos completos para favorecer la diversidad en el conjunto de datos. Esto es debido a que el trabajo de anotar requiere mucho esfuerzo por parte de los anotadores, al etiquetar a nivel de oraciones en lugar de textos completos se alcanza una cobertura de diferentes eventos reportados en diferentes artículos en lugar de múltiples menciones del mismo evento mencionado múltiples veces en diferentes partes de un mismo artículo. Esta decisión de trabajar con oraciones en lugar de textos completos no constituye una simplificación de la tarea. Solo cambia la tarea por una diferente, con ventajas y desventajas propias. Por un lado, al tratarse de textos cortos los modelos pueden resultar más sencillos, no siendo necesario resolver largas dependencias entre partes muy alejadas del texto. Por otro lado, se tiene el problema de que no está presente todo el contexto. Puede haber información útil para la tarea de OED que puede no estar simplemente porque formaba parte de una oración previa. En este sentido, esta tarea así definida, comparte desafíos con la tarea de clasificación de textos cortos (*short-text classification*).

Dado que la tarea OED no está limitada a un conjunto predefinido de eventos cualquier evento del mundo real en curso reportado en un artículo periodístico podría ser considerado un evento en curso válido. En consecuencia, cada término del conjunto de datos fue etiquetado como *event-trigger* o *non-event-trigger* de eventos en curso, dando lugar a una tarea de clasificación binaria. Como se mencionó previamente, es importante distinguir aquellos eventos y situaciones que están en curso al momento en que la noticia es reportada de aquellos eventos que sucedieron tiempo atrás y son mencionados más tarde cuando el evento ya no está sucediendo. De la misma forma hay que distinguir eventos que están en curso en el mundo real, de eventos futuros, hipotéticos o eventos que no sucedieron. En el conjunto de datos utilizado aquí se anotan como *event-triggers* de evento en curso únicamente a aquellos eventos del mundo real que estén en curso. Basándose en este criterio, palabras que en otros contextos típicamente serían consideradas eventos pueden ser etiquetadas como *non-event triggers* debido a que no están en curso al momento en que son reportadas en el artículo periodístico. Esto es ilustrado por el ejemplo presentado en la Sección 3.1 con los diferentes ejemplos donde aparece la palabra “crisis”.

Para la creación del conjunto de datos se desarrolló una herramienta de aprendizaje activo (*active learning*) para asistir durante el proceso de anotación del conjunto de entrenamiento, pero no durante el proceso de anotación del conjunto de validación. La herramienta utilizada es un prototipo inicial del modelo RNN utilizado para predicción

de eventos (descrito en la Sección 3.4.2) usado para sugerir las etiquetas de las oraciones aún no etiquetadas por los usuarios. Cada oración fue presentada a los cuatro usuarios encargados del etiquetado junto con las correspondientes sugerencias generadas por el sistema. Los cuatro usuarios entonces debían llegar a un acuerdo respecto a qué palabras eran *event triggers* de eventos en curso, y cuáles no. Las sugerencias del sistema podían ser cambiadas en su totalidad: los *event triggers* sugeridos podían ser desmarcados y los *non-event-triggers* podían ser marcados como *event-triggers*.

Todo el proceso de construcción del conjunto de datos requirió de quince sesiones de aproximadamente 2 horas cada una. Como todo el proceso fue llevado a cabo usando una política de consenso, no se puede calcular, y por ende no se reporta, una métrica del nivel de acuerdo entre los anotadores.

Dado que inicialmente no había ninguna instancia etiquetada para entrenamiento, el proceso comenzó en frío (*cold-start*), esto es, las predicciones iniciales del modelo no eran mejor que predicciones aleatorias. A medida que nuevas instancias eran etiquetadas, el modelo era reentrenado. El proceso de reentrenamiento se ejecutaba cada 50 nuevas instancias etiquetadas.

Es importante resaltar que, para evitar sesgar la decisión de los usuarios al etiquetar el conjunto de test, el sistema presentaba a los anotadores las oraciones sin sugerencias durante todo el proceso de etiquetado de dicho conjunto. Los usuarios, al igual que antes, debían llegar a un consenso respecto a qué términos constituían *event-triggers* de eventos en curso y cuáles no, pero esta vez partiendo de oraciones sin sugerencias. Dado que el proceso de anotación fue tan costoso en términos de recursos humanos, el conjunto de datos generado y el conjunto de test resultantes son relativamente pequeños comparados con conjuntos de datos ya existentes. Sin embargo, para obtener resultados confiables, test estadísticos fueron aplicados sobre los resultados reportados en la Sección 3.5 para garantizar que las hipótesis puestas a prueba durante este trabajo sean estadísticamente significativas.

3.4.2. El Modelo RNN

Se diseñó, implementó y entrenó una RNN para la tarea de OED. Se llevaron a cabo varios experimentos combinando diferentes hiperparámetros, atributos y arquitecturas.

En esta sección, se repasan las diferentes configuraciones probadas y la intuición detrás de cada decisión de diseño tomada.

Atributos. Para llevar a cabo la detección de eventos en curso a partir de textos escritos en lenguaje natural, se plantea la hipótesis de que la información sintáctica, semántica y gramatical es necesaria. Para representar la semántica de cada *token* se utiliza la representación *Word2vec* [MCCD13] (W). Notemos que se eligió *Word2vec* en lugar de otras representaciones tradicionales (como por ejemplo *FastText* [BGJM17]) para permitir la comparación con el modelo *baseline* del estado del arte, ya que *Word2vec* es la representación utilizada en dicho modelo [NG15]. Las representaciones *Word2vec* son *embeddings* de 300 dimensiones preentrenados. Para estos experimentos se utilizan estas representaciones de base, pero las mismas son ajustadas a medida que el modelo es entrenado para la tarea de OED (la representación para los distintos *tokens* es actualizada junto con la tarea). También se prueban tensores sensibles al contexto que son provistos por la librería *spaCy* (Sp). Estos tensores son vectores de 96 dimensiones que se obtienen del estado interno de los modelos neuronales presentes en la librería de NLP *spaCy* al preprocesar textos[5]. Estos tensores no son ajustados durante el entrenamiento, por lo que la representación permanece fija durante todo el proceso.

Para codificar información sintáctica, se utilizó la librería de *spaCy* para identificar el árbol de dependencias (*dependency tree*) (D). Para cada *token*, se extrajo la relación de dependencia con el *token* cabeza (*head*) del árbol de dependencias. Usando esa información, se entrenó un *embedding* de 10 dimensiones usando la capa de *embeddings* de *Keras*. Esta capa comienza con pesos aleatorios y la representación es ajustada incrementalmente al entrenarse para la tarea de OED. La información gramatical fue codificada utilizando dos *embeddings* distintos, ambos construidos a partir de la herramienta para etiquetar partes de discurso (*Part-Of-Speech*) (POS). Se utilizaron dos versiones de esta herramienta de etiquetamiento provista por *spaCy*: la versión simplificada (P) y la versión detallada (T). Como se hizo para el caso de la información sintáctica, se utilizó una capa de *embeddings* de *Keras* de 10 dimensiones para representar la información, que es ajustada (entrenada) junto con el resto del modelo para la tarea de OED.

Se usó el reconocedor de entidades (*Named Entity Recognizer* (NER)) incluido en la librería de *spaCy* para extraer información de entidades que van a ser utilizadas en la tarea de OED en forma de otro atributo (E). Cada palabra fue representada con una etiqueta de

[5]https://spacy.io/

dos partes. La primera parte es una de las tres posibles letras de la notación IOB (dentro de entidad (I), fuera de entidad (O), comienzo de entidad (B)) y la segunda parte es el tipo de entidad. Como se hizo para los otros atributos (T, P, D), se utilizó una capa de *embeddings* de *Keras* para transformar un vector *one-hot* en un vector de 10 dimensiones. La capa es ajustada (entrenada) con el resto del modelo durante el entrenamiento para la tarea de OED.

Los *embeddings* contextuales, como *ELMo* [PNI+18] y *BERT* [DCLT18], fueron propuestos para solucionar el problema de las semánticas mezcladas. Esto sucede cuando un *embedding* no sensible al contexto tiene una única representación para una palabra con diferentes significados de acuerdo al contexto. Ambos significados están entonces mezclados en una sola representación, limitando la usabilidad de la representación. Estos nuevos *embeddings* contextuales han permitido alcanzar desempeños nunca vistos en una gran variedad de tareas de NLP. En este capítulo se plantea que contar con una representación sensible al contexto es crucial para la tarea de OED, permitiendo una mejora en el desempeño del modelo propuesto en comparación con los modelos *baseline* que no tienen este tipo de atributos. Para ilustrar esta hipótesis, consideremos las siguientes dos oraciones: "The firm had to fire employees" y "Fire burns a home near Brainerd airport" (se utilizan ejemplos de oraciones en inglés porque es el idioma del conjunto de datos usado en este trabajo). En ambas oraciones se utiliza la palabra "fire" escrita exactamente de la misma manera, pero con significados totalmente distintos. En las representaciones *Word2vec* ambas apariciones de la palabra tendrán asociado el mismo vector, lo que podría limitar el desempeño de un modelo de OED. Esto se vuelve especialmente problemático cuando la misma palabra puede corresponder o no a un evento en curso dependiendo del contexto (ver ejemplos de la Sección 3.1). Con esta motivación en mente, se incorporan *embeddings* contextuales preentrenados, en particular *BERT embeddings* (B), como un atributo adicional del modelo para la OED. Este vector se computa sumando las últimas 4 capas del modelo *BERT*, dando lugar a un vector de 768 dimensiones.

Un problema similar surge cuando la palabra es utilizada con una semántica similar (por ejemplo, las diferentes apariciones de la palabra "crisis" en los ejemplos de la Sección 3.1) pero dependiendo del contexto puede estar describiendo un evento en curso o no. En los ejemplos de la sección mencionada, se muestra que la palabra "crisis" siempre es usada con el mismo significado: "dificultad, situación complicada, difícil o inestable", pero en algunos casos, dependiendo el contexto, se puede referir a una crisis en curso, o a

Abr.	Tamaño	Descripción
W	300	*Embedding* de Palabras Preentrenado (*Word2vec*).
P	10	*Embedding* de las etiquetas generadas con *Part-Of-Speech tag* (versión simplificada).
T	10	*Embedding* de las etiquetas generadas con *Part-Of-Speech tag* (versión detallada)
D	10	*Embedding* de la etiqueta generada con el *Dependency parser*.
E	10	*Embedding* de la etiqueta de Entidades.
Sp	96	*Embedding* de palabra contextual (*spaCy*).
B	768	*Embedding* Contextual preentrenado de palabras (*BERT*).
S	768	*Embedding* Contextual preentrenado de oraciones (basado en *BERT*).

Tabla 3.1: **Atributos** usados en el modelo RNN propuesto.

otro tipo (crisis futura, o hipotética o crisis pasada). Siguiendo esta intuición, se agregó a cada *token* un vector contextual de 768 dimensiones representando un resumen de toda la oración. Notemos que cada *token* dentro de la misma oración tendrá el mismo vector de oración asociado. La intuición para usar un *embedding* de oración para cada *token* es debido a que, de acuerdo a la experiencia de los anotadores, usualmente una oración está dando información del pasado, hablando del futuro o un futuro posible (eventos no considerados en este trabajo) o la oración puede estar reportando un evento en curso (que es el objeto de este trabajo); es infrecuente que una oración mezcle estos dos tipos posibles de eventos. Basándose en esta experiencia de los anotadores, en este trabajo se incorpora un atributo más, que es una representación (*embeddings*) de oración (*S*) construida sumando los *embeddings BERT* de cada *token* de toda la oración. Los atributos *B* y *S* no son modelados con capas de *embeddings* de *Keras* y no son entrenados con el modelo durante la tarea de OED. Un resumen de los ocho atributos descritos previamente para el modelo RNN es presentado en la Tabla 3.1.

Se utiliza transferencia de aprendizaje (*transfer learning*) implícitamente en varios de los atributos del modelo. Primero, al usar *embeddings* de palabras que fueron preentrenados con grandes colecciones de datos de manera no supervisada (*Word2vec*, *spaCy*, *BERT*) se está incorporando información semántica extraída de esas grandes colecciones. Segundo, al incorporar una herramienta de NLP como un *Part-Of-Speech tagger* y un *Dependency tagger* se está incorporando información gramatical y sintáctica que proviene de los conjuntos de datos etiquetados que se utilizaron para entrenar estas herramientas.

Arquitectura. Se eligió una arquitectura RNN basada en celdas *Long-Short Term Memory* (LSTM) para explotar las dependencias entre los *tokens* siguientes y previos al

token actual que está siendo clasificado. Los datos de entrada para cada *token* son los ocho atributos (*embeddings*) previamente mencionados, estos son concatenados para formar un vector de 1962 dimensiones. Una capa de *Dropout* es agregada luego de esta capa de entrada. Posterior a la capa de *Dropout*, se agregó una capa de celdas LSTM Bidireccionales (Bi-LSTM) de 15 unidades ocultas. Seguida a la capa de Bi-LSTM se agregó una capa densa de una única unidad oculta para generar la predicción. La arquitectura del modelo RNN propuesto se describe en la Figura 3.2. Aunque esta arquitectura es la utilizada durante los experimentos en el conjunto reservado para test, arquitecturas adicionales (con otra cantidad de capas de Bi-LSTM y unidades ocultas) fueron investigadas. Los resultados de estos experimentos con diferentes arquitecturas se pueden encontrar en el Apéndice B.

Hiperparámetros. Para las capas de *embeddings* de *Keras* se utilizó la configuración predefinida, solo cambiando la inicialización aleatoria por los pesos preentrenados de los *embeddings Word2vec*. Se utilizó $p = 0,1$ para la capa de Dropout y se configuró la capa de Bi-LSTM con la configuración por defecto, agregando solo regularización L1L2 con valores de $0,001$ tanto para L1 como para L2. Para los experimentos finales sobre el conjunto reservado para test se utilizó la arquitectura de una capa de Bi-LSTM con 15 unidades ocultas porque fue la arquitectura con mejor desempeño durante los estudios preliminares reportados en el Apéndice B. Finalmente se utilizó la función *sigmoid* para la activación de la capa densa final.

Estadísticas sobre el conjunto de datos generado y sus atributos se presentan en las Tablas 3.2 y 3.3. Detalles adicionales sobre cómo fue construido el conjunto de datos pueden ser encontrados en el sitio oficial del conjunto de datos(`https://cs.uns.edu.ar/~mmaisonnave/resources/ED_data/`).

3.4.3. Modelos *Baseline*

Como no existen modelos preexistentes que hayan sido probados sobre el conjunto de datos utilizado, no existen métricas o resultados del área de ED que puedan ser directamente comparables con este trabajo. Por esta razón, se replicó un modelo de ED del estado del arte para ser usado como *baseline* así como también se creó un modelo basado en una técnica clásica (SVM) para ser usado como un segundo *baseline*. Dado que el foco de este trabajo es la tarea de detección de eventos (ED), en particular la tarea de OED,

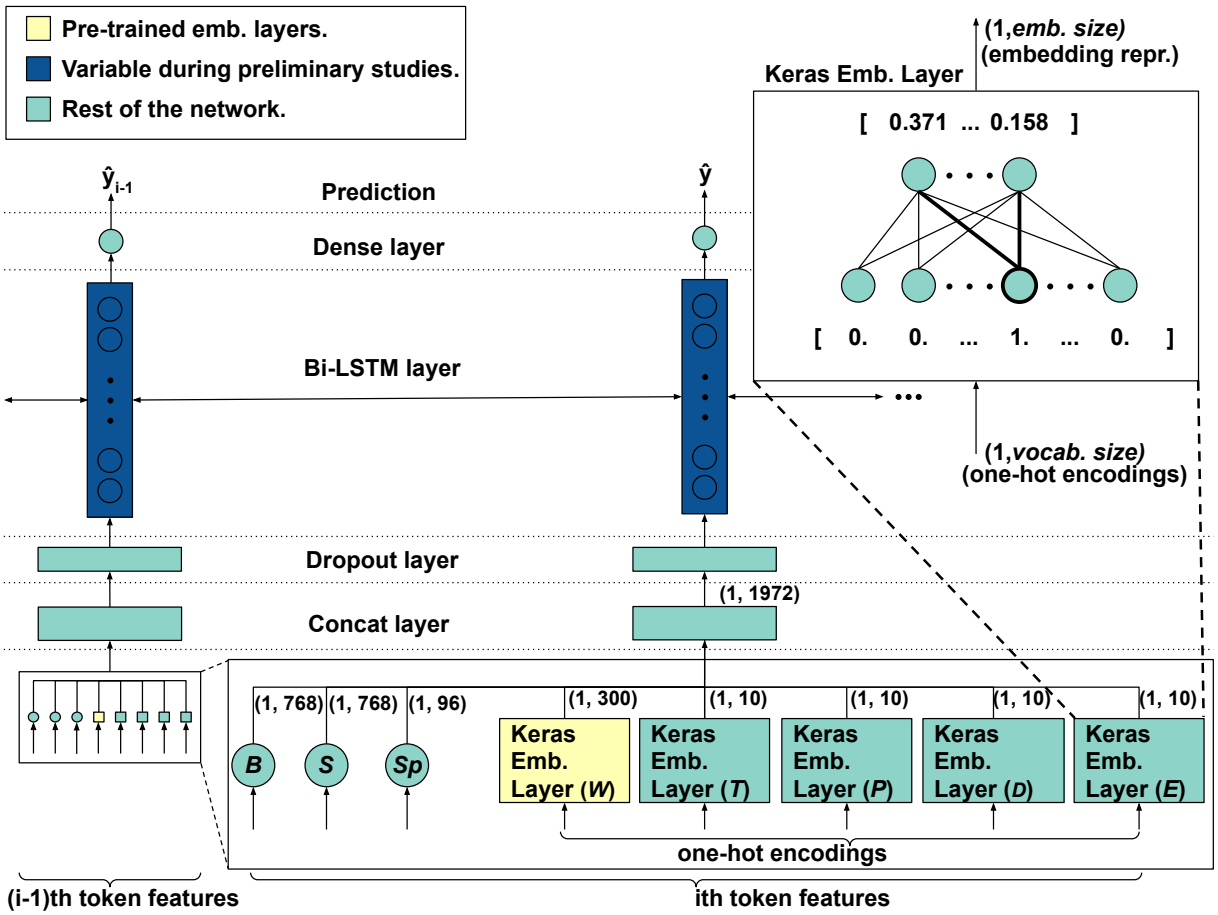


Figura 3.2: **Esquema de la arquitectura del modelo RNN propuesto**. La arquitectura se detalla de abajo hacia arriba. Primero se muestra una representación de los ocho atributos de entrada, para el *token* i-ésimo y el $(i-1)$-ésimo. Para cada *token*, 3 de los atributos de entrada son representaciones vectoriales (B, S, Sp); los otros 5 son codificaciones *one-hot* que son suministradas como entrada a 5 capas de *embeddings* de *Keras*. Cada una de estas capas de *embeddings* están implementadas como capas densas como se muestra en la esquina superior derecha. Una de estas capas arranca con pesos preentrenados (W), las otras comienzan con pesos aleatorios. Luego de las capas de *embeddigs*, los ocho vectores son concatenados en un solo vector de 1.972 dimensiones. A este vector se le aplica una capa de *Dropout*. Durante los estudios preliminares se probaron diferente cantidad de capas Bi-LSTM y unidades ocultas siguiendo a la capa de *Dropout*. Para los experimentos sobre el conjunto de datos reservado para test se utilizó solo una capa de Bi-LSTM con 15 unidades ocultas (como se muestra en la figura). La capa final es una capa densa, posterior a la de Bi-LSTM, que tiene una unidad oculta. Esta última capa es la que computa la predicción.

se evaluaron modelos para ED y no para ED+EE. Vale remarcar que no se lo compara con modelos existentes para OED porque dicha tarea es definida por primera vez en este trabajo. Basándose en estas consideraciones, y por su simplicidad y desempeño compara-

Conjunto de Datos Completo (entrenamiento/validación + test)	
Cantidad de Oraciones	2.200
Tamaño del Vocabulario	8.647
Tamaño del Vocabulario de Entidades (E)	34
Tamaño del Vocabulario generado con el *Part-Of-Speech Simplified* (P)	16
Tamaño del Vocabulario generado con el *Dependency Parser (D)*	47
Tamaño del Vocabulario generado con el *Part-Of-Speech Detailed (T)*	47

Tabla 3.2: Estadísticas acerca de los vocabularios del conjunto de datos manualmente anotado para la tarea de OED.

ble con otros modelos del estado del arte de ED, se eligió el modelo de Redes Neuronales Convolucionales (*Convolutional Neural Network* (CNN)) presentado en [NG15] para ser usado como *baseline* de comparación (además del modelo SVM). A pesar de haber sido propuesto en 2015, este modelo es competitivo con otros métodos más recientes del estado del arte. Por ejemplo, dicho modelo presenta un F1-score de solo $4,1$ puntos porcentuales por debajo del modelo presentado en [NG18]. De acuerdo a los autores, el modelo *baseline* elegido presenta un F1-score de 69 % para la tarea de ED usando las anotaciones de entidades reales (*gold-standard entity annotations*), mientras que el modelo presentado en [NG18] reporta un F1-score de $73,1$ % sobre el mismo conjunto de datos (ACE 2005 Corpus) usando las mismas anotaciones de entidades. Los autores no reportan en [NG18] los resultados del modelo con las entidades predichas. Sin embargo, los autores de [NG15] sí reportan los resultados de su modelo, que es el *baseline* de este trabajo, utilizando las entidades predichas en lugar de las entidades *gold-standard*. Para las entidades predichas, reportan un F1-score de $67,6$ %.

Métrica	entrenamiento/validación		test	
	Total	Promedio por Oración	Total	Promedio por Oración
Cantidad de *tokens*	76,629	38,31	7,382	36,91
Cantidad de palabras	67,032	33,52	6,442	32,21
Cantidad de Entidades	11,502	5,75	950	4,75
Cantidad de Eventos	5,119	2,56	416	2,08

Tabla 3.3: Número total de *tokens*, palabras, entidades y eventos encontrados en el conjunto de datos manualmente anotado para la tarea de OED.

Como se mencionó previamente, para presentar comparaciones adicionales se decidió agregar también como *baseline* un modelo predictivo basado en un enfoque clásico. Para esto, se utilizó el clasificador SVM de la librería *scikit-learn*[6] versión 0.22, usando las representaciones *Word2vec* (W) como atributos del modelo. Se utilizó la configuración por defecto de dicho clasificador y se probaron cuatro *kernels* diferentes (*linear*, *polynomial*, *sigmoid* y *RBF*).

El modelo CNN de [NG15] fue replicado tan fielmente como fue posible. Sin embargo, aunque se intentó replicarlo exactamente como estaba reportado en el trabajo original, para poder adaptarlo al conjunto de datos de este trabajo, se tuvieron que hacer cambios menores. Más aún, como el código no estaba disponible, algunos detalles de implementación pueden diferir del modelo original. Algunas decisiones debieron tomarse sobre aspectos y configuraciones del modelo que no estaban explícitos en el trabajo original. Por ejemplo, se tuvo que decidir qué herramienta de NLP usar para la extracción de entidades. En nuestro trabajo, se utilizó la librería *spaCy* para todas las tareas de NLP. En el resto de esta sección, se describe en detalle el modelo CNN usado. También se describen algunos cambios menores y decisiones que se tuvieron que tomar, explicando la intuición detrás de cada decisión.

Atributos del modelo CNN. Como se hizo en el trabajo original, se usaron tres atributos de entrada para el modelo *baseline* (CNN). Primero, se usaron *embeddings Word2vec* (los mismos que se usaron en el modelo RNN propuesto) (W). Segundo, se usaron *embeddings* de entidades (E). Para construir estos *embeddings*, se utilizó el reconocedor de entidades de la librería de *spaCy* y la capa de *embeddings* de *Keras*; el reconocedor se utilizó para obtener la información categórica (las entidades) y la capa de *Keras* se utilizó para transformar esta información categórica en vectores. Como no hay mención de una herramienta de NLP especifica en el trabajo original, se eligió el etiquetador de entidades de *spaCy* por su buen desempeño en múltiples tareas de NLP. Los atributos W y E se construyen de la misma forma que para el modelo propuesto (RNN). El último atributo usado para el modelo base es un *embedding* de posiciones (Po), que representa la posición relativa de cada palabra con respecto al *token* actual bajo clasificación. La capa de *embeddings* de palabra (W) comienza con la representación aprendida de los *embeddings* pre-entrenados *Word2vec*, y las otras dos capas arrancan con pesos aleatorias, como se

[6]https://scikit-learn.org/stable/index.html

reportó en [NG15]. Las tres capas son actualizadas por descenso de gradiente junto con el entrenamiento para la tarea de OED.

Arquitectura del modelo CNN. La arquitectura usada en [NG15] es una arquitectura de una sola capa convolucional, seguida por una capa de *Max-Pooling*, seguida por una capa de *Dropout*, con una última capa densa para la predicción. Los atributos de entrada son tres vectores recuperados de tres tablas de *look-up*. Cada palabra, etiqueta de entidad y posición de cada vocabulario tiene una representación en las tablas *look-up*. Las representaciones de estas tablas son ajustadas (entrenadas) durante el entrenamiento usando descenso por el gradiente. Se replicó la misma arquitectura y comportamiento, reemplazando las tablas por capas de *embeddings* de *Keras* que cumplen el mismo propósito: almacenar y proveer representaciones vectorizadas y ajustar las representaciones mientras se entrena para la tarea de clasificación. El resto de la red se mantiene igual a la reportada en [NG15]. En la Figura 3.3 se presenta un esquema de la arquitectura del modelo *baseline* (CNN).

Hiperparámetros del modelo CNN. Aunque muchos de los hiperparámetros se mantienen como en el trabajo original, dado que el conjunto de datos usado es distinto al usado en [NG15], se debió ajustar el tamaño de la ventana para que se adapte mejor al conjunto de datos aquí usado. Dado que, en el conjunto de datos usado para este trabajo, cada ítem de dato es una oración y no un documento completo como en el conjunto de datos ACE 2005 (el conjunto usado en [NG15]), se debió utilizar un tamaño de ventana menor. Para encontrar un buen tamaño de ventana se probaron varios tamaños menores al usado en el trabajo original (tamaño de ventana 31): 1, 3, 5, 11, 21 y también por comparación se probó el tamaño original, 31. Se utilizaron el mismo número de filtros para la capa convolucional (150), y el mismo tamaño de filtros (2, 3, 4 y 5) que en el trabajo original. Se utilizó la función *sigmoid* como función de activación para la capa densa final para poder usar las mismas métricas usadas en el modelo RNN. Se utilizó, así como en el trabajo original, un tamaño de *batch* de 50, con una probabilidad para la capa de *Dropout* de $0,5$. Se utilizó entropía cruzada binaria como función de coste (*binary cross-entropy loss function*) y el optimizador de Adam.

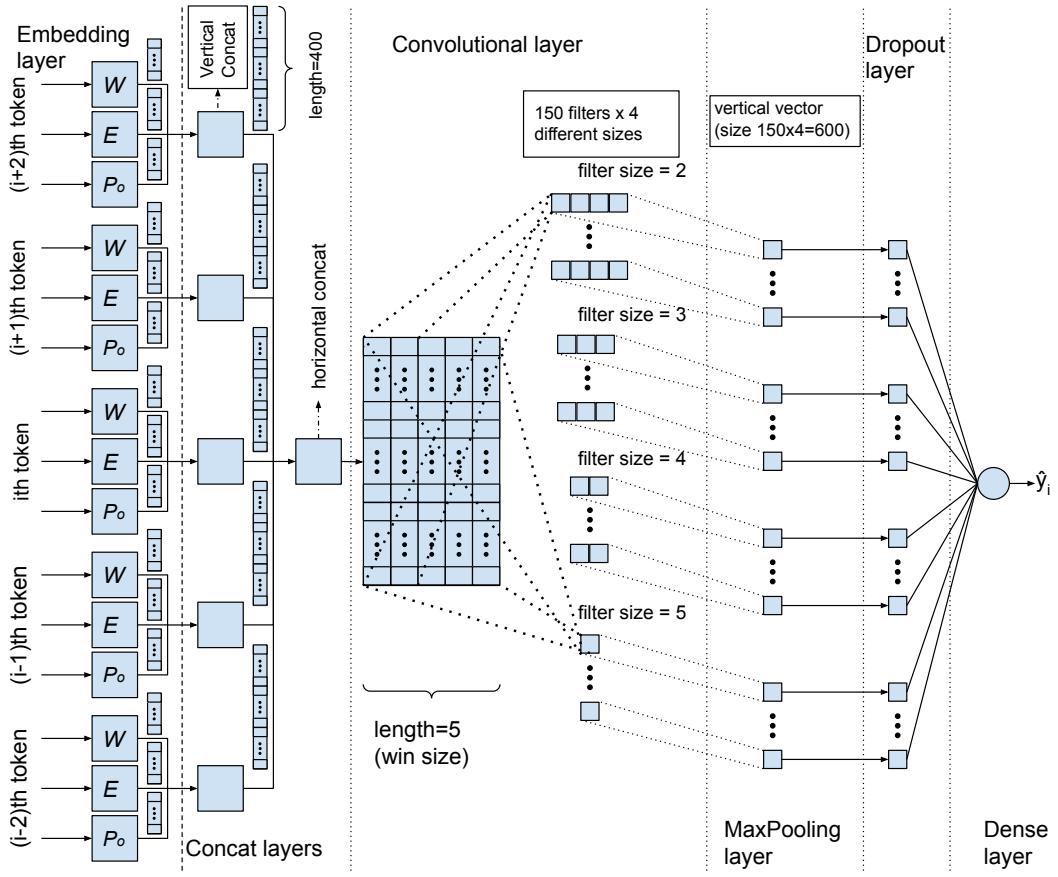


Figura 3.3: **Esquema de la arquitectura del modelo *baseline* (CNN)** con tamaño de ventana 5. La descripción se detalla de izquierda a derecha. (a) Atributos de entrada. Una representación de los tres atributos de entrada diferentes (*embeddings* de palabra W, *embeddings* de entidad E y *embeddings* de posición Po) para los cinco *tokens* en la ventana. Cada entrada es una representación *one-hot* de información categórica. (b) Capas de *embeddings*. Las representaciones *one-hot* son la entrada a capas de *embeddings* de *Keras*. (c) Capa de concatenación. Se concatena el *embedding* resultante en una vector de 400 dimensiones que representa toda la información extraída de cada *token*. Otra capa de concatenación apila los vectores de 400 dimensiones para cada una de las cinco palabras de la ventana, generando una matriz. (d) Una capa convolucional. Se aplican 150 filtros para cada uno de los diferentes tamaños de filtro (2, 3, 4 y 5), dando un total de 600 filtros aplicados. (e) Capa de Max-Pooling. Una capa de Max-Pooling se aplica a la salida de la capa convolucional. (f) Capas de *Dropout* y Densa. Finalmente, las últimas dos capas son una capa de *Dropout* a la salida de la de Max-Pooling y una capa densa a continuación de la de *Dropout*. La capa densa final es la última capa de la red y es la que realiza la predicción.

3.5. Resultados y Discusión

En esta sección, se presentan los resultados y discusiones para las diferentes variantes del modelo propuesto (RNN) y los modelos *baselines* (CNN y SVM). Para cada variante de cada modelo (excluyendo los modelos SVM), se separó el conjunto de entrenamiento+validación en diferentes particiones de entrenamiento y validación. Para cada variante de cada modelo, se entrenó el modelo hasta que no se vieron mejoras en el desempeño de la métrica F1-score sobre el conjunto de validación durante 400 épocas seguidas (*early stopping* con coeficiente de paciencia 400). Se utilizaron semillas consecutivas desde 1 hasta 10 para garantizar reproducibilidad. Se seleccionan los pesos de la instancia (*checkpoint*) con el mejor desempeño sobre el conjunto de validación (de todos los pesos generados durante todas las épocas) para ser usados para predicción sobre el conjunto de test. La instancia (*checkpoint*) con el mejor desempeño sobre el conjunto de validación es seleccionada para hacer la predicción sobre el conjunto de test. En el caso de los modelos basados en SVM, como no fue necesario *early stopping* (porque el modelo no es estocástico), se usó el conjunto de entrenamiento+validación todo junto para entrenamiento. Por el mismo motivo, para los modelos basados en SVM solo se realizó un entrenamiento del modelo sobre el conjunto entrenamiento+validación y se reportan las métricas sobre el conjunto de test usando los cuatro *kernels* distintos. También, por lo antes mencionado, para los modelos basados en SVM se reportan las métricas del único modelo en lugar de valores promediados de varios modelos con diferentes semillas (como es el caso para los modelos RNN y CNN). Es importante resaltar que se utilizó el mismo conjunto reservado para test para reportar el desempeño de todos los modelos. Dicho conjunto de datos nunca fue utilizado para ninguna etapa de entrenamiento.

En esta sección, se presentan y discuten los resultados del desempeño de cada modelo CNN y RNN sobre el conjunto de validación (el usado para *early stopping*) y el conjunto reservado para test. De la misma forma, en esta sección, se presentan y discuten los resultados del desempeño de cada modelo basado en SVM sobre el conjunto de entrenamiento+validación y el conjunto reservado para test. Se reportan resultados para un total de quince variantes de modelos distintos. Para seleccionar estas variantes se realizaron varios experimentos preliminares. Estos experimentos preliminares y sus discusiones son presentados en el Apéndice B. Para cada variante basada en redes neuronales, se reportó el promedio de las métricas de 10 ensayos con las 10 semillas consecutivas. Las métricas reportadas son sensibilidad (sens), especificidad (spec) y la media armónica entre estos

dos valores, esto es, el F1-score (F1). No se reporta la exactitud del modelo (*accuracy*), ya que, al tratarse de un conjunto de datos altamente desbalanceado ($93,41\,\%$ de los *tokens* son no eventos), la exactitud es una métrica que tiene una interpretación no tan directa y puede derivar en conclusiones erróneas.

Para analizar en profundidad los resultados obtenidos en términos de F1-score, se reportan los intervalos de confianza (IC) con un nivel de confianza del $95\,\%$ y el p-valor de un test estadístico t (*t-test*) entre el modelo de cada fila de una tabla contra el mejor modelo de esa tabla. Por ejemplo, en la primera fila de la Tabla 3.4, se reporta el p-valor de un t-test a una sola cola entre el F1-score del modelo de esa fila (modelo 1) y el mejor modelo de esa tabla (modelo 7). Se utilizó la métrica F1-score obtenida de evaluar el modelo sobre el conjunto reservado para test, no la obtenida de los conjuntos de validación.

Dado que los modelos SVM no son estocásticos, solo se realizó un entrenamiento del modelo sobre todo el conjunto de entrenamiento+validación y se computaron las métricas sobre el conjunto de entrenamiento usado (el conjunto entrenamiento+validación) y sobre el conjunto reservado para el test. Todo esto se realizó para cada uno de los cuatro *kernels* reportados. Debido a que solo se reporta un modelo por cada kernel (no múltiples modelos con diferentes semillas) solo se reporta un único valor por métrica por modelo (y no un promedio como para los modelos basados en redes neuronales). Debido a esto no se pueden realizar ni reportar test estadísticos.

3.5.1. Experimentos Realizados sobre el Conjunto de Test

Para el modelo RNN propuesto, se utilizaron siete conjuntos de atributos distintos (modelos 1 al 7). Para cada uno de estos modelos se realizaron diez repeticiones usando diferentes semillas. Las métricas promediadas a lo largo de las diez repeticiones se reportan en la Tabla 3.4. Aunque se probaron diferentes números de capas Bi-LSTM y diferente cantidad de unidades ocultas, sobre el conjunto reservado para el test solo se usó la arquitectura que obtuvo el mejor desempeño durante los estudios preliminares, esto es, la arquitectura con una sola capa Bi-LSTM con 15 unidades ocultas. Los resultados de esta arquitectura se reportan en la Tabla 3.4.

Se eligieron estas 7 variantes para los atributos de entrada de los modelos RNN por dos razones principales. Primero, para probar la hipótesis de que los *embeddings* de palabras

contextuales (B) tienen un impacto positivo significativo en el desempeño. Esta hipótesis, además de haber sido puesta a prueba con estos experimentos, también se la puso a prueba durante los experimentos preliminares (discutidos en el Apéndice B). La caída significativa en el desempeño (de $11,7$ puntos porcentuales) de incluir a excluir los *embeddings* de palabras contextuales (modelo 1 versus modelo 2) es un indicativo de la importancia de este atributo para la tarea de OED. En segundo lugar, se realizaron el resto de los experimentos (modelos 3 a 7) a manera de estudio de ablación para evaluar el impacto de los atributos propuestos sobre el desempeño general. Se estudió la contribución de cada atributo durante los estudios preliminares, y durante los estudios en el conjunto reservado para test se buscó más evidencia para apoyar la hipótesis de que varios atributos tenían un impacto insignificante en el desempeño una vez que otros estaban presentes.

Estos resultados permitieron simplificar el modelo y así obtener incluso mejores resultados. El modelo 7, el cual excluye todos los atributos excepto los *embeddings* contextuales de palabra y oración, es el modelo con el mejor desempeño. Este resultado muestra que los otros atributos pasan a ser irrelevantes para la tarea de OED en presencia de estos dos, y que agregarlos solo añade complejidad y por ende ruido al modelo. Los tests estadísticos entre el modelo 7 y los otros, que se pueden apreciar en la columna de los p-valores de la Tabla 3.4, muestran que las diferencias son todas significativas con un nivel de confianza de al menos $90\,\%$ (p-valores menores que $0,1$).

Del modelo *baseline* CNN se realizaron 10 repeticiones para cada uno de los diferentes tamaños de ventana. En este estudio se analizan dos tamaños de ventana 1 y 11 (otros tamaños de ventana se analizan en estudios preliminares disponibles en el Apéndice B). Para cada uno de los dos tamaños de ventana se hicieron experimentos incluyendo y excluyendo los *embeddings* de entidad, pero siempre usando los *embeddings* de palabra. Para tamaño de ventana 1, se excluye el *embedding* de posición. El resultado de estos experimentos se presentan en la Tabla 3.5. El mejor modelo *baseline* CNN (modelo 8) muestra un F1-score de $0,575$ sobre el conjunto de test, que es significativamente menor que el desempeño de todos los modelos RNN propuestos, excepto el que excluye los *embeddings* contextuales de palabras (modelo 2).

Para comparar los modelos RNN y el mejor modelo *baseline* CNN, se evaluó el p-valor de un t-test a una cola entre el F1-score promedio entre los mejores y peores modelos RNN (modelo 2 y 7, respectivamente) contra los dos mejores modelos CNN (modelo 8 y 9). El modelo 7 es significativamente mejor que todos los otros modelos, presentando un

Modelo	Atributos	validación			test				
		$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	F1 IC	p-valor
1	all	0,696	0,945	0,712	0,667	0,931	0,676	± 0,026	0,024
2	all-$\{B\}$	0,545	0,957	0,591	0,522	0,948	0,559	± 0,046	0,000
3	all-$\{T,P,D\}$	0,677	0,949	0,697	0,654	0,935	0,666	± 0,031	0,012
4	all-$\{T,P,D,W\}$	0,703	0,948	0,718	0,686	0,935	0,690	± 0,018	0,076
5	all-$\{T,P,D,Sp\}$	0,689	0,945	0,706	0,672	0,931	0,683	± 0,013	0,007
6	all-$\{T,P,D,E\}$	0,686	0,949	0,706	0,662	0,936	0,675	± 0,023	0,012
7	all-$\{T,P,D,Sp,W,E\}$	0,726	0,943	**0,734**	0,706	0,928	**0,704**	± 0,012	—

Tabla 3.4: Desempeño promedio de los modelos basados en **RNN** sobre el conjunto de validación y el conjunto de test para cada una de las siete variantes que utilizan diferentes conjuntos de atributos y siempre la misma arquitectura: una capa Bi-LSTM con 15 unidades ocultas. Los diferentes conjuntos de atributos usados en cada modelo se indican a través de los atributos eliminados del conjunto total de atributos (all), donde todo el conjunto de atributos es $\{T, P, D, Sp, W, E, B, S\}$. Se adoptó esta notación para simplificar la interpretación de los experimentos como un estudio de ablación (comparando el modelo con todos los atributos contra modelos que tienen alguno de los atributos eliminados). El conjunto de entrenamiento es usado para entrenar y el de validación para hacer *early stopping*, mientras que el conjunto de test está reservado solo para evaluar las métricas (sin ser usado para entrenar en ninguna etapa). Las métricas en el conjunto de entrenamiento son omitidas del presente análisis y no se reportan.

p-valor de $0,0$ para todos los t-test. El t-test entre los modelos 8 y 9 da un p-valor de $0,384$. Este valor elevado indica que la diferencia de desempeño entre estos dos modelos no es estadísticamente significativa, lo que apunta al hecho de que los *embeddings* de entidad (E) tienen un impacto despreciable cuando los otros atributos están presentes. La diferencia en desempeño entre el peor modelo RNN (modelo 2) con respecto a los mejores modelos CNN (modelos 8 y 9) no es estadísticamente significativa (p-valor de $0,264$ y $0,344$, respectivamente).

Para el modelo *baseline* basado en SVM se realizó solo un entrenamiento sobre el conjunto de entrenamiento+validación y se computaron las métricas sobre este conjunto de datos y el conjunto reservado para el test para cada uno de los *kernels* usados: *linear*, *polynomial*, *sigmoid* y *RBF*. Los resultados de estos experimentos se presentan en la Tabla 3.6. El mejor modelo SVM (modelo 15) muestra un F1-score de $0,564$ en el conjunto reservado para el test. Este modelo tiene un desempeño significativamente menor que todos los modelos basados en RNN excepto el modelo que excluye los *embeddings* contextuales

de palabra (modelo 2). Por otro lado, su desempeño es levemente inferior al obtenido por los dos mejores modelos basados en CNN (modelos 8 y 9).

3.5.2. Discusión

En esta sección, se revisa y discute el comportamiento de cuatro modelos representativos elegidos. Los modelos seleccionados para esta discusión son el modelo 2 (el modelo RNN con todos los atributos excepto los *embeddings* contextuales de palabra), el modelo 7 (el mejor modelo RNN), el modelo 8 (mejor modelo CNN) y el modelo 15 (mejor modelo SVM).

Se plantea la hipótesis de que los *embeddings* contextuales de palabra son un factor crucial para obtener un aumento de desempeño en el modelo propuesto. Para probar esta hipótesis, en el modelo 2, se excluyen solo los *embeddings* contextuales de palabra para medir el impacto de estos sobre el desempeño. Como el modelo no tiene este atributo, se espera un desempeño pobre en comparación con el modelo 7 (que tiene solo los *embeddings* contextuales de palabra y de oración). El mal desempeño obtenido por el modelo 2, que es menor incluso que los mejores modelos *baselines*, provee evidencia que apoya nuestra hipótesis. Más aún, el desempeño superador alcanzado por el modelo 7 (que consiste solo de *embeddings* contextuales B y S como atributos) indica que la inclusión de otros atributos (W, Sp, T, P, D, E) no resulta útil para la tarea de OED una vez que los *embeddings* contextuales están incluidos. También es interesante notar que el modelo 15,

Modelo	Tam. Ventana	Atributos	validación			test				
			$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	F1 IC	p-valor
8	1	$\{W,E\}$	0,477	0,971	0,570	0,499	0,968	**0,575**	± 0,031	—
9	1	$\{W\}$	0,472	0,972	0,565	0,493	0,969	0,569	± 0,031	0,384
10	11	$\{W,E,Po\}$	0,507	0,963	0,596	0,326	0,960	0,394	± 0,028	0,000
11	11	$\{W,Po\}$	0,499	0,965	0,589	0,345	0,942	0,406	± 0,035	0,000

Tabla 3.5: Desempeño promedio de los modelos basados en **CNN** sobre el conjunto de validación y el conjunto de test para las cuatro variantes (dos conjuntos de atributos distintos para cada uno de los dos tamaños de ventana estudiados). El conjunto de entrenamiento es usado para entrenar y el de validación para hacer *early stopping*, mientras que el conjunto de test está reservado solo para evaluar las métricas (sin ser usado para entrenar en ninguna etapa). Las métricas en el conjunto de entrenamiento son omitidas del presente análisis y no se reportan.

Modelo	Kernel	entrenamiento + validación			test		
		sens	*spec*	$F1$	*sens*	*spec*	$F1$
12	linear	0,247	0,995	0,395	0,231	0,994	0,375
13	polynomial	0,494	0,993	0,660	0,380	0,980	0,548
14	RBF	0,459	0,993	0,628	0,346	0,984	0,512
15	*sigmoid*	0,372	0,957	0,536	0,401	0,949	**0,564**

Tabla 3.6: Desempeño de los modelos basados en **SVM** sobre el conjunto de entrenamiento+validación y el conjunto de test para cada una de las cuatro variantes (una por cada *kernel* usado). Como SVM tiene un entrenamiento determinístico no hay necesidad de early stopping, por lo que no fue necesario hacer diferentes particiones de entrenamiento y validación. Se utilizaron ambos conjuntos juntos (entrenamiento+validación) para la etapa de entrenamiento. También por la naturaleza determinística de SVM no fueron necesarias múltiples repeticiones y por eso se reportan las métricas de esa única repetición, no se reportan promedios.

que tiene un *kernel sigmoid*, obtiene un F1-score de $0,564$ sobre el conjunto de test, superando incluso algunos modelos basados en redes neuronales (modelos 2, 10 y 11).

Se derivan dos conclusiones principales de los resultados discutidos. Primero, en una tarea como la OED, que se entrena y evalúa sobre un conjunto relativamente pequeño de datos, el uso de atributos pre-entrenados en conjuntos grandes de datos es crucial para obtener un aumento de desempeño. Estas conclusiones se pueden extraer del hecho de que los *embeddings* contextuales basados en BERT (B y S) y los *embeddings* de palabra *Word2vec* (W) tuvieron un impacto positivo muy significativo en el desempeño de los modelos 7 y 15, respectivamente. Esto provee evidencia adicional para apoyar el ya conocido hecho de que el uso de transferencia de aprendizaje (*transfer learning*) en tareas con pocos datos puede contribuir significativamente a mejorar el desempeño. Segundo, en la presencia de un conjunto de datos pequeño, la elección de atributos y en particular el uso de *transfer learning* es más importante que tener un modelo grande o complejo. Esto se deduce del hecho de que un modelo clásico (modelo 15) que se apoya en el uso de atributos preentrenados tiene desempeño comparable a modelos más complejos basados en redes neuronales.

3.6. Conclusiones

La contribución principal de este capítulo es la definición de la tarea OED y una extensa evaluación experimental de la misma, que combina diferentes modelos y atributos. El mejor de los modelos propuestos, basado en una arquitectura RNN y usando *embeddings* contextuales basados en *BERT* de palabra y de oración como atributos, muestra una mejora de $12,9$ puntos porcentuales en el F1-score con respecto al mejor modelo *baseline* en el conjunto de datos reservado para test. Considerando que algunos de los enfoques para resolver ED solo mejoran nuestro *baseline* por $4,1$ puntos porcentuales [NG18], se tiene evidencia suficiente para suponer que el modelo propuesto alcanza desempeños competitivos con el estado del arte, incluso superando modelos más avanzados. Vale la pena mencionar que el código y datos para replicar nuestro modelo son puestos a disposición de manera *online*.

Dos conclusiones principales se derivan del extenso análisis presentado en este capítulo. Primero, que los *embeddings* contextuales resultaron más adecuados para la tarea de OED que otros *embeddings* y atributos analizados. En particular, algunos atributos mostraron tener un impacto insignificante en el desempeño ante la presencia de otros atributos en el modelo. Por ejemplo, información gramatical (capturada por *Part-Of-Speech taggers* y *Dependency Parser taggers*), información de entidades y *embeddings* no contextuales mostraron no tener impacto positivo una vez que los *embeddings* contextuales estaban incluidos en el modelo. La ausencia de un impacto positivo no implica necesariamente que los atributos no tengan información útil para la tarea, sino que también se puede deber a que esta información ya está (al menos parcialmente) capturada por los *embeddings* contextuales. Además, el análisis de modelos *baselines* clásicos mostró que, en este contexto de conjuntos de datos relativamente pequeños, a veces es más importante utilizar los atributos correctos antes que tener un modelo grande o complejo.

La segunda conclusión que se deriva de los resultados de este capítulo es que a pesar de que el modelo propuesto es más adecuado para texto (por estar basado en RNNs) que los *baselines* (basados en CNNs y SVMs), la mayor diferencia en desempeño está dada por los atributos usados, en particular por los *embeddings* contextuales *BERT* y no tanto por el modelo usado.

Otra importante contribución es la construcción de un conjunto de datos público para la tarea de OED. El proceso de anotación del conjunto de datos fue asistido por una

herramienta de aprendizaje activo. El conjunto de datos resultante es particularmente útil para la tarea de OED que se concentra en eventos en curso exclusivamente. Difiere de otros conjuntos de datos para ED y EE existentes ya que no depende de una taxonomía fija de eventos predefinida.

Como parte del trabajo futuro se propone evaluar el impacto de los *embeddings* contextuales en otros dominios de ED (usando otros modelos y conjuntos de datos).

Capítulo 4

Aprendizaje Causal y su Aplicación a Textos

Resumen

El creciente volumen de información textual disponible abre nuevas posibilidades para el análisis de diferentes episodios del mundo real (análisis del precedente, desarrollo y secuelas de una crisis o guerra, entre otros). La extracción de variables relevantes a estos episodios, y su posterior vinculación con relaciones causa-efecto, son de gran interés para permitir a los expertos que están analizando el dominio, entender o explicar los diferentes eventos que acontecieron durante ese periodo, o incluso asistirlos en la predicción de posibles desenlaces en episodios en curso. El presente capítulo constituye la etapa final del trabajo de extracción de relaciones causales a partir de medios de noticias digitales. Dicho trabajo se divide en dos grandes etapas: (i) la extracción de variables relevantes al episodio a partir de los textos de artículos periodísticos y (ii) el aprendizaje de la estructura causal a partir de las variables extraídas en (i). En capítulos previos se examina la tarea (i) desde dos posibles ópticas: extracción de términos relevantes y extracción de eventos relevantes. Para estas tareas se utiliza una técnica de pesaje de términos (FDD_β) y un modelo predictivo para detectar eventos en curso, respectivamente. El presente capítulo ofrece un extenso análisis comparativo de diferentes técnicas de aprendizaje de estructuras causales para su posible aplicación a la tarea (ii). Las contribuciones principales de este capítulo son: (a) la formulación completa del marco de trabajo (*framework*) para obtener estructuras causales para expertos a partir de artículos periodísticos (pasos (i) y (ii));

(b) la presentación de un extenso análisis comparativo de técnicas de aprendizaje de estructura causal en series de tiempo que compara 9 técnicas del estado del arte en 64 conjuntos de datos sintéticos y un conjunto de datos reales sobre demanda eléctrica en el Gran Buenos Aires (GBA) (c) la presentación de un caso de estudio de la aplicación del *framework* completo a texto. El análisis comparativo presentado (b) es el primero que compara nueve técnicas de distintas áreas de las ciencias (computación, econometría, sistemas complejos), permitiendo sacar conclusiones originales sobre los métodos y su aplicación a diferentes conjuntos de datos. Por otra parte, el *framework* presentado (a) y evaluado (c) es el primero que combina tantas herramientas diferentes para resolver distintos aspectos de la tarea global y que muestra resultados prometedores para continuar en esa dirección para la elaboración de estructuras causales a partir de textos que serán de gran interés para expertos que quieren entender un dominio o escenario.

4.1. Introducción

La inferencia de la existencia y cuantificación de efectos causales detrás de fenómenos observados es uno de los focos principales de muchos de los esfuerzos científicos [RBB+19]. Por ejemplo, durante muchos años se estudió el efecto causal entre fumar cigarrillos y el desarrollo de cáncer de pulmón en el individuo fumador [DH50]. De manera similar, a mediados del 1700 James Lind delineó el primer experimento aleatorizado (*randomized experiment*) para develar causa-efecto entre el consumo de cítricos y la recuperación del escorbuto [Lin57]. A mediados del 1800, John Snow descubrió que el agua contaminada con materia fecal causaba el cólera [Sno56].

Durante muchos años las herramientas para sacar conclusiones a partir de datos observados eran principalmente estadísticas, con poco desarrollo de perspectivas o formalismos causales. De acuerdo a Pearl el debate sobre si fumar causa cáncer de pulmón (que se extendió desde 1950 hasta 1964) podría haber sido más corto si los científicos hubieran tenido disponible una teoría de causalidad más formal [PM18]. Las herramientas estadísticas típicamente usadas no permiten sacar conclusiones causales (correlación no implica causalidad). Esto transforma la inferencia y cuantificación de efectos causales en un problema mucho más difícil sin una teoría formal de casualidad. Por ejemplo, la correlación entre fumar y el desarrollo de cáncer se podía medir de los datos sin mucha ambigüedad y sin embargo esto no implicaba causalidad. Uno de los argumentos más importantes en

contra de la teoría de que fumar causa cáncer era la posible existencia de factores no medibles que causen deseo por fumar y cáncer de pulmón, sugiriendo que el fumar no causaba cáncer, sino que estaban correlacionados por una causa en común.

La importancia de poder responder este tipo de preguntas, que se encuentran en el centro de los esfuerzos científicos dio lugar a un creciente interés por definir herramientas y formalismos para poder determinar y medir efectos causales. Múltiples marcos de trabajo para inferencia y razonamiento causal han surgido desde entonces [Pea09, PJS17, SGSH00]. Estos esfuerzos se pueden dividir en dos grandes categorías: (i) inferencia causal y (ii) razonamiento causal. El primero pretende partir de datos e inferir el modelo causal que dio origen a los datos, mientras que el segundo parte de un modelo causal ya definido y lo utiliza para contestar preguntas de razonamiento causal (o incluso preguntas de índole estadístico).

Para entender la diferencias entre un modelo causal y un modelo puramente estadístico hay que entender cómo se relacionan entre sí y qué preguntas permite contestar uno y cuáles el otro. Un resumen de estas relaciones se puede ver en la Figura 4.1 adaptada de [PJS17]. Un aspecto importante de la relación entre estos dos, es que el modelo causal subsume al estadístico, pero no al revés. Esto es, el modelo causal provee un entendimiento mayor del proceso generativo de los datos de lo que un modelo puramente estadístico puede proveer. Con el modelo causal se pueden contestar las mismas preguntas que con el modelo estadístico y potencialmente algunas más. En específico, mientras que ambos modelos pueden contestar preguntas observacionales, solo el modelo causal permite potencialmente contestar preguntas sobre intervenciones y preguntas contrafactuales.

Las preguntas intervencionales son cruciales, por ejemplo, para la creación de políticas y toma de decisiones sobre tratamientos médicos, y son inherentemente distintas a las preguntas observacionales. Por ejemplo, consideremos que se tiene una base de datos de pacientes que recibieron dos posibles tratamientos, A o B (siendo el A mejor que el B). La pregunta de cuál es la probabilidad de recuperarse dado que un paciente recibió el tratamiento A, no es lo mismo que preguntar cuáles son las probabilidades de recuperarse dado que efectivamente se intervino en el sistema y se le dio el tratamiento A a un paciente. El tratamiento A en los datos recolectados puede tener una mala proporción de recuperados versus no recuperados simplemente porque los médicos que atendieron a esos pacientes asignaron el tratamiento A (que es mejor que el B) a todos los pacientes difíciles (porque eran los que más necesitaban el tratamiento A). Por otra parte, intervenir

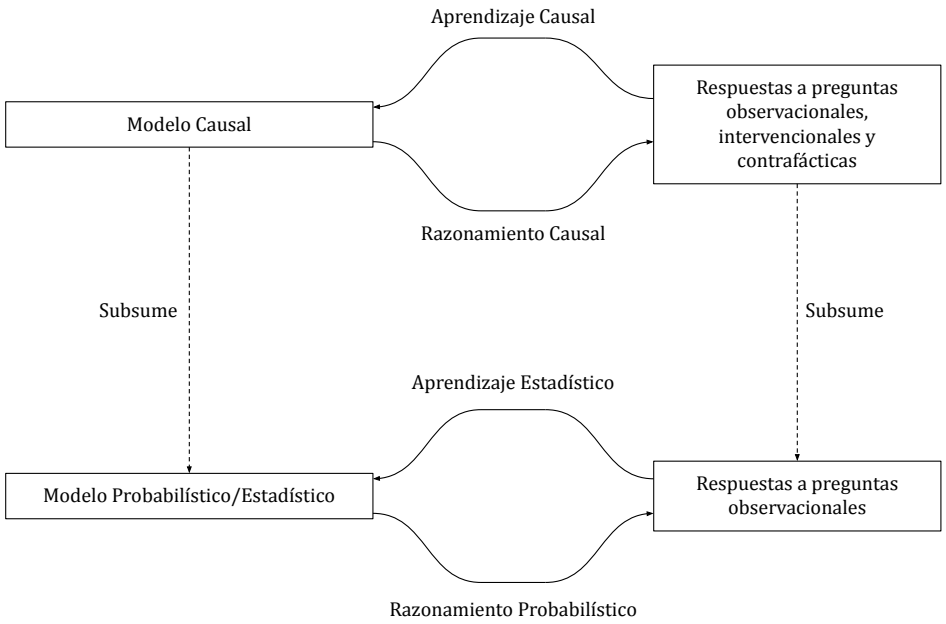


Figura 4.1: Diferencias y relaciones entre un modelo causal y un modelo estadístico/probabilístico, Figura adaptada de [PJS17]. En la parte superior se puede ver cómo información observacional, intervencional y contrafáctica puede ser usada para aprender un modelo casual, o un modelo causal puede usarse para consultar esta información. A esto se lo llama aprendizaje causal y razonamiento causal, respectivamente. Análogamente en la parte inferior se puede observar que de datos observados se puede construir un modelo estadístico/probabilístico, o si ya se cuenta con dicho modelo se lo puede usar para responder preguntas observacionales. A esto lo llamamos aprendizaje estadístico y razonamiento probabilístico, respectivamente.

en el sistema y asignar el tratamiento A a un paciente no hace que sea automáticamente un caso difícil, ya que se lo asignó independientemente de todas las demás variables del sistema (o en otras palabras, de manera aleatoria). Entonces en términos formales $P(R|T = A) < P(R|do(T = A))$, donde $P(R|T = A)$ es la probabilidad de recuperarse dado que se observa que el paciente recibió el tratamiento A, y $P(R|do(T = A))$ es la probabilidad de recuperarse de un paciente dado que se interviene en el sistema y se lo asigna al tratamiento A.

De manera similar, entender los mecanismos causales nos permite entender cuándo las preguntas observacionales y las intervencionales tienen el mismo o distinto resultado. Por ejemplo dado el gráfico de dispersión de la Figura 4.2, se puede sospechar de una relación lineal entre X e Y y realizar una regresión lineal. Dicha regresión nos daría una forma de estimar el valor de Y dado el valor observado de X. Sin embargo, la regresión lineal no nos permite estimar el efecto de una intervención en X sin conocer el modelo

causal. Si X causa Y en el sentido que el cómputo del valor de Y está afectado por el valor de X (por ejemplo cómo está modelado en la ecuación 4.1), entonces el efecto de una intervención se puede medir con el uso de la regresión lineal. Sin embargo si el verdadero modelo generativo es el descrito por la ecuación 4.2, en ese caso al intervenir en X el valor de Y no cambia (en este modelo Y causa a X, y no viceversa). Por ende, la regresión de Y en función de X no serviría para estimar el efecto de una intervención.

En resumen, los datos observacionales presentados en la Figura 4.2 permiten determinar la distribución probabilística observada de ambas variables (la cual está descrita en la ecuación 4.3). Sin embargo, esos datos y esa distribución pueden ser generados por distintos modelos generativos de datos (por ejemplo modelos 4.1 y 4.2), y es importante recuperar esta información para poder recuperar el efecto de una intervención. No siempre es posible determinar el verdadero modelo generativo de los datos a partir de un conjunto de datos observacionales. Ese es uno de los grandes desafíos del área de inferencia causal y muchas veces requiere de supuestos adicionales. Por otro lado, las preguntas contrafácticas son aquellas que consultan sobre mundos posibles que no ocurrieron. Dado el estado actual del mundo, ¿Tomás no se habría recuperado si no le hubieran dado el tratamiento A? Al igual que para las preguntas intervencionales, con datos puramente observacionales no siempre es posible recuperar el modelo causal necesario para responder este tipo de preguntas.

$$\begin{cases} X := N_X & N_X \sim \mathcal{N}(0;1) \\ Y := 2X + N_Y & N_Y \sim \mathcal{N}(0;1) \end{cases} \tag{4.1}$$

$$\begin{cases} Y := N'_Y & N'_Y \sim \mathcal{N}(0;\sqrt{5}) \\ X := 0,4Y + N'_X & N'_X \sim \mathcal{N}(0;\sqrt{0,2}) \end{cases} \tag{4.2}$$

$$P(X,Y) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix} \; ; \begin{pmatrix}1 & 2\\2 & 5\end{pmatrix}\right) \tag{4.3}$$

Una vez obtenido el modelo causal (ya sea de manera automática o construido por un experto), este tiene la capacidad de explicar las relaciones entre las variables y poder potencialmente responder preguntas observacionales, intervencionales y contrafácticas. Esto hace que se genere un gran interés por parte de la comunidad para diseñar técnicas para aprender estos modelos de manera automática a partir de los datos. Adicionalmente, estos modelos causales permitirían la construcción de modelos de aprendizaje automático

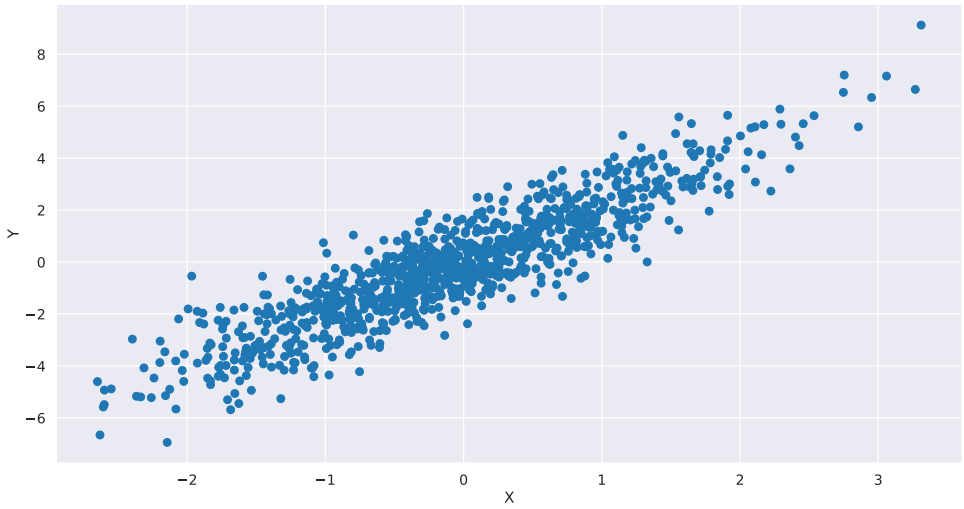


Figura 4.2: Gráfico de dispersión que muestra un conjunto de datos observacionales con distribución probabilística descripta por la ecuación 4.3. Los datos pueden haber sido generados con los procesos generativos 4.1 o 4.2. Sin información adicional no es posible determinar cuál de los dos es el proceso generativo real de los datos y por ende no es posible determinar si $X \to Y$ o si $Y \to Y$. Más aún, si no se cuenta con el supuesto de que todas las causas en común son conocidas (*causal sufficiency*) entonces puede haber una tercera variable que cause a ambas variables (*confounder*) y que no exista causalidad directa de X a Y ni de Y a X. Los supuestos e información adicional sobre el proceso generativo sobre los datos a veces son fundamentales para poder determinar si existe causalidad y la dirección de la misma.

(*machine learning*) más robustos ante cambios en la distribución de los datos que pueden aparecer en problemas del mundo real [SLB+21]. Esto es, en los modelos de *machine learning* tradicional se asume que los datos están independientemente e idénticamente distribuidos (i.i.d.), esto es, en general no están pensados para predecir bajo cambios en la distribución (intervenciones). Por esto es que aprender modelos de *machine learning* con perspectiva causal permitiría modelar explícitamente las intervenciones posibles y hacer sistemas de predicción más robustos ante cambios en la distribución que ocurren normalmente en problemas del mundo real. Más aún, la construcción de modelos predictivos basados en estructuras causales permitiría aprender mecanismos causales independientes transferibles entre modelos, permitiendo así la transferencia de aprendizaje (*transfer learning*) entre diferentes tareas abordadas con técnicas de *machine learning* [SLB+21].

Los datos a partir de los cuales se pretende recuperar el modelo causal pueden tener distintas características, y dependiendo de esas características involucrar diferentes dificultades para la recuperación de dicho modelo causal. Por ejemplo, los datos pueden ser de corte transversal, o de series de tiempo. Los datos también pueden ser puramente obser-

vacionales, pueden contener intervenciones conocidas obtenidas a través de experimentos controlados [KSv+14, SPP+05] o pueden ser conjuntos de datos sintéticos con intervenciones [CY13]. Por otro lado, en algunos trabajos [EM07] se extraen estructuras causales a partir de una combinación de datos observacionales e intervencionales con intervenciones desconocidas [SPP+05].

Este capítulo revisa nueve propuestas de aprendizajes de estructura causal a partir de datos observacionales de series de tiempo y compara sus desempeños para sesenta y seis conjuntos de datos con diferentes características. A partir del análisis realizado se toman las cuatro técnicas con mejor desempeño para ser incluidas como parte del *framework* de recuperación de estructuras causales a partir de textos de artículos periodísticos. Para maximizar la precisión de los vínculos obtenidos por la herramienta, se construye el algoritmo de descubrimiento causal utilizando las cuatro mejores técnicas con votación unánime (*ensemble* de técnicas). Un caso de estudio es llevado a cabo para mostrar el potencial del *framework* completo para cumplir el objetivo de obtener una representación causal de un episodio del mundo real que les permita a expertos tener un mejor entendimiento del mismo.

En la Sección 4.2 se revisan conceptos base, presentando las diferentes granularidades posibles para un modelo causal y se introduce el modelo causal que se busca capturar en este capítulo (modelo causal gráfico), así como también las diferentes técnicas del estado del arte relevadas y comparadas para realizar la tarea del aprendizaje del modelo causal gráfico. En la Sección 4.3, se revisan todos los conjuntos de datos de las cuatro fuentes de datos usadas para el análisis comparativo de las técnicas relevadas: (i) 56 conjuntos de datos sintéticos (categorizados en 5 escenarios) generados con la herramienta de simulación *TETRAD* [SSG+98], (ii) 8 experimentos sintéticos del conjunto *nonlinear-VAR* obtenidos de la plataforma de evaluación comparativa (*benchmarking*) de técnicas de causalidad *CauseMe* [RBB+19] (*CauseMe*), (iii) un conjunto de datos reales de demanda de energía eléctrica en el aglomerado urbano del Gran Buenos Aires (GBA) proporcionada por la empresa proveedora de energía eléctrica *CAMMESA*[1] y (iv) un conjunto de datos de series de tiempo de menciones de términos y eventos extraídos del conjunto de datos de noticias del *New York Times* [San08]. Una descripción detallada del proceso para generar las series de tiempo a partir de los textos es dada en esa sección.

[1] https://cammesaweb.cammesa.com/

En la Sección 4.4 se presentan y discuten los resultados de aplicar las técnicas relevadas sobre los datos de origen sintético (las primeras dos fuentes) (i, ii). Luego, en las Secciones 4.5 y 4.6, se revisan los resultados de aplicar las técnicas sobre los datos de origen real: *CAMMESA* y *The New York Times*, respectivamente. El *framework* completo de extracción de causalidad es delineado en este capítulo a partir de herramientas presentadas en capítulos anteriores. La generación completa de los datos a partir de los textos es explicada en la Sección 4.3 y las estructuras causales resultantes de aplicar las técnicas de causalidad estudiadas sobre estos datos son reportadas en la Sección 4.6. Finalmente, en la Sección 4.7 se discuten las conclusiones generales del capítulo y se presentan los posibles trabajos futuros que se desprenden del presente trabajo.

4.2. Conceptos Base y Trabajos Relacionados

En la Tabla 4.1 adaptada de [SLB+21] se pueden observar tres tipos de modelos causales y sus características en comparación con un modelo puramente estadístico. Este último modelo es el que se encuentra en la última fila de la tabla y es el más sencillo de los presentes en la tabla. La complejidad de los modelos va creciendo a medida que se sube en las filas llegando al modelo más completo, el modelo mecánico/físico que consiste en un conjunto de ecuaciones diferenciales acopladas que modelan los mecanismos físicos responsables de la evolución de las variables en el tiempo [SLB+21].

Para muchos autores ([PJS17, SLB+21]) el vínculo entre los modelos causales y los modelos estadísticos está dado por el principio de sentido común de Reichenbach: si dos variables X e Y están correlacionadas, entonces o (1) $X \to Y$, o (2) $Y \to X$, o (3) están vinculados por una causa común, o (4) una combinación de las tres opciones. Por lo tanto, según dicho principio, la tarea de detección de causalidad se la puede pensar como la tarea de, dado un vínculo de correlación, determinar de cuál de las cuatro opciones se trata.

Como se puede observar, el modelo estadístico es el único que, sin supuestos adicionales, puede ser aprendido a partir de datos observacionales (datos sin intervenciones que cambien la distribución). Esto es, usando datos i.i.d. El signo de pregunta en los otros modelos indica que la respuesta no es afirmativa o negativa, sino que depende del contexto particular del problema. Con supuestos adicionales puede ser posible, pero sin éstos puede que no lo sea.

Modelo	Predecir en i.i.d	Predecir ante intervenciones	Responder contrafactuales	Aprender de datos i.i.d.
Mecánico/Físico	yes	yes	yes	?
Estructural Causal	yes	yes	yes	?
Gráfico Causal	yes	yes	no	?
Estadístico	yes	no	no	yes

Tabla 4.1: Tabla adaptada de [SLB+21], que resume los distintos tipos de modelos discutidos al principio de la Sección 4.2 y sus características. Los modelos son reportados de más complejo (en la primera fila) a menos complejo (en la última fila). Se puede ver cómo el modelo más sencillo (estadístico) puede ser aprendido de los datos i.i.id. pero solo permite contestar preguntas observacionales (en i.i.d.). Por otro lado, a medida que se gana en complejidad, por ejemplo, con los modelos Gráficos o los Estructurales, se pueden contestar preguntas adicionales: intervencionales y contrafactuales, respectivamente. El modelo más complejo es el mecánico/físico que consiste en un conjunto de ecuaciones diferenciales acopladas que modelan los mecanismos físicos responsables de la evolución de las variables en el tiempo. El aprendizaje causal se ubica entre medio de los dos extremos (modelo estadístico y modelo mecánico/físico) tratando de recuperar o bien un modelo gráfico causal o uno estructural causal.

De la Tabla 4.1 se puede observar que todos los modelos permiten predecir en i.i.d., esto es, estimar el resultado de preguntas observacionales. Sin embargo, para respuestas del tipo intervencional es necesario tener al menos el modelo gráfico causal, y para responder preguntas contrafactuales es necesario tener al menos el modelo estructural causal. Teniendo un modelo causal de filas superiores se puede simplificar y obtener un modelo de filas inferiores. Esto es, por ejemplo, a partir del modelo estructural causal se puede extraer el modelo gráfico causal o el modelo estadístico, pero no se puede ir en la dirección opuesta (de abajo hacia arriba). Esto es, no se puede obtener el modelo estructural gráfico a partir de un modelo de filas inferiores (modelo gráfico causal o modelo estadístico).

El modelo mecánico/físico es el estándar de oro (*ground truth*) de la causalidad, conteniendo toda la información posible sobre el fenómeno físico/mecánico. Por otro lado, el modelo estadístico es el modelo más sencillo que no tiene en cuenta las relaciones causales entre las variables, solo permiten predecir cuando las condiciones experimentales no cambian y la distribución se mantiene (sin intervenciones). El modelado causal se encuentra entre estos dos extremos [SLB+21].

Un modelo estructural causal (SCM por sus siglas en inglés) está compuesto por (i) un conjunto de variables exógenas, (ii) un conjunto de variables endógenas, (iii) un

conjunto de funciones que relacionan estas variables y (iv) una distribución de probabilidad que se desprende del conjunto de funciones [Pea09]. En la práctica se lo representa como un conjunto de asignaciones que sirven para determinar el valor de cada variable en función de otras (pudiendo ser estas exógenas o endógenas). Un ejemplo de modelo SCM son los SCMs de la ecuación 4.1 y la ecuación 4.1. Las variables exógenas (N_X,N_Y,N'_X,N'_Y) las variables endógenas (X,Y) y la distribución de probabilidad que describen se pueden deducir de este conjunto de asignaciones. Las relaciones causales se pueden leer directamente de un SCM. Cuando una variable es función de otra, entonces estamos en la situación en la que la variable dependiente causa la variable independiente, esto es, en el ejemplo del SCM definido por la ecuación 4.1 se puede ver que Y es causado por X, mientras que en el SCM definido por la ecuación 4.1 Y causa X.

Un modelo gráfico causal sobre un conjunto de variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ consiste de un grafo $\mathcal{G}$ y una colección de funciones $f(X_j, Pa(X_j)_{\mathcal{G}})$ cuya integral da 1. Donde $Pa(X_j)_{\mathcal{G}}$ es el conjunto de variables padres de X_j (variables que tienen arcos salientes cuyo destino es X_j). Las funciones del modelo causal inducen una distribución probabilística $P_{\boldsymbol{X}}$ sobre $\boldsymbol{X}$:

$$P_{\boldsymbol{X}} = p(X_1, \ldots, X_d) = \prod_{j \neq k} f(X_j, Pa(X_j)_{\mathcal{G}}) \tag{4.4}$$

Los nodos del grafo $\mathcal{G}$ representan las variables de interés y sus arcos representan una relación de causalidad, donde el arco va desde la causa al efecto. Es importante resaltar que en el grafo los arcos representan relaciones de causalidad directa, esto es si X causa a Y, e Y causa a Z, es cierto que un cambio en X produce un efecto en Z, pero este efecto se puede explicar a través de Y, por ende, no es una relación directa y no debería haber una flecha de X a Z. Como se mencionó previamente a partir de un modelo estructural causal se puede obtener un modelo causal gráfico, ya que el primero contiene estrictamente más información que el segundo [PJS17].

En este capítulo se analizan diferentes técnicas del estado del arte para aprendizaje de modelos gráficos causales a partir de datos observacionales (i.i.d.) de series de tiempo. En la Figura 4.3b se puede observar una representación de la tarea a realizar en este capítulo. En contraste con la tarea de aprendizaje de estructura causal a partir de datos de corte transversal (Figura 4.3a), en el aprendizaje de estructuras a partir de series de tiempo se tiene un gráfico que se "desenrolla" en el tiempo con flechas que se pueden clasificar en contemporáneas y no contemporáneas. Las **flechas contemporáneas** son aquellas que parten en un determinado instante de tiempo y afectan (se dirigen a) variables en ese

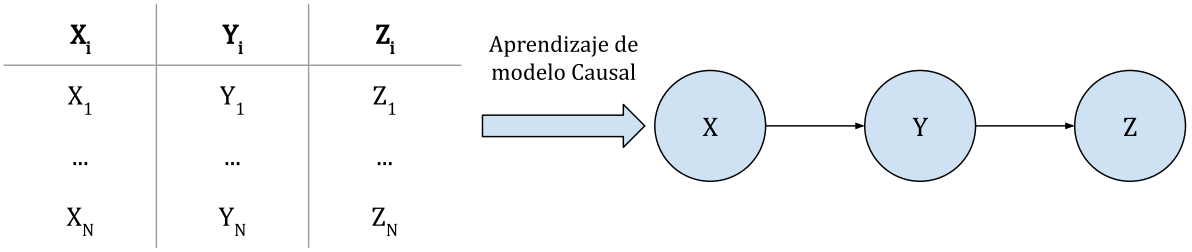
mismo instante de tiempo. Por ejemplo, en la Figura 4.3b, la única relación contemporánea que existe es de Z a Y y se representa con las flechas de la forma $Z_{t-i} \to Y_{t-j}$ con $i = j$. Las **flechas no contemporáneas** tienen la misma forma que las anteriores pero con $i < j$. En el caso de la Figura 4.3b existen dos relaciones no contemporáneas, de X a X (relación autorregresiva (AR)) con un periodo de una unidad de tiempo (AR de orden 1 ($AR(1)$)), y la relación de Y a X con un período de 2. Diferentes herramientas de aprendizaje causal pueden enfocarse en encontrar un tipo o ambos tipos de relaciones. Esto es, existen herramientas de aprendizaje causal que solo apuntan a encontrar las relaciones no contemporáneas y otras herramientas que apuntan a descubrir las dos.

En el presente capítulo se pretende obtener solo los arcos no contemporáneos del grafo causal $\mathcal{G}$ a partir de datos de series de tiempo. En esta capítulo no se calculan las funciones $f(X_j, Pa(X_j)_{\mathcal{G}})$. Aunque teniendo el grafo $\mathcal{G}$ se las puede computar como la distribución de probabilidad condicional de cada variable dado sus padres, el cálculo de dichas funciones no es el objetivo de este trabajo y por ende no se computan ni reportan. Los grafos obtenidos, por una cuestión de interpretación, no son reportados "desenrollados" sino que se muestra un grafo resumido donde las variables no están rezagadas y cada arco de $X \to Y$ se tiene que interpretar como que valores pasados de X causan valores futuros de Y.

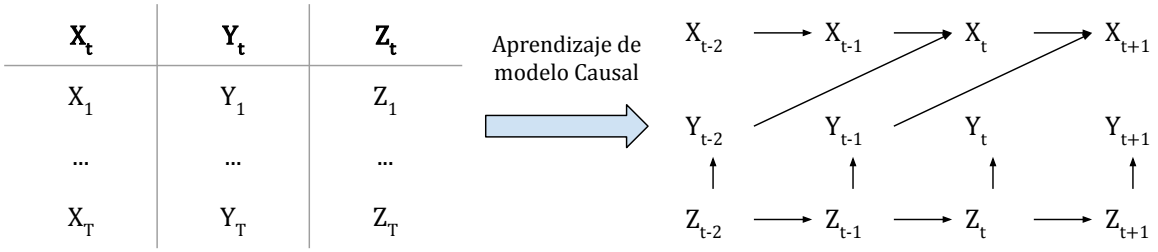
4.2.1. Modelos Comparados

Como se mencionó previamente se estudian nueve diferentes técnicas en el marco de aprendizaje de estructuras causales a partir de series de tiempo en un contexto i.i.d. Estas técnicas son: (i) *BigVAR* [NMB17], (ii) *Direct-LiNGAM* [SIS+11], (iii) *ICA-LiNGAM* [SHHK06], (iv) *Lasso-Granger* [Gra69, Tib96], (v) *PC* [SG91], (vi) *PCMCI* [RNK+19], (vii) *SIMoNe* [CSG+08], (viii) *Transfer Entropy* [Sch00] y (ix) *VAR* [Sim80].

Para poner las técnicas en contexto se las divide en tres categorías principales: (1) basadas en independencias, (2) basadas en modelos estructurales causales restringidos y (3) basadas en modelos autorregresivos. En lo que resta de esta sección se contextualizan las técnicas relevadas, explorando las bases teóricas y supuestos sobre los que están apoyados y se revisan otras técnicas de las mismas categorías que por motivos que se detallan no formaron parte del estudio.



a



b

Figura 4.3: Representación gráfica del proceso de aprendizaje causal para el caso de datos de corte transversal (a) y de datos de series de tiempo (b). En (a) se observan "mediciones" de los valores de X, Y y Z para N individuos o entidades. Utilizando esos datos se construye un modelo causal entre las variables. La dimensión del tiempo no forma parte de los datos ni del grafo resultante. En (b) se observan T mediciones de las variables X_t, Y_t, Z_t a lo largo del tiempo para el mismo individuo o entidad. A partir de estos datos, usando aprendizaje causal, se obtiene un grafo causal "desenrollado" en el tiempo que muestra las relaciones entre los valores de las variables en distintos momentos del tiempo. Se puede observar una relación contemporánea ($Z_{t-i} \rightarrow Y_{t-j}$ para $i = j$). También se pueden ver dos relaciones autorregresivas de orden uno ($Z_{t-i} \rightarrow Z_{t-j}$ para $i = j + 1$ y $X_{t-i} \rightarrow X_{t-j}$ para $i = j + 1$) y una relación no contemporánea ($Y_{t-i} \rightarrow X_{t-j}$ para $i = j + 2$).

Las **técnicas basadas en independencias** (1) se apoyan en dos supuestos fundamentales: *propiedad de Markov para grafos dirigidos* y *faithfulness* [KF09]. Asumiendo estos dos supuestos como válidos se tiene una correspondencia uno a uno entre las independencias condicionales del grafo (d-separaciones [KF09]) y las independencias condicionales que se pueden estimar en los datos observacionales a partir de test de estadísticos. Al existir esta correspondencia se puede estimar la existencia de arcos en el grafo causal que originó los datos solo con realizar test estadísticos (test de independencia condicional) sobre el conjunto de datos observacional y sin tener información previa de la estructura

causal. Por ejemplo, si solo se tienen dos variables X e Y, llevando a cabo un test de independencia entre las variables se puede saber si estamos en la situación donde X e Y son independientes. Si ese es el caso ya sabemos que el grafo causal está compuesto por dos nodos ({X,Y}) y el conjunto de arcos vacío ({}). Si X e Y son dependientes reducimos el conjunto de arcos posible a una de dos opciones: {X $\rightarrow$ Y} o {Y $\rightarrow$ X}. Aunque no siempre es posible determinar el grafo original (como el caso anterior donde nos pueden quedar dos o más opciones posibles), se ha demostrado que es posible identificar la clase equivalente de *Markov* [KF09]. Esto es, la clase de todos los grafos que comparten el mismo esqueleto (arcos sin orientar) y *colliders* (conjuntos de tres nodos con la estructura: $X \rightarrow Y \leftarrow Z$). La estructura *colliders* puede ser recuperada de los datos por las independencias condicionales que exhiben, esto es, los *colliders* presentan un conjunto de independencias que cuando son analizadas con test estadísticos dan como resultado una única posible estructura y no múltiples alternativas.

Al utilizar técnicas basadas en independencias se obtiene la clase equivalente de *Markov*, la cual puede tener flechas sin orientar (se obtiene correctamente el esqueleto del grafo, pero no necesariamente se obtiene la dirección de todas las flechas). Como en este trabajo se analizan causalidades no contemporáneas se puede usar la dirección del tiempo para orientar los arcos que queden sin orientar. Esto es, como la causa tiene que suceder antes que el efecto, se sabe que las flechas no pueden ir hacia atrás en el tiempo. Usando ese criterio se obtiene un grafo totalmente dirigido a partir de las técnicas basadas en independencia usadas. Para el presente trabajo se utilizan dos técnicas basadas en independencias: *PC* [SG91] (v) y *PCMCI* [RNK$^+$19] (vi). Se utiliza para tal efecto el paquete *TIGRAMITE*[2] [RNK$^+$19] que tiene ambas técnicas implementadas. Para analizar las independencias condicionales presentes en los datos observados se utiliza el test de correlaciones parciales presente en el mismo paquete de software (*ParCor*). Dicho test estima las correlaciones parciales mediante una regresión lineal computada con mínimos cuadrados ordinarios y un test de correlación lineal de Pearson distinto de cero en los residuos. Otros test de independencias condicionales no lineales no son incluidos por su tiempo de cómputo prohibitivo.

El algoritmo *PC* [SG91] comienza con el grafo no dirigido completo, y luego reduce el conjunto de arcos primero probando test de independencia condicional de orden cero (conjunto condicional vacío), luego de nuevo con test de independencia condicional de

[2]https://github.com/jakobrunge/tigramite

orden uno (conjunto condicional de una sola variable), y así siguiendo. Luego de este procedimiento ya se tiene el esqueleto del grafo (estructura final con todos los arcos no dirigidos). En una segunda etapa, el algoritmo *PC* orienta algunos de estos arcos utilizando las propiedades de las estructuras *colliders*. En el presente capítulo se utiliza una adaptación del algoritmo *PC* presentada en [RNK+19] diseñada específicamente para series de tiempo, que computa un grafo totalmente dirigido pero que no tiene en cuenta relaciones contemporáneas.

La técnica *PCMCI* [RNK+19] está también basada en el enfoque de test de independencias condicionales. De acuerdo a los autores, es una adaptación para series de tiempo altamente interdependientes. El algoritmo consiste de dos pasos: (i) la aplicación del algoritmo *PC* para identificar todos los posibles conjuntos de padres relevantes para cada variable X_t^j, notado como $\hat{Pa}(X_t^j)$ (el conjunto de nodos directamente conectados de acuerdo al algoritmo *PC*); y luego (ii) la aplicación del test de independencia condicional momentáneo (MCI por sus siglas en inglés) que definido de la siguiente forma nos permite probar si $X_{t-\tau}^i \to X_t^j$:

$$MCI : X_{t-\tau}^i \perp\!\!\!\perp X_t^j | \hat{Pa}(X_t^j) \backslash \{X_{t-\tau}^i\}, \hat{Pa}(X_{t-\tau}^i) \tag{4.5}$$

De este modo, usando MCI se condiciona en los padres de X_t^j y en los padres rezagados (*time-shifted*) de $X_{t-\tau}^i$. Los dos pasos del algoritmo sirven para lo siguiente: la aplicación de *PC* permite descubrir el conjunto de *Markov* de cada nodo, descartando variables irrelevantes sobre las que no es necesario condicionar, obteniendo de resultado un conjunto de padres candidatos para ser usados en el segundo paso del algoritmo *PCMCI*, el test MCI. Este segundo paso sirve para abordar el problema de controlar los falsos positivos para series altamente interdependientes. Por ejemplo, para probar $X_{t-2}^1 \to X_t^3$ usar el conjunto de padres de X_t^3 ($\hat{Pa}(X_t^3)$) es suficiente para detectar causas en común o relaciones indirectas. Condicionar también sobre el conjunto de padres de X_{t-2}^1 ($\hat{Pa}(X_{t-2}^1)$) permite controlar por la autocorrelación, lo que posibilita controlar la tasa de falsos positivos al nivel esperado de acuerdo a los resultados presentados por los autores.

Las **técnicas basadas en modelos estructurales causales restringidos** (2) utilizan supuestos adicionales sobre la forma funcional de las relaciones causales para ganar identificabilidad. Por ejemplo, como se mencionó en la Sección 4.1, dos modelos generativos distintos (4.1 o 4.2) pueden haber generado los datos de la figura 4.2, y solo con los datos es imposible determinar cuál de los dos modelos es el que produjo dichos datos. Se

puede plantear (1) un modelo donde X sea función de Y y (2) otro donde Y sea función de X y ambos van a explicar correctamente los datos. Sin embargo, si la función real que explica los datos no es invertible (restricción adicional), generando y comparando ambos modelos ((1) y (2)) es posible detectar cuál es el modelo real que generó los datos observados.

Por ejemplo, para el caso de modelos lineales no gaussianos (LiNGAM por sus siglas en inglés), es posible analizar la asimetría entre la causa y el efecto para poder determinar cuál es la causa y cuál el efecto. En la Figura 4.4, se puede observar un conjunto de datos generados a partir del modelo causal real $X \to Y$, donde $Y = f(X)$ con f siendo una función lineal con ruido uniforme. Los mismos datos son mostrados de dos formas, como Y en función de X (izquierda) y como X en función de Y (derecha). El modelo correcto es el de la izquierda (modelo generativo real de los datos), y se puede observar que la regresión lineal computada tiene residuos homogéneos y correctamente distribuidos. Por otra parte, en el modelo de la derecha, donde se trata de computar $X = f(Y)$, se puede observar que por las características del modelo y por no tratarse del modelo real, se tiene una regresión lineal cuyos residuos están relacionados con la variable Y. Esto sugiere que el modelo causal correcto de estos datos es $X \to Y$ y no $Y \to X$. Usando esta estrategia se pueden usar solo datos observacionales para determinar causas y efectos, siempre y cuando se tengan restricciones adicionales sobre el modelo (en este caso que sea lineal no gaussiano). Existen otras combinaciones de restricciones que nos permiten romper la simetría entre causa y efecto para detectar causalidad usando solo datos observacionales. Un resumen de estas configuraciones para ruido gaussiano se puede observar en la Tabla 4.2 adaptada de [PJS17].

En el presente capítulo se analizan y reportan resultados para dos técnicas basadas en modelos estructurales restringidos:*ICA-LiNGAM* [SHHK06] y *Direct-LiNGAM* [SIS$^+$11]. En 2006 la técnica *ICA-LiNGAM* de aprendizajes de estructuras causales fue propuesta utilizando la asimetría entre causa y efecto que presentan los modelos lineales no gaussianos (ver Figura 4.4). Sin embargo, de acuerdo a los autores, *ICA-LiNGAM* presentaba varios problemas: (i) el algoritmo podría no converger a la solución correcta en un número finito de pasos si el estado inicial no era el adecuado o si el tamaño del paso no era seleccionado adecuadamente para las versiones del método que buscaban la solución a través del método del gradiente. (ii) Algunos pasos del algoritmo no eran invariantes ante problemas de escala, por ende, podrían tener problemas de desempeño dependiendo de

		Restricciones	Identificabilidad
Modelo Estructural Causal	$X_i = f_i(Pa(X_i), N_i)$	ninguna	No
Modelo de Ruido Aditivo	$X_i = f_i(Pa(X_i)) + N_i$	no lineal	Si
Modelo Causal Aditivo	$X_i = \sum_{k \in Pa(X_i)} f_{ik}(X_k) + N_i$	no lineal	Si
Modelo Lineal Gaussiano	$X_i = \sum_{k \in Pa(X_i)} \beta_{ik} X_k + N_i$	lineal	No
Modelo Lineal Gaussiano (con varianzas de error iguales)	$X_i = \sum_{k \in Pa(X_i)} \beta_{ik} X_k + N_i$	lineal	Si

Tabla 4.2: Combinaciones de restricciones para la forma funcional y el impacto que tienen sobre la identificabilidad del vínculo causal usando solo datos observacionales. Esta tabla es adaptada de [PJS17] y son todas para configuraciones con ruido gaussiano.

la escala o la desviación estándar de las variables, especialmente para cuando hay una gran variedad de escalas. Por estos motivos en 2011 Shimizu junto con otros autores de los cuales varios trabajaron en *ICA-LiNGAM*, propusieron un nuevo algoritmo llamado *Direct-LiNGAM*. De acuerdo a los autores, este nuevo algoritmo propuesto garantiza que va a converger a la solución en una cantidad finita de pasos igual al número de variables si los datos siguen estrictamente el modelo [SIS+11]. En el presente capítulo se analiza el desempeño de ambas técnicas en varios conjuntos de datos.

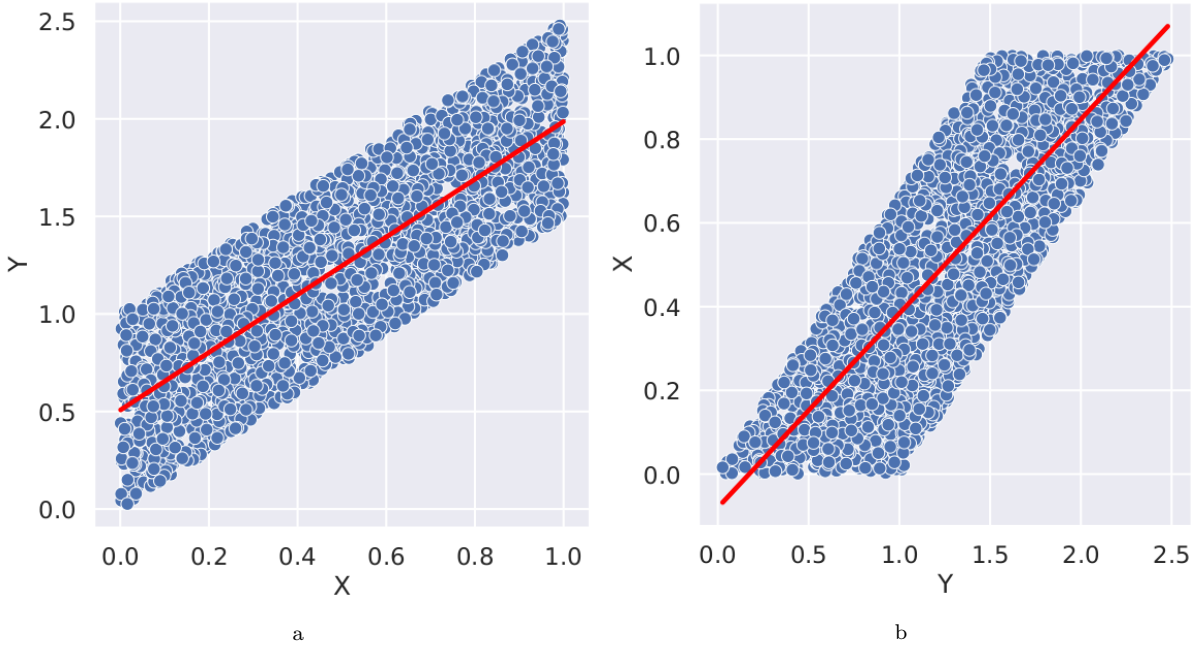


Figura 4.4: Dos gráficos de dispersión que muestran el mismo conjunto de datos invirtiendo los ejes: primero X en el eje horizontal e Y en el vertical (a), y luego al revés (b). Los datos fueron generados a partir de un modelo causal real $X \to Y$, donde $Y = f(X)$ con f siendo una función lineal con ruido uniforme. En esta figura se puede apreciar la asimetría entre la causa y el efecto. Al tratar de modelar los datos con una regresión lineal $Y \sim f(X)$ (a) se pueden observar residuos uniformemente distribuidos. Por otra parte, al modelar los datos con la regresión lineal $X \sim f(Y)$ (b) se obtienen residuos no uniformes (dependientes de Y). Las técnicas de aprendizaje de estructuras causales basados en modelos funcionales restringidos aprovechan este tipo de asimetría para detectar el modelo correcto, en este caso el (a).

Las tercera y última categoría de las técnicas analizadas en este capítulo son las **técnicas basadas en modelos autorregresivos** (3). Este categoría es exclusiva a series de tiempo, y se basa en tratar de ver cómo valores del pasado de una variable X dan información única (no presente en otras variables) para predecir o explicar valores del futuro de otra variable Y; si esto sucede se hipotetiza que $X \to Y$. Esta idea está resumida en los dos principios en los que se basa la técnica propuesta por Clive Granger en 1969

denominada Causalidad de Granger [Gra69]: (1) la causa tiene que preceder al efecto, (2) la causa produce cambios únicos en el efecto, por lo que el pasado de la causa aporta información única a la tarea de predecir o explicar el efecto. En este capítulo se analizan cinco técnicas que, apoyándose en esos dos principios, son usadas para inferencia de estructuras causales en series de tiempo: (1) *Lasso-Granger* [Gra69, Tib96], (2) *Transfer Entropy* [Sch00], (3) *VAR* [Sim80], (4) *BigVAR* [NMB17] y (5) *SIMoNe* [CSG+08].

La propuesta *Lasso-Granger* está basada en la aplicación de la técnica de regularización lasso (*least absolute shrinkage and selection operator*) como técnica de selección de variables, seguida de la aplicación de la técnica de causalidad de Granger entre pares de variables. Para cada variable X_i se realiza un primer paso de selección de variables, para esto se utiliza lasso para modelar una regresión lineal penalizada con X_i como variable dependiente, y todas las variables del sistema rezagadas como variables independientes. Todas las variables acompañadas por coeficientes significativos (distintos de cero) son utilizados como variables padres candidatas de X_i. Finalmente se realiza el test de causalidad de Granger de a pares entre X_i y todas las variables candidatas X_j, para determinar si $X_j \rightarrow X_i$. Se computa la causalidad de Granger usando 4 rezagos y se establece un link causal entre las variables cuando el p-valor asociado al estadístico F está por debajo de $0,05$.

La técnica *Transfer Entropy* [Sch00] está basada en los mismos dos principios que el test de Granger pero en lugar de utilizar regresión lineal utiliza el concepto de transferencia de información, el cual está basado en nociones de teoría de la información. Esto es, para probar si X_j causa a X_i ($X_j \rightarrow X_i$), la técnica analiza la transferencia de información del pasado de X_j hacia valores futuros de X_i. Si el pasado de la primera variable aporta información única (no presente en el pasado de X_i) para predecir valores futuros de X_i se dice que X_j causa a X_i. La técnica *Transfer Entropy* se puede entender como una extensión no paramétrica del test de causalidad de Granger [ONSH20]. En [BBS09] se mostró que estas dos técnicas son equivalentes para procesos Gaussianos. Aunque la técnica de causalidad de Granger fue definida concretamente usando un test F en los coeficientes de una regresión lineal, utilizando los dos principios de la técnica y otro tipo de modelos de predicción se pueden definir otras técnicas de detección de causalidad. Este es el caso para la técnica *Transfer Entropy* que reemplaza la regresión lineal por el concepto de transferencia de información. De manera similar, en [Hma20] se define una técnica de extracción de causalidad basada en Granger pero que reemplaza la regresión

lineal por una red neuronal.

La técnica Vector Autorregresivo (*VAR*) [Sim80] se puede utilizar para capturar la relación entre múltiples variables a lo largo del tiempo. Es la extensión del modelo autorregresivo univariado de orden p (*AR(p)*) donde una variable es modelada usando p rezagos de sí misma. Dicho modelo se define como sigue:

$$X_t = c + \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + \varepsilon_t \tag{4.6}$$

En este modelo $\beta_1, \ldots, \beta_p$ son las constantes que acompañan a las variables rezagadas, c es una constante y ε_t es ruido blanco. La extensión de este modelo para múltiples variables es el modelo *VAR(p)* donde se describe la evolución de k variables, llamadas endógenas, a lo largo del tiempo. Un modelo *VAR(p)* con k variables se lo puede describir con una ecuación similar a la Ecuación 4.6 pero donde cada X_i es un vector de k dimensiones. También se lo puede describir en términos de variables de una sola dimensión usando múltiples ecuaciones. Por ejemplo, un modelo *VAR(p)* con $k = 2$ se lo puede describir como:

$$\begin{cases} X_{1,t} = c_1 + \beta_1 X_{1,t-1} + \cdots + \beta_p X_{1,t-p} + \alpha_1 X_{1,t-1} + \cdots + \alpha_p X_{1,t-p} + \varepsilon_{1,t} \\ X_{2,t} = c_2 + \gamma_1 X_{1,t-1} + \cdots + \gamma_p X_{1,t-p} + \delta_1 X_{1,t-1} + \cdots + \delta_p X_{1,t-p} + \varepsilon_{2,t} \end{cases} \tag{4.7}$$

Como se puede observar, un modelo *VAR(p)* representa cada variable con una regresión lineal de todas las variables endógenas del sistema rezagadas (de 1 hasta p), de forma que cada variable se modela usando el pasado de todas las demás variables del sistema. En este modelo $\alpha_i, \beta_i, \gamma_i, \delta_i, c_1$ y c_2 son constantes y $\varepsilon_{1,t}$ y $\varepsilon_{2,t}$ son variables de ruido blanco. Los modelos *VAR(p)* son ampliamente usados en diferentes áreas como economía y ciencias naturales. Aunque no son siempre utilizados con el objetivo de estudiar causalidad. Guiados por los principios (1) y (2) en los que se basa la causalidad de Granger podemos pensar al modelo *VAR* como una extensión multivariada del test de causalidad de Granger. En este capítulo se estudia la técnica *VAR* para la extracción de causalidad al estilo Granger, donde de todos los coeficientes del modelo $(\alpha_i, \beta_i, \gamma_i, \delta_i)$ consideramos que los que son significativos (p-valor $< 0{,}05$) indican la existencia de un vínculo causal de la variable independiente a la dependiente. Por ejemplo, si en la Ecuación 4.7 el coeficiente γ_1 da significativo entonces se considera que $X_{1,t-1}$ causa $X_{2,t}$.

Las técnicas *BigVAR* [NMB17] y *SIMoNe* [CSG+08] están basadas en un modelo *VAR(p)* pero tienen la característica adicional que tienen factores de penalización o regularización para obtener modelos más estables e interpretables (y con más coeficientes

con valor cero). De esta manera, estas técnicas generan modelos con menos *overfitting* y menos sensibles al ruido inherente de los datos. En este contexto de extracción de causalidad es posible obtener modelos con mayor precisión en la recuperación de los arcos (potencialmente afectando a la cobertura), ya que efectos pequeños serán penalizados y llevados a cero (potencialmente descartando efectos espurios pequeños detectados por las técnicas). Al igual que para la técnica *VAR*, se consideran como efectos causales a todos aquellos coeficientes significativos encontrados en las ecuaciones del modelo. A diferencia de la técnica *VAR* sin penalizar (en el cual se determinaba la significancia usando el p-valor de un test estadístico), para *BigVAR* y *SIMoNe* se consideró coeficiente significativo a todo aquel que tenga magnitud distinta de cero (que no haya sido penalizado de tal forma de alcanzar el valor cero).

Es importante notar que las técnicas basadas en modelos regresivos permiten detectar la dirección de la causalidad apoyándose en los principios de la causalidad de Granger, por ende, solo es posible detectar causalidades rezagadas (no contemporáneas). Por otra parte, las técnicas basadas en modelos estructurales causales restringidos permiten detectar la direccionalidad incluso para efectos contemporáneos o para conjuntos de datos de corte transversal porque asumen restricciones adicionales a la forma funcional que permiten romper la simetría entre la causa y el efecto (y así obtener la dirección de la causa al efecto sin necesidad del tiempo). Por otro lado, las técnicas basadas en independencias, solo en algunos casos permiten distinguir efectos causales contemporáneos. Para muchos arcos la dirección no es conocida dando lugar a un grafo parcialmente dirigido. Para las técnicas basadas en independencias es posible determinar alguna de las direcciones faltantes del grafo utilizando los mismos principios aplicados para la causalidad de Granger. Esto es, si existe un vínculo causal no dirigido entre dos variables en distintos instantes de tiempo (relación no contemporánea), usando la intuición de Granger, de que la causa tiene que preceder al efecto, se puede orientar la flecha en la dirección del paso del tiempo. En este trabajo se utiliza una versión de *PC* [SG91] adaptada para series de tiempo que computa un grafo totalmente dirigido utilizando el criterio antes mencionado (solo considerando relaciones no contemporáneas).

La extracción de causalidad también se ha abordado desde la perspectiva de sistemas complejos [SMY+12, MAC14, HA10]. De esta literatura surgen una serie de herramientas de detección de causalidad que se apoyan en el modelo de espacio de estados, siendo el mayor representante de esta categoría la técnica *Convergent Cross Mapping*

(CCM) [SMY+12]. Diversos autores mencionan que las técnicas tradicionales de causalidad asumen separabilidad, lo cual resulta inadecuado para algunos contextos [HLSP17]. En particular los autores, en [SMY+12], aseguran que la técnica de causalidad de Granger no es adecuada para sistemas complejos, ya que asume la propiedad de separabilidad, requerimiento que no siempre se cumple en este dominio. Esto es, la relación causa efecto no siempre está correctamente diferenciada, muchas veces la relación entre variables o sistemas presentan simultaneidad (el comportamiento del predador afecta a la presa y viceversa). Las técnicas de extracción de causalidad de esta categoría fueron desarrolladas para el dominio de sistemas complejos. Dicho dominio tiene algunas particularidades, como por ejemplo los sistemas suelen estar altamente interconectados, vinculados de forma determinística y con comportamientos caóticos (pequeñas variaciones en las condiciones iniciales provocan cambios arbitrariamente grandes en la evolución del sistema). Ya que todos los conjuntos de datos trabajados en este capítulo no se corresponden con este dominio, esta categoría de técnicas no es considerada.

Se puede considerar la existencia de una cuarta categoría de métodos de extracción de causalidad que tampoco es analizada en este capítulo porque han sido propuestos para escenarios donde no solo se tienen datos observacionales sino que además se cuenta con datos intervencionales, esto es, se tienen observaciones del sistema bajo diferentes intervenciones, las cuales pueden ser conocidas o desconocidas [PBM16, HDPM18, EM07, CY13]. Estas técnicas no son consideradas ya que no se cuenta con este tipo de datos para el presente trabajo.

Si bien en el presente trabajo se presentan técnicas de extracción de causalidad a partir de series de tiempo, vale la pena mencionar que existe una gran cantidad de herramientas y literatura dedicada a extraer causalidad directamente de textos [ZLZ+16, FHK+20, PK07, LLZR21, KBR91, KCN00, RDM12, ZWM+17, GM02, Gar97, DSDN18]. En estos escenarios los vínculos causales están explícitos en el texto y se busca la creación de herramientas que los detecten y extraigan. Por ejemplo, en la oración: “La Crisis Financiera Global de 2008 se desató de manera directa debido al colapso de la burbuja inmobiliaria en los Estados Unidos en el año 2006”, se podría detectar y extraer la relación causa-efecto entre la burbuja inmobiliaria del 2006 y la crisis financiera del 2008. Una desventaja de este enfoque es que este tipo de herramientas requieren que el reportero que escribe la noticia o el texto conozca la relación causal y la deje explícita en el texto (directamente con la palabra “causa” o con palabras que indiquen la causa “el desarrollo de X desembocó

en la ocurrencia de Y"). Adicionalmente el reportero puede estar dando su opinión sobre las posibles causas, pero no necesariamente se trata de una representación correcta de la situación. Más aún, puede estar hablando de un posible efecto causal refiriéndose a una causa que aún no sucedió: "Si las empresas especulan una subida del precio de un factor importante en su proceso productivo y deciden subir paulatinamente sus precios causarán inflación".

En [ZWM+17] los autores usan conectores causales para detectar pares (x, y) causales en el texto (como por ejemplo "x because y" y "x leads to y"). Luego para obtener patrones generales de alto nivel generalizan los sustantivos a sus hiperónimos utilizando *WordNet* [Mil95] y los verbos a sus clases con *VerbNet* [KS05] (por ejemplo, "kill" pertenece a la clase "murder-42.1"). De esta manera logran obtener patrones generales de causalidad que no necesariamente estaban explícitos en los textos pero que se pueden deducir. Por ejemplo, se plantean como objetivo transformar un par causal como "*a massive 8.9-magnitude earthquake hit northeast Japan on Friday* → *a large amount of houses collapsed*" en un par causal más general no presente en el texto: "*earthquake hit* → *house collapse*" a través de generalizaciones con *WordNet* y *VerbNet*. Con estos pares de causalidad construyen una red causal y una representación distribuida de la misma (embedding) que argumentan sirve para tareas posteriores. En su caso la usan para predicción de movimientos de los precios de la bolsa. Esta estrategia sirve para agregar flexibilidad y capacidad de generalización a la extracción de causalidad de texto, pero aún se requieren menciones de causalidad explícitas en textos y se necesitan herramientas sofisticadas para lograr detectar y extraer esas menciones.

En [RDM12] construyen redes causales a partir de textos con el objetivo de predecir. El algoritmo que presentan, *Pundit*, extrae pares causales de textos de noticias no estructurados buscando por patrones causales gramaticales ("*because*", "*due to*", "*lead to*", y otros). Luego generalizan estos pares usando conocimiento del mundo que proviene de diferentes ontologías. El grafo está compuesto por conceptos de *Wikipedia*, *ConceptNet* [LS04], *WordNet* [Mil95], *Yago* [SKW07] y *OpenCyc*. Por otra parte, para las relaciones entre los conceptos (por ejemplo "CapitalOf) usan el proyecto *LinkedData* [BHBL11]. El objetivo es poder predecir a través de la generalización de eventos y sus relaciones. Por ejemplo, detectar "Earthquake hits [Country Name]" causa "Red Cross help sent to [Capital of Country]" (a través de múltiples acontecimientos de este tipo de evento), lo que permitiría predecir, ante la ocurrencia de un terremoto en un país, el envío de ayuda por parte

de la Cruz Roja a la Capital del país.

Un trabajo relacionado al realizado en este capítulo es el de [BCFS19]. En dicho trabajo construyen un grafo causal midiendo cómo la ocurrencia de una palabra puede influenciar la ocurrencia de otras en el futuro. Utilizan el concepto de Causalidad de Granger para medir la capacidad de una palabra de influenciar la ocurrencia de otra. En su trabajo realizan un trabajo de filtrado de palabras ad-hoc, filtrando palabras demasiado frecuentes (si aparece en más del 50 % de los días) o demasiado infrecuentes (si aparecen en menos de 100 artículos). También utilizan entidades como parte de su vocabulario, las cuales son recuperadas con un reconocedor de entidades (NER). Por último, utilizan un algoritmo para reconocer *triggers* de eventos [Ahn06] para ser incorporados al vocabulario. A este vocabulario le agregan bigramas frecuentes. Una de las principales limitaciones de este trabajo es que trabajar a nivel de términos y bigramas aislados puede conducir a un vocabulario difícil de interpretar en una red causal. En este trabajo los autores no reportan ninguna red causal como resultado, sino que muestran la utilidad de la herramienta para hacer predicciones del precio de la bolsa de valores. Una representación semánticamente significativa de cada variable es necesaria para poder mostrar un grafo causal con nodos y relaciones causales interpretables. Otra limitación es el uso de una única herramienta para la parte de extracción de causalidad (Causalidad de Granger). La tarea de descubrimiento causal es una tarea con una gran complejidad y no existe una única técnica que se adecue a todos los dominios. Por tal motivo es importante el análisis de múltiples herramientas y la adopción de la o las mejores opciones para el dominio de trabajo.

En el presente trabajo se utiliza la técnica de pesaje de términos relevantes a un dominio presentada en el Capítulo 2 (FDD_β) para obtener términos relevantes (unigramas, bigramas y trigramas). Luego se detectan *event-triggers* de eventos en curso utilizando el modelo para tal fin presentado en el Capítulo 3. Estos *event-triggers* son luego representados utilizando una representación vectorial que tiene en cuenta todo el contexto, y posteriormente mostrados utilizando una representación que permite visualizar el *event-trigger* y todo el contexto, dando lugar a descripciones de eventos más interpretables. Por último, los términos relevantes y los eventos completos son unidos en un solo conjunto de datos para la construcción de la estructura causal. Esta construcción de la estructura causal es llevada a cabo a través de un *ensemble* de cuatro técnicas de descubrimiento causal (*PC*, *PCMCI*, *Direct-LiNGAM* y *VAR*).

4.3. Conjuntos de Datos

Para los experimentos realizados en el presente capítulo se utilizaron datos de cuatro fuentes distintas: (1) *TETRAD* [SSG+98], (2) *CauseMe* [RBB+19], (3) *CAMMESA*[3] y (4) *The New York Times* [San08]. La primera fuente se trata de una herramienta de simulación con la cual se generaron 56 conjuntos de datos sintéticos con diferentes características. La segunda fuente es una plataforma de *benchmarking* de técnicas de extracción de causalidad con varios conjuntos de datos disponibles. De esta plataforma se utilizaron los 8 conjuntos de datos correspondientes a los experimentos *nonlinear-VAR*. Estas dos primeras fuentes (1) y (2), son conjuntos de datos sintéticos. Por otro lado, las otras dos fuentes se corresponden con datos observados del mundo real. La fuente (3) es la Compañía Administradora del Mercado Mayorista Eléctrico Sociedad Anónima (*CAMMESA*) que puso a disposición un conjunto de datos con mediciones de demanda de energía eléctrica en el área metropolitana de la ciudad de Buenos Aires (Gran Buenos Aires (GBA)) junto con mediciones de variables climáticas para el mismo área geográfica. La fuente (4) son los textos completos del corpus del *The New York Times*. A partir de estos se construye un conjunto de datos de series de tiempo de menciones de términos y eventos en curso detectados en dichos textos. En la presente sección se describe en detalle cómo están constituidos cada uno de los conjuntos de datos y sus diferentes características. Estos conjuntos de datos son presentados en ese orden ((1), (2), (3) y (4)) para respetar el orden en el que se reportan los resultados en las Secciones 4.4, 4.5 y 4.6

4.3.1. Fuente #1: *TETRAD*

TETRAD [SSG+98] es una aplicación de escritorio escrita en Java que permite crear, estimar o buscar modelos causales. A partir de estos modelos se pueden realizar tests y predicciones. Adicionalmente permite crear modelos causales aleatorios y a partir de estos generar conjuntos de datos. Esta última funcionalidad es la que se utilizó para el presente trabajo. Dado que esta herramienta permitía de manera flexible modificar varios de los parámetros de creación de los modelos y datos simulados, se aprovechó la herramienta para generar diversas configuraciones para probar diferentes aspectos de las técnicas de aprendizaje de estructura causal estudiadas. Las diversas configuraciones probadas se pueden categorizar en cuatro escenarios distintos: (i) variar el número de nodos (**N**) del

[3] https://cammesaweb.cammesa.com/

modelo causal, (ii) variar el tamaño de la serie de tiempo (**T**), (iii) variar la cantidad de variables latentes (**H**), y (iv) variar la cantidad de rezagos presentes en el modelo causal real (**L**). Para todos los escenarios se fijaron todos los parámetros excepto el que estaba siendo analizado el cual se lo variaba dentro de un rango de valores elegidos. Se reportan los parámetros fijos y los rangos utilizados en la Tabla 4.3.

Se utilizó la configuración predeterminada de *TETRAD* para todos los parámetros excepto para los que varían (N, T, H, L) y para el valor máximo del rango de coeficientes (de 0,7 se lo cambió a 0,5). Los valores mínimos y máximos de los coeficientes son los valores a partir de los cuales se muestrean los coeficientes asociados a los vínculos causales. En el caso aquí presentado se muestrean de manera uniforme aquellos correspondientes al intervalo (-0,5; -0,2) ∪ (0,2; 0,5). Utilizando un parámetro de 0,7 se obtenían series no estacionarias con facilidad (las series explotaban hacia infinito) y por esta razón se utilizó un valor máximo menor (0,5). Una descripción detallada de cada parámetro se puede encontrar en los menús flotantes de cada parámetro en la aplicación *TETRAD*[4]. Una descripción detallada de cada uno de los cuatro escenarios es dada a continuación.

Para el **escenario #1** se utilizaron nueve valores distintos para la cantidad de nodos ($N \in \{6, 9, 12, 15, 18, 21, 24, 27, 30\}$). Para poder medir el desempeño de las diferentes técnicas ante un número creciente de variables se mantuvieron el resto de los parámetros en una configuración sencilla: cantidad de variables ocultas igual a cero ($H = 0$), cantidad de rezagos incluidos en el modelo igual a uno ($L = 1$) y longitud de la serie se la fijó a 1.069 ($T = 1.069$). Esta longitud para la serie fue elegida para concordar con el conjunto de datos obtenido del *New York Times* que consiste en de 1069 semanas (desde enero 1987 hasta junio 2007). Aunque finalmente no se usó esta frecuencia para este conjunto de datos (se usó frecuencia mensual para los datos extraídos del *New York Time*), se mantuvo la configuración $T = 1.069$ ya que igual constituye un tamaño adecuado para los presentes experimentos.

Para el **escenario #2** se utilizaron siete diferentes valores para la longitud de la serie, $T \in \{100, 500, 1.000, 2.000, 3.000, 4.000, 5.000\}$. Se utilizó $N = 30$, y al igual que para el escenario #1 se utilizó $H = 0$, $L = 1$ y $T = 1.069$.

Para el **escenario #3**, además de las variables observadas se agregaron variables no observadas que podían introducir complejidad a la tarea de descubrimiento causal. Además de tener 20 variables observadas ($N = 20$) se crearon siete conjuntos de datos

[4]https://www.ccd.pitt.edu/tools/

con diferente cantidad de variables no observadas (ocultas), $H \in 0, 2, 4, 6, 8, 10, 12$. Al igual que para los escenarios anteriores se usaron $L = 1$ y $T = 1.069$.

Para el **escenario #4** se varió la cantidad de variables rezagadas incluidas en el modelo causal real. El objetivo del presente capítulo es detectar correctamente las relaciones causales no contemporáneas de exactamente una unidad de tiempo en el pasado (relaciones del tipo $X_{j,t-1} \rightarrow X_{i,t}$). Sin embargo la herramienta crea también vínculos contemporáneos($X_{j,t} \rightarrow X_{i,t}$), y en este escenario, vínculos causales con mayor distancia ($X_{j,t-\tau} \rightarrow X_{i,t}$ con $\tau \in \{2, 3, 4, 5\}$). Lo que se pretende analizar a partir de este escenario es la capacidad de las técnicas de encontrar las relaciones causales directas correctas con distancia uno, a pesar de tener correlaciones adicionales originadas por arcos adicionales que no son los buscados ($X_{j,t-\tau} \rightarrow X_{i,t}$ con $\tau \in \{0, 2, 3, 4, 5\}$). La cantidad de nodos se la fijó en diez ($N = 10$).

En resumen, se varió T en el escenario 2, pero para todos los demás se lo fijó a $T = 1.069$ para que tuviera la misma dimensión que el conjunto de datos originado con *The New York Times* con frecuencia semanal. Por otro lado, los valores de H y L fueron variados en los escenarios 3 y 4 respectivamente, pero en todos los demás, por simplicidad, se los fijó en $H = 0$ y $L = 1$. Por último, se varió N en el escenario 1, pero para todos los demás se usó el N más grande posible. Es importante destacar que no se podían elegir valores de N arbitrariamente grandes en la herramienta, ya que con un valor grande de N se tiene una gran cantidad de arcos y se generan situaciones problemáticas en las que las series tienen comportamientos no estacionarios (explotan hacia infinito).

La estructura causal real de los conjuntos de datos generados a partir de TETRAD puede ser representada usando Grafos Acíclicos Dirigidos (DAG por sus siglas en inglés). Para cada uno de los cuatro escenarios se utilizaron las dos posibles configuraciones para construcción de DAG provistas por *TETRAD*: *Random Foward DAG* (*RFDAG*) y *Scale-free DAG* (*SFDAG*). La primera estrategia crea un DAG aleatoriamente agregando arcos hacia adelante (arcos que no apunten a antecesores de la variable), donde los arcos son insertados de a uno. Por otro lado, la estrategia *SFDAG* crea un DAG cuyas variables tienen un grado de conectividad que sigue una ley de potencias. Se utilizan y reportan resultados para ambas configuraciones de construcción de DAG.

Habiendo **nueve** configuraciones posibles para el escenario 1, **siete** para el escenario 2, **siete** para el escenario 3 y **cinco** para el escenario 4, se tiene un total de 28 configuraciones. Para cada una de estas configuraciones se crea un conjunto de datos usando la estrategia

Descripción del Parámetro	Valor
Valores mínimo y máximo del rango de los coeficientes	(0,2; 0,5)
Valores mínimo y máximo del rango de la covarianza	(0; 0)
Grado máximo del grafo	100
Varianza del ruido de medición aditivo	0
Cantidad de rezagos incluidos en el modelo (L)	{1, 2, 3, 4, 5}
Cantidad de variables ocultas incluidas en el modelo (H)	{0, 2, 4, 6, 8, 10, 12}
Cantidad de variables observadas incluidas en el modelo (N)	{6, 9, 12, 15, 18, 21, 24, 27, 30}
Longitud de la serie de tiempo (T)	{100, 500, 1.000, 2.000, 3.000, 4.000, 5.000}
Para grafos *scale-free*, el parámetro alfa	0,05
Para grafos *scale-free*, el parámetro beta	0,9
Para grafos *scale-free*, el parámetro delta_in	3
Para grafos *scale-free*, el parámetro delta_out	2
Valores mínimo y máximo del rango de varianza	(1; 3)
Coeficientes negativos	Si
Covarianza negativa	Si
Estandarizar datos	No

Tabla 4.3: Descripción de los parámetros usados en la herramienta de simulación de datos *TETRAD* para generar los 56 conjuntos de datos reportadas. Se reportan en esta tabla tanto los parámetros fijos como los variables (siendo estos últimos los que están entre llaves). Por ejemplo, se varió T en el escenario 2 utilizando los valores entre llaves reportados en esta tabla, pero para todos los demás se lo fijó a $T = 1.069$. Por otro lado, los valores de H y L fueron variados en los escenarios 3 y 4 respectivamente (usando los valores entre llaves), pero en todos los demás, por simplicidad, se los fijó en $H = 0$ y $L = 1$. Por último, se varió N en el escenario 1 (usando los valores entre llaves), pero para todos los demás se usó el N más grande posible. Siendo estos N: $N = 30$, $N = 20$ y $N = 10$ para los escenarios 2, 3 y 4, respectivamente.

RFDAG y otro usando *SFDAG*. Finalmente se tiene un total de 56 conjuntos de datos diferentes para la fuente #1 (*TETRAD*).

4.3.2. Fuente #2: *CauseMe*

Para tener variedad en datos de origen sintético, y para agregar conjuntos de datos existentes (no solo los datos creados especialmente para este trabajo) se agregaron datos provenientes de la plataforma para evaluación comparativa (*benchmarking*) *CauseMe*[5] [RBB+19]. A la fecha en la que esta tesis fue escrita, había en dicha plataforma 19 conjuntos de datos disponibles para su descarga. Cualquiera de estos puede ser descargado

[5] causeme.net

y probado con un algoritmo de recuperación de estructura causal. Es importante resaltar que los arcos correctos (*ground truth*) no están disponibles para descargar. En su lugar, los resultados deben ser subidos a la plataforma para que allí se calculen las métricas de desempeño sobre la técnica propuesta.

En la documentación de *TETRAD* se aclara que los coeficientes son muestreados de una distribución uniforme, pero —de acuerdo a lo que se pudo observar— no se presenta la forma funcional con la que se computan los valores simulados. Como la forma funcional no está descrita, se asume que debe ser sencilla y por ende se asume lineal. Para agregar variedad a los conjuntos de datos sintéticos se toma de *CausaMe* el cojunto *nonlinear-VAR*[6].

De acuerdo a la descripción del conjunto de datos usado, los datos presentan tres desafíos usualmente encontrados en estos procesos estocásticos: autocorrelación, relaciones rezagadas en el tiempo y no linealidad. Se los combina con desafíos para las herramientas estadísticas/computacionales: dimensionalidad alta y series de tiempo cortas. El máximo rezago en las relaciones causales es 5 (distancia máxima entre la causa y el efecto).

El conjunto de datos *nonlinear-VAR* consiste de ocho repeticiones con diferentes parámetros. Cuatro repeticiones utilizando longitud de serie 300 ($T = 300$), para 3, 5, 10 y 20 nodos ($N \in \{3, 5, 10, 20\}$). Y cuatro repeticiones utilizando longitud de serie 600 ($T = 600$), para 3, 5, 10 y 20 nodos ($N \in \{3, 5, 10, 20\}$).

4.3.3. Fuente #3: *CAMMESA*

La tercera fuente de datos es la Compañía Administradora del Mercado Mayorista Eléctrico Sociedad Anónima (*CAMMESA*), una compañía argentina encargada de operar el mercado eléctrico mayorista de Argentina. Dicha compañía suministró para este trabajo los datos de la demanda de energía eléctrica en la zona metropolitana de la ciudad de Buenos Aires (Gran Buenos Aires (GBA)) para el periodo enero-2012 hasta diciembre-2018. Dentro de *CAMMESA* se trata de resolver el problema de predecir la demanda de energía eléctrica que va a haber en cada zona en el corto, mediano y largo plazo, para solicitar el suministro correspondiente a las compañías generadoras de energía eléctrica. Para realizar estas predicciones *CAMMESA* recopila información del clima y otros indicadores relevantes (por ejemplo, Estimador mensual de actividad económica (EMAE)). A

[6]https://causeme.uv.es/model/nonlinear-VAR/

partir de los datos suministrados por *CAMMESA* (que son de frecuencia horaria) se crea **un** conjunto de datos de tipo serie de tiempo con frecuencia diaria durante el periodo enero-2012 hasta diciembre-2018, teniendo un total de 2.558 observaciones y nueve variables. Las nueve variables describen diferentes aspectos de la zona analizada (GBA): (1) la demanda de energía eléctrica (DemGBA), (2) La temperatura promedio (Temp), (3) la componente vx del viento (vx), (4) la componente vy del viento (vy), (5) Irradiancia Horizontal Global (GHI), (6) la humedad (Hum), (7) la presión (Pres), (8) la sensación térmica (Ster) y (9) una variable que representa si el día actual es laborable o no (Wrk). Todas las variables son de tipo real excepto la última que se trata de una variable binaria.

El conjunto de datos provista por *CAMMESA* con frecuencia diaria contiene 2.558×9 datos (TxN), pero no contiene información respecto a la estructura causal real del problema. Sin embargo, analizando la naturaleza de las variables y conversando con los expertos de la compañía llegamos a las siguientes conclusiones respecto a la estructura causal real (*ground truth*).

***Ground truth CAMMESA*, simplificación**. Debido a que no se contaba con expertos en climatología no se tiene información detallada del modelo causal real respecto a las variables climatológicas entre sí. Por simplicidad se toman como referentes de las variables climáticas la humedad y la temperatura (que se entiende son las más vinculadas con la variable de interés, la demanda). Las demás variables no son consideradas para la *ground truth* ya que sumarlas no aportaba al conocimiento de este mismo y cada nueva variable climatológica agrega muchos arcos posibles sobre los que se tiene poca información. Esto quiere decir que para los experimentos y para reportar las métricas de desempeño solo se usan las dos variables climatológicas consideradas, la demanda y si es día laborable ($\{\text{Wrk}, \text{Hum}, \text{Temp}, \text{DemGBA}\}$).

***Ground truth CAMMESA*, arcos inexistentes**. Para construir la *ground truth* se parte de algunas relaciones que por sentido común se asumen incorrectas, esto es, arcos que no deberían ser encontrados por las herramientas de aprendizaje causal. Por ejemplo, (1) no debería existir un vínculo causal desde ninguna variable climática hacia la variable Wrk, el clima no afecta a que un día sea laborable o deje de serlo. Tampoco debería afectar la demanda de energía eléctrica a la variable Wrk. Análogamente, (2) se entiende que la demanda de energía eléctrica no puede afectar a ninguna variable climática ni tampoco determinar que un día sea feriado. Por último, (3) que un día sea feriado no puede afectar a ninguna variable climática. En resumen, sabemos que hay siete arcos incorrectos que no

deberían ser encontrados por las técnicas de descubrimiento causal. Estos arcos incorrectos están dibujados en rojo en la Figura 4.5. En dicha Figura la *ground truth* está representada con el grafo "desenrollado" (abajo) y con el grafo resumido (atemporal) (arriba). En el grafo de abajo los arcos deben leerse como no contemporáneos, esto es, una flecha de la forma $x \to y$ representa $x_{t-\tau} \to y_t$ con $\tau \neq 0$.

***Ground truth CAMMESA*, arcos existentes posibles**. De acuerdo a las discusiones con expertos de la empresa se asume un vínculo causal entre algunas variables climatológicas y la demanda de energía eléctrica en el Gran Buenos Aires. En particular, se sospecha un vínculo con la temperatura y la humedad ($\text{Temp}_{t-1} \to \text{DemGBA}$ y $\text{Hum}_{t-1} \to \text{DemGBA}$). También, de la misma discusión, se asume que si el día es laborable o no tiene un impacto en la demanda ($\text{Wrk}_t \to \text{DemGBA}_t$). Más aún, según los expertos, si el día anterior fue feriado tiene un impacto en la demanda (no es lo mismo la demanda un lunes, que un miércoles, que un viernes, o que un martes luego de un feriado) ($\text{Wrk}_{t-1} \to \text{DemGBA}_t$). Adicionalmente, si bien se desconoce con exactitud la relación entre las dos variables climáticas consideradas (Hum y Temp), se asume que se influencian entre sí en el tiempo ($\text{Hum}_{t-1} \to \text{Temp}_t$ y $\text{Temp}_{t-1} \to \text{Hum}_t$).

Por último se considera que las variables climáticas y la demanda tienen una fuerte inercia en sus valores, mostrando un comportamiento autorregresivo ($Temp_{t-1} \to \text{Temp}_t$, $\text{Hum}_{t-1} \to \text{Hum}_t$ y $\text{DemGBA}_{t-1} \to \text{DemGBA}_t$). Adicionalmente, se asume que la variable feriado tiene una componente autorregresiva, aunque menor que para las tres variables anteriores ($\text{Wrk}_{t-1} \to \text{Wrk}_t$). Esto se debe a que los días hábiles se tienden a agrupan juntos y los días no laborables también. En una semana regular sin feriados, luego de un día hábil (lunes, martes, miércoles, jueves o viernes) hay una probabilidad de 4/5 de que el día siguiente sea día hábil nuevamente y solo 1/5 de que sea día no laborable (solo si el día actual es viernes).

Un resumen de los arcos correctos esperados se puede ver en la Figura 4.5. Los arcos correctos están representados en negro del lado derecho de la Figura. En la parte superior está representado el grafo atemporal, y en la parte de abajo está representado el grafo "desenrollado". Como se puede observar, los arcos autorregresivos no son representados en la representación atemporal ya que no es el objetivo de este trabajo capturar relaciones causales de variables hacia sí mismas ($x_{t-1} \to x_t$).

La aplicación de técnicas de descubrimiento causal a este conjunto de datos es presentada y analizada en la Sección4.5. Las técnicas son comparadas en base a la *ground truth*

aquí discutida. Diferentes transformaciones de los datos para la eliminación de ciclos son consideradas para sumar al análisis comparativo de las técnicas de descubrimiento causal.

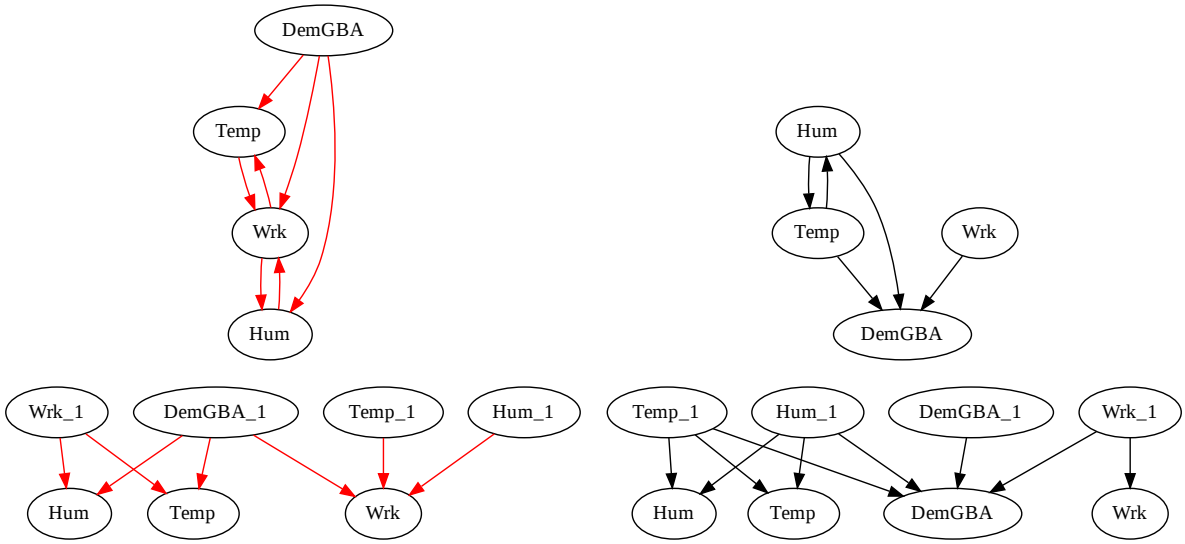


Figura 4.5: Descripción gráfica de la estructura causal real (*ground truth*) del conjunto de datos obtenido de la empresa *CAMMESA*. A la izquierda (a y c) están representados los arcos incorrectos en rojo. A la derecha (b y d) están representados en negro los arcos correctos. La *ground truth* se muestra en dos formatos distintos (pero codificando la misma información): abajo (c y d) se muestran los grafos "desenrollados" en el tiempo, mostrando cómo se afectan causalmente las variables de un instante al siguiente (como se trabaja con frecuencia diaria, de un instante al siguiente hay 24 horas de diferencia). Como las flechas contemporáneas no son buscadas, las mismas no se analizan ni reportan en este trabajo. En la parte superior de la figura (a y b) se muestra el grafo atemporal resumido, donde cada arco representa una relación de causalidad de un instante al otro. Esto es, la flecha Hum $\rightarrow$ Temp en (b) se debe leer como $\text{Hum}_{t-1} \rightarrow \text{Temp}_t$. Como se puede observar, los arcos autorregresivos no son representados en la representación atemporal ya que no es el objetivo de este trabajo capturar relaciones causales de variables hacia sí mismas ($x_{t-1} \rightarrow x_t$). Para obtener esta estructura de arcos incorrectos y arcos correctos se analizó la naturaleza de las variables y se conversó con los expertos de la compañía.

4.3.4. Fuente #4: *The New York Times*

La cuarta fuente de datos es el *The New York Times Annotated Corpus*[7] [San08]. El objetivo de la utilización de textos en el presente trabajo es generar un prototipo de una herramienta que permita a los expertos entender y sacar conclusiones sobre eventos y otras variables del mundo real que hayan sido reportados en artículos de noticias. La

[7] https://catalog.ldc.upenn.edu/LDC2008T19

herramienta tiene por objetivo detectar y mostrar variables relevantes a un dado dominio de manera semiautomática, y mostrar posibles vínculos causales entre esas variables para permitir a los expertos un mejor entendimiento del dominio analizado. En el presente trabajo se muestra un caso de uso de la herramienta aplicada no a un dominio específico sino a una entidad geopolítica (GPE por sus siglas en inglés) particular, en este caso Irak. Para elegir una GPE con una buena cantidad de menciones se utilizó un detector de entidades (*NER*) de la librería *spaCy* para detectar todas las menciones de dichas entidades en todo el corpus del *New York Times* en el periodo enero-1987 a junio-2007. Se detectaron las siguientes 10 GPEs como las más mencionadas en el corpus (se muestran ordenadas de las más mencionadas a las menos mencionadas): `New York`, `the United States`, `Manhattan`, `Washington`, `Iraq`, `New York City`, `America`, `China`, `Brooklyn` y `New Jersey`. Se prefirió elegir una GPE fuera del país de origen del corpus ya que sobre el mismo país probablemente haya mayor cantidad de noticias y más variadas que sobre países extranjeros. Por este motivo se eligió el primer GPE externo a Estados Unidos, Irak (con 180.206 menciones en un total de 170.497 oraciones distintas).

A partir del texto completo de las noticias de *The New York Times* en el período enero-1987 a junio-2007 se filtra por GPE, solo considerando como dominio de interés las oraciones con menciones explícitas del país Irak. A partir de estos textos se construye un conjunto de datos de tipo series de tiempo con variables de interés con el objetivo de aplicar sobre ella las técnicas de descubrimiento de estructura causal. Para la construcción de este conjunto de datos se consideraron dos tipos de variables de interés: términos mencionados en el texto (que pueden ser unigramas, bigramas o trigramas) y menciones de eventos en curso del mundo real. Para el primero se utiliza la técnica de pesaje de términos FDD_β presentada en el Capítulo 2 y para el segundo tipo de variables se utiliza el modelo de detección de eventos en curso presentado en el Capítulo 3. Es importante mencionar que el filtro de GPE y todo el trabajo posterior es a nivel de oraciones y no de artículos completos para que sea compatible con el trabajo del Capítulo 3 donde se trabaja con la detección de eventos a nivel de oración y no de artículos.

Las variables de ambos tipos (términos y eventos en curso) son detectadas en el texto de manera independiente, y a cada mención de cada variable se le asocia el mes y año de publicación del artículo donde aparece reportada. De esta manera se generan dos conjuntos de datos preliminares, uno con menciones de términos (con M términos) y otro con menciones de eventos (con P eventos). Como se utiliza la misma frecuencia y

el mismo periodo de tiempo (frecuencia mensual en el periodo enero-1987 a junio-2007) ambos conjuntos tienen la misma longitud (246 meses) y por ende estos pueden ser unidos. Esta unión resulta en el conjunto de datos final de dimensiones $246 \times (P + M)$, donde $(P + M)$ es la cantidad de variables y 246 es la longitud de las series. A continuación, se describe el proceso a través del cual se generan los dos conjuntos de datos preliminares, el de menciones de términos $(246 \times M)$ y el de menciones de eventos en curso $(246 \times P)$.

Conjunto de datos de menciones de términos. Para poder construir este conjunto de datos de términos se debió detectar repeticiones de unigramas, bigramas y trigramas para construir un vocabulario y luego utilizar la técnica de pesaje de términos FDD_β definida en el Capítulo 2 para obtener un conjunto reducido de términos altamente relevantes sobre el cual elegir los M términos finales. Como la técnica de pesaje de términos a ser utilizada es supervisada, se necesita un conjunto de textos relevantes al contexto y otro conjunto de textos irrelevantes para poder estimar la relevancia de cada término para el contexto dado. En este caso el contexto es el país Irak, por ende, se usa como conjunto relevante las 170.497 oraciones que mencionan explícitamente al GPE, y se selecciona un conjunto de la misma cantidad de oraciones de forma aleatoria de las 63.734.239 oraciones restantes del corpus (las que no mencionan al GPE Irak). El conjunto final tiene 340.994 oraciones, donde la mitad mencionan al país de interés y la otra mitad no. Notar que el término "`Iraq`" va a estar presente en todas las oraciones relevantes y no estará en las irrelevantes, alcanzando un poder tanto descriptivo como discriminativo de $1,0$.

A partir de este conjunto de datos se construye un vocabulario de términos (unigramas, bigramas y trigramas) presentes en el corpus. Se descartan automáticamente términos que no aparezcan ni una vez en el conjunto de relevantes (ya que tienen poder descriptivo y discriminativo cero). Se descartan también de forma automática aquellos términos que aparezcan en menos del $0,1\,\%$ del corpus (341 menciones o menos), ya que esa cantidad de menciones no alcanzaría para construir una serie de tiempo con suficientes datos como para aplicar las técnicas de causalidad. Por último, se filtran también aquellos términos que representen cantidades o signos de puntuación, así como también *stopwords*. A la totalidad de los términos restantes del vocabulario se les aplica FDD_β utilizando el mejor β encontrado durante los experimentos del Capítulo 2 para la estimación de relevancia de términos asignados por usuarios, esto es $\beta = 0,477$. En la Tabla 4.4 se muestran tres listas con los 10 unigramas, 10 bigramas y 10 trigramas con mayor FDD_β de todo el vocabulario utilizado. Como el objetivo de la aplicación de la técnica FDD_β era descubrir

nuevos términos relevantes al dominio Irak, no se consideró "`Iraq`" como una palabra relevante a ser descubierta y por ende se la descartó de la lista de unigramas.

La detección de variables es una tarea que se la considera no totalmente automatizable, ya que para diferentes usuarios la definición de variable relevante puede variar. Por este motivo es que si bien las herramientas propuestas en este capítulo detectan variables de manera automática, se espera que el usuario del sistema sea el que finalmente decida, a partir de una lista reducida de variables posibles, si incorpora o no a cada una de estas. Para el caso de los términos, se espera que dada la lista de 30 palabras presentadas en la Tabla 4.4, el usuario decida qué variables le parecen relevantes para ser usadas en el próximo paso de detección de estructura causal. Vale aclarar que el usuario podría ajustar el valor de β o la cantidad de términos a visualizar (en este caso 30) para tener más o diferentes opciones para elegir. A modo de ejemplo en este trabajo se eligen los 10 términos ($M = 10$) que se consideran más relevantes para el estudio: ('`weapons`', '`mass`', '`destruction`'), ('`Persian`', '`Gulf`', '`war`'), ('`United`', '`Nations`', '`Security`'), ('`Iraq`', '`invasion`', '`Kuwait`'), ('`chemical`', '`biological`', '`weapons`'), ('`military`', '`action`', '`Iraq`'), ('`United`', '`States`'), ('`war`', '`Iraq`'), ('`Saddam`', '`Hussein`') y ('`Bush`', '`administration`'). Varios trigramas son elegidos por tener una semántica fácil de interpretar y con mucha relevancia para el dominio. Por otra parte otros son ignorados por encontrarse representadas en otros. Por ejemplo ('`President`', '`Saddam`', '`Hussein`') subsume a los trigramas ('`Saddam`', '`Hussein`', '`Iraq`') y a ('`Saddam`', '`Hussein`'). Por ende se usa el término ('`Saddam`', '`Hussein`') que a la vez es el de mayor FDD_β entre los tres. En la Tabla 4.4 además de mostrarse los diez unigramas, diez bigramas y diez trigramas con mayor FDD_β, también se indica con un asterisco los diez términos seleccionados previamente mencionados.

Finalmente, el conjunto de datos de menciones de términos se construye contando la frecuencia de aparición de cada uno de estos términos a lo largo de los 246 meses del periodo abarcado por el corpus de *The New York Times*. Las dimensiones finales del conjunto de datos de términos son de 246×10. En la Tabla 4.6 se reporta estadística descriptiva sobre el conjunto de datos de términos creado.

Conjunto de datos de menciones de eventos. Para poder construir un conjunto de datos de eventos en curso mencionados en el corpus del *New York Times* se realizaron los siguientes cuatro pasos. Primero, se utiliza el modelo presentado en el Capítulo 3 [MDT+21] para extraer los eventos en curso de todas las oraciones del corpus. Segundo,

Unigrama	DISCR	DESCR	FDD_β	Bigrama	DISCR	DESCR	FDD_β	Trigrama	DISCR	DESCR	FDD_β
war	0,975	0,154	0,490	('United', 'States')*	0,896	0,086	0,325	('President', 'Saddam', 'Hussein')	0,998	0,014	0,073
United	0,920	0,144	0,460	('war', 'Iraq')*	1,000	0,071	0,292	('weapons', 'mass', 'destruction')*	0,996	0,012	0,061
American	0,902	0,131	0,432	('United', 'Nations')	0,975	0,065	0,270	('Persian', 'Gulf', 'war')*	0,995	0,011	0,056
said	0,623	0,181	0,428	('Saddam', 'Hussein')*	0,994	0,041	0,189	('Saddam', 'Hussein', 'Iraq')	1,000	0,010	0,052
Mr.	0,660	0,146	0,399	('Mr.', 'Bush')	0,935	0,034	0,159	('United', 'Nations', 'Security')*	0,989	0,006	0,031
Bush	0,949	0,100	0,368	('President', 'Bush')	0,965	0,033	0,154	('Nations', 'Security', 'Council')	0,989	0,006	0,031
States	0,894	0,086	0,325	('Security', 'Council')	0,989	0,028	0,134	('Iraq', 'invasion', 'Kuwait')*	1,000	0,006	0,031
military	0,954	0,079	0,312	('Persian', 'Gulf')	0,986	0,022	0,108	('Iraq', 'invaded', 'Kuwait')	1,000	0,005	0,027
Hussein	0,991	0,069	0,285	('Bush', 'administration')*	0,980	0,022	0,108	('chemical', 'biological', 'weapons')*	1,000	0,004	0,022
Iraqi	0,985	0,067	0,279	('Mr.', 'Hussein')	0,991	0,021	0,105	('military', 'action', 'Iraq')*	1,000	0,004	0,022

Tabla 4.4: En esta tabla se presentan los 10 mejores unigramas, los 10 mejores bigramas y los 10 mejores trigramas de acuerdo a la técnica FDD_β con $beta = 0,477$, siendo este valor de β el que obtuvo el mejor desempeño como estimador de relevancia de términos para un dominio por parte de usuarios (análisis presentado en el Capítulo 2). Para cada término (unigrama, bigrama, trigrama) se reporta su poder descriptivo, su poder discriminativo y el puntaje de FDD_β obtenido. Un usuario potencial podría elegir el β y ajustar la cantidad de términos que se visualizan (en este caso 30) para luego manualmente elegir cuáles utilizar para el análisis causal. A modo de caso de uso se eligen los 10 que resultan más interesantes para el análisis posterior causal, estos 10 términos son marcados con un asterisco.

como cada evento es distinto (con diferentes contextos alrededor del *trigger*), para poder compararlos se construyó una representación vectorial para cada uno de ellos. Tercero, como cada representación vectorial es única, para poder agrupar menciones equivalentes (mismo evento con diferentes palabras) se aplicó una técnica de agrupamiento para encontrar K grupos de menciones de eventos. Por último, el cuarto paso fue seleccionar de los K grupos los P más relevantes para ser utilizados como las variables del conjunto de datos. Cada grupo seleccionado es un evento distinto (variable) y cada instancia dentro del grupo constituye una mención de dicho evento.

Al igual que para el conjunto de datos de términos, se utiliza la totalidad del periodo cubierto por el *New York Times* (enero-1987 a junio-2007) usando frecuencia mensual (246 meses en total), resultando en un conjunto de datos de menciones de eventos con dimensión $246 \times P$. A continuación, se describen en detalle los cuatro pasos antes mencionados que permiten obtener las series de tiempo a partir de los textos completos del *New York Times*.

Paso #1, detección de menciones de eventos en curso. Al igual que para el conjunto de datos de menciones de términos se utiliza un dominio de interés para ilustrar un posible caso de uso de la herramienta. Nuevamente se utilizan todas las oraciones que mencionan al GPE "`Iraq`". Se utiliza el modelo de detección de eventos en curso presentado en el Capítulo 3 sobre las 170.497 oraciones que contienen al país de interés. Se detectan un total de 498.560 menciones de eventos en total (un promedio de 2,92 eventos por oración).

Paso #2, construcción de una representación vectorial para cada mención. A partir de cada una de las 498.560 menciones de eventos en curso, se define la tarea de agrupar menciones del mismo evento en un mismo grupo para poder construir la serie de tiempo de menciones de cada evento en el tiempo. Para poder construir los grupos se plantea el desafío de construir una representación vectorial para cada mención de tal manera que permita la comparación entre eventos. Se espera que eventos semánticamente similares estén cercanos en esta representación. Para poder comparar menciones de eventos se considera que no solo el *event-trigger* es importante sino todo su contexto. Por esta razón, en este trabajo se introduce la representación de la frase del evento (Event-Phrase Embedding Representation (EPER)), que se define como una suma de representaciones *GloVe* [PSM14] con un decaimiento cuadrático. Esto es, se considera como la parte más importante del evento al *event-trigger* y por tal motivo la representación *GloVe* de esta palabra es incluida en la representación sin penalización. Por otra parte, cada *token* a la

izquierda y a la derecha hasta terminar la oración, dan lugar a representaciones *GloVe*, las que son sumadas a la representación ajustadas por un coeficiente de penalización de tal manera que cuanto más lejos del *trigger* se encuentra el token representado, más fuerte es la penalización. Esta penalización es de orden cuadrático. La definición formal de la EPER se puede ver en la ecuación presentada a continuación:

$$\text{EPER}(e_k, P) = \sum_{w_i \in P} \frac{1}{(|k-i|+1)^2} \cdot \text{GloVe}(w_i) \tag{4.8}$$

En esta fórmula se tiene una frase P compuesta de palabras w_i definida como sigue: $P = w_1, w_2, \ldots, w_n$. Siendo el *event-trigger* e_k la palabra w_k, para algún k, $1 \leq k \leq n$. De esta manera se agrega la representación de cada palabra de la oración, pero con un factor de penalización que crece cuadráticamente con la distancia al *event-trigger*. Así, se tiene una representación de 300 dimensiones de cada mención de evento que tiene en cuenta principalmente el *trigger* y en menor medida todas las palabras del contexto. Al finalizar la construcción de esta representación se tienen 498.560 vectores de 300 dimensiones, uno para cada mención de evento en curso detectado en las 170.497 oraciones que mencionan a Irak.

Paso #3, aplicación de una técnica de agrupamiento para detectar menciones del mismo evento. Finalmente, teniendo una representación vectorial para cada mención de evento se procede a realizar el agrupamiento planteado para poder unir los 498.560 vectores en K grupos donde cada grupo representa el mismo evento semántico y cada instancia dentro del grupo representa una mención. El primer desafío que se plantea es la elección de la técnica de agrupamiento adecuada. Para el presente problema se emplea la clásica y ya bien establecida técnica de agrupamiento *KMeans* [Llo82].

El segundo desafío que se presenta es la elección del valor de K (cantidad de clusters) adecuado para obtener la granularidad adecuada en cada grupo. Esto es, lo suficientemente grande como para tener múltiples menciones del mismo evento, sin comprometer la cohesividad del grupo y que se terminen agrupando eventos distintos. Para poder elegir el K se procedió a utilizar la técnica gráfica *Elbow* [Tho53] para analizar el resultado de aplicar agrupamiento para diferentes valores de K. Se utiliza como métrica de cohesión de los grupos la distancia a los centroides al cuadrado (inercia). Se toma un conjunto de valores tentativos de K ($K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 50, 100, 500, 1.000, 5.000, 6.000, 7.000, 8.000, 9.000, 10.000, 50.000, 100.000\}$) y se grafica la inercia obtenida por cada

agrupamiento para esos valores de K. Debido a la gran cantidad de instancias, el costo computacional de la técnica *KMeans* y la cantidad de valores de K a probar, para este análisis, no se pudo usar la técnica *KMeans* tradicional. En su lugar se utilizó la técnica *MiniBatch KMeans* [Scu10], la cual presenta modificaciones a la técnica original de tal modo que permite alcanzar soluciones similares a un costo computacional mucho menor. El resultado de este procedimiento se puede ver en la Figura 4.6.

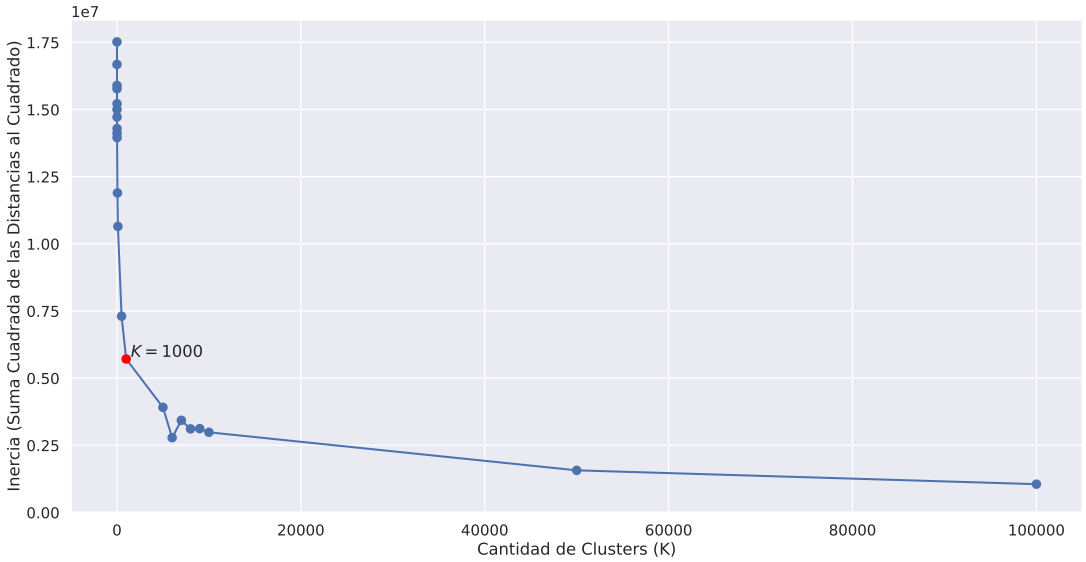


Figura 4.6: Visualización de la inercia de aplicar la técnica de agrupamiento *MiniBatch KMeans* para diferentes valores de K. La inercia se mide como la suma cuadrada de las distancias al centroide. A mayor K, mayor cantidad de grupos, y más cerca está el centroide de cada grupo a las instancias dentro del mismo, por ende, alcanzando una inercia menor. En el extremo se tendría un grupo por instancia, con un centroide que coincide con dicha instancia, en cuyo caso se tendría inercia cero. Se observa una caída pronunciada del valor de inercia hasta $K = 1.000$ (valor dibujado en rojo). Luego, cada aumento en K no tiene un beneficio grande en disminución de inercia. Por este motivo se selecciona $K = 1.000$ como el valor óptimo de grupos a usar.

Como se esperaba para pocos grupos (K pequeños) la distancia a los centroides es mayor. En el extremo, para $K = 1$ con un solo centroide en el centro de las 498.560 instancias, se tiene la máxima suma de distancias cuadradas, siendo esta igual a 17.516.440,97. En el otro extremo, con K igual a la cantidad de instancias, se tiene que cada instancia es su propio grupo, y por ende cada centroide coincide con esa instancia. Esto da lugar a una inercia de cero. Se puede observar que hasta $K = 1.000$ se tenía una pendiente pronunciada, donde cada crecimiento en el valor de K generaba una gran diferencia en la inercia resultante. A partir de ese valor se puede observar como la pendiente se achata.

Por este motivo, se concluye que el punto de inflexión más grande está en $K = 1.000$ y se continúa con ese valor de K para los próximos pasos. Ya con el valor de K definido, para construir el conjunto de datos final se utiliza el algoritmo *KMeans* tradicional (ya no su optimización *MiniBatch Kmeans*) con el valor de K elegido ($K = 1.000$). La agrupación resultante tiene un valor de inercia de 5.341.146,00. Un histograma con la cantidad de menciones en cada grupo se reporta en la Figura 4.7.

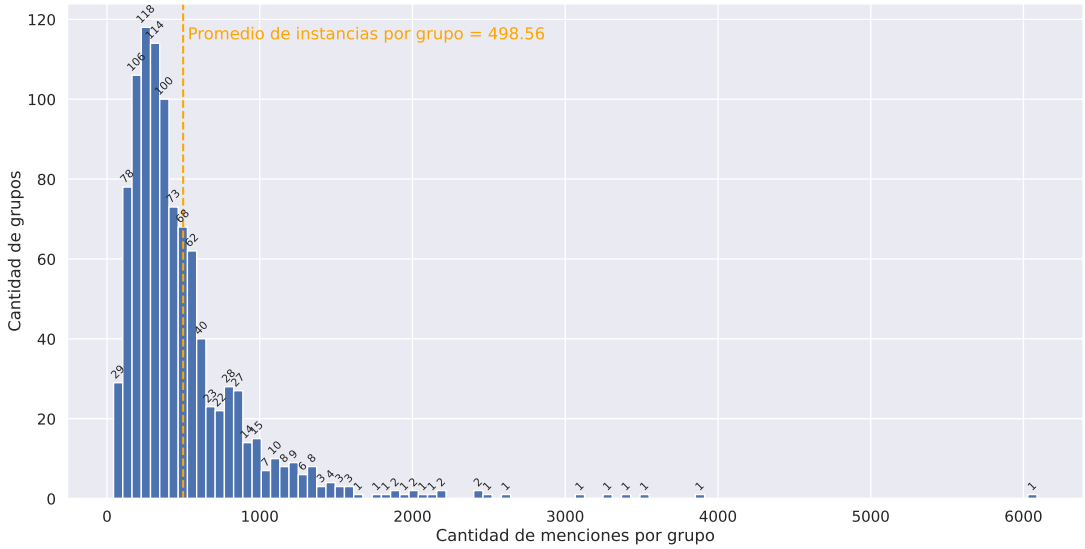


Figura 4.7: Histograma que muestra la dispersión de los tamaños de los grupos obtenidos al aplicar *KMeans* con $K = 1.000$. Como se puede observar, el promedio de instancias por grupo es de casi 500 instancias. Sin embargo, se pueden observar algunos grupos muy poblados: un grupo con más de 6.000 instancias, seis grupos con más de 3.000 instancias en cada uno. Por otra parte, hay 23 grupos con 103 elementos o menos y 107 con 163 elementos o menos.

Paso #4, selección de los P grupos más relevantes y construcción del conjunto de datos de eventos. A partir del agrupamiento realizado por la técnica *KMeans* con $K = 1.000$ se analizan los 1.000 grupos resultantes en términos de la cohesión y cantidad de menciones por grupo (p_i). El objetivo es seleccionar grupos con alta cantidad de menciones para poder aplicar las técnicas de descubrimiento causal, y con alto nivel de cohesión para preferir grupos con una semántica específica y bien definida (de preferencia que se refieran a un único evento del mundo real bien definido). La cohesión, en este trabajo, se computa utilizando la distancia cuadrada promedio al centroide para cada grupo (d_i). Usando este valor, la cohesión del grupo (c_i) se define como $c_i = 1/(d_i + 1)$. De esta manera se trata de tomar grupos que maximicen ambos valores: c_i y p_i. Para encontrar grupos que tengan un buen balance de ambos indicadores se plantea maximizar una fun-

ción (g) que calcule la media armónica entre estos dos valores. Para poder comparar los indicadores en la misma escala se los divide por el máximo de cada escala ($\max(c_i)$ y $\max(p_i)$) para llevarlos al rango $[0, 1]$. Luego la función g queda definida como sigue:

$$g(c_i, p_i) = 2 \times \frac{c_i/\max(c_i) \times p_i/\max(p_i)}{c_i/\max(c_i) + p_i/\max(p_i)} \tag{4.9}$$

Se analizaron diferentes umbrales (u) para la función g ($g(c_i, p_i) > u$) con $u \in [0,10;\ 0,40]$ con un paso de $0,01$. Por ejemplo para los valores de u de $0,10$, $0,20$ y $0,30$, solo 374, 85 y 28 grupos, respectivamente, obtuvieron un valor de g por encima del umbral fijado. Al buscar solo los grupos que estén por encima del umbral se prioriza al mismo tiempo (y por igual) la cohesividad y el tamaño del grupo de forma tal que los pequeños o poco cohesivos son descartados. Utilizando esta estrategia se realizó un filtrado de grupos de tal forma de obtener un conjunto reducido de grupos (20 o menos) altamente relevantes sobre los cuales elegir manualmente los más interesantes. Para obtener estos 20 grupos se buscó el mínimo umbral a partir del cual solo 20 instancias o menos son seleccionadas (umbral $u = 0,35$).

Al igual que para el conjunto de datos de términos, se considera que la herramienta debe hacer un trabajo automático para filtrar variables relevantes de un gran conjunto hasta obtener una cantidad manejable por un usuario (en este caso 20), y que sea el usuario quien elija las variables finales que quiere incluir en el modelo. Para este trabajo como caso de estudio se analizan manualmente los veinte grupos resultantes y se seleccionan seis, los cuales se identifican con las etiquetas: `C109`, `C165`, `C201`, `C249`, `C269` y `C550`. Las etiquetas de los grupos son creadas utilizando el número de grupo obtenido del algoritmo *KMeans* anteponiendo la letra C a dicho número (resultando en las siguientes posibles etiquetas: `C000`, `C001`, ..., `C998` y `C999`).

En la Figura 4.8 se pueden observar los 1.000 grupos en el plano donde en el eje horizontal se tienen la cohesión y en el eje vertical se tiene el tamaño de los grupos. Se dibuja la línea negra que representa el umbral $u = 0,35$ y en rojo los 20 grupos que quedan por encima del umbral. Los grupos representados con una cruz roja forman parte de los seis manualmente elegidos mientras que los marcados con círculos rojos son los que están por encima del umbral pero no fueron seleccionados. Se etiqueta cada grupo por encima del umbral con su nombre de grupo (en gris los no seleccionados y negro los seleccionados). Los 980 puntos azules por debajo de la línea son los grupos descartados automáticamente por no tener una buena combinación de cohesión y tamaño (están por debajo del umbral).

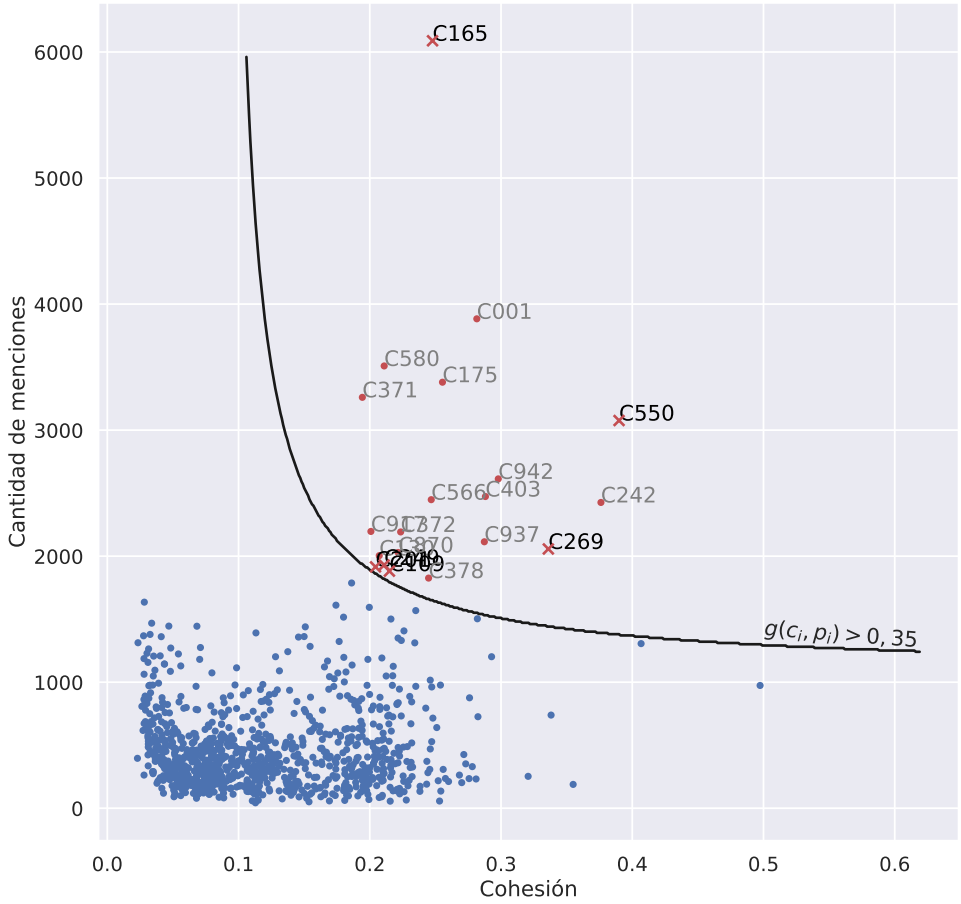


Figura 4.8: Gráfico de dispersión donde se visualizan los 1.000 grupos obtenidos con *KMeans* con $k = 1.000$. Se comparan los grupos en términos de cohesión y cantidad de menciones (instancias) dentro del grupo. La cohesión de cada grupo representa la similitud promedio de cada instancia del grupo con respecto a su centroide. Cuanto más grande el valor de cohesión mayor similitud entre las instancias dentro del grupo. Se buscan los mejores veinte grupos en términos de mayor cohesión y mayor cantidad de instancias. Para obtener los grupos que tengan un buen balance de estas dos métricas se computa la función g que representa la media armónica de los valores normalizados de estas métricas para cada instancia. Se mueve el umbral hasta encontrar exactamente 20 grupos por encima del umbral establecido para la función g (línea sólida negra), los 20 grupos elegidos se marcan en rojo (con cruz y círculo). De los veinte grupos se seleccionan manualmente seis (marcados con una cruz) para el análisis causal posterior. Para los veinte grupos en rojo se muestra la etiqueta del grupo, en negro para los seis manualmente seleccionados y en gris para los demás. Los puntos azules son los otros 980 grupos que **no** tienen un valor por encima del umbral ($g \leq 0,35$).

Para la selección manual de los grupos, como no es posible analizarlos en su totalidad (porque cada uno de los 20 grupos tiene cientos de menciones de eventos), se utilizó una representación visual de textos clásica: nubes de palabras. Se construyó una nube

de palabras para cada uno de los veinte grupos utilizando los mismos coeficientes de penalización usados para calcular los EPER (Ecuación 4.8). Esto es, para cada grupo se tomó cada mención de cada evento involucrado y se incluyó la totalidad de las palabras de la oración que contiene el evento, pero penalizando su importancia de acuerdo a qué tan lejos está del *event-trigger* de la mención. El *event-trigger* siempre es incluido sin penalización y cada palabra del contexto es incluida también en la nube de palabras, pero penalizada de manera cuadrática de acuerdo a la distancia al *event-trigger*. Las representaciones resultantes para los seis grupos elegidos se pueden ver en la Figura 4.9. En la Tabla 4.7 se reportan valores de estadística descriptiva del conjunto de eventos creado.

Como se puede observar en las nubes de palabras, se eligieron grupos que tuvieran una semántica clara y definida, y que puedan asociarse a eventos del mundo real. Por ejemplo, el grupo `C109` parece corresponderse con reportes de muertes durante la Guerra de Irak, tanto soldados (de ambos bandos) como civiles. Por otra parte, el grupo `C165` parece corresponderse con reportes de oposición a la guerra con Irak. El grupo `C201` parece tratarse principalmente de reportes de ataques terroristas, el grupo `C249` parece concentrarse en ataques o acciones militares (aparentemente por parte de Estados Unidos). El grupo `C269` trata sobre la invasión a Kuwait por parte de Irak, mientras que el grupo `550` reúne menciones de la guerra en Irak en general. Una descripción de los grupos es presentada en la Tabla 4.5.

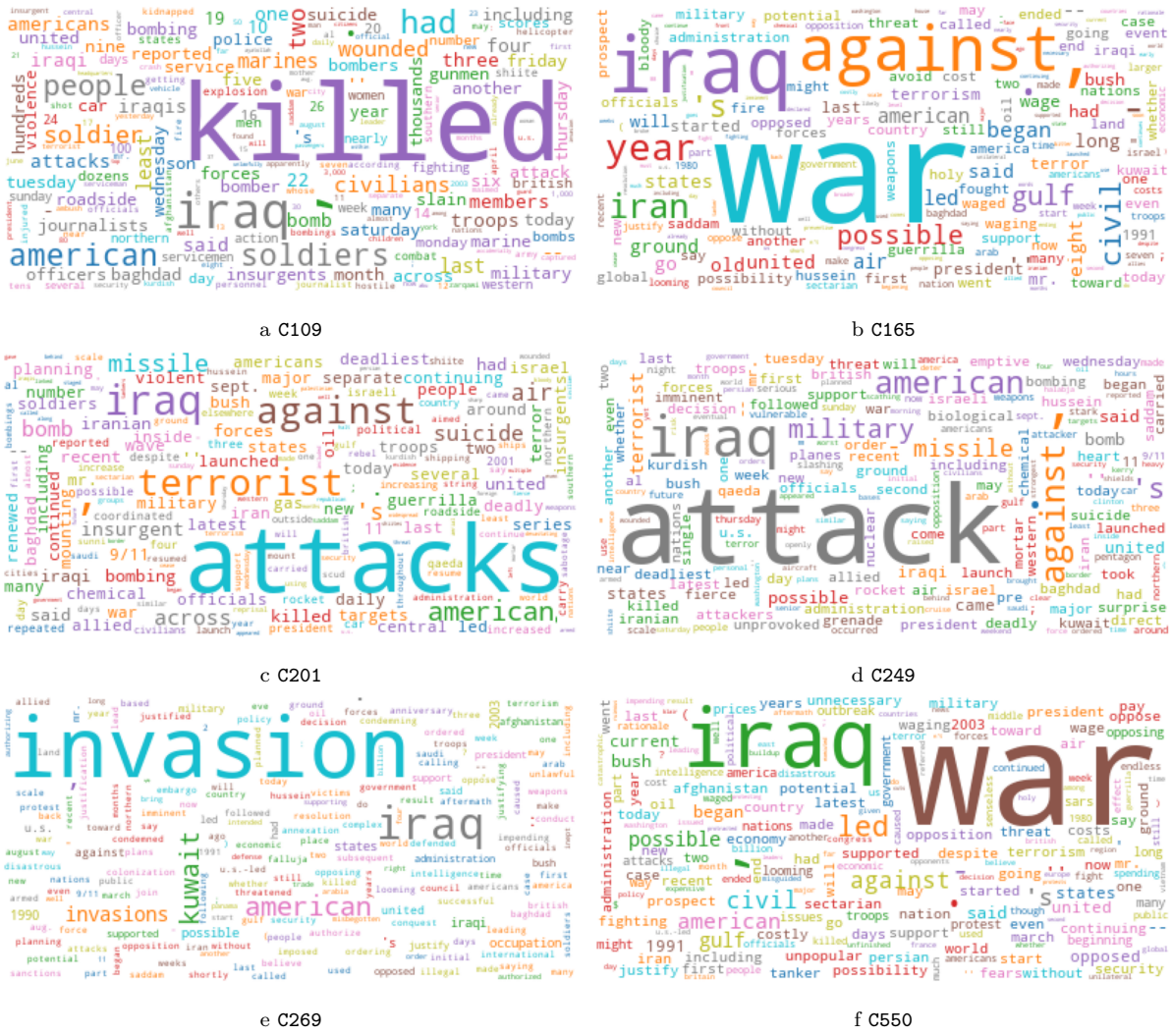
Conjunto de datos de final. Finalmente, luego de utilizar la técnica de pesaje de términos del Capítulo 2 para construir el conjunto de datos de términos (de dimensiones 246×10) y el modelo de detección de eventos en curso del Capítulo 3 para construir un conjunto de datos de eventos (de dimensiones 246×6) se construye un conjunto de datos final. Dado que el conjunto de datos de términos y el conjunto de datos de eventos en curso comparten la frecuencia (mensual) y el periodo de tiempo (enero-1987 a junio-2007), solo hay que juntar las variables en un solo conjunto de datos final (de dimensiones 246×16). Vale la pena mencionar que el propósito de utilizar las herramientas de los Capítulos 2 y 3 no es llegar a un único conjunto de datos con 16 variables, sino dar a un usuario potencial un conjunto manejable de variables, para que este usuario pueda seleccionar manualmente las variables que quiere utilizar para el aprendizaje de estructura causal. Si bien se pueden incorporar muchas más variables, es necesario que cada usuario elija las variables que considere relevantes para que la estructura causal resultante sea analizable de manera

Etiqueta de grupo	Descripción
C109	Reportes de muertes de soldados y civiles (términos salientes: killed, Iraq, american, soldiers, civilians)
C165	Menciones de la guerra de Irak, con una componente negativa (términos salientes: against, war, iraq)
C201	Reportes de ataques terroristas (términos salientes: attacks, terrorist, iraq)
C249	Menciones de acciones militares (términos salientes: attack, iraq, military, missile, against
C269	Menciones de la invasión de Kuwait (términos salientes: invasion, iraq, kuwait, invasions, american)
C550	Menciones de la Guerra de Irak (términos salientes: war, iraq, led, 2003)

Tabla 4.5: Descripción de seis eventos extraídos del corpus del *New York Times* para ser usados como variables del *framework* de recuperación de estructuras causales a partir de textos. Cada uno de estos eventos (o variables) consiste de múltiples menciones en diferentes textos del mismo evento. Por ejemplo, el grupo C109 consiste de múltiples menciones del mismo tipo de evento: reportes de muertes de soldados o civiles. Estos grupos son construidos a través de una técnica de agrupamiento (*KMeans*) la cual agrupó menciones de eventos semánticamente similares en el mismo grupo. Se toman seis de los 1.000 grupos formados, priorizando grupos con alta cohesión y mayor cantidad de menciones. Esto es, se necesitan variables que tengan una semántica bien definida y con muchas menciones.

sencilla y no esté superpoblada con demasiados nodos. En este trabajo en particular se redujo un gran vocabulario a 30 posibles términos. A la vez, se redujeron 1.000 eventos (grupos) a solo 20. En el presente trabajo se toman esas 50 variables sugeridas y se eligen 16 variables relevantes para mostrar un posible caso de uso de la herramienta.

Todos los conjuntos de datos sintéticos obtenidas de las fuentes *TETRAD* y *CauseMe*, como así también el conjunto de datos de demanda de energía eléctrica en el GBA (origen *CAMMESA*) se utilizan para evaluar las diferentes técnicas de aprendizaje de estructuras causales con el objetivo de elegir las mejores para ser usadas luego en los conjuntos de datos creados a partir de los textos del *New York Times*. Primero se presentan los resultados del análisis en los datos sintéticos en la Sección 4.4. Posteriormente se presenta el análisis sobre el conjunto de datos de *CAMMESA* en la Sección 4.5. Por último, en la Sección 4.6 tanto el conjunto de datos de términos (246×10) como el de eventos (246×6) como la



a C109 b C165

c C201 d C249

e C269 f C550

Figura 4.9: Seis nubes de palabras que describen a los seis grupos (*clusters*) elegidos manualmente a partir de las veinte opciones que fueron seleccionadas por tener mejor balance de cohesión y cantidad de menciones (los 20 por encima de la línea negra en la Figura 4.8). Estos seis grupos son los grupos Nro. 109, 165, 201, 249, 269 y 550 del total de 1.000 grupos construidos con *KMeans* ($K = 1.000$) (numerados desde el Nro. 0 hasta el 999). Como se puede observar existe una semántica identificable en cada grupo: (a) reporte de muertes debido a la Guerra en Irak (tanto civiles como soldados), (b) reportes de sentimientos en contra de la Guerra, (c) ataques terroristas, (d) ataques militares americanos contra Irak, (e) invasión de Kuwait por parte de Irak y (f) menciones de la Guerra de Irak.

unión de ambos (246×16) son utilizados para aprendizaje de estructura causal usando las herramientas que parecen más apropiadas para el dominio. De esta manera se demuestra el prototipo completo de una herramienta causal que parte de información textual, extrae variables de interés y muestra vínculos causales entre ellas.

	('United', 'States')	('war', 'Iraq')	('Saddam', 'Hussein')	('Bush', 'administration')	('weapons', 'mass', 'destruction')	('Persian', 'Gulf', 'war')	('United', 'Nations', 'Security')	('Iraq', 'invasion', 'Kuwait')	('chemical', 'biological', 'weapons')	('military', 'action', 'Iraq')
mean	2699,73	49,37	98,07	144,33	21,80	33,76	20,39	4,05	7,33	2,84
std	562,88	105,56	159,61	214,57	35,47	48,38	18,38	13,94	12,12	8,34
min	1757	0	0	0	0	0	0	0	0	0
25 %	2291,25	2	15	0	4	9,25	8	0	1	0
50 %	2579	7	37	1	10	19	16	1	3	0
75 %	3008,75	59	104,5	365,75	25	35	25	3	8	2
max	4643	1007	1015	727	213	377	117	164	87	65

Tabla 4.6: Estadística descriptiva del conjunto de datos de términos (unigramas, bigramas y trigramas) generado. Se reporta (de abajo para arriba): el promedio, el desvío estándar, el valor mínimo, los valores de los percentiles 25 %, 50 % y 75 %, y el valor máximo. El conjunto consiste de diez variables (términos) con una longitud de serie de 246 meses (frecuencia mensual). Para cada variable en cada instante de tiempo se reporta la cantidad de menciones de dicha variable en ese instante de tiempo (mes).

	C109	**C165**	**C201**	**C249**	**C269**	**C550**
mean	7,65	24,75	7,78	7,85	8,36	12,50
std	13,62	53,79	12,16	14,68	14,17	36,48
min	0	0	0	0	0	0
25 %	0	3	0	0	0	0
50 %	1	8,5	2	2	1	1
75 %	5	24,75	11,75	9	13	9,75
max	77	534	67	118	89	357

Tabla 4.7: Estadística descriptiva del conjunto de datos de eventos en curso generado. Se reporta (de abajo hacia arriba): el promedio, el desvío estándar, el valor mínimo, los valores de los percentiles 25 %, 50 % y 75 %, y el valor máximo. El conjunto consiste de seis variables (eventos) con una longitud de serie de 246 meses (frecuencia mensual). Para cada variable en cada instante de tiempo se reporta la cantidad de menciones de dicha variable en ese instante de tiempo (mes).

4.4. Aplicación a Datos Sintéticos

En la presente sección se presentan los resultados y discusiones obtenidas del análisis comparativo realizado de las técnicas de descubrimiento causal aplicadas a datos de origen sintéticos. Estos son, los datos generados a partir de la herramienta *TETRAD* y los obtenidos de la plataforma *CauseMe*. Primero, las nueve técnicas del estado del arte para descubrimiento causal, mencionadas en la Sección 4.2, son analizadas en los 56 conjuntos de datos generados con *TETRAD*. Cada uno de estos conjuntos tienen diferentes características porque son generados usando diferentes configuraciones de la herramienta: variando cantidad de variables, longitud de la serie, cantidad de variables ocultas, cantidad de rezagos en el modelo causal real y variando la estrategia de generación de grafos (*RFDAG* o *SFDAG*). Los resultados y discusiones de aplicar estas nueve técnicas sobre los conjuntos de datos generados con *TETRAD* se reportan en la Sección 4.4.1. La gran cantidad y diversidad de los conjuntos de datos generados a partir de esta herramienta permitieron sacar conclusiones fuertemente fundamentadas sobre el desempeño de las nueve técnicas ante distintos escenarios. Posteriormente, utilizando estos experimentos se seleccionan las cuatro mejores técnicas para ser usadas en los experimentos realizados sobre el resto de los conjuntos de datos sintéticos: los obtenidos de la plataforma *CauseMe*. Los datos obtenidos de dicha plataforma agregan aún más diversidad al análisis al incluir relaciones causales no lineales. Los resultados de aplicar las cuatro mejores técnicas de descubrimiento causal sobre las ocho bases de datos obtenidas de *CauseMe* son reportados en la Sección 4.4.2. En la misma sección se presenta una discusión sobre dichos resultados.

4.4.1. Análisis Comparativo en *TETRAD*

En la presente sección se reportan los resultados de aplicar las nueve técnicas del estado del arte reportadas en la Sección 4.2 (*BigVAR*, *Direct-LiNGAM*, *ICA-LiNGAM*, *Lasso-Granger*, *PC*, *PCMCI*, *SIMoNe*, *Transfer Entropy* y *VAR*) sobre el conjunto de datos sintético originado con la herramienta de simulación de datos *TETRAD*. Utilizando dicha herramienta de simulación se crean 56 conjuntos de datos diferentes con diferentes configuraciones de acuerdo a lo descrito en la Sección 4.3.1 para ser utilizados como parte del presente análisis comparativo para estudiar la diferencia entre las técnicas de causalidad presentadas. Se agrega una décima técnica de causalidad que agrega arcos de manera aleatoria. Esta técnica de referencia, a la que se denomina *Random*, se incorpora

para poder establecer un desempeño mínimo a superar por las demás técnicas. Para construir los grafos causales la técnica *Random* analiza cada posible arco de un grafo totalmente conexo y aleatoriamente con proporción 50-50 decide si incorpora el arco al modelo causal resultante o no.

Como se mencionó en la Sección 4.3.1, los 56 conjuntos de datos se pueden dividir en cuatro escenarios de acuerdo a la configuración utilizada para generarlos. A su vez cada escenario pudo haber sido generado utilizando una de dos posibles técnicas de construcción de grafos acíclicos dirigidos (DAG por sus siglas en inglés): *scale-free DAG* (*SFDAG*) o *random forward DAG* (*RFDAG*). Finalmente, los conjuntos de datos se dividen en ocho categorías: escenarios 1 a 4 para *SFDAG* y escenarios 1 a 4 para *RFDAG*. Se reportan los resultados para cada una de esas ocho categorías en la presente sección.

En la Figura 4.10 se muestran los resultados para cuatro categorías: escenarios 1 y 2 tanto para *SFDAG* (derecha) como para *RFDAG* (izquierda). En la Figura 4.11 se muestran los resultados de las otras cuatro categorías: escenarios 3 y 4 tanto para *SFDAG* (derecha) como para *RFDAG* (izquierda). Debido a la gran cantidad de resultados a reportar, de este análisis se excluye la precisión y la cobertura (que son analizados posteriormente). Solo se reporta para cada una de las ocho categorías el F1-score obtenido por las diez técnicas para cada una de las diferentes variaciones de parámetros. Por ejemplo, para el escenario 1 usando *RFDAG* se reportan los nueve valores de F1-score obtenidos (uno por cada N utilizado) para un total de diez técnicas (noventa valores de F1-score son reportados en total). Estos 90 valores son reportados en el gráfico en la esquina superior izquierda de la Figura 4.10 ("Escenario 1 - *RFDAG*").

A continuación, en la Figura 4.12, se incluye al presente análisis una comparación de las técnicas en términos de las métricas precisión, cobertura y F1-score promedio para cada escenario para cada configuración de DAG (*SFDAG* (derecha) y *RDDAG* (izquierda)). En la figura se puede observar seis gráficos de barra. Se reportan en la primera fila los gráficos de barra con los valores de precisión para todas las técnicas en los cuatro escenarios usando *RFDAG* y luego los valores de precisión para todas las técnicas en los cuatro escenarios para *SFDAG*. Análogamente, en la segunda fila se reportan los valores promedios de cobertura, primero para *RFDAG* y luego para *SFDAG*. Finalmente, en la última fila se reportan los valores promedios de F1-score para ambas estrategias de construcción de DAG (usando el mismo orden que para las dos anteriores). Además de los valores promedio se reportan los intervalos de confianza usando nivel de confianza de 95 %. Por ejemplo,

se computan los valores de precisión, cobertura y F1-score para cada uno de los nueve valores de N usando la técnica *BigVAR* para el escenario 1 usando *RFDAG*, se promedian los nueve valores de precisión, cobertura y F1-score y se los reporta en la primera barra de los gráficos que están en la columna de la izquierda. Estos tres valores son representados con tres barras, una por gráfico, y se les agrega a cada una la visualización del intervalo de confianza.

Por último, en la Figura 4.13, se reporta nuevamente la precisión, cobertura y F1-score promedio pero en este caso sin distinguir entre escenarios. Esto es, se reporta en la columna izquierda la precisión, cobertura y F1-score promedio para cada técnica en la totalidad de los conjuntos de datos *RFDAG* (28 conjuntos). Análogamente, en el lado derecho se reportan las mismas métricas promediadas para los 28 conjuntos *SFDAG*. Además de los valores promedio se reportan los intervalos de confianza usando nivel de confianza de 95 %. Por ejemplo, se computan los valores de precisión, cobertura y F1-score usando la técnica *BigVAR* en todos los escenarios usando *RFDAG*, se promedian los valores de precisión, cobertura y F1-score y se los reporta en la primera barra de los gráficos que están en la columna de la izquierda. Estos tres valores son representados con tres barras, una por gráfico, y se les agrega a cada una la visualización del intervalo de confianza.

Para el cómputo de todas las métricas (precisión, cobertura y F1-score) no se utiliza como *ground truth* la totalidad de los arcos de la estructura real, ya que esta puede contener vínculos causales contemporáneos o vínculos con distancia mayor a un intervalo de tiempo. Estos dos tipos de arcos escapan al análisis aquí presentado. Esto es, los arcos contemporáneos no son encontrados por varias de las técnicas de causalidad, por ende son descartados del análisis y por simplicidad (sin disminuir la complejidad de la tarea) solo se buscan vínculos causales de distancia uno. Esto no simplifica la tarea, ya que vínculos con mayor distancia existen (en el escenario 4) y pueden complicar la tarea de encontrar los vínculos de distancia uno correctos.

Discusión de los resultados obtenidos sobre los conjuntos de datos generados con *TETRAD*. Como se puede ver en la Figura 4.10, para **el escenario 1** (donde se varía la cantidad de nodos (N)) se puede ver una diferencia marcada entre las cinco técnicas con mejor desempeño (*BigVAR*, *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*) y las cuatro técnicas con peor desempeño (*ICA-LiNGAM*, *Lasso-Granger*, *SIMoNe* y *Transfer Entropy*). Se puede observar que para pocos nodos algunas técnicas tuvieron un desempeño muy distinto al obtenido para muchos nodos. Por ejemplo, para ***RFDAG*** (izquierda)

con $N = 6$, *ICA-LiNGAM* obtuvo F1-score= 0, mientras que, para el mismo N y mismo DAG, *SIMoNe* tuvo un desempeño inusualmente bueno comparado con su desempeño para otros valores de N. Esto se puede explicar debido al impacto del azar. Al tener menos nodos, y por ende menos arcos, cada decisión de la técnica (cada error o acierto) impacta mucho más en el desempeño que cuando el N es más grande (porque, al haber pocos nodos, cada arco representa una proporción mayor del total). Por este motivo se considera que los valores de F1-score para $N < 12$ pueden ser ruidosos y no son buenos representantes del desempeño general de la técnica.

Aunque los valores de N pequeños no son buenos representantes del desempeño, también existen técnicas que reportan valores poco estables de desempeño para todos los valores de N. Por ejemplo podemos ver que *ICA-LiNGAM* para el escenario 1 con *RFDAG* comienza con F1-score= 0 (peor que *Random*) y luego tiene su mejor desempeño en $N = 15$ para luego presentar una caída en el desempeño para $N > 15$. De manera similar, esa misma técnica comienza con un mal desempeño para *SFDAG*, mejora apenas por encima de *Random* y luego su desempeño vuelve a bajar por debajo de *Random* para $N = 15$ (el valor de N que para el otro tipo de DAG obtuvo su mejor desempeño). Se puede concluir que *ICA-LiNGAM* no solo tiene un mal desempeño global, sino que también es poco consistente.

Para el caso del escenario 1 con *RFDAG* se puede ver que, en su mayoría, las técnicas tienen buena estabilidad en el desempeño ante variaciones en el valor del N. Esto se observa en general, excepto para para valores pequeños de N y para las técnicas *PC* y *BigVAR* que para $N = 30$ muestran una clara caída en el valor del F1-score. Por otro lado **para el caso del escenario 1 con *SFDAG*** se puede observar que las cinco mejores técnicas siguen siendo las mismas pero con menos diferencia en desempeño con las demás, y esta diferencia se hace aún menor para $N = 6$ y $N = 9$ (por lo previamente discutido). Se puede ver nuevamente que *BigVAR* presenta una caída en el desempeño para $N = 30$, mientras que las demás técnicas mantienen mayor estabilidad en el valor de F1-score ante variaciones en el N.

Para el escenario 2 usando *RFDAG* (Figura 4.10) se puede observar una diferencia menos pronunciada entre las mejores técnicas y las peores. Nuevamente las técnicas *Direct-Lingam*, *PCMCI*, *PC* y *VAR* se ubican entre las mejores. La técnica *BigVAR* pasó de tener un buen desempeño para el escenario 1, a tener un desempeño muy poco estable y peor que *Random* para algunos valores de N ($N = 30$ y $N = 27$). Se puede observar

que las técnicas *SIMoNe*, *Transfer Entropy* y *Lasso-Granger* tienen un desempeño bajo, comparable a la técnica *Random*. **Para el caso de *SFDAG*** se puede observar una mayor diferencia entre las mismas cuatro mejores técnicas y las demás. *BigVAR* nuevamente presenta un desempeño poco consistente, comenzando con buen desempeño, cayendo a un mínimo para $T = 1.000$ y volviendo a subir casi hasta obtener el mismo F1-score que las mejores técnicas. Se puede ver que en algunas ocasiones esta técnica alcanza buenos desempeños, pero no de manera consistente. Como no es posible identificar un conjunto de características que deban cumplir los datos para que *BigVAR* tenga un desempeño consistentemente bueno, no es una técnica que se considere como entre las mejores, sino que es considerada de bajo desempeño como *ICA-LiNGAM*, *Lasso-Granger*, *Transfer Entropy* y *SIMoNe*. En términos generales, el tamaño de la serie de datos (T) no parece tener un impacto significativo (ni positivo ni negativo) en el desempeño de las técnicas en general.

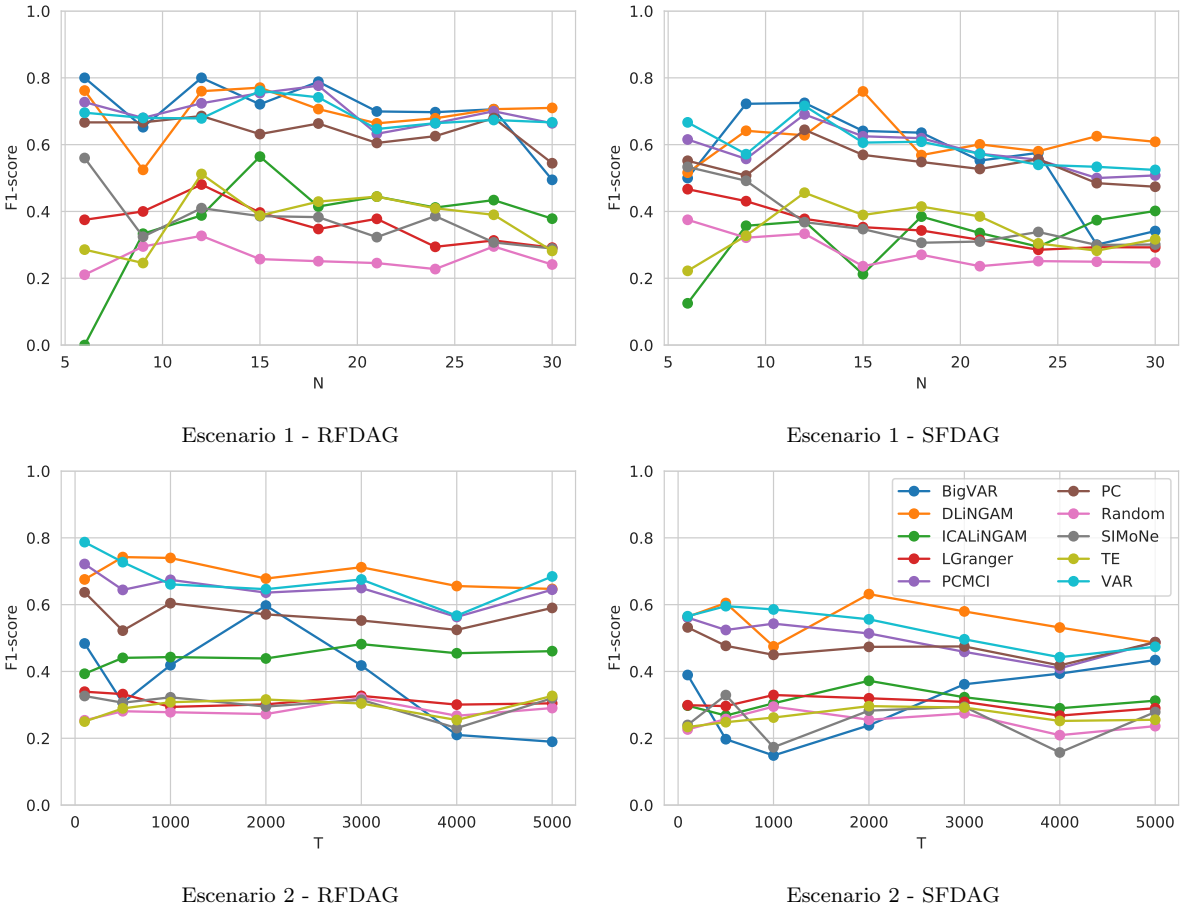
Se reportan los resultados **para el escenario 3** en la Figura 4.11, donde se agregan por primera vez variables no observadas (H), variando la cantidad de las mismas. **Para el caso de *RFDAG*** se puede observar que *Transfer Entropy* tiene un desempeño peor que *Random*. También se puede ver, una vez más, la inestabilidad de *BigVAR* y como las mismas cuatro técnicas siguen siendo las que tienen mejor desempeño. Los resultados sugieren que existe una cierta estabilidad de las técnicas hasta $H = 8$, punto a partir del cual todas las técnicas (excepto *BigVAR*) comienzan a tener peor desempeño. **Para el caso de *SFDAG*** se puede observar valores de F1-score menores que para el caso de *RFDAG*, sugiriendo que los conjuntos construidos con *SFDAG* son más complejos de resolver para las técnicas. Por otro lado, no se observa una tendencia clara de parte de todas las técnicas de pérdida de desempeño al aumentar el valor de H (como era el caso para *RFDAG*). Sin embargo, se observa una tendencia a la baja para algunas técnicas, como ser el caso de *Direct-LiNGAM* o *VAR* que tienen una tendencia a la baja de desempeño al aumentar H. Nuevamente *BigVAR* presenta un desempeño poco consistente pero esta vez con muchos valores peores que *Random* (para $H \in \{2, 4, 8, 10\}$). Las cuatro mismas mejores técnicas siguen teniendo una diferencia por encima de las otras cinco, pero no tan marcada en este escenario para *SFDAG*.

Se reportan los resultados **para el escenario 4** en la Figura 4.11, donde el modelo causal además de tener relaciones causales de un instante al siguiente (distancia uno ($L = 1$)), como todos los escenarios anteriores, ahora se agregan vínculos causales a mayor distancia en el tiempo ($L \in \{1, 2, 3, 4, 5\}$). **Para el caso de *RFDAG*** se puede observar

nuevamente que para valores bajos de L, especialmente para $L = 1$ (el valor usado en los escenarios anteriores), existe una diferencia entre las mejores cuatro técnicas y las demás (a excepción de *BigVAR* que en este caso tuvo desempeño comparable al de las cuatro mejores técnicas). Sin embargo, a medida que aumenta el L se observa una clara pérdida de desempeño por parte de las mejores técnicas, alcanzándose un desempeño relativamente similar para todas las técnicas. Por último, **para el caso de *SFDAG***, se puede observar cómo las cuatro mejores técnicas comienzan con valores casi equivalentes de F1-score para $L = 1$, y su desempeño cae rápidamente al crecer el valor de L. Una vez más se observa que *BigVAR*, que suele tener esporádicos buenos desempeños para algunas configuraciones en algunos conjuntos de datos, presenta un desempeño inconsistente y peor que *Random* para varios valores de L. Para este escenario y tipo de DAG se puede observar que hubo un peor desempeño en términos generales, donde varias técnicas tuvieron desempeños cercanos o peores a *Random* (incluso *PC* y *Direct-LiNGAM* que pertenecen al grupo de las cuatro mejores técnicas).

En resumen, en términos generales se concluye que las técnicas que tuvieron mejor desempeño son *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*, siendo de estas cuatro *PC* la que tiene peor desempeño en términos generales. De las otras tres no hay una técnica que sea superior para todos los escenarios y los parámetros. En términos generales no parece haber un gran impacto del valor de N y de T excepto para algunos casos puntuales. Por otro lado, al crecer el valor de H para *RFDAG*, hay evidencia que sugiere una caída de desempeño. En el caso de *SFDAG* esta caída existe para algunas técnicas pero dicha caída no es pronunciada. Para el caso de L en *RFDAG* se puede observar una caída para el valor de $L = 5$ con respecto a todos los demás. Mientras que este no es el caso para *SFDAG*. En términos globales las técnicas tuvieron peor desempeño para *SFDAG* que para *RFDAG*, más aún el escenario más difícil fue el 4 con la configuración *SFDAG*. *BigVAR* fue la técnica con peor consistencia, alcanzado de forma alternada los mejores y peores resultados, sin respetar ningún patrón evidente de características que marcasen un mejor o peor desempeño.

Un análisis similar se desprende de analizar **el desempeño en términos de las métricas precisión, cobertura y F1-score** disponibles en la Figura 4.12. Por ejemplo, se puede ver que **para el caso de *RFDAG* los mejores valores de precisión** son alcanzados por las mejores cuatro técnicas y *BigVAR*. Este valor alto de precisión para *BigVAR* es esperable ya que al ser un modelo VAR penalizado se espera que tenga mayor-



Escenario 1 - RFDAG

Escenario 1 - SFDAG

Escenario 2 - RFDAG

Escenario 2 - SFDAG

Figura 4.10: Desempeño, medido en términos de F1-score, de las nueve técnicas del estado del arte y el modelo de referencia (*baseline*) *Random* sobre 32 conjuntos de datos sintéticos generados con la herramienta de simulación de *TETRAD*. Estos conjuntos de datos se dividen en cuatro: (i) escenario 1 construido usando el enfoque *RFDAG*, (ii) escenario 1 construido usando el enfoque *SFDAG*, (iii) escenario 2 usando el enfoque *RFDAG* y (iv) escenario 2 usando el enfoque *SFDAG*. Estas cuatro categorías (i, ii, iii, iv) tienen 9, 9, 7 y 7 conjuntos de datos cada una, respectivamente. El escenario 1 mantiene la misma configuración, pero modificando la cantidad de nodos (N). El escenario 2 mantiene la misma configuración, pero variando la longitud de la serie (T). Se puede observar que para la métrica considerada las mejores técnicas son *Direct-LiNGAM* (DLiNGAM), *PCMCI*, *PC* y *VAR*. Mientras que las peores técnicas son: *Lasso-Granger* (LGranger), *SIMoNe*, *Transfer Entropy* (TE), e *ICA-LiNGAM*. Se puede ver que la técnica *BigVAR* tiene un desempeño poco consistente. En términos de los parámetros considerados (N y T) no se observan diferencias significativas en desempeño. Solo se observa desempeños menos consistentes para N pequeño en *SIMoNe* (con un desempeño inusualmente alto con $N = 6$), *ICA-LiNGAM* (con F1-score cero para $N = 6$) y *Direct-LiNGAM* (con desempeño inusualmente pequeño para $N = 6$).

mente coeficientes cero (penalizados) y solo unos pocos coeficientes significativos distintos de cero (muy precisos). Se puede observar que *ICA-LiNGAM* tiene buenos valores de

precisión promedio, aunque inferiores a los de las mejores cuatro técnicas y con intervalos de confianza mayores en la mayoría de los casos. Finalmente se puede observar que el escenario cuatro es el más difícil, ya que todas las técnicas obtienen valores de precisión más bajos en comparación y hay menos distancia entre las mejores y las peores técnicas. Un escenario similar se da para los valores de **precisión obtenidos con *SFDAG***, donde las mismas cuatro técnicas y *BigVAR* sobresalen en precisión, y donde se puede ver un escenario 4 con menores valores de precisión indicando la dificultad de dicho escenario. Se observa también, en líneas generales, peor desempeño en términos de precisión para la tarea usando *SFDAG* comparado con el desempeño usando *RFDAG*. La técnica con los peores valores de precisión es *Random*, aunque es igualada en mal desempeño por varias técnicas en varias configuraciones.

Para el caso de la **cobertura promedio**, se puede observar que para el **escenario 1 en *RFDAG*** existe una clara diferencia entre las cuatro mejores técnicas y las demás. En este caso se puede ver el bajo desempeño de *BigVAR* (lo cual es esperable ya que por construcción prioriza precisión) y además se puede observar que tiene un intervalo de confianza grande, indicando baja consistencia en sus resultados. Se puede ver, por ejemplo, por parte de técnicas como *Lasso-Granger*, que si bien mostraron mal desempeño en términos de F1-score, para el caso de cobertura el desempeño es mejor. Dicha técnica alcanzó valores de cobertura similares a los de las cuatro mejores técnicas, confirmando que es una técnica que prioriza la cobertura obteniendo una mala precisión. **Para el caso de *SFDAG*** se observa un comportamiento similar de las técnicas en términos de cobertura promedio: las cuatro mejores técnicas alcanzaron los mejores valores de cobertura, la técnica *BigVAR* obtuvo valores considerablemente bajos, y la técnica *Lasso-Granger* una vez más tuvo valores de cobertura altos, casi comparables con las cuatro mejores técnicas. Otra técnica que sobresale para algunos escenarios es *SIMoNe*, que también obtiene buenos valores de cobertura promedio en varias configuraciones. La técnica que tiene los peores valores de cobertura es *ICA-LiNGAM*.

Por último, de los **valores de F1-score promedio** reportados en la Figura 4.12 se extraen conclusiones similares a las obtenidas de los valores de F1-score no promediados reportados en 4.10 y 4.11. Esto es, se puede ver el desempeño superior de las cuatro técnicas **para el caso de *RFDAG***. También se puede ver un desempeño bueno por parte de *BigVAR* pero con intervalos de confianza mayores (debido a su inconsistencia). Adicionalmente se puede ver que *ICA-LiNGAM* obtiene, en términos generales, desempeños

malos y presenta intervalos de confianza grandes. Se puede ver que la técnica *Random* es la que tiene peor desempeño excepto para el escenario 3 (donde *Transfer Entropy* tiene el peor desempeño). Adicionalmente, si bien *Random* es la peor técnica, se puede observar que no tiene una diferencia significativa con algunas técnicas en algunos escenarios. Por ejemplo, para el escenario 4 debido al gran intervalo de confianza de *ICA-LiNGAM* no se puede asegurar que sus desempeños sean estadísticamente distintos, indicando que no hay una diferencia significativa entre *Random* e ICA-LiNGAM para este escenario con *RFDAG*. Por último, **para el caso de *SFDAG*** en el escenario 4, se puede ver que las mejores cuatro técnicas siguen manteniendo resultados consistentes mientras *BigVAR* muestra una gran caída de desempeño y un gran aumento en los intervalos de confianza. La dificultad de este escenario se vuelve evidente ya que la diferencia entre las mejores y las peores técnicas se hace menor. Finalmente, para las técnicas con bajo desempeño se puede observar un valor de F1-score promedio muy cercano a *Random* (o peor para algunas técnicas en algunos escenarios).

Un tercer análisis se desprende de la Figura 4.13 donde se reporta nuevamente la **precisión, cobertura y F1-score promedio pero en este caso sin distinguir entre escenarios**. Se puede observar que tanto **para *RFDAG* como para *SFDAG*** se tiene que las mejores cinco técnicas en términos de precisión (de mejor a peor) son: *BigVAR*, *Direct-LiNGAM*, *PCMCI*, *VAR* y *PC*. En términos de cobertura se puede ver que **para *RFDAG*** las mejores cuatro técnicas son: *VAR*, *PCMCI*, *Direct-LiNGAM* y *PC*. **Para el caso de *SFDAG*** las cuatro mejores técnicas son las mismas, solo invirtiendo el orden de *Direct-LiNGAM* con *PC*. Por último, en términos de F1-score se puede apreciar que, **tanto para *RFDAG* como para *SFDAG***, las mejores cuatro técnicas son (de mejor a peor): *Direct-LiNGAM*, *PCMCI*, *VAR* y *PC*. En esta Figura se puede apreciar un intervalo de confianza muy grande para la técnica *BigVAR* dando, nuevamente, indicios de sus inconsistencias.

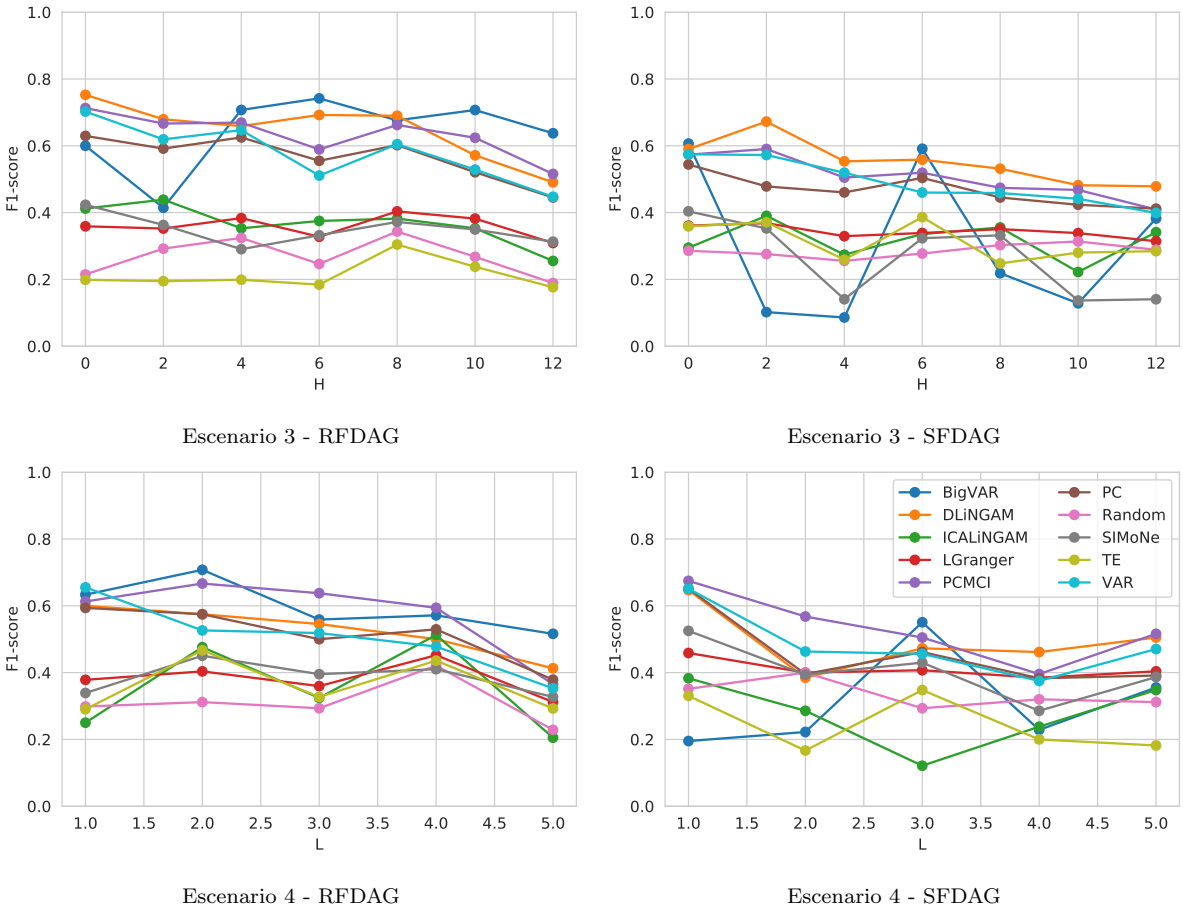
También se puede observar que, si bien *Random* fue la peor técnica para ambas configuraciones (***RFDAG* y *SFDAG***), las técnicas con peor desempeño (*BigVAR*, *ICA-LiNGAM*, *Lasso-Granger*, *SIMoNe* y *Transfer Entropy*) **en la configuración *SFDAG*** no tuvieron un desempeño significativamente mejor que este modelo de referencia (*baseline*). Esto da indicios de que las técnicas no tienen buen desempeño en general y de que el escenario *SFDAG* es más difícil que el *RFDAG*, esto se vuelve especialmente evidente al notar que la diferencia entre las peores y las mejores técnicas es menor para esta

configuración.

Para el caso de *RFDAG*, la técnica *Random* no fue significativamente distinta a *ICA-LiNGAM*, *Lasso-Granger*, *SIMoNe* y *Transfer Entropy*. El bajo desempeño por parte de estas cuatro técnicas y *BigVAR* puede ser explicado por diferentes motivos. Por ejemplo, para el caso de *ICA-LiNGAM*, una serie de limitaciones fueron descritas por los autores en publicaciones posteriores donde propusieron a *Direct-LiNGAM* como una mejora a ICA-LiNGAM [SIS+11]. Estas limitaciones son mencionadas en la Sección 4.2). Para el caso de *Transfer Entropy*, la técnica puede estar limitada por tratarse de una técnica que analiza causalidad de a pares (lo cual deja fuera del modelo variables potencialmente relevantes para el descubrimiento causal). Similarmente, para el caso de *Lasso-Granger*, si bien se aplica una etapa de selección de variables con la técnica lasso, el proceso posterior de descubrimiento causal es utilizando la técnica de causalidad de Granger de a pares, que puede tener las mismas limitaciones que *Transfer Entropy*. Por último, las técnicas basadas en modelos VAR penalizados, *BigVAR* y *SIMoNe*, tienen valores de cobertura bajos con respecto a las mejores técnicas, probablemente por la componente de penalización que afecta negativamente a esta métrica. Aunque ambas técnicas tienen valores de cobertura bajos, la diferencia entre estas técnicas es en su desempeño en términos de precisión. *BigVAR* tiene valores altos de precisión y bajos de cobertura, consistentes con una técnica que encuentra mayormente arcos correctos y penaliza fuertemente a los que sean probablemente negativos. Mientras que *SIMoNe* tiene valores bajos para ambas métricas, esto quiere decir que la técnica recupera muchos arcos incorrectos (baja precisión) y el proceso de penalización afecta a la cobertura dando como resultado una técnica que no resulta útil para el descubrimiento causal.

La variedad y cantidad de los datos generados a partir de la herramienta *TETRAD* permite sacar conclusiones fundamentadas sobre cuáles técnicas aportan al descubrimiento causal en este dominio y para este tipo de conjuntos de datos (y cuales técnicas tienen un desempeño cercano a *Random*). A partir de este estudio, como se pudo observar un desempeño consistentemente bueno por parte de las técnicas *Direct-LiNGAM*, *PCMCI*, *PC* y *VAR*, esas cuatro técnicas son seleccionadas para continuar el análisis comparativo de herramientas causales. Para sumar a este análisis se incorporan los conjuntos de datos obtenidos de la plataforma *CauseMe* y de la compañía *CAMMESA*. Sobre esos conjuntos de datos se comparan las cuatro mejores técnicas para sacar conclusiones sobre su desempeño. El objetivo es elegir la o las mejores para ser utilizadas como parte del *framework*

de recuperación de estructuras causales a partir de textos. Esto es, poder utilizar la o las mejores técnicas sobre los conjuntos de datos originadas a partir del *New York Times* que son descritos en 4.3.4.



Escenario 3 - RFDAG

Escenario 3 - SFDAG

Escenario 4 - RFDAG

Escenario 4 - SFDAG

Figura 4.11: Desempeño, medido en términos de F1-score, de las nueve técnicas del estado del arte más el modelo de referencia (*baseline*) *Random* sobre 24 conjuntos de datos sintéticos generados con la herramienta de simulación de *TETRAD*. Estos conjuntos de datos se dividen en cuatro: (i) escenario 3 usando el enfoque *RFDAG*, (ii) escenario 3 usando el enfoque *SFDAG*, (iii) escenario 4 usando el enfoque *RFDAG* y (iv) escenario 4 usando el enfoque *SFDAG*. Estas cuatro categorías (i, ii, iii, iv) tienen 7, 7, 5 y 5 conjuntos de datos cada una, respectivamente. El escenario 3 mantiene la misma configuración, pero modificando la cantidad de variables ocultas (H). El escenario 4 mantiene la misma configuración, pero variando la cantidad de variables rezagadas del modelo causal real (L). Para el escenario 3 se puede ver que las mejores cuatro técnicas siguen siendo las mismas: *Direct-LiNGAM* (DLiNGAM), *PCMCI*, *PC* y *VAR*. Por otra parte, para el mismo escenario se observa que las cuatro peores técnicas son *Lasso-Granger* (LGranger), *SIMoNe*, *Transfer Entropy* (TE) e *ICA-LiNGAM*. Se puede ver que la técnica *BigVAR* tiene un desempeño poco consistente, especialmente para *SFDAG*. Para el escenario 4 *RFDAG* y *SFDAG* las cuatro mejores técnicas comienzan con mejor desempeño, pero la diferencia disminuye al aumentar el L o el H. Para el escenario 4 con *SFDAG* la diferencia entre las mejores y peores técnicas es mucho menor. Se observa que, en general, *SFDAG* es una tarea más difícil que *RFDAG*, donde hay menos diferencia entre las mejores y peores técnicas. También se puede observar que los parámetros analizados (H y L) también tienen un impacto en el desempeño (a mayor valor menor desempeño).

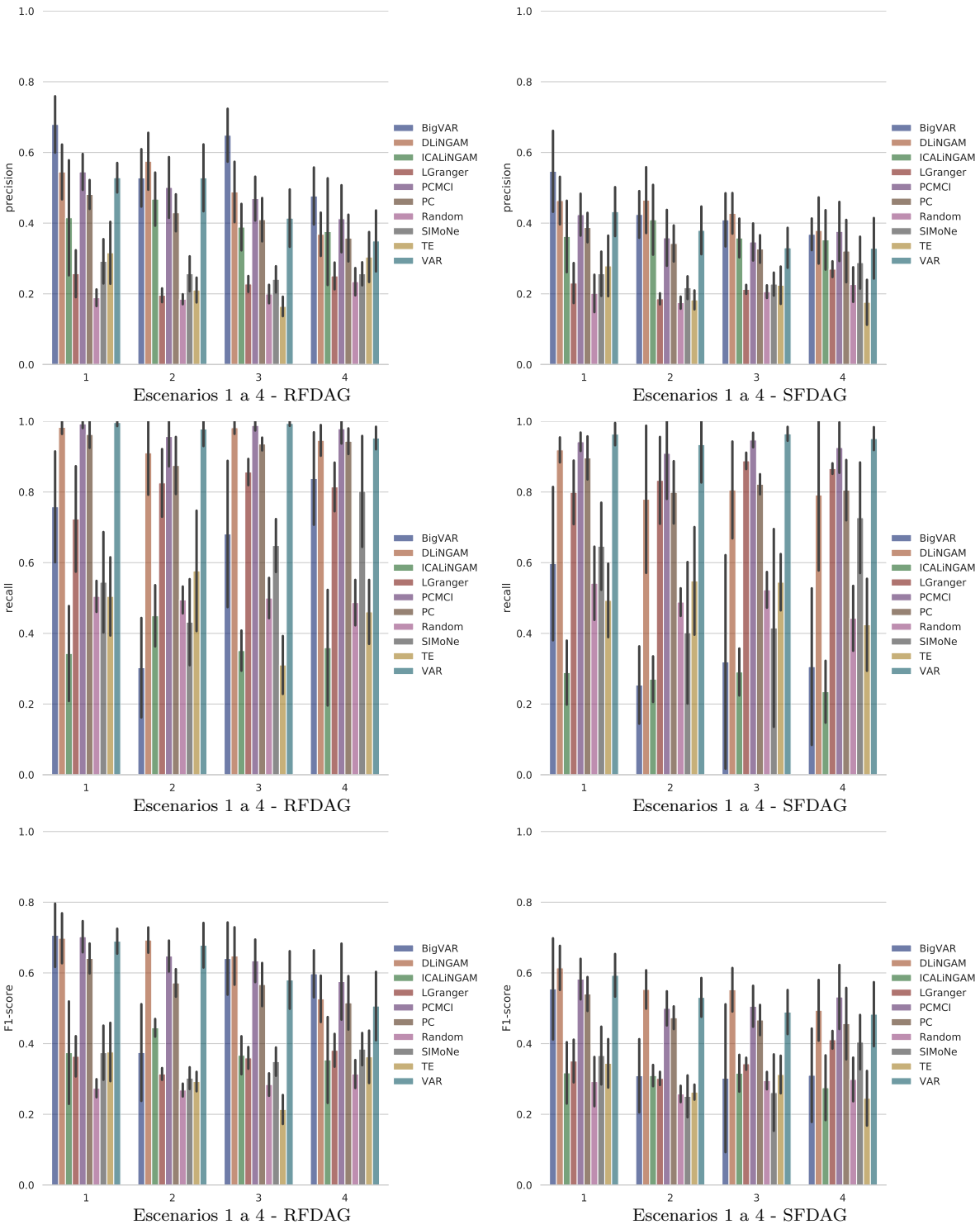


Figura 4.12: Desempeño en términos de precisión (arriba), cobertura (medio) y F1-score (abajo) de las nueve técnicas del estado del arte analizadas y el modelo de referencia (*baseline*) *Random* sobre los cuatro escenarios generados con *TETRAD*. Los conjuntos de datos se separan en 28 conjuntos de datos generados usando *RFDAG* (columna izquierda) y 28 conjuntos de datos usando *SFDAG* (columna derecha). En el eje horizontal de cada gráfico de barras se separan los resultados para cada uno de los cuatro escenarios. Se reporta para cada métrica de cada escenario el promedio de la métrica para ese escenario y el intervalo de confianza para el mismo (con un nivel de confianza de 95 %).

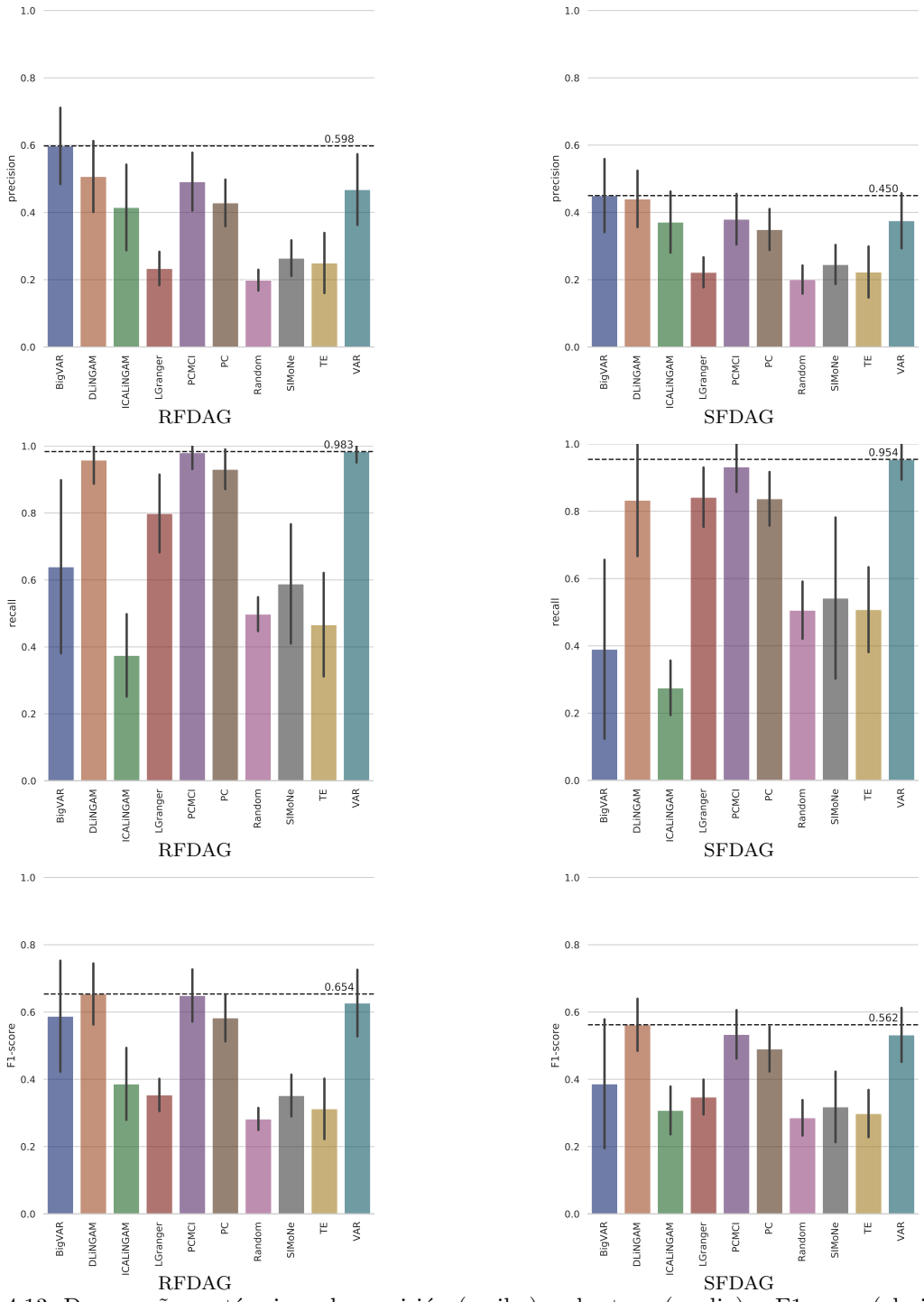


Figura 4.13: Desempeño en términos de precisión (arriba), cobertura (medio) y F1-score (abajo) de las nueve técnicas del estado del arte analizadas y el modelo de referencia (*baseline*) *Random* sobre todos los conjuntos de datos generados con *TETRAD*. Para computar cada una de las tres métricas se promedian los desempeños obtenidos para cada técnica promediando para todos los escenarios (1 al 4) para *RFDAG* (izquierda) y lo mismo para *SFDAG* (derecha). Entonces para cada métrica usando *RFDAG* se calcula un promedio de 28 conjuntos de datos, y otros 28 para *SFDAG*. Se reportan junto con los promedios, los intervalos de confianza con nivel de confianza de 95 %.

4.4.2. Aplicación a *CauseMe*

En la presente sección se reportan los resultados de aplicar cuatro técnicas de recuperación causal sobre ocho conjuntos de datos recuperados de la plataforma de *benchmarking* *CauseMe*. Se seleccionan cuatro del total de nueve técnicas del estado del arte basándose en los experimentos realizados sobre los conjuntos de datos generados con *TETRAD*, los métodos seleccionados son: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*. Adicionalmente, se reporta la técnica *Random* (construida de la misma forma que para *TETRAD*) a modo de modelo de referencia (*baseline*). Estas cinco técnicas son aplicadas sobre los ocho conjuntos de datos correspondientes al conjunto *nonlinear-VAR* que se encuentra disponible en *CauseMe*. Estos ocho experimentos están construidos de manera similar, pero variando el tamaño de la longitud de la serie ($T \in \{300, 600\}$) y la cantidad de nodos ($N \in \{3, 5, 10, 20\}$). Una descripción completa de los ocho conjuntos de datos se presenta en la Sección 4.3.2. La *ground truth* de los experimentos no se encuentra disponible para descargar de la plataforma, en su lugar es necesario subir el resultado de las técnicas a la plataforma y la misma retorna las métricas de desempeño (tasa de falsos positivos, tasa de verdaderos positivos y F-measure con $\beta = 0,5$). Utilizando las métricas provistas se computa y reportan las mismas métricas que para los experimentos en *TETRAD*: precisión, cobertura y F1-score.

En la Figura 4.14 se reportan los valores de precisión, cobertura y F1-score para los ocho escenarios. La columna de la izquierda contiene los resultados de las métricas para los cuatro conjuntos de datos con longitud de serie trescientos ($T = 300$), mientras que la columna de la derecha contiene los resultados de las métricas para los cuatro conjuntos de datos con longitud seiscientos ($T = 600$). Para ambas columnas primero se muestra la precisión, en la siguiente fila se muestra la cobertura y en la última el F1-score. Para cada subfigura se muestra el resultado obtenido por cada una de las cinco técnicas para cada uno de los cuatro valores de N.

Discusión de los resultados sobre los conjuntos de datos obtenidos de *CauseMe*. Se puede observar, tanto para $T = 300$ como para $T = 600$, una caída en la precisión por parte de todas las técnicas al aumentar el N (incluso por parte de *Random*), aunque la pendiente de esta caída varía de técnica a técnica. *Direct-LiNGAM* es la única técnica que al crecer el N mantiene un nivel de precisión con menos caída. Esto habla positivamente de la estabilidad de los arcos encontrados por esta herramienta, sugiriendo que es una buena técnica si se tiene por objetivo solo encontrar arcos correctos

(maximizar precisión). Se puede observar que la técnica *PC* es la que tiene peor precisión seguida por *PCMCI*. Como esta última técnica construye los arcos a partir de *PC* es esperable que tenga mejor precisión (toma los arcos de *PC* como el conjunto de padres potenciales y los refina con la técnica *MCI*). Aunque no hay mucha diferencia en términos de precisión para $T = 300$ comparado con $T = 600$, algunas técnicas parecen tener una leve mejoría al aumentar el T, aunque no hay suficientes datos como para saber si es una diferencia significativa. Se puede observar que tanto para $T = 300$ como para $T = 600$ las técnicas se ubican igual en términos de precisión: primero *Direct-LiNGAM*, luego *VAR*, luego *PCMCI* y por último *PC*.

En términos de cobertura, se observa que la técnica *Random* obtiene los resultados más altos junto con *PC*. Por la forma en la que está construida, la técnica *Random* va a tener en promedio la mitad de arcos del grafo totalmente conexo (ya que por cada posible arco del grafo completo aleatoriamente elige si considerarlo causal o no usando una proporción 50-50). Por este motivo, si el grafo causal real consiste de unos pocos arcos y la técnica *Random* toma arcos como correctos con una proporción alta, se va a maximizar la cobertura y minimizar la precisión. Se puede observar que tanto para $T = 300$ como para $T = 600$ las técnicas se ubican igual en términos de cobertura: primero *PC*, luego *VAR*, luego *Direct-LiNGAM* y por último *PCMCI*. A diferencia de la precisión, la cobertura no se ve afectada por el aumento en el N.

En términos de F1-score, se puede ver que tanto para $T = 300$ como para $T = 600$ las técnicas comienzan ordenadas de mejor a peor de la siguiente manera: *PC*, *VAR*, *Random*, *Direct-LiNGAM* y *PCMCI*. Al aumentar el N el desempeño de las técnicas es afectado de formas distintas, siendo *Random* la más afectada en términos de desempeño y *Direct-LiNGAM* la menos afectada. Para $N = 20$ se puede ver que *Direct-LiNGAM* pasó a ser la mejor, seguida por *VAR*, *PC*, *PCMCI* y en último lugar *Random*.

Vale la pena mencionar que los desempeños absolutos en términos de F1-score son diferentes a los obtenidos en *TETRAD* porque las métricas de desempeño son computadas de maneras distintas. Para el caso de *CauseMe* se computan las métricas de desempeño usando todos los arcos de la *ground truth*, esto incluye arcos contemporáneos (si es que hay) y arcos con distancias en el tiempo mayores a uno (si es que hay). Estos dos tipos de arcos no son los objetivos del presente estudio y no son considerados y, por ende, para computar las métricas de los experimentos de *TETRAD* no son considerados. Como la plataforma *CauseMe* no proporciona la *ground truth* sino que proporciona las métricas

obtenidas por los métodos, no es posible computar las métricas como están computadas en *TETRAD*. Por este motivo se reportan las métricas obtenidas a partir de la plataforma (que tienen en consideración todos los arcos). Por este motivo, por ejemplo, una cobertura perfecta $(1, 0)$ no es posible, ya que las técnicas propuestas solo se concentran en una porción de los arcos que existen en la estructura real (los arcos no contemporáneos con distancia uno).

Del presente análisis se puede observar que la tarea de recuperación causal tiene muchas dimensiones a tener en cuenta, y que dependiendo del dominio trabajado y el objetivo que se tenga, el análisis y las conclusiones pueden variar. Por ejemplo, que la técnica *Random* haya tenido mejor F1-score que *Direct-LiNGAM* y *PCMCI* para $N = 3$ da indicio que hay que ser cuidadoso al momento de elegir las métricas y la forma en la que se estudian las técnicas. *Random* obtuvo mejor F1-score al maximizar cobertura (agregando arcos con proporción 50-50) y aunque tuvo la peor precisión que todas las demás técnicas, obtuvo valores de F1-score mejores que técnicas que hacían análisis de causalidad reales (no que elegían arcos aleatoriamente). Entonces, dependiendo del dominio, la métrica F1-score podría no ser la mejor debiendo optar por la técnica F-measure con un β que priorice la precisión. Más aún, en lugar de utilizar métricas de recuperación de información como falsos positivos, falsos negativos, precisión u otros, es importante conocer y plantear el uso (en caso de ser necesario) de otras métricas como distancia estructural de Hamming [AdC03] o distancia estructural de intervenciones [PB15].

Estos resultados muestran que no existe una única técnica mejor que todas, ya que dependiendo el dominio y las características de los datos una técnica que era la mejor puede dejar de serlo. Por ejemplo *PCMCI* tuvo el segundo mejor F1-score promedio en los conjuntos de *TETRAD* mientras que en *CauseMe* tuvo consistentemente peores resultados que las otras tres técnicas para casi todos los escenarios con todas las configuraciones (excepto para $N = 3$ donde fue mejor que *Direct-LiNGAM* por un pequeño margen). Análogamente, *Direct-LiNGAM* fue la peor técnica en términos de F1-score para $T = 300$ y $T = 600$ con $N = 3$. Sin embargo, para $N = 20$ pasó a ser la mejor técnica de las cuatro. Por este motivo, para seguir profundizando en el análisis de las técnicas, se incluye un tercer dominio que es el conjunto de datos de demanda de energía eléctrica provisto por *CAMMESA*. Sobre este conjunto de datos se estudian nuevamente las misma cuatro mejores técnicas: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*.

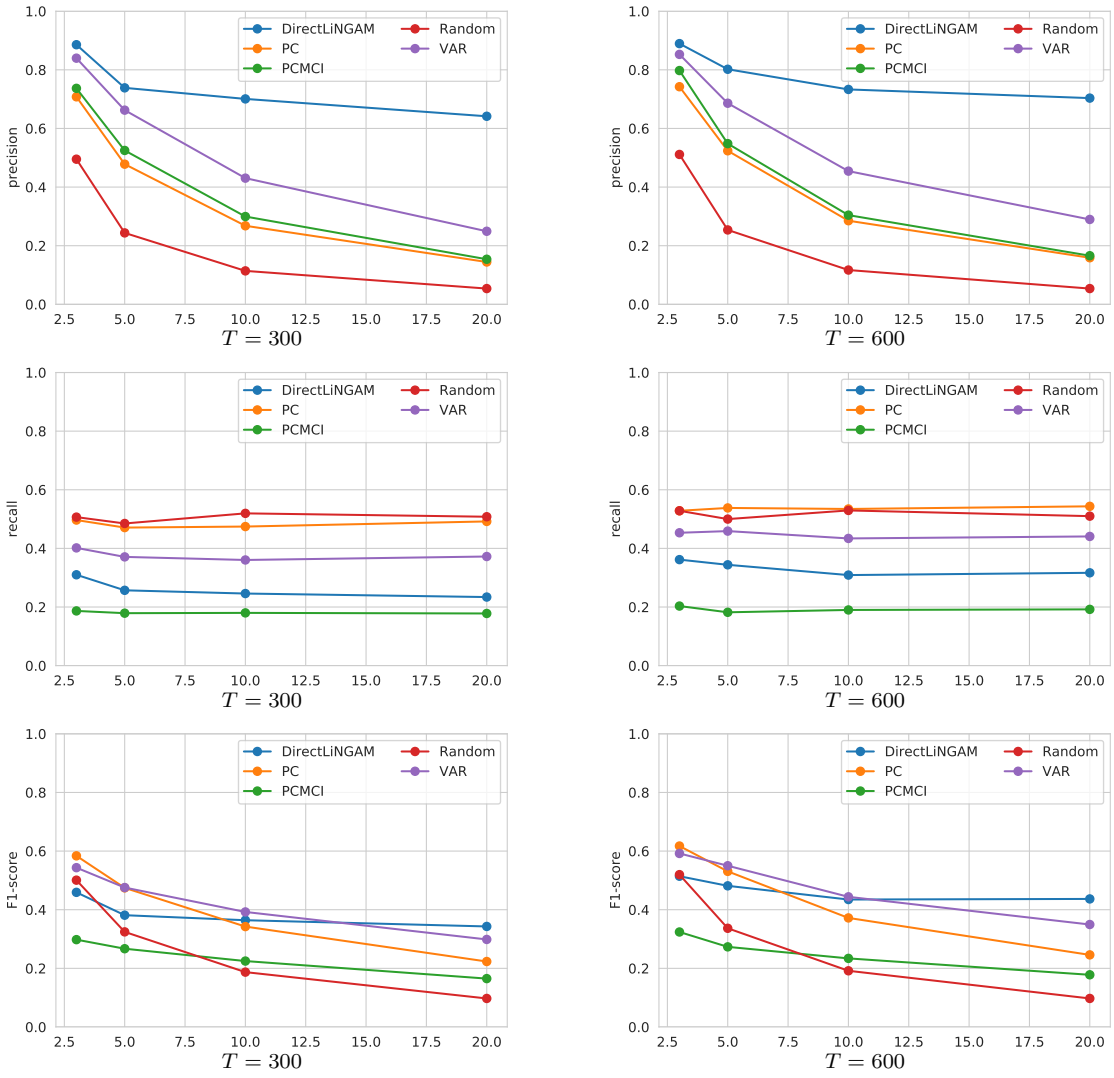


Figura 4.14: Resultado de aplicar técnicas seleccionadas del estado del arte (*Direct-LiNGAM*, *PCMCI*, *PC*, y *VAR*), y el modelo de referencia (*baseline*) *Random*, sobre los 8 conjuntos de datos obtenidos de la plataforma de *CauseMe*. Estos conjuntos de datos se corresponden con ocho experimentos del conjunto de datos *nonlinear-VAR* detallado en la plataforma antes mencionada. Estos experimentos varían la longitud de la serie (T), con $T = 300$ en la primera columna y $T = 600$ en la segunda. Dentro de estas dos longitudes de series se varía la cantidad de nodos ($N \in \{3, 5, 10, 20\}$) (eje horizontal). Para todas estas técnicas y conjuntos de datos se reporta, de arriba para abajo, la precisión, la cobertura y la media armónica de estos dos valores (F1-score) con respecto a los arcos reales. Se puede observar que para $N = 20$ y para ambos valores de T, las técnicas se ordenan en desempeño en términos de F1-score de la siguiente manera (de mejor a peor): *Direct-LiNGAM*, *VAR*, *PC* y *PCMCI*. Para $N = 3$, donde hay menos oportunidad para demostrar la capacidad de cada técnica (solo hay que decidir sobre unos pocos arcos) y donde cada error cuenta más (es una proporción mayor del total de arcos), las técnicas *PCMCI* y *Direct-LiNGAM* no demuestran ser mejores que *Random*. Sin embargo, *Direct-LiNGAM* muestra una mayor estabilidad en desempeño al aumentar el N.

4.5. Aplicación a Datos de Demanda Eléctrica

En la presente sección se reportan los resultados de aplicar las cuatro mejores técnicas de recuperación causal sobre el conjunto de datos de demanda de energía eléctrica provista por la compañía administradora del mercado mayorista eléctrico de argentina (*CAMMESA*). Una descripción detallada del conjunto de datos se puede encontrar en la Sección 4.3.3. Al igual que para *CauseMe*, se seleccionan cuatro del total de nueve técnicas del estado del arte basándose en los experimentos realizados sobre los conjuntos de datos de *TETRAD*. Los métodos seleccionados son: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*.

Los resultados de aplicar estas cuatro técnicas al conjunto de datos de *CAMMESA* se pueden ver en la primera fila de la Figura 4.15. De izquierda a derecha se ven los grafos resultantes de aplicar *Direct-LiNGAM*, *VAR*, *PCMCI* y *PC* sobre dicho conjunto de datos. Se representan en rojo los arcos incorrectos y en negro los correctos de acuerdo a la *ground truth* discutida en la Sección 4.3.3. Como se puede observar las técnicas tuvieron un mal desempeño en términos de precisión. De los 7, 9, 8, 8 arcos recuperados por las técnicas *Direct-LiNGAM*, *VAR*, *PCMCI*, *PC*, respectivamente, se recuperaron 3, 5, 4 y 4 arcos incorrectos (respectivamente). Esto da como resultado los siguientes valores de precisión para las técnicas *Direct-LiNGAM*, *VAR*, *PCMCI* y *PC* respectivamente: $0,571$; $0,444$; $0,500$ y $0,500$. En términos de cobertura se puede ver que las cuatro técnicas recuperaron los mismos cuatro arcos correctos (del total de cinco), llegando a una cobertura de $0,800$. Se puede notar que en todos los casos fallaron en encontrar el arco Hum $\rightarrow$ Temp.

Los bajos niveles de precisión obtenidos por las cuatro técnicas en este conjunto de datos pueden ser explicados por la naturaleza cíclica de los datos. Al tratarse de datos de demanda eléctrica y variables climatológicas de varios años con frecuencia diaria se presentan varios ciclos en los datos. Primero, la temperatura presenta un ciclo anual que se corresponde con las estaciones del año (presentando, cada año, temperaturas más altas en verano y más bajas en invierno). Segundo, la serie de la demanda de energía eléctrica (que se vincula fuertemente con la temperatura) presenta un ciclo similar anual en respuesta al ciclo de la temperatura (temperaturas muy altas o muy bajas incentivan el consumo, generando un pico en invierno y otro en verano todos los años). Tercero, la demanda tiene una componente cíclica semanal. Esto es, cada siete días se puede esperar que se repita un patrón (la demanda de un lunes es similar a la del lunes anterior y la demanda de un domingo es similar a la del domingo anterior). Se tiene la hipótesis de que, debido a la presencia de esos ciclos (que agregan correlaciones entre diferentes variables

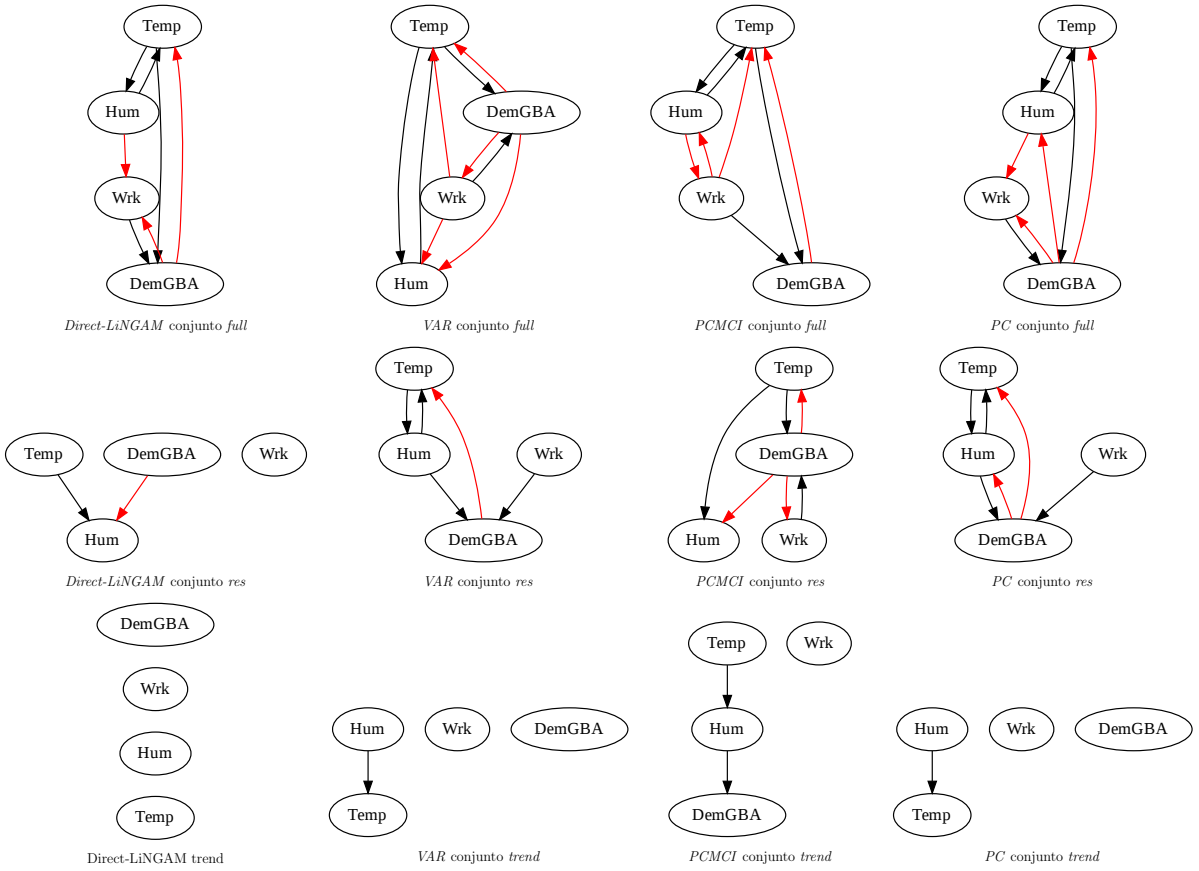


Figura 4.15: Grafos causales resultantes de aplicar las técnicas *Direct-LiNGAM*, *VAR*, *PCMCI* y *PC* sobre el conjunto de datos provisto por *CAMMESA* (*full*) (arriba) y dos transformaciones del mismo: *res* (medio) y *trend* (abajo). La primera transformación consiste de reemplazar las variables Temp, Hum y DemGBA por los residuos de modelar esas mismas variables como una regresión lineal de 20 variables que dan información de la estación, mes del año y día de la semana. Al modelar las variables como lo no explicado por esas 20 variables (los residuos de la regresión), se eliminan componentes cíclicos. El conjunto *trend*, consiste de las mismas variables, pero luego de aplicarles un filtro de descomposición estacional mediante promedios móviles para solo mantener la componente de la tendencia. Estas dos transformaciones de los datos son aplicadas a las variables reales, no a la variable binaria Wrk. Se puede ver un gran nivel de precisión para el conjunto *trend*, pero comprometiendo el desempeño en términos de cobertura. Se observa un grafo poco informativo para *full*, dónde se encuentra casi el grafo totalmente conectado, obteniendo alta cobertura, pero mala precisión. Por otro lado, el conjunto *res* resulta un buen intermedio a los escenarios obtenidos por los otros dos conjuntos de datos.

en diferentes instantes de tiempo) la capacidad de extracción causal se puede ver afectada debido a que se vuelve posible predecir valores futuros de variables a partir de valores pasados de variables no causales. Esta capacidad adicional de predicción puede hacer que las técnicas de extracción de causalidad detecten vínculos causales entre variables que no

los tienen, generando así una pérdida en precisión.

Para contrarrestar este fenómeno se aplicaron dos técnicas distintas para la eliminación de los componentes cíclicos en los datos. Primero se aplica el filtro de descomposición estacional del paquete *statsmodels*[8] versión 0.10.2 usando frecuencia siete, para eliminar el ciclo semanal. De esta descomposición solo se tomó la componente de la pendiente (descartando el ciclo). A este conjunto de datos se lo denomina *trend*. Como segunda técnica de filtrado de ciclo se construyeron veinte variables auxiliares binarias para capturar las componentes cíclicas del conjunto de datos. Estas variables son: *primavera*, *verano*, *otoño*, *lunes*, *martes*, *miércoles*, *jueves*, *viernes*, *sábado*, *febrero*, *marzo*, *abril*, *mayo*, *junio*, *julio*, *agosto*, *septiembre*, *octubre*, *noviembre* y *diciembre*. Se realiza la regresión de cada una de las variables continuas (Hum, Temp y DemGBA) con respecto a las variables auxiliares.

$$\begin{aligned} \text{Hum} &= \beta_0 + \beta_1 \text{primavera} + \beta_2 \text{verano} + \cdots + \beta_{20} \text{diciembre} + \varepsilon_{\text{Hum}} \\ \text{Temp} &= \beta_0 + \beta_1 \text{primavera} + \beta_2 \text{verano} + \cdots + \beta_{20} \text{diciembre} + \varepsilon_{\text{Temp}} \\ \text{DemGBA} &= \beta_0 + \beta_1 \text{primavera} + \beta_2 \text{verano} + \cdots + \beta_{20} \text{diciembre} + \varepsilon_{\text{DemGBA}} \end{aligned} \tag{4.10}$$

El objetivo es modelar las componentes cíclicas de cada variable en término de estas variables auxiliares. Finalmente, cada una de las tres variables es reemplazada por los residuos de la regresión lineal de sí misma contra las variables auxiliares. Ya que los residuos resultantes capturan todo lo no explicado por las variables que representan el momento del año, se espera que los residuos contengan la información de la tendencia de la serie (y no de sus componentes cíclicas). Utilizando esta técnica se construye el conjunto de datos filtrado *res*.

En la segunda y tercera fila de la Figura 4.15 se pueden ver los resultados de aplicar las cuatro técnicas a los conjuntos de datos *res* y *trend*, respectivamente. Se puede observar para ambos conjuntos una gran caída en cantidad de arcos incorrectos recuperados, en muchos casos aumentando notablemente la precisión. Por ejemplo, para la última fila, correspondiente al conjunto *trend*, se puede apreciar que al no encontrar ningún arco incorrecto se tiene una precisión perfecta (todos los arcos marcados como relevantes son correctos, es decir, no hay falsos positivos). Por otra parte, para *res* se encuentran algunos arcos incorrectos, pero aun así se tiene un incremento de la precisión para las técnicas *VAR* y *PC*, que pasaron de $0,444$ y $0,500$ de precisión a tener $0,800$ y $0,667$, respectivamente. Respecto a la cobertura, se puede observar que para el conjunto de datos *trend* todas

[8] https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.seasonal_decompose.html

las técnicas empeoraron los valores para esa métrica. Por otra parte, para el conjunto *res* se mantuvo constante para *PC* y *VAR*, pero para las otras dos técnicas la cobertura empeoró.

Un resumen de los valores de precisión, cobertura y F1-score se pueden ver en la Tabla 4.8. Como se puede apreciar, considerando todos los conjuntos de datos, la técnica con mejor precisión fue *VAR*. En términos de cobertura, teniendo en cuenta todos los conjuntos de datos, *PCMCI*, *PC* y *VAR* obtuvieron el mismo valor de cobertura promedio. Por último, considerando todos los conjuntos de datos, la técnica *PCMCI* tuvo el mejor F1-score promedio.

En estos experimentos se puede ver, nuevamente, que de acuerdo a las características de los datos y del dominio, y de los objetivos que se persiguen, la metodología, las métricas y las técnicas deben ser ajustadas adecuadamente. No hay una sola técnica ni métrica mejor para todos los dominios. Por ejemplo, la técnica *Direct-LiNGAM* obtuvo el mejor F1-score para el conjunto de datos *full* (0,667) pero el peor para los datos *res* (0,286). La elección de las técnicas y la metodología va a depender de las características de los datos y de si el objetivo es obtener un conjunto pequeño de arcos correctos o si el objetivo es obtener buena cobertura para un posterior filtrado de falsos positivos. Estos resultados también muestran que, dependiendo de la naturaleza de los datos y los objetivos, un paso de preprocesamiento de los datos puede o no ser necesario.

Por último, como los conjuntos de datos obtenidos de *CAMMESA* eran pocos (uno más dos transformaciones de datos) y con pocas variables (solo cuatro variables consideradas) se pudo visualizar los grafos causales resultantes (algo que no fue posible para los datos sintéticos por la gran cantidad de conjuntos de datos y de variables). Del análisis de los grafos surge nuevamente la importancia de las métricas: mientras el conjunto *full* tuvo el mejor desempeño en términos de F1-score promedio, resulta evidente que no se trata de un conjunto informativo para ese conjunto de datos (ya que los grafos son casi totalmente conectados, y muchos arcos son incorrectos). Por otro lado, al aplicar las técnicas al conjunto *res*, se obtuvo peor desempeño en términos de F1-score promedio, pero los grafos causales resultantes son más informativos.

Guiado por estas conclusiones, no se elige una única técnica para ser usada como parte del *framework* de extracción de relaciones causales a partir del texto, sino que las cuatro técnicas son utilizadas en la Sección 4.6 utilizando una estrategia de votación por consenso (*ensemble* de técnicas) para maximizar la precisión de los arcos extraídos.

	Precisión				Cobertura				F1-score			
	full	***trend***	***res***	**promedio**	***full***	***trend***	***res***	**promedio**	***full***	***trend***	***res***	**promedio**
PCMCI	0,500	1,000	0,500	0,667	0,800	0,400	0,600	**0,600**	0,615	0,571	0,545	**0,577**
PC	0,500	1,000	0,667	0,722	0,800	0,200	0,800	**0,600**	0,615	0,333	0,727	0,559
DLiNGAM	0,571	0,000[9]	0,500	0,357	0,800	0,000	0,200	0,333	0,667	0,000	0,286	0,317
VAR	0,444	1,000	0,800	**0,748**	0,800	0,200	0,800	**0,600**	0,571	0,333	0,800	0,568
Promedio	0,504	**0,750**	0,617	0,624	**0,800**	0,200	0,600	0,533	**0,617**	0,310	0,590	0,505

Tabla 4.8: Desempeños en términos de precisión, cobertura y F1-score de aplicar las técnicas *Direct-LiNGAM* (*DLiNGAM*), *VAR*, *PCMCI* y *PC* sobre el conjunto de datos provisto por *CAMMESA* (*full*) y dos transformaciones del mismo: *res* y *trend*. La primera transformación consiste de reemplazar las variables Temp, Hum y DemGBA por los residuos de modelar esas mismas variables como una regresión lineal de 20 variables que dan información de la estación, mes del año y día de la semana. Al modelar las variables como lo no explicado por esas 20 variables (los residuos de la regresión), se eliminan componentes cíclicos. El conjunto *trend*, consiste de las mismas variables, pero luego de aplicarles un filtro de descomposición estacional mediante promedios móviles para solo mantener la componente de la tendencia. Estas dos transformaciones de los datos son aplicadas a las variables reales, no a la variable binaria Wrk. Se puede observar la alta cobertura obtenida con el conjunto *full* a costa de una mala precisión. Se observa el escenario inverso para *trend* (alta precisión, baja cobertura). Por otra parte, *res* representa un escenario intermedio entre los otros dos. Se puede observar que, si bien el conjunto *full* parece el mejor en términos de F1-score, no necesariamente indica que este sea el mejor conjunto de datos. Como se ve en la Figura 4.15, el grafo obtenido con *full* no es informativo ya que es casi el grafo totalmente conectado. Debido a la gran cantidad de arcos en la *ground truth* de *CAMMESA*, una representación de los datos que maximiza la cobertura se ve beneficiada, esto no implica que esa representación sea la más indicada para esos datos.

4.6. Aplicación a Textos de Artículos Periodísticos

En la presente Sección se reportan los resultados de aplicar las cuatro mejores técnicas de recuperación causal sobre los tres conjuntos de datos de textos extraídos del *New York Times*: (1) el conjunto de datos de términos extraídos usando la técnica de pesaje de términos FDD_β, (2) el conjunto de datos de eventos en curso detectados de los textos y (3) el conjunto de dato que combina los dos anteriores. Una descripción detallada de los conjuntos de datos se puede encontrar en la Sección 4.3.4. Al igual que para los experimentos sobre *CauseMe* y *CAMMESA*, se seleccionan cuatro del total de nueve

[9]La técnica *Direct-LiNGAM* (*DLiNGAM*) para el conjunto de datos obtenido de *CAMMESA* con la transformación *trend* encontró cero arcos causales, por ende el valor de la precisión queda indefinido (0/0). Para evitar usar este valor indefinido, a esa técnica con ese conjunto, se le asigna el valor de precisión cero.

técnicas del estado del arte basándose en los experimentos previos realizados. Los métodos seleccionados son: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*. Estas cuatro técnicas son usadas en conjunto para crear una única técnica que consiste de la aplicación de las cuatro anteriores con una política de votación unánime. Esto es, la técnica solo considera un arco como correcto si ese arco fue detectado simultáneamente por las cuatro técnicas utilizadas. A esta técnica, en este trabajo, se la denomina *ensemble*.

Los resultados de la aplicación de la técnica *ensemble* al conjunto de datos de términos (1) y al conjunto de datos de eventos en curso (2) se pueden ver en la Figura 4.16 (arriba a la izquierda y arriba a la derecha, respectivamente). En la misma figura (abajo) se reporta el grafo resultante de aplicar la técnica *ensemble* al conjunto de datos que combina eventos en curso y términos (3).

De analizar el grafo resultante de aplicar el *ensemble* al conjunto de datos de términos (Figura 4.16, arriba a la izquierda), se desprenden las siguientes conclusiones. Algunas relaciones causales resultan de interpretación directa y se pueden presuponer correctas. Por ejemplo, las menciones de acciones militares sobre Irak que hayan precedido a la guerra en dicho país. Otros arcos no tienen una interpretación tan directa como el caso de United Nations Security (Council) y su relación causal con la guerra en Irak, con Saddam Hussein o con la guerra del golfo pérsico. Sin embargo, teniendo en cuenta que en este análisis causal la causa precede al efecto, el arco (‘`United`’, ‘`Nations`’, ‘`Security`’) → (‘`war`’, ‘`Iraq`’) puede estar capturando los esfuerzos por parte de esa entidad para que “cumplan con sus obligaciones de desarme” a través de resoluciones como la 1441, esfuerzos que precedieron a la guerra. Otros vínculos como (‘`weapons`’, ‘`mass`’, ‘`destruction`’) → (‘`chemical`’, ‘`biological`’, ‘`weapons`’) son más difíciles de interpretar, pero es posible que se deban a que cada vez que surgía el tema de posibles programas de armas de destrucción masiva que funcionaban en Irak en ese momento, se hablaba después de posibles programas de armas biológicas. De igual manera, el vínculo (‘`United`’, ‘`Nations`’, ‘`Security`’) → (‘`Persian`’, ‘`Gulf`’, ‘`war`’), no puede interpretarse de manera directa (ya que las causas de una guerra no se pueden resumir a una sola interacción con una sola entidad o institución), pero se puede considerar que se refieren a la resolución 678 del *United Nations Security Council*, en la cual se le da una fecha límite a Irak para retirarse de Kuwait, evento que precedió a la Guerra del Golfo.

De analizar el grafo resultante de aplicar el *ensemble* al conjunto de datos de eventos en curso (Figura 4.16, arriba a la derecha), se desprenden las

siguientes conclusiones. Las variables `C550` y `C165` están vinculadas con una relación causal en ambas direcciones (`C550` $\leftrightarrow$ `C165`). Ambas variables están compuestas mayoritariamente por menciones de la guerra de Irak, pero la primera tiene una fuerte componente de menciones de sentimientos en contra de la guerra (against) de acuerdo a las nubes de palabras de la Figura 4.9 (para ver la descripción completa de los grupos ver Tabla 4.5). Que eventos relacionados a la guerra reportados en los artículos periodísticos causen sentimientos en contra de la guerra es un vínculo causal esperable. El sentido contrario es más difícil de analizar, pero puede tener que ver con la dinámica de la forma en la que se reportaban noticias de la guerra. Por ejemplo, es posible que cuando la conversación sobre la guerra (y el estar en contra o no) está más presente en lo cotidiano, se vuelve más probable que los medios reporten noticias relacionadas a la guerra. Aunque esta es una posible teoría, también puede ser que esto se deba simplemente a lo similares que son los eventos y que suelen ser mencionados en conjunto, lo cual indicaría que es una limitación de las técnicas al confundir co-ocurrencia con causalidad. Los vínculos causales `C269` $\rightarrow$ `C109` $\leftarrow$ `C201` son de interpretación más sencilla. Las variables representan la invasión a Kuwait por parte de Irak, reportes de muertes de soldados y civiles por parte de diferentes países y reportes de ataques terroristas, respectivamente. El hecho de que la invasión a Kuwait y los ataques terroristas causen posteriores reportes de muertes de civiles y soldados es una relación causal fácil de interpretar y que se puede suponer correcta.

Por último, **de analizar el grafo resultante de aplicar el *ensemble* a la combinación de los conjuntos de datos de eventos en curso y de términos (Figura 4.16, abajo), se desprenden las siguientes conclusiones**. Primero, se puede ver cómo variables de distintos tipos (eventos y términos) se vinculan causalmente en el grafo resultante de aplicar la técnica *ensemble*, aun siendo que estas variables se construyeron de maneras distintas y presentando diferentes características (diferentes frecuencias medias y desvíos estándares, ver Tablas 4.7 y 4.6). Esto daría evidencia para afirmar que el *framework* de recuperación de estructuras causales puede manejar diferentes tipos de variables extraídas del texto e incluso variables exógenas al texto que se pueden incluir si están en el mismo periodo de tiempo y frecuencia (como, por ejemplo, agregar precios de acciones de la bolsa, o precios de materias primas o indicadores socioeconómicos de países, entre otros).

Respecto a las relaciones causales extraídas, se puede observar vínculos causales con una semántica bien definida. Por ejemplo, el vínculo entre la posible existencia de armas

de destrucción masiva en Irak como uno de los posibles justificativos para iniciar acciones militares (‘weapons’, ‘mass’, ‘destruction’) → (‘military’, ‘action’,‘Iraq’) dando inicio a la guerra en Irak (‘military’, ‘action’,‘Iraq’) → (‘war’, ‘Iraq’). Este tipo de relaciones, si bien no deberían ser interpretadas automáticamente como una relación causal real, brinda información sobre la forma en la que son reportados ciertos eventos en los textos de artículos de noticias utilizados. Su extracción puede ofrecer información sobre pares de eventos con fuerte co-ocurrencia donde uno precede al otro, y permitiría un análisis sobre el contexto a la luz de esta información. Por otro lado, siendo que la variable C109 representa reportes de muertes de civiles y soldados, el vínculo causal (‘war’, ‘Iraq’) → C109, es una relación causal que se puede presuponer correcta, ya que es razonable considerar la existencia de la guerra como la causa de los reportes de bajas de civiles y soldados. Los vínculos causales C201 → C109 ← C269, que relacionan causalmente los ataques terroristas y la invasión a Kuwait con reportes de civiles y soldados muertos, representan un vínculo que también se puede presuponer correcto. Este vínculo causal ya había sido encontrado por el *ensemble* al ser aplicado en el conjunto de datos de eventos. Análogamente, el vínculo (‘weapons’, ‘mass’, ‘destruction’) → (‘chemical’, ‘biological’, ‘weapons’), ya había sido encontrado y discutido durante los experimentos con el *ensemble* sobre el conjunto de datos de términos.

Como se mencionó previamente, no todos los arcos encontrados son necesariamente correctos, pero proveen a un posible experto con información sobre qué variables están fuertemente interconectadas (y posiblemente causalmente vinculadas) en un dominio. Esta herramienta puede ser de gran utilidad a un experto que está tratando de entender un dominio que puede ser muy complejo y contener una gran cantidad de variables y relaciones. Adicionalmente, al transformar los textos de lenguaje natural a información estructurada de variables relevantes (series de tiempo), se provee una gran flexibilidad para combinar información del mundo real reportada en las noticias con diferentes variables que ya tienen formato de serie de tiempo (precios de materias primas, precios de acciones de la bolsa, indicadores socioeconómicos, entre otros), lo cual permitiría enriquecer el dominio y dar información extra a los expertos.

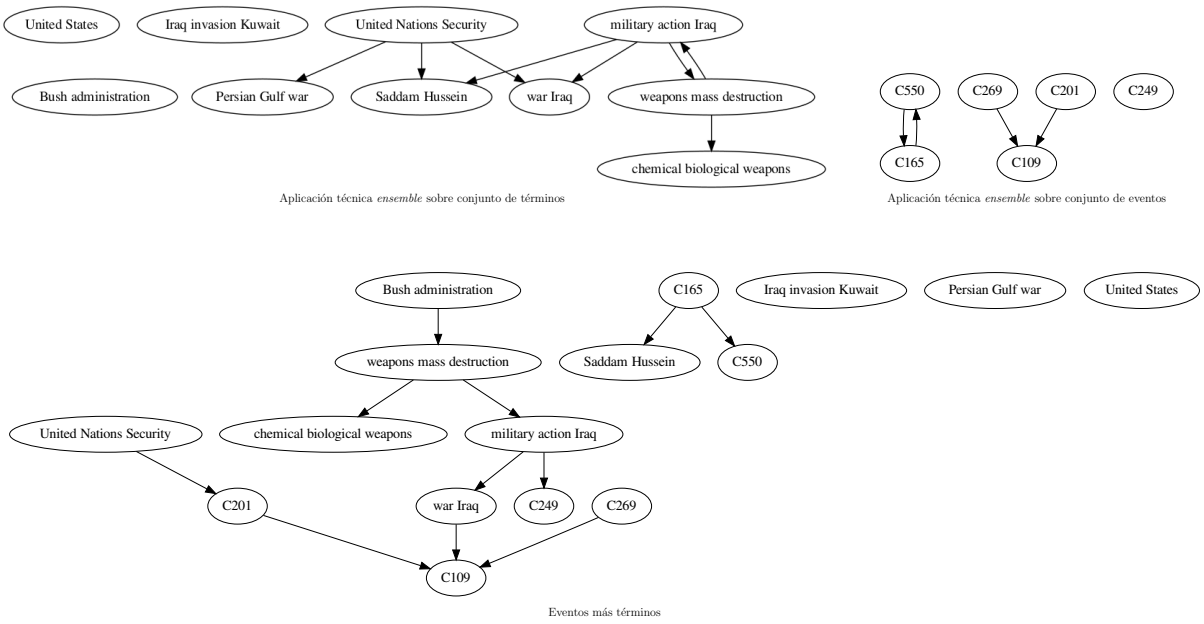


Figura 4.16: Grafos causales resultantes de aplicar la técnica *ensemble* sobre el conjunto de datos de términos (arriba a la izquierda), eventos en curso (arriba a la derecha) y el conjunto de datos construido combinando los anteriores (abajo). Ambos conjuntos de datos (el de términos y el de eventos) son extraídos del corpus del *New York Times*. La técnica *ensemble* consiste de la aplicación de las cuatro técnicas *Direct-LiNGAM*, *PCMCI*, *PC* y *VAR*, decidiendo sobre cada arco con una política de votación unánime. Los nodos del conjunto de datos de eventos están identificados con el número de grupo (cluster). Cada grupo contiene menciones (instancias) de un mismo evento (semántica similar). El grupo `C109` se corresponde con reportes de muertes durante la Guerra de Irak, tanto soldados (de ambos bandos) como civiles. El grupo `C165` se corresponde con reportes de opinión sobre la guerra de Irak. El grupo `C201` son menciones de eventos de ataques terroristas. El grupo `C249` son menciones en ataques o acciones militares. El grupo `C269` trata sobre la invasión a Kuwait. Finalmente, el grupo `550` reúne menciones de la guerra en Irak en general. Una descripción de estos grupos es presentada en la Tabla 4.5. Se puede ver la capacidad del *framework* de recuperar variables relevantes al dominio, y la capacidad del mismo de encontrar vínculos causales interpretables entre las mismas. A partir de estos vínculos se puede estudiar la precedencia y el impacto de una variable hacia otra.

4.7. Conclusiones

En el presente capítulo se presentó un extenso análisis de nueve técnicas de descubrimiento causal, utilizando 64 conjuntos de datos sintéticos (56 creados con *TETRAD*, 8 descargados de *CauseMe*) y dos reales (los conjuntos de datos provistos por la empresa *CAMMESA* y los creados a partir de los textos del *New York Times*). Las nueve técnicas son puestas a prueba en los 56 conjuntos de datos creados con *TETRAD*, posteriormente, se seleccionan las mejores cuatro para continuar en los demás conjuntos, siendo las

cuatro mejores técnicas: *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*. El *framework* de descubrimiento causal a partir de textos de artículos periodísticos es formalmente definido en este capítulo y un caso de uso es presentado usando los datos generados a partir del *New York Times*. Dicho *framework* consiste de una etapa de selección de variables de textos en la cual dos tipos distintos de variables son extraídos (términos y eventos en curso). Esta primera etapa de selección de variables es llevada a cabo con las herramientas propuestas en los Capítulos 2 y 3. Posteriormente una técnica de descubrimiento causal es aplicada sobre dichas variables, la técnica es una combinación de las cuatro mejores siguiendo una política de votación unánime, a la cual se la denomina *ensemble*.

En los experimentos realizados sobre los conjuntos de datos generados con *TETRAD*, se puede observar una gran disparidad de desempeños. Se puede ver que hay tres técnicas que en muchos casos tienen un desempeño no mucho mejor, e incluso a veces peor, que *Random*. Estas tres técnicas son *Lasso-Granger*, *Transfer entropy* e *ICA-LiNGAM*, las cuales no tuvieron una diferencia significativa en términos de F1-score comparado con *Random*. Para el caso de *Transfer Entropy* se puede explicar debido a que esta técnica tiene la limitante de analizar causalidad de a pares (potencialmente dejando variables relevantes fuera del análisis, generando problemas de variables ocultas). Por otra parte, *ICA-LiNGAM* tiene limitaciones que fueron descritas por los mismos autores en publicaciones posteriores: problemas de convergencia a la solución correcta si el estado inicial no es el adecuado y sensibilidad a la escala de los datos. Para el caso de *Lasso-Granger*, si bien dicha técnica tiene un proceso previo de selección de variables usando lasso, sufre de la misma limitación que *Transfer Entropy*, esto es, solo considera causalidad de a pares (usando la técnica de causalidad de Granger de a pares de variables). Por las limitaciones de estas técnicas al ser aplicadas a estos contextos multivariados, las mismas no fueron consideradas para los experimentos posteriores en los demás conjuntos de datos (*CauseMe*, *CAMMESA*, *New York Times*).

Por otro lado, las técnicas que construyen un modelo VAR penalizado (*BigVAR* y *SIMoNe*), obtuvieron resultados poco consistentes (grandes intervalos de confianza) y en muchos casos, la técnica *SIMoNe* tuvo desempeño peor que la técnica *Random*. Más aún *SIMoNe*, en términos promedio, no tuvo un desempeño significativamente distinto a *Random*. En contraposición, la técnica *VAR* (sin penalización) obtuvo desempeños que la ubican entre las mejores cuatro técnicas. Esto demuestra que los procesos de penalización no fueron adecuados para el proceso de descubrimiento causal. Para el caso de

BigVAR, aumentó el desempeño en términos de precisión, pero obtuvo un mal desempeño en términos de cobertura. Dependiendo de los objetivos de la tarea, una técnica de este estilo puede ser considerada, aunque su desempeño fue en términos generales poco consistente. Por otro lado, *SIMoNe* obtuvo bajos valores de cobertura, probablemente debido a la penalización, pero también un desempeño bajo en términos de precisión. Por lo cual tanto la inferencia como la penalización de esta técnica no fueron buenas, resultando en muchos falsos negativos (tal vez por problemas de penalización) y falsos positivos (por limitaciones de la técnica para el descubrimiento de vínculos causales correctos). Ninguna de las dos técnicas basadas en modelos VAR penalizados fueron consideradas para los experimentos posteriores. Finalmente, las cuatro técnicas seleccionadas para los experimentos en los demás conjuntos de datos (*CauseMe*, *CAMMESA* y *New York Times*) son *Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*.

Se puede observar que, estas cuatro mejores técnicas, se ordenan de mayor a menor en términos de F1-score de la siguiente manera: primero *Direct-LiNGAM* con el mejor F1-score, luego *PCMCI*, *VAR* y por último *PC*. Aunque las técnicas compartan este orden en términos promedio, se puede observar de los intervalos de confianza que no hay una diferencia significativa en sus desempeños para la métrica F1-score.

En los experimentos realizados sobre los conjuntos de datos obtenidos de *CauseMe*, se puede observar nuevamente, que varias técnicas en ciertas configuraciones no pudieron superar a *Random*. Para el caso de pocas variables, donde cada arco equivocado tiene proporcionalmente un impacto mayor en las métricas, se puede ver que la técnica *Random* fue mejor en término de F1-score, que *PCMCI* (para $N = 3$ y $N = 5$) y que *Direct-LiNGAM* (para $N = 3$). Estos resultados se mantienen tanto para $T = 300$ como para $T = 600$. La técnica *Random*, logró valores altos de cobertura con respecto a las otras técnicas a través de seleccionar aleatoriamente qué arcos incluir con una proporción de aceptación alta (50-50). De esta manera, esta técnica logró superar en términos de F1-score a otras sin aportar información causal real. Esto pone en evidencia la importancia de las métricas en el descubrimiento causal. Siendo que se puede obtener cobertura perfecta con una técnica que siempre da como resultado el grafo totalmente conectado, en muchos casos hay que priorizar la precisión, u obtener un balance de ambas métricas pero dándole mayor ponderación a la precisión (por ejemplo, usando F-measure con un $\beta < 1$). También se pudo ver que técnicas que fueron las mejores en otros conjuntos de datos pasaron a tener desempeños peores que *Random* en algunas configuraciones. Por ejemplo,

PCMCI pasó de ser la segunda mejor técnica en *TETRAD* a ser peor que *Random* en términos de F1-score para $N = 3$ y $N = 5$, tanto con $T = 300$ como con $T = 600$.

Los experimentos en los conjuntos de datos obtenidos de *CauseMe* permiten llegar a conclusiones similares a las obtenidas de *TETRAD*. Primero, las cuatro mejores técnicas siguen teniendo desempeños similares, no existe una técnica que haya tenido un desempeño notablemente mejor o peor que otras. También se pudo ver que dependiendo del conjunto de datos puede que las técnicas tengan pequeñas variaciones en los desempeños, dando evidencia para suponer que no hay una única técnica que sea mejor para todos los conjuntos de datos y todas las configuraciones. Se puso en evidencia la importancia de usar las métricas correctas. De otro modo se pueden juzgar técnicas que no aportan información causal como *Random* (que no realiza descubrimiento causal real), como mejores que otras que sí han demostrado capacidad de descubrir relaciones causales. Teniendo en cuenta que muchas veces el desempeño de las técnicas depende del conjunto de datos y que no necesariamente hay una mejor, se continúa con las cuatro técnicas para los experimentos en los conjuntos de datos reales: el provisto por la empresa *CAMMESA* y el generado a partir de los textos del *New York Times*.

Los experimentos realizados sobre los conjuntos de datos provistos por la empresa *CAMMESA* tenían por objetivo agregar perspectivas diferentes al análisis. Primero, se agrega al análisis los resultados de aplicar las técnicas a un conjunto de datos real con formas funcionales, ruidos y características desconocidas. Adicionalmente, a diferencia de los conjuntos de datos sintéticos de los cuales se tenían 64 conjuntos de datos, para *CAMMESA* solo se tenía un único conjunto de datos, al cual se le aplicaron dos transformaciones resultando en tres conjuntos de datos basados en una misma fuente. Tener pocas variables y pocos conjuntos de datos permitió, por primera vez, reportar e inspeccionar en detalle los grafos resultantes (no solo concentrarse en las métricas de desempeño). Analizar el mismo conjunto de datos, pero sometido a diferentes transformaciones (filtrados de ciclos), permitió incorporar una dimensión más al análisis: la importancia del correcto preprocesamiento de los datos.

Se puede observar que, para el conjunto de datos sin transformar, se obtienen, para todas las técnicas, altos valores de cobertura, pero unos valores de precisión muy bajos, con una gran cantidad de arcos incorrectos. Un escenario de este estilo puede ser beneficioso si se busca maximizar la cobertura para un filtrado de precisión posterior (tal vez con un usuario involucrado en esta segunda etapa). Pero si se busca maximizar la precisión, los

otros dos conjuntos de datos son mejores (los dos conjuntos que se obtienen de aplicar transformaciones a los datos). En el otro extremo, la transformación que surge de aplicar un filtro de descomposición para obtener la tendencia de los datos (*trend*), obtiene precisión perfecta reportando cero arcos incorrectos encontrados, pero a costa de tener valores de cobertura muy bajos. Nuevamente se vuelve evidente la importancia de las métricas, observándose que en términos de F1-score promedio, el conjunto de datos sin transformar tiene los mejores resultados. Sin embargo, al analizar en detalle los grafos resultantes se puede observar que estos grafos no aportan mucha información causal ya que son grafos casi totalmente conectados. Estos grafos tienen poco valor para un usuario que tendría que analizar un grafo casi totalmente conectado para tratar de sacar conclusiones causales. En este caso, al guiarse por la métrica F1-score, uno puede llegar a la conclusión de que no transformar los datos es la mejor opción, pero depende mucho del objetivo que se persigue (alta cobertura o alta precisión).

Una vez más se puede apreciar que no hay una técnica que sea consistentemente superadora en términos de F1-score para todos los conjuntos de datos. En promedio la técnica con peor F1-score es *Direct-LiNGAM*, la cual tuvo el mejor desempeño en términos de F1-score tanto para los conjuntos generados con *TETRAD* como los obtenidos de *CauseMe*. Esto puede deberse a que las fuertes restricciones que impone la técnica (linealidad con ruidos no gaussianos) no se cumplen en un conjunto de datos real como el de *CAMMESA*. Al no haber una técnica superadora se plantea el uso de las cuatro técnicas para el conjunto de datos generados del *New York Times*. Como se trata de un conjunto de datos en el que la *ground truth* no es conocida, y que se tiene un gran número de variables, se prioriza la precisión por encima de la cobertura. Por este motivo, en lugar de utilizar las cuatro técnicas por separado y analizar su desempeño, se propone la creación de una única técnica que consiste en aplicar las cuatro mejores técnicas con una política de votación unánime. A esta técnica se la denomina *ensemble*.

Para el caso de uso del *framework* realizado sobre el conjunto de datos construido a partir del *New York Times* se puede observar que a pesar de haber aplicado la técnica *ensemble*, que maximiza la precisión, se tiene un gran número de arcos resultantes para analizar. De los resultados y el análisis presentado en la Sección 4.6 se puede ver el potencial del *framework* para capturar variables altamente relevantes a un dominio, las cuales en muchos casos capturaban eventos del mundo real reportados en las noticias que son de suma importancia para el dominio. Adicionalmente, se vio

la capacidad de la técnica de proporcionar vínculos entre estas variables que, con una correcta interpretación por parte de un experto, le permitirían a este mismo entender mejor los sucesos relacionados al dominio que acontecieron (o están aconteciendo).

Adicionalmente, se vio cómo la técnica de recuperación causal fue capaz de encontrar relaciones significativas entre variables que fueron construidas de formas distintas. Esto es, relaciones entre variables tipo evento en curso (construidas a partir de un modelo de aprendizaje automático implementado para la detección de dichas variables) y variables tipo términos (extraídas a partir de la herramienta de pesaje de términos propuesta, FDD_β). Estos dos tipos de variables fueron construidas de formas distintas y presentan escalas distintas (por ejemplo, diferente media y desvío estándar), e incluso ante variables heterogéneas como estas, se pudieron reconstruir relaciones causales relevantes. Esto da indicios para asumir que el *framework* propuesto posee una gran flexibilidad, ya que podría incorporar variables de distintos tipos (siempre y cuando todas las variables utilizadas puedan ser representadas como series de tiempo comprendidas dentro del mismo periodo de tiempo y con igual frecuencia). Esto permitiría, por ejemplo, incorporar variables financieras como el precio de acciones de empresas relevantes, o precios de materias primas importantes para el dominio, o indicadores socioeconómicos del país, o los países, relevantes al dominio.

El desarrollo completo del presente caso de uso del *framework* demostró la capacidad del mismo para obtener representaciones resumidas de contextos complejos a través de la representación de variables y relaciones relevantes entre las mismas. El *framework*, demostró un gran potencial para ser usado como parte de una herramienta de visualización interactiva, donde diferentes parámetros se pueden configurar para seleccionar más o menos variables. Por ejemplo cambiar los valores de β de la técnica FDD_β para seleccionar variables más descriptivas o más discriminativas. Así como también es posible configurar diferentes aspectos de la recuperación de eventos (diferentes valores de K, o diferentes umbrales mínimos de cohesión y cantidad de instancias por grupo) para obtener mayor o menor cantidad de variables de tipo evento sobre las cuales elegir cuáles se visualizan y cuáles no. Adicionalmente, se puede incorporar interacción para el paso de descubrimiento causal, cambiando la política de votación *ensemble*, o agregando o sacando técnicas del mismo.

Si bien la propuesta demostró ser viable, aún existen algunos puntos que se pueden mejorar para obtener representaciones más útiles, completas y fáciles de visualizar. Por

ejemplo, la representación de los grupos de menciones del mismo evento en el grafo se realizó a través de la etiqueta de identificación del grupo (`CXXX`). Para poder interpretar cada grupo se procedió a la construcción de nubes de palabras para cada evento, de tal forma que la representación se concentró principalmente en el *event-trigger* y en menor medida en el resto de los términos de la oración, ponderando como más relevantes a aquellos términos más cercanos al *trigger*. Esta representación de nube de palabras no resultó adecuada para incorporar al grafo causal (cada nodo evento tendría que haber sido una nube de palabras dificultando la legibilidad del grafo). Una posible dirección para mejorar la representación de grafo causal es cambiar la etiqueta de grupo (`CXXX`) por un resumen textual del grupo a través de alguna técnica de resumen automática (*Text summarization*).

Otra mejora posible consiste en complementar la etapa de recuperación causal con herramientas de NLP que recuperan menciones de causalidad explícita del texto. Estás técnicas se puede aplicar sin modificación al *framework*, ya que estas recuperan relaciones causales entre palabras (o conjuntos de palabras) dentro del mismo texto, y como las variables del *framework* son también conjuntos de palabras (unigramas, bigramas, trigramas o *event-triggers* de eventos en curso), las relaciones de causalidad podrían ser directamente agregadas al grafo causal o podrían ser usadas como un voto adicional para el *ensemble* para incrementar la precisión del mismo.

En resumen, el caso de uso presentado fue efectivo en mostrar la viabilidad de la propuesta y en señalar posibles direcciones de trabajo futuro en las cuales se puede mejorar el mismo. Por ejemplo, a través de los trabajos futuros previamente mencionados, que incluyen: (1) la incorporación de más variables de distintos tipos, (2) la construcción de una herramienta interactiva para modificar los diferentes parámetros y visualizar el grafo resultante, (3) mejorar la representación de los nodos mediante la implementación de una herramienta para resumir los grupos de menciones de eventos y (4) agregar la funcionalidad de extraer pares causales directamente del texto a través de herramientas de NLP.

Capítulo 5

Conclusiones y Trabajo a Futuro

Esta tesis presenta un estudio y discusión acerca de cómo definir relaciones de causalidad y llevar a cabo inferencias causales. En particular, se investiga la capacidad de los modelos causales para responder a preguntas a las cuales los modelos puramente estadísticos/probabilísticos no pueden contestar (preguntas intervencionales y contrafactuales). Más aún, en [SLB+21] se mencionan los beneficios de incorporar una perspectiva causal al área de aprendizaje automático, entre ellos una mayor robustez y una menor vulnerabilidad a ataques adversarios (*adversarial attacks*). En [Ahe16] se discute el creciente interés por investigar y entender la estructura del sistema financiero y los correspondientes canales de propagación de riesgos, mostrando que el análisis de redes ha resultado ser una herramienta crucial para modelar el complejo sistema de interconexiones de este tipo de estructuras. Por otro lado, en [Var14], el autor discute las oportunidades, aún no enteramente explotadas, de una posible sinergia entre aprendizaje automático y econometría para la inferencia causal.

El presente trabajo ha sido motivado por la necesidad de generar herramientas para entender la estructura de sistemas complejos, como el sistema financiero. Existe un contexto favorable para este trabajo, caracterizado por un creciente número de trabajos orientados a diseñar herramientas de inferencia causal que aún no han sido explotadas en la práctica de la forma en la que aquí se propone. Un aspecto original del trabajo aquí presentado es que desarrolla un enfoque interdisciplinario combinando las áreas de procesamiento de lenguaje natural, aprendizaje automático y econometría, apoyándose en la creciente disponibilidad de gran cantidad de datos de tipo textual. Con estas motivaciones en mente, en la presente tesis se presentó y analizó un marco de trabajo (*framework*) para

la detección y selección de eventos del mundo real y otras variables relevantes de textos de noticias con el objetivo de aprender modelos causales a partir de ellas. El objetivo final fue el estudio de la viabilidad de una herramienta para ayudar a expertos a entender y explicar el desarrollo de eventos en escenarios complejos y sus interconexiones a través de la presentación de un modelo causal del dominio.

El *framework* se desarrolló en dos etapas principales: (1) la construcción y selección de variables y (2) la inferencia de la estructura causal que vincula dichas variables. A su vez la etapa (1) se divide en dos subetapas, en las cuales distintos tipos de variables son extraídas: (1.a) las extracción de variables de tipo términos (unigramas, bigramas y trigramas) y la etapa (1.b) donde variables de tipo eventos son extraídas del texto. Las tareas (1.a), (1.b) y (2) son las presentadas en los Capítulos 2, 3 y 4, respectivamente. A continuación se presentan las conclusiones obtenidas como resultado de los experimentos realizados para cada una de esas etapas. Dichas conclusiones se presentan en el mismo orden utilizado para la tesis ((1.a), (1.b), y (2)).

Para la selección de términos para ser usados como variables (1.b) se define la técnica de ponderación de términos FDD_β en el Capítulo 2. En dicho capítulo se analiza su comportamiento y se evalúa su desempeño para diferentes tareas. En particular, se muestra la utilidad de la técnica propuesta para la tarea de identificación de términos específicos a un dominio (estimación de puntajes de relevancia elegidos por usuarios). Esta tarea representa el objetivo principal para el cual se aplica la técnica de pesaje de términos propuesta dentro del *framework*. Esto es, la técnica FDD_β es usada dentro del *framework* para estimar puntajes de relevancia para cada término dentro de un dominio. De esta manera se la utiliza para la selección de variables (solo tomando los K términos con mayor FDD_β).

Las contribuciones del capítulo donde se introduce la técnica FDD_β son las siguientes. Primero se introduce formalmente la definición de la técnica y se presenta un extenso estudio sobre el análisis y el impacto del parámetro β en el comportamiento de la técnica FDD_β. La segunda contribución es la presentación de un extenso estudio comparativo que involucra la técnica propuesta y dieciocho técnicas del estado del arte y tradicionales (tanto supervisadas como no supervisadas). El estudio presenta evidencia de que, a pesar de su simplicidad, la técnica FDD_β obtiene resultados competitivos o superadores con respecto a las demás técnicas, incluso en comparación con técnicas basadas en conceptos más complejos como de teoría de la información o estadística. La tercera contribución

es una nueva perspectiva acerca de la cuestión de cómo construir consultas de múltiples términos basándose en técnicas supervisadas de pesaje de términos. Esta perspectiva surge de los resultados reportados en el Capítulo 2 que demuestran la ventaja de la técnica FDD_β para la construcción de consultas complejas explorando diferentes objetivos (utilizando su parámetro ajustable). Finalmente, la cuarta contribución es la creación y publicación de un conjunto de datos utilizado durante los experimentos en dicho capítulo. Los análisis presentados en dicho capítulo demuestran la utilidad de la técnica presentada como estimador de la relevancia de términos para un tópico, permitiendo la selección de los términos de un dominio de interés para el usuario. Más aún, como el parámetro ajustable es de sencilla interpretación, brinda la posibilidad de que el usuario lo defina contando así con más grados de libertad al momento de seleccionar términos, permitiendo apuntar a diferentes objetivos.

Para la construcción de variables tipo evento (1.b) fue necesario definir la tarea de detección de eventos en curso (*Ongoing Event Detection* (OED)), para luego, construir y evaluar un modelo predictivo para dicha tarea. La definición y los experimentos son presentados en el Capítulo 3. La contribución principal de dicho capítulo es la definición de la tarea OED y una extensa evaluación experimental de la misma, que combina diferentes modelos y atributos. Como parte de las contribuciones de dicho capítulo se encuentra la construcción y publicación de un conjunto de datos específicamente creado para la tarea. Como la tarea y el conjunto de datos se definen específicamente para el *framework* propuesto, y como el modelo predictivo mostró un desempeño superador al de los modelos del estado del arte presentados, se obtuvo evidencia que indicaría que el modelo diseñado resulta adecuado para formar parte del *framework*. Una discusión sobre los diferentes modelos, atributos y arquitecturas usados es presentada en la sección de conclusiones de dicho capítulo (Sección 4.7). Durante dichos experimentos se pudo observar que, para la tarea de detectar eventos en curso, el contexto es sumamente relevante (muchas veces para notar si es evento en curso o no es necesario tener en cuenta el contexto). Por este motivo, la utilización de *embeddings* contextuales, en particular *BERT* [DCLT18], resultó crucial para alcanzar un desempeño superador a los modelos de referencia (*baseline*). Finalmente, también se pudo observar la importancia del *transfer learning* para tareas como la de OED donde el conjunto de datos es reducido. En este contexto, utilizar *embeddings* preentrenados en grandes conjuntos de datos (tales como *Word2Vec* [MSC+13] o *BERT* [DCLT18]) resultó decisivo para obtener buenos desempeños. En muchos casos, la elección correcta de los atributos y la inclusión de atributos preentrenados mostró tener mayor impacto

que la elección de los hiperparámetros de los modelos o incluso que la elección del tipo de modelo.

Para la etapa de recuperación de estructuras causales (2), en el Capítulo 4, se probaron nueve técnicas del estado del arte para dicha tarea. Un extenso análisis de las nueve técnicas de descubrimiento causal es presentado en dicho capítulo, utilizando 64 conjuntos de datos sintéticos (56 creados con *TETRAD* [SSG+98], 8 descargados de *CauseMe* [RBB+19]) y dos reales (un conjunto de datos provistos por la empresa *CAMMESA* y otro creado a partir de los textos del *New York Times* [San08]). Las nueve técnicas son puestas a prueba en los 56 conjuntos de datos creados con *TETRAD*. Las mejores cuatro técnicas se seleccionaron para seguir siendo evaluadas en los demás conjuntos. Estas cuatro técnicas son: *Direct-LiNGAM* [SIS+11], *PC* [SG91], *PCMCI* [RNK+19] y *VAR* [Sim80]. El *framework* de descubrimiento causal a partir de textos de artículos periodísticos es formalmente definido en dicho capítulo y un caso de uso es presentado usando los datos generados a partir de artículos tomados del corpus *New York Times* [San08].

En los experimentos realizados sobre los conjuntos de datos generados con ***TETRAD***, se puede observar un bajo desempeño por parte de cuatro técnicas en términos de F1-score (*Lasso-Granger* [Gra69, Tib96], *Transfer entropy* [Sch00], *ICA-LiNGAM* [SHHK06] y *SIMoNe* [CSG+08]) y un desempeño a veces bueno, pero poco consistente por parte de *BigVAR* [NMB17]. Estas técnicas, por su bajo desempeño, no son consideradas para los experimentos en los demás conjuntos de datos (*CauseMe*, *CAMMESA* y *New York Times*). Solo las cuatro mejores técnicas (*Direct-LiNGAM*, *PC*, *PCMCI* y *VAR*) son utilizadas para los experimentos posteriores. El desempeño de estas técnicas en términos de F1-score promedio permite ordenarlas de la siguiente forma: la mejor es *Direct-LiNGAM*, seguida de *PCMCI*, *VAR* y por último por *PC*. Aunque en términos promedio se pueda especificar qué técnicas son mejores que otras, de los resultados se puede ver que no existe una única técnica que sea consistentemente mejor que todas las demás para todos los conjuntos de datos analizados.

Los experimentos realizados sobre los ocho conjuntos de datos de ***CauseMe*** permiten arribar a conclusiones similares a las obtenidas de los experimentos en *TETRAD*. Primero, las cuatro técnicas mantuvieron un desempeño considerablemente bueno sin que exista una técnica que exhiba un desempeño que la ubique por encima de todas las demás para todos los conjuntos de datos. Se puso en evidencia la importancia de utilizar métricas adecuadas, ya que la técnica *Random*, que elige arcos aleatorios con una proporción 50-50

logró obtener buenos valores de cobertura y así superar en F1-score a varias otras técnicas. Para evitar esta clase de problemas se puede, por ejemplo, elegir métricas que ponderen la precisión por encima de la cobertura.

Luego de los experimentos en datos sintéticos, se propusieron y llevaron a cabo experimentos sobre el conjunto de datos reales provisto por ***CAMMESA***. El objetivo de estos experimentos es analizar las técnicas en conjuntos de datos en los que las formas funcionales, los ruidos y otras características sean desconocidas. Más aún, como se trata de solo un conjunto de datos (más dos conjuntos de datos construidos con transformaciones del primero) y unas pocas variables (solo cuatro) fue posible hacer un análisis pormenorizado de cada uno de los grafos causales resultantes (no solo guiándose por los valores de precisión, cobertura y F1-score). El análisis de cada uno de los grafos permitió nuevamente discutir la importancia de elegir bien las métricas, ya que algunos grafos tuvieron buen desempeño en términos de F1-score, pero al ser analizados se pudo notar que se trata de grafos poco informativos (grafos casi completamente conectados, que dan poca información y que solo alcanzan un buen valor de F1-score por tener alto valor de cobertura). El análisis del desempeño de las técnicas en el conjunto de datos de *CAMMESA* con diferentes transformaciones también pone en evidencia la importancia del preprocesamiento de los datos para ciertos dominios (especialmente en dominios donde los datos presentan ciclos).

Una vez más se puede ver que las técnicas tienen comportamientos diferentes de acuerdo al conjunto de datos sobre las que se las aplica. Por ejemplo, en términos de F1-score la técnica *Direct-LiNGAM* estuvo entre las que alcanzaron los mejores desempeños en los conjuntos de datos sintéticos (*TETRAD* y *CauseMe*) pero fue la peor en el conjunto de datos de *CAMMESA*. Esto podría estar indicando que las restricciones funcionales impuestas por la técnica (linealidad y ruidos no gaussianos) no siempre se cumplen en datos reales como es el caso de *CAMMESA*. Debido a que los experimentos señalan que no existe una técnica mejor que todas las demás, y debido a que el *framework* es aplicado en un contexto de muchas variables y con una *ground truth* desconocida, se prioriza la elección de una técnica que maximiza la precisión. Por este motivo es que se utiliza una técnica que combina las cuatro mejores (*Direct-LiNGAM*, *PCMCI*, *PC* y *VAR*) con una estrategia de votación unánime (*ensemble*).

Sobre el *framework* completo. De los resultados y el análisis presentado del caso de uso del *framework* completo en el Capítulo 4 se puede observar la capacidad del mismo

para recuperar y seleccionar variables altamente relevantes a un dominio. Se reconoce también el potencial de la herramienta propuesta para reportar vínculos causales relevantes y con un significado semántico bien definido. Estas características del *framework* demuestran la viabilidad del mismo como herramienta para presentar un resumen de variables y vínculos relevantes a un potencial usuario, asistiéndolo en la toma de decisiones y en el entendimiento del dominio.

Adicionalmente, se puede observar el potencial del *framework* para permitir la combinación de variables extraídas del texto con otro tipo de variables (tales como indicadores económicos o socioeconómicos, así como también precios de bienes o servicios). Esta gran flexibilidad del *framework* lo vuelve atractivo para extender su implementación agregando funcionalidades, como por ejemplo, nuevas variables extraídas de texto u otras fuentes y nuevas técnicas de descubrimiento causal (que pueden estar basadas en series de tiempo o en NLP). En el mismo capítulo donde se introduce el caso de uso, se discute cómo ciertos parámetros del *framework*, por sus semánticas claras, permiten que un usuario interactúe con ellos de tal forma de obtener resultados que se ajusten aún mejor a los objetivos perseguidos. Esto permite obtener una herramienta más dinámica que cumple aún mejor su objetivo de brindar un conjunto de variables, eventos y entidades relevantes a un dominio y sus posibles interconexiones causales. Una descripción detallada del trabajo futuro posible (como la inclusión de nuevos tipos de variables, técnicas de descubrimiento causal e interactividad) son descritos en la Sección 4.7.

Si bien el *framework* propuesto demostró resolver los objetivos para los cuales fue planteado, la propuesta presenta algunas limitaciones menores. La herramienta que implemente el *framework* necesariamente va a estar restringida al conjunto de datos con el que se trabaje, reflejando solamente las variables seleccionadas presentes en el texto y los vínculos causales que se deriven de ella. Pero si eventos o variables relevantes están excluidos de los textos (accidental o intencionalmente) el grafo obtenido no va a reflejar la totalidad de las variables relevantes. Análogamente, si variables relevantes no son omitidas pero sí algunas menciones de las mismas, tal vez sea posible detectarlas pero no así los vínculos causales correctos entre ellas.

Otra de las limitaciones de la propuesta es que, debido a que la herramienta se compone de varias partes, cada una con su propio grado de error, la correctitud del grafo causal resultante se ve afectada por posibles inexactitudes que resulten de cada parte del *framework*. Más aún, el paso de recuperación de estructuras causales depende de la

correcta aplicación de los dos pasos previos de construcción y selección de variables. Por último, la propuesta debe ser considerada como una herramienta más para un usuario, no como un modelo que produce respuestas absolutas. Si bien en muchos casos la herramienta produce vínculos causales claros y con semántica definida, en muchos casos es necesaria la lectura y exploración de los datos por parte de un usuario para poder sacar las conclusiones correctas. Teniendo en cuenta esto, la herramienta puede ser extendida a través de la incorporación de interactividad de tal forma que permita al usuario explorar los textos de los cuales se extraen las variables y sus vínculos causales para sacar sus propias conclusiones cuando la visualización del grafo no es suficiente.

En resumen, más allá de sus limitaciones, los experimentos realizados en el contexto de esta tesis permiten demostrar la viabilidad de una posible herramienta que, siguiendo el marco de trabajo (*framework*) aquí propuesto, pueda recuperar un conjunto de variables relevantes a un dominio a partir de textos y construir una estructura de red causal vinculándolas. Esta herramienta permitiría a un experto entender las entidades, eventos u otras variables involucradas en un dominio y sus interacciones. Adicionalmente, el *framework* tal como es propuesto ofrece una gran flexibilidad en término de las variables que pueden ser incluidas a futuro y las herramientas de descubrimiento casual a ser usadas. Por estos motivos, múltiples propuestas de trabajo futuro se desprenden del presente trabajo.

Producción Científica

Publicaciones en Revistas

La técnica FDD_β presentada en el Capítulo 2 fue propuesta y analizada desde diferentes perspectivas en dos publicaciones de revista.

- Maisonnave, M., Delbianco, F., Tohmé, F. A., and Maguitman, A. G. A flexible supervised term-weighting technique and its application to variable extraction and information retrieval. *Inteligencia Artificial 22*, 63 (Feb. 2019), 61–80.
- Maisonnave, M., Delbianco, F., Tohmé, F., and Maguitman, A. Assessing the behavior and performance of a supervised term-weighting technique for topic-based retrieval. *Information Processing & Management 58*, 3 (2021), 102483.

Manuscritos en Revisión

La definición de la tarea de detección de eventos en curso usada en el Capítulo 3, fue presenta en un manuscrito que se encuentra en revisión. En el mismo, se presenta un análisis de diferentes tipos de modelos predictivos para la tarea, y se evalúa el uso de distintos tipos de arquitecturas y atributos. Se presentan modelos de referencia (*baseline*) adaptados de tareas similares y se introduce un modelo superador con respecto a dichos modelos de referencia.

- Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A., and Milios, E. Detecting ongoing events using contextual word and sentence embeddings. *arXiv preprint arXiv:2007.01379* (2021).

Trabajos en Conferencias

Dos trabajos y una comunicación fueron presentados en conferencias durante el desarrollo de la presente tesis. El primer trabajo involucrando la técnica FDD_β fue presentado en el XIX Simposio Argentino de Inteligencia Artificial en el año 2018. Una estrategia alternativa de detección de eventos se presentó como comunicación en el congreso *Dr. Antonio Monteiro XV* en el año 2019. Por último, una versión preliminar del *framework* fue presentada en el simposio *Document Engineering* en el año 2020.

- Maisonnave, M., Delbianco, F., Tohmé, F. A., and Maguitman, A. G. A supervised term-weighting method and its application to variable extraction from digital media. In *XIX Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 47 (CABA, 2018)* (2018), p. 40–53.
- Maisonnave, M. Detección de textos similares a través de una técnica de agrupamiento basada en densidad. In *Communication at the XV Dr. Antonio Monteiro Congress, Bahía Blanca, Argentina* (2019).
- Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A. G., and Milios, E. E. Assessing causality structures learned from digital text media. In *Proceedings of the ACM Symposium on Document Engineering 2020* (New York, NY, USA, 2020), DocEng '20, Association for Computing Machinery.

Conjuntos de Datos

Para los experimentos realizados durante esta tesis se diseñaron y construyeron dos conjuntos de datos que se encuentran disponibles en la plataforma *Mendeley Data*. El primero de ellos es "*Economic relevant news from the Guardian*" [1], que contiene 1.789 textos de artículos periodísticos del portal de noticias británico *The Guardian* recuperados a través de su API. Cada uno de los textos de este conjunto fue analizado manualmente y clasificado en dos posibles categorías: (i) relevante para la economía o (ii) irrelevante para la economía. Dicho conjunto se utilizó para evaluar varios aspectos de la técnica FDD_β, entre ellos, su desempeño para estimar la relevancia de términos para tópicos (siendo en

[1] https://cs.uns.edu.ar/~mmaisonnave/resources/ERN_data/

este caso el tópico el dominio económico). Este conjunto de datos se publicó en 2019 en el repositorio de *Mendeley Data*. El segundo conjunto de datos se denomina "Event detection dataset" [2] y consiste de 2.200 fragmentos de texto (aproximadamente una oración cada uno). Cada fragmento de texto fue manualmente analizado en búsqueda de *event-triggers* de eventos en curso. Cada token de cada fragmento fue anotado como *event-trigger* o como *non-event-trigger*. Este segundo conjunto de datos fue utilizado para evaluar el desempeño de los diferentes modelos de detección de eventos en curso presentados en el Capítulo 3 y fue publicado en el repositorio de *Mendeley Data* en el año 2020. Información detallada sobre la construcción de ambos conjuntos de datos se puede encontrar *Mendeley* o en las páginas oficiales de dichos conjuntos.

- Maisonnave, M., Delbianco, F., Tohmé, F., and Maguitman, A. Economic relevant news from the guardian. Mendeley Data, V3 - http://dx.doi.org/10.17632/yt8j2f3hpp.3, 2019.
- Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A., and Milios, E. Event detection dataset. Mendeley Data, V1 - http://dx.doi.org/10.17632/7d54rvzxkr.1, 2020.

[2] https://cs.uns.edu.ar/∼mmaisonnave/resources/ED_data/

Apéndice A

Apéndice del Capítulo: Selección de Variables

Resumen

Para profundizar en el análisis del parámetro β en la técnica FDD_β, se analizó qué términos (unigramas, bigramas o trigramas) son los que obtienen el mayor F_1 en el conjunto de entrenamiento para cada una de las 20 categorías del conjunto de datos *20NG*. Las tablas del presente Apéndice contienen los mejores términos seleccionados usando el mejor FDD_β para diferentes rangos de β y su desempeño en términos de precisión, cobertura y F_1 en las particiones de entrenamiento y test del conjunto de datos *20NG*.

Término	je suis autre	Malcolm Lee	Private activity	Kent	Christian	write	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,3]$	$[0,4;\ 0,5]$	$[0,6;\ 0,7]$	$[0,8;\ 2,6]$	$[2,7;\ 5,8]$	$[5,9;\ 9,9]$
Cobertura	$0,02/0,03$	$0,06/0,06$	$0,09/0,06$	$0,12/0,07$	$0,27/0,29$	$0,81/0,70$	$0,97/0,98$
Precisión	$1,00/1,00$	$1,00/1,00$	$0,67/0,62$	$0,44/0,45$	$0,20/0,25$	$0,04/0,05$	$0,03/0,04$
F_1	$0,05/0,05$	$0,12/0,12$	$0,16/0,10$	$0,19/0,12$	$0,23/0,27$	$0,08/0,09$	$0,07/0,08$

talk.religion.misc

Término	hide line	polygon	animation	graphic	image	program	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,1]$	$[0,2;\ 0,3]$	$[0,4;\ 0,7]$	$[0,8;\ 2,0]$	$[2,1;\ 2,2]$	$[2,3;\ 9,9]$
Cobertura	$0,02/0,00$	$0,05/0,06$	$0,08/0,06$	$0,18/0,17$	$0,21/0,20$	$0,25/0,15$	$0,89/0,84$
Precisión	$1,00/0,00$	$0,93/0,86$	$0,81/0,59$	$0,45/0,51$	$0,37/0,34$	$0,18/0,12$	$0,05/0,05$
F_1	$0,04/0,00$	$0,10/0,11$	$0,14/0,11$	$0,26/0,25$	$0,27/0,25$	$0,21/0,13$	$0,09/0,10$

comp.graphics

Término	strong crypto	encryption	key	to	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,9]$	$[1,0;\ 3,4]$	$[3,5;\ 4,5]$	$[4,6;\ 9,9]$
Cobertura	$0,02/0,02$	$0,30/0,27$	$0,42/0,43$	$0,95/0,91$	$0,98/0,96$
Precisión	$1,00/1,00$	$0,97/0,90$	$0,52/0,55$	$0,06/0,06$	$0,05/0,06$
F_1	$0,04/0,04$	$0,45/0,41$	$0,46/0,48$	$0,11/0,11$	$0,10/0,11$

sci.crypt

Término	Mario Lemieux	NHL	hockey	team	game	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,2]$	$[0,3;\ 0,6]$	$[0,7;\ 1,2]$	$[1,3;\ 4,5]$	$[4,6;\ 9,9]$
Cobertura	$0,02/0,02$	$0,19/0,23$	$0,24/0,24$	$0,44/0,43$	$0,51/0,49$	$0,89/0,88$
Precisión	$1,00/1,00$	$0,99/1,00$	$0,95/0,98$	$0,54/0,49$	$0,42/0,38$	$0,05/0,05$
F_1	$0,03/0,03$	$0,32/0,38$	$0,38/0,39$	$0,48/0,46$	$0,46/0,43$	$0,10/0,09$

rec.sport.hockey

Término	New Mutants	well offer	sale	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,3]$	$[0,4;\ 3,1]$	$[3,2;\ 9,9]$
Cobertura	$0,02/0,02$	$0,07/0,07$	$0,27/0,28$	$0,71/0,69$
Precisión	$1,00/1,00$	$0,90/0,93$	$0,50/0,53$	$0,04/0,04$
F_1	$0,03/0,03$	$0,14/0,13$	$0,35/0,37$	$0,07/0,08$

misc.forsale

Tabla A.1: Mejores términos aprendidos del conjunto de entrenamiento para diferentes rangos de β y su desempeño como consultas en los conjuntos de entrenamiento/test (Conjunto de datos *20NG* - Parte I).

Término	call MS	IDE	controller	card	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,4]$	$[0,5;\ 0,7]$	$[0,8;\ 2,5]$	$[2,6;\ 9,9]$
Cobertura	$0,02/0,02$	$0,11/0,12$	$0,16/0,19$	$0,29/0,32$	$0,93/0,90$
Precisión	$1,00/1,00$	$0,84/0,73$	$0,60/0,69$	$0,30/0,34$	$0,05/0,05$
F_1	$0,05/0,05$	$0,20/0,20$	$0,25/0,30$	$0,30/0,33$	$0,10/0,10$

comp.sys.ibm.pc.hardware

Término	FirstClass	IIsi	Apple	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,2]$	$[0,3;\ 2,4]$	$[2,5;\ 9,9]$
Cobertura	$0,02/0,03$	$0,07/0,07$	$0,25/0,26$	$0,93/0,87$
Precisión	$1,00/1,00$	$0,97/1,00$	$0,68/0,63$	$0,05/0,05$
F_1	$0,03/0,06$	$0,14/0,13$	$0,36/0,37$	$0,10/0,09$

comp.sys.mac.hardware

Término	Benedikt Rosenau	Gregg Jaeger	Jon Livesey	atheist	say	write	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,1]$	$[0,2;\ 0,2]$	$[0,3;\ 1,9]$	$[2,0;\ 2,1]$	$[2,2;\ 8,1]$	$[8,2;\ 9,9]$
Cobertura	$0,03/0,03$	$0,06/0,08$	$0,09/0,10$	$0,21/0,19$	$0,55/0,53$	$0,89/0,86$	$0,98/0,99$
Precisión	$1,00/1,00$	$1,00/1,00$	$0,89/0,94$	$0,63/0,64$	$0,08/0,08$	$0,07/0,06$	$0,05/0,04$
F_1	$0,06/0,06$	$0,11/0,15$	$0,16/0,18$	$0,32/0,29$	$0,15/0,14$	$0,12/0,11$	$0,09/0,08$

alt.atheism

Término	Mark Ira Kaufman	Israeli	Israel	write	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,5]$	$[0,6;\ 2,9]$	$[3,0;\ 3,9]$	$[4,0;\ 9,9]$
Cobertura	$0,02/0,03$	$0,27/0,26$	$0,34/0,34$	$0,78/0,78$	$0,97/0,99$
Precisión	$1,00/1,00$	$0,97/0,96$	$0,83/0,84$	$0,06/0,07$	$0,05/0,06$
F_1	$0,03/0,07$	$0,43/0,41$	$0,48/0,48$	$0,12/0,13$	$0,10/0,11$

talk.politics.mideast

Término	Ford Probe	Mustang	car	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,1]$	$[0,2;\ 5,5]$	$[5,6;\ 9,9]$
Cobertura	$0,02/0,03$	$0,05/0,03$	$0,61/0,61$	$0,94/0,96$
Precisión	$1,00/1,00$	$0,95/0,86$	$0,59/0,63$	$0,05/0,06$
F_1	$0,03/0,06$	$0,10/0,06$	$0,60/0,62$	$0,10/0,11$

rec.autos

Tabla A.2: Mejores términos aprendidos del conjunto de entrenamiento para diferentes rangos de β y su desempeño como consultas en los conjuntos de entrenamiento/test (Conjunto de datos *20NG* - Parte II).

Término	Ryan C Scharfy	Clayton Cramer	government	people	article	write	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,9]$	$[1,0;\ 1,6]$	$[1,7;\ 3,2]$	$[3,3;\ 3,8]$	$[3,9;\ 4,9]$	$[5,0;\ 9,9]$
Cobertura	$0,02/0,03$	$0,12/0,14$	$0,29/0,23$	$0,52/0,49$	$0,71/0,72$	$0,80/0,77$	$0,97/0,97$
Precisión	$1,00/1,00$	$1,00/0,95$	$0,19/0,14$	$0,10/0,09$	$0,07/0,06$	$0,06/0,05$	$0,04/0,04$
F_1	$0,04/0,05$	$0,21/0,25$	$0,23/0,17$	$0,17/0,16$	$0,12/0,11$	$0,11/0,10$	$0,08/0,08$

talk.politics.misc

Término	wood stave inside	firearm	gun	not	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,4]$	$[0,5;\ 3,1]$	$[3,2;\ 3,2]$	$[3,3;\ 9,9]$
Cobertura	$0,02/0,01$	$0,17/0,17$	$0,37/0,32$	$0,81/0,80$	$0,96/0,97$
Precisión	$1,00/1,00$	$0,91/0,81$	$0,57/0,47$	$0,06/0,05$	$0,05/0,04$
F_1	$0,04/0,03$	$0,29/0,27$	$0,45/0,38$	$0,11/0,09$	$0,10/0,08$

talk.politics.guns

Término	expose event	window manager	Motif	X	to	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,1]$	$[0,2;\ 0,4]$	$[0,5;\ 3,2]$	$[3,3;\ 8,9]$	$[9,0;\ 9,9]$
Cobertura	$0,02/0,02$	$0,07/0,10$	$0,14/0,16$	$0,35/0,37$	$0,87/0,87$	$0,88/0,87$
Precisión	$1,00/1,00$	$0,98/0,96$	$0,83/0,88$	$0,47/0,52$	$0,05/0,06$	$0,05/0,06$
F_1	$0,04/0,04$	$0,14/0,18$	$0,25/0,27$	$0,40/0,43$	$0,09/0,11$	$0,09/0,11$

comp.windows.x

Término	Chuck Rogers	bike	DoD	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 1,0]$	$[1,1;\ 3,8]$	$[3,9;\ 9,9]$
Cobertura	$0,02/0,02$	$0,42/0,45$	$0,44/0,41$	$0,93/0,93$
Precisión	$1,00/1,00$	$0,95/0,98$	$0,84/0,84$	$0,05/0,05$
F_1	$0,03/0,03$	$0,58/0,61$	$0,58/0,55$	$0,10/0,10$

rec.motorcycles

Término	Internal Medicine	Gordon Banks	doctor	to	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,3]$	$[0,4;\ 2,0]$	$[2,1;\ 2,3]$	$[2,4;\ 9,9]$
Cobertura	$0,03/0,01$	$0,12/0,07$	$0,19/0,20$	$0,91/0,92$	$0,95/0,95$
Precisión	$1,00/1,00$	$0,97/1,00$	$0,69/0,66$	$0,05/0,06$	$0,05/0,06$
F_1	$0,05/0,02$	$0,22/0,13$	$0,30/0,31$	$0,10/0,11$	$0,10/0,11$

sci.med

Tabla A.3: Mejores términos aprendidos del conjunto de entrenamiento para diferentes rangos de β y su desempeño como consultas en los conjuntos de entrenamiento/test (Conjunto de datos *20NG* - Parte III).

Term	File Manager	late driver	Windows	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,1]$	$[0,2;\ 4,2]$	$[4,3;\ 9,9]$
Cobertura	$0,02/0,02$	$0,03/0,03$	$0,48/0,42$	$0,90/0,85$
Precisión	$0,93/1,00$	$0,83/1,00$	$0,59/0,54$	$0,05/0,05$
F_1	$0,03/0,04$	$0,05/0,05$	$0,53/0,47$	$0,10/0,09$

comp.os.ms-windows.misc

Término	Andy Byler	James Kiefer	Christ	God	to	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,1]$	$[0,2;\ 0,5]$	$[0,6;\ 4,2]$	$[4,3;\ 4,3]$	$[4,4;\ 9,9]$
Cobertura	$0,02/0,02$	$0,04/0,03$	$0,30/0,24$	$0,53/0,49$	$0,94/0,97$	$0,97/0,96$
Precisión	$1,00/1,00$	$1,00/1,00$	$0,67/0,57$	$0,50/0,43$	$0,06/0,06$	$0,06/0,05$
F_1	$0,05/0,04$	$0,07/0,05$	$0,41/0,34$	$0,51/0,46$	$0,11/0,11$	$0,11/0,10$

soc.religion.christian

Término	Michael Adams	orbit	space	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,8]$	$[0,9;\ 2,6]$	$[2,7;\ 9,9]$
Cobertura	$0,04/0,06$	$0,20/0,20$	$0,30/0,29$	$0,96/0,97$
Precisión	$1,00/1,00$	$0,93/0,90$	$0,43/0,41$	$0,05/0,05$
F_1	$0,08/0,11$	$0,33/0,33$	$0,36/0,34$	$0,10/0,10$

sci.space

Término	acid battery	circuit	use	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 1,5]$	$[1,6;\ 2,7]$	$[2,8;\ 9,9]$
Cobertura	$0,02/0,01$	$0,15/0,12$	$0,51/0,51$	$0,93/0,95$
Precisión	$1,00/1,00$	$0,76/0,75$	$0,08/0,08$	$0,05/0,05$
F_1	$0,04/0,02$	$0,25/0,20$	$0,14/0,13$	$0,10/0,09$

sci.electronics

Término	bat average	pitcher	pitch	game	be
Rango de β	$[0,0;\ 0,0]$	$[0,1;\ 0,4]$	$[0,5;\ 1,0]$	$[1,1;\ 3,2]$	$[3,3;\ 9,9]$
Cobertura	$0,02/0,01$	$0,18/0,15$	$0,23/0,18$	$0,38/0,45$	$0,92/0,93$
Precisión	$1,00/1,00$	$0,98/1,00$	$0,77/0,90$	$0,31/0,36$	$0,05/0,05$
F_1	$0,04/0,02$	$0,30/0,26$	$0,35/0,31$	$0,34/0,40$	$0,10/0,10$

rec.sport.baseball

Tabla A.4: Mejores términos aprendidos del conjunto de entrenamiento para diferentes rangos de β y su desempeño como consultas en los conjuntos de entrenamiento/test (Conjunto de datos *20NG* - Parte IV).

Apéndice B

Apéndice del Capítulo: Detección de Eventos en Curso

B.1. Estudios Preliminares para los Modelos *Baseline*

B.1.1. Resultados

En esta sección, se describen los estudios preliminares para el modelo *baseline* basado en CNN. Basándose en estos estudios, se eligieron las cuatro variantes finales usadas en el conjunto reservado para test (conjunto no usado durante los estudios preliminares). Se utilizaron estos estudios preliminares para determinar el mejor tamaño de ventana para ser usado en el modelo *baseline*.

Para cada variante de cada modelo, se entrenó el modelo hasta que no se vieron mejoras en el desempeño de la métrica F1-score sobre el conjunto de entrenamiento durante 200 épocas seguidas (*early stopping* con coeficiente de paciencia 200). Por cada uno de los modelos del estudio preliminar se realizaron cinco repeticiones. Se usaron cinco semillas consecutivas de 1 a 5 para cada modelo para garantizar la reproducibilidad. Para cada modelo se reportaron las métricas promedio de las cinco repeticiones. Las métricas usadas son sensibilidad, especificidad y la media armónica de estas dos (F1-score) en el conjunto de entrenamiento y validación. No se reporta la exactitud del modelo (*accuracy*), ya que, al tratarse de un conjunto de datos altamente desbalanceado ($93,41\,\%$ de los *tokens* son

no eventos), la exactitud es una métrica que tiene una interpretación no tan directa y puede derivar en conclusiones erróneas. Para analizar en detalle el resultado de la métrica F1-score en los conjuntos de validación, se computaron dos métricas adicionales. Primero, se calculó el intervalo de confianza (IC) usando un nivel de confianza de 95 %. Segundo, se computó el p-valor de un t-test entre cada modelo de cada tabla contra el mejor modelo de dicha tabla. Se seleccionó el mejor modelo en términos del valor del F1-score en el conjunto de validación. Por ejemplo, el p-valor de $0,410$ en la cuarta fila de la Tabla B.1 indica que no hay suficiente evidencia para rechazar la hipótesis nula de que el modelo 4 tiene un F1-score estadísticamente diferente al del mejor modelo de la tabla (modelo 1) en el conjunto de validación.

Se realizaron experimentos usando seis tamaños de ventana diferentes para el modelo *baseline* CNN. Aunque en el trabajo original los autores usaron un tamaño de ventana de 31, en este trabajo se debieron realizar experimentos para determinar si este tamaño de ventana es el más apropiado para la tarea que se estaba resolviendo. La necesidad de explorar tamaños de ventana alternativos (en particular, más pequeños) se debió a que el conjunto de datos usado es diferente al del trabajo original. En dicho trabajo se usaban artículos completos, mientras que en el presente trabajo se usaron fragmentos de textos (oraciones). Por ende, cada ítem de dato en la tarea aquí desarrollada era considerablemente más pequeño. Por esto es que se exploraron tamaños de ventana más pequeños, así evitando tener que usar excesivo relleno en cada ítem de dato (*padding*). Por ejemplo, en una oración de tamaño 31, solo el *token* del medio no necesitaría *padding*, siendo necesario para todos los otros *tokens*. Esta cantidad excesiva de *padding* agregaría ruido al modelo e incrementaría su costo computacional con poca ganancia en desempeño (ya que el *padding* no agrega información útil). Dada esta situación se decidió realizar experimentos preliminares para explorar diferentes tamaños de ventana posible, siempre considerando tamaños menores al usado en el trabajo original.

En la Tabla B.1, se presentan los resultados de los seis modelos analizados para los seis tamaños de ventana estudiados (1, 3, 5, 11, 21 y 31). Dado que el *embedding* de posición es usado para representar la posición relativa del *token* dentro de la ventana, para el caso de tamaño de ventana 1 no es necesario representar esta información, por lo que el *embedding* de posición es excluido de este modelo (modelo 1). Excepto para el caso del *embedding* de posición del modelo 1, todos los modelos incluyen todos los demás atributos considerados para los modelos *baseline*.

B.1.2. Discusión

Se derivan las siguientes conclusiones de estos experimentos preliminares. Como se mencionó antes, se plantea la hipótesis de que ventanas largas no son adecuadas para este contexto donde los ítems de datos son oraciones y no textos completos. Se encuentra evidencia para apoyar esta hipótesis a partir de los resultados de los modelos 5 y 6. Para estos modelos con ventanas grandes, se obtiene una especificidad de $1,0$ y una sensibilidad cercana a cero y por ende el F1-score cae a casi cero. Por esta razón no se eligieron estos tamaños de ventanas para los experimentos finales sobre el conjunto de datos reservado para el test.

El modelo con tamaño de ventana 1 (modelo 1), alcanza el mejor desempeño con un F1-score promedio de $59,4\,\%$ en el conjunto de validación. Por ende, se selecciona este modelo para ser usado en el conjunto reservado para test. También se selecciona el modelo 4 (aunque no sea el mejor) por su alto desempeño en el conjunto de validación. Más aún, el test estadístico da como resultado evidencia insuficiente para rechazar la hipótesis nula (p-valor $> 0,05$) de que el modelo 1 tiene un desempeño (en términos de F1-score) estadísticamente diferente a este modelo. Aunque los modelos 2 y 3 tienen un desempeño relativamente bueno ($56,2\,\%$ y $55,8\,\%$, respectivamente), el análisis estadístico muestra que se puede rechazar la hipótesis nula. Por lo tanto, se tiene evidencia para considerar que el modelo 1 es estadísticamente mejor que los modelos 2 y 3.

En resumen, el mejor modelo de estos experimentos preliminares es el modelo 1, con una ventana de tamaño 1. El modelo 4, con una ventana tamaño 11, tiene un desempeño comparable y el análisis estadístico mostró que no hay una diferencia significativa de desempeño en términos del F1-score en la validación. Por ende, se eligieron estos dos tamaños de ventana para realizar experimentos sobre el conjunto de datos reservado para el test.

B.2. Estudios Preliminares para los Modelos Propuestos (RNN)

En esta sección, se describen los estudios preliminares llevados a cabo para evaluar diferentes variantes del modelo propuesto (RNN). A partir de estos experimentos, se

Experimentos con modelos CNN sobre el conjunto de validación: Evaluando el Tamaño de la Ventana										
Modelo	Tam. Ventana	Atributos	entrenamiento			validación				
			$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	F1 IC	p-valor
1	1	$\{W,E\}$	$0,608$	$0,980$	$0,692$	$0,502$	$0,970$	**0,594**	$\pm\, 0,019$	—
2	3	$\{W,E,Po\}$	$0,749$	$0,986$	$0,808$	$0,466$	$0,975$	$0,562$	$\pm\, 0,015$	$0,003$
3	5	$\{W,E,Po\}$	$0,816$	$0,988$	$0,858$	$0,464$	$0,975$	$0,558$	$\pm\, 0,023$	$0,006$
4	11	$\{W,E,Po\}$	$0,811$	$0,979$	$0,852$	$0,500$	$0,965$	$0,590$	$\pm\, 0,043$	$0,410$
5	21	$\{W,E,Po\}$	$0,015$	$1,000$	$0,022$	$0,002$	$0,999$	$0,002$	$\pm\, 0,002$	$0,000$
6	31	$\{W,E,Po\}$	$0,001$	$1,000$	$0,001$	$0,001$	$1,000$	$0,001$	$\pm\, 0,002$	$0,000$

Tabla B.1: Desempeño promedio del modelo CNN para seis tamaños de ventana diferente. Para cada tamaño de ventana se hicieron cinco repeticiones y se reporta la sensibilidad (sens), la especificidad (spec) y el F1-score (F1) promedio obtenidos en los conjuntos de entrenamiento y validación. Para el conjunto de validación, también se reporta el intervalo de confianza con un nivel de significancia de 95 % para el F1-score. También para cada modelo de cada fila se reporta el p-valor de un t-test entre el F1-score en la validación de ese modelo y el mejor modelo de la tabla para la misma métrica. Un p-valor bajo provee evidencia para rechazar la hipótesis de que ambos modelos tienen el mismo F1-score, indicando que el mejor modelo es estadísticamente distinto al modelo bajo comparación.

En estos experimentos, se probaron seis tamaños de ventana diferentes para determinar cuál es el mejor para usar sobre el conjunto reservado para el test. El modelo *baseline* obtuvo el mejor desempeño con tamaño de ventana 1. Se encontró evidencia que sugiere que ese modelo es estadísticamente mejor que los modelos con tamaño de ventana 3, 5, 21 y 31. Aunque el modelo con tamaño de ventana 1 obtuvo mejores resultados que el modelo con tamaño 11, no se pudo encontrar evidencia estadística que sugiera que estos modelos son distintos. Por ende, se usan estos dos tamaños de ventana (1 y 11) para el conjunto de datos reservado para test.

eligieron siete variantes finales para ser usadas sobre el conjunto reservado para test. Se condujeron dos estudios preliminares. Uno para determinar el mejor número de capas Bi-LSTM y unidades ocultas, y otro para estudiar los atributos usados.

Para cada variante de cada modelo, se entrenó el modelo hasta que no se vieron mejoras en el desempeño de la métrica F1-score sobre el conjunto de entrenamiento durante 200 épocas seguidas (*early stopping* con coeficiente de paciencia 200). Para cada uno de los modelos en los estudios preliminares, se realizaron cinco repeticiones y se reporta el promedio de esas cinco repeticiones para cada modelo. Se utilizaron cinco semillas consecutivas de 1 al 5 para garantizar la reproducibilidad. Se reporta para cada modelo el promedio de las métricas: sensibilidad, especificidad y la media armónica de estas dos (F1-score); todas las métricas son calculadas tanto sobre el conjunto de entrenamiento como el de validación. Se excluye de este análisis la métrica exactitud (*accuracy*), ya que, al tratarse de un conjunto de datos altamente desbalanceado la exactitud es una métrica que tiene una interpretación no tan directa y puede derivar en conclusiones erróneas. Para analizar en profundidad la métrica F1-score en el conjunto de validación se computan dos valores adicionales sobre dicha métrica. Primero, se calculó el intervalo de confianza (IC) usando un nivel de confianza de 95 %. Segundo, para cada modelo de cada fila se reporta el p-valor de un t-test entre el F1-score en la validación de ese modelo y el mejor modelo de la tabla para la misma métrica. Por ejemplo, el p-valor de $0,160$ en la quinta columna de la Tabla B.2 indica que no hay suficiente evidencia para rechazar la hipótesis nula de que el F1-score del modelo de la quinta columna es estadísticamente diferente al F1-score del mejor modelo de la tabla (modelo 6). Ambos F1-score fueron calculados sobre el conjunto de validación. Se define al mejor modelo en términos de la métrica F1-score obtenida en el conjunto de validación.

B.2.1. Resultados del Primer Experimento Preliminar sobre el Modelo Propuesto (RNN)

Se experimentó con el modelo propuesto usando ocho arquitecturas diferentes (variando el número de capas Bi-LSTM y unidades ocultas). Se utilizó siempre una cantidad de unidades ocultas en orden descendiente, con la primera capa teniendo la mayor cantidad y la última la menor cantidad de unidades ocultas. Esta configuración sigue la intuición de que cada capa debería tomar la salida de la capa anterior y construir menos y más elabo-

radas representaciones de la salida con mayor nivel de abstracción. Durante la descripción de estos trabajos preliminares, se refiere a cada arquitectura como una lista ordenada de unidades ocultas, donde el primer número es la cantidad de unidades ocultas de la primera capa Bi-LSTM, el segundo número la cantidad de unidades ocultas de la segunda capa Bi-LSTM, y así siguiendo. Por ejemplo, la arquitectura ⟨100,15,5⟩ es una arquitectura de tres capas Bi-LSTM, con 100 unidades ocultas en la primera capa, 15 en la siguiente y 5 en la última. Excepto por las capas Bi-LSTM el resto de la estructura de la red es idéntica para los ocho modelos.

En la Tabla B.2, se presenta el resultado del modelo propuesto para ocho configuraciones de capas Bi-LSTM con las cuales se experimentó. El primer modelo tiene tres capas Bi-LSTM con 100, 15 y 5, en la primera, segunda y tercera capa, respectivamente. El segundo y tercer modelo tienen dos capas Bi-LSTM cada uno, con diferente cantidad de unidades ocultas (el segundo modelo tiene 15 y 5 unidades ocultas, mientras que el tercero tiene 5 y 2). Los restantes cinco modelos son todos de una sola capa, con diferente número de unidades ocultas. Se presentan los modelos del más complejo al más simple. Los modelos 4 a 8 tienen 200, 50, 15, 7 y 1 unidades ocultas, respectivamente. Como se mencionó previamente, el resto de la arquitectura del modelo es igual para las ocho variantes.

B.2.2. Discusión

Los resultados obtenidos por los modelos multicapa 1 y 2 son similares ($69,2\,\%$ y $68,2\,\%$, respectivamente) y considerablemente peores que los obtenidos por el modelo 6 ($72,9\,\%$), el cual es un modelo considerablemente más simple. Los modelos 3 y 4 son los dos modelos con peor desempeño, con un F1-score en validación de $66,8\,\%$ y $66,0\,\%$, respectivamente. Estos resultados muestran la importancia de encontrar un buen balance entre complejidad y simplicidad. El modelo de 3 capas (modelo 1) obtiene buen desempeño, pero no tan bueno como el alcanzado por los modelos de una sola capa (modelos 5 a 8), que son considerablemente más simples. Sin embargo, modelos demasiado simples, como el modelo 3, que tiene un pequeño número de unidades ocultas también tiene un desempeño pobre.

El mejor modelo es el modelo 6, el cual es un modelo de una sola capa con quince unidades ocultas, alcanzando un buen balance entre simplicidad y complejidad. Más aún,

los cuatro mejores modelos son todos modelos de una sola capa, con 50 unidades ocultas o menos (50, 15, 7 y 1). De estos resultados obtenidos por estos modelos, se observa una relación inversa entre complejidad y desempeño. Modelos más simples obtienen mejores desempeños. Estos resultados confirman la intuición de que modelos muy grandes y complejos requieren un conjunto de datos muy grande para aprender relaciones de alto nivel y evitar el sobreajuste (*overfitting*). Por eso, en el contexto de ED y OED, donde los conjuntos de datos son del orden de los cientos de documentos, los sistemas complejos tienden a tener un mal desempeño. Siguiendo esta intuición, que está respaldada por los resultados obtenidos, se elige la arquitectura del modelo 6 para el segundo estudio preliminar y para los experimentos sobre el conjunto reservado para test.

B.2.3. Resultados del Segundo Experimento Preliminar sobre el Modelo Propuesto (RNN)

Se realizó un segundo experimento preliminar para estudiar el desempeño del modelo propuesto al variar los atributos incluidos en el modelo. Estos experimentos se realizaron de una manera similar a un estudio de ablación. Primero se midió el desempeño del modelo con todos los atributos, y después por cada modelo o hipótesis que se quiso probar, se creó un modelo con diferente cantidad de atributos y se comparó su desempeño con el modelo que tenía todos los atributos.

El objetivo de este estudio preliminar fue el de evaluar el impacto de cada atributo en el desempeño general del modelo. Esto permitió eliminar del modelo aquellos atributos que no estuvieran aportando a la tarea, los cuales potencialmente estaban agregando ruido y perjudicando el desempeño. Se realizaron cinco repeticiones por cada conjunto de atributos usado utilizando la mejor arquitectura encontrada durante el estudio preliminar anterior. El resultado de estos experimentos se presenta en la Tabla B.3, donde la primera fila reporta el desempeño del modelo con todos los atributos, seguido por el mismo modelo con uno o más atributos eliminados.

Se probaron nueve diferentes configuraciones de atributos de entrada, incluyendo el modelo que contiene todos los atributos (*all*) (modelo 1). Para evaluar el impacto de los *embeddings* contextuales se evaluaron tres modelos diferentes (modelos 2 a 4). El modelo 2 incluye todos los atributos excepto por los *embeddings* contextuales de palabra (*all*-$\{B\}$). El modelo 3 incluye todos los atributos menos los *embeddings* contextuales de oración *all*-

Modelos propuesto RNN sobre el conjunto de validación: Evaluando Diferentes Arquitecturas											
Modelo	Arquitectura	Atributos	entrenamiento			validación					
			$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	F1 IC	p-valor
1	⟨100,15,5⟩	*all*	0,722	0,988	0,742	0,654	0,962	0,692	± 0,037	0,026
2	⟨15,5⟩	*all*	0,732	0,984	0,748	0,648	0,958	0,682	± 0,019	0,002
3	⟨5,2⟩	*all*	0,723	0,981	0,741	0,628	0,964	0,668	± 0,053	0,015
4	⟨200⟩	*all*	0,741	0,982	0,758	0,630	0,956	0,660	± 0,043	0,004
5	⟨50⟩	*all*	0,729	0,982	0,747	0,685	0,950	0,704	± 0,059	0,160
6	⟨15⟩	*all*	0,757	0,977	0,765	0,709	0,952	**0,729**	± 0,026	—
7	⟨7⟩	*all*	0,749	0,983	0,763	0,681	0,948	0,698	± 0,039	0,052
8	⟨1⟩	*all*	0,763	0,977	0,771	0,694	0,945	0,708	± 0,026	0,070

Tabla B.2: Desempeño promedio del modelo propuesto (RNN) para las ocho diferentes arquitecturas. La arquitectura es representada con una lista que representa el número de unidades ocultas en cada capa. La arquitectura ⟨15,5⟩ es una red con dos capas Bi-LSTM con 15 y 5 unidades ocultas en la primera y segunda capa, respectivamente. Las capas previas y posteriores a las de Bi-LSTM son fijas para las ocho variantes del modelo propuesto. Por cada arquitectura, se realizaron cinco repeticiones y se reportan los promedios de las métricas sensibilidad (sens), especificidad (spec) y la media armónica de ambas (F1-score (F1)). Todas las métricas son reportadas tanto sobre los conjuntos de entrenamiento como de validación. Para la métrica F1-score en el conjunto de validación también se reporta el intervalo de confianza (IC) con un nivel de significancia de 95 %. También se reporta el p-valor de un t-test a una sola cola de cada modelo de cada fila contra el mejor modelo de la tabla. El test se realiza entre los valores de F1-score de los dos modelos sobre el conjunto de validación. Un p-valor bajo provee evidencia para rechazar la hipótesis de que los modelos tienen el mismo F1-score, indicando que el mejor modelo de la tabla es estadísticamente mejor. En estos experimentos, se probaron ocho arquitecturas diferentes (cambiando el número de capas y de unidades ocultas) para el modelo propuesto (RNN).

$\{S\}$), que son los atributos construidos sumando los *embeddings* contextuales de palabra (B) para toda la oración. Finalmente, el modelo 4 tiene tanto los *embeddings* contextuales de oración como los de palabra eliminados (*all*-$\{B,S\}$).

También se evaluaron dos modelos diferentes para medir el impacto del uso de la información gramatical. Primero, el modelo 5, el cual tenía la información del *Part-Of-Speech tagger* eliminada, tanto la versión simplificada como la versión detallada (*all*-$\{P,T\}$). Segundo, el modelo 6, donde se elimina la información del *dependency parser tagger* (*all*-$\{T\}$).

Finalmente, se evaluaron 3 modelos adicionales para evaluar el impacto de los tres atributos faltantes: *embeddings* de entidad (E), *embeddings* no contextuales de palabra *Word2vec* (W), y *embeddings* contextuales de palabra de *spaCy* (Sp). Estos modelos son los modelos 7, 8 y 9, respectivamente.

B.2.4. Discusión

Se observa que el desempeño cae considerablemente (caída de $15,1$ puntos porcentuales en F1-score) para el modelo que no contiene los *embeddings* contextuales basados en *BERT* (modelo 2) en comparación con el modelo con todos los atributos (modelo 1). De manera similar, el modelo sin los *embeddings* contextuales de oración basados en *BERT* (modelo 3) tiene una caída de $10,3$ puntos porcentuales. Por otro lado, el desempeño del modelo sin ambos atributos (modelo 4) tiene un desempeño similar al modelo donde solo se eliminan los *embeddings* contextuales de palabra (modelo 3). Un análisis estadístico muestra que, para los tres modelos, la hipótesis de que tienen el mismo desempeño que el mejor modelo en términos de F1-score en la validación, puede ser rechazada con un nivel de confianza del 95 %. Los dos análisis individuales de *embeddings* contextuales basados en *BERT* muestran un impacto significativo en el desempeño e indican un impacto más grande para los *embeddings* contextuales de palabras por sobre los *embeddings* de oración.

Los resultados para el modelo sin la información del *dependency parser tagger* (D) (modelo 6) son similares a los obtenidos por el modelo con todos los atributos (modelo 1). Estos resultados indican que el modelo no logra aprovechar ese atributo para mejorar el desempeño. Al eliminar la información del *Part-Of-Speech tagger* (T, P) (modelo 5), se observó una pequeña baja en el desempeño (de $3,1$ puntos porcentuales). Este resultado sugiere que estos atributos pueden ser útiles para la tarea de OED. Se exploró más en

Modelos propuesto RNN sobre el conjunto de validación: Evaluando los Atributos de Entrada										
Modelos	Arquitectura	Atributos	entrenamiento			validación				
			$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	$\overline{sens}$	$\overline{spec}$	$\overline{F1}$	F1 IC	p-valor
1	$\langle 15 \rangle$	*all*	0,757	0,977	0,765	0,709	0,952	0,729	$\pm$ 0,026	0,326
2	$\langle 15 \rangle$	*all*-{B}	0,707	0,980	0,728	0,528	0,963	0,578	$\pm$ 0,072	0,001
3	$\langle 15 \rangle$	*all*-{S}	0,671	0,995	0,710	0,571	0,971	0,626	$\pm$ 0,061	0,003
4	$\langle 15 \rangle$	*all*-{B,S}	0,781	0,997	0,790	0,594	0,935	0,637	$\pm$ 0,023	0,000
5	$\langle 15 \rangle$	*all*-{T,P}	0,752	0,976	0,761	0,667	0,957	0,698	$\pm$ 0,037	0,026
6	$\langle 15 \rangle$	*all*-{D}	0,747	0,980	0,758	0,718	0,944	0,733	$\pm$ 0,017	0,430
7	$\langle 15 \rangle$	*all*-{E}	0,761	0,977	0,769	0,697	0,950	0,721	$\pm$ 0,034	0,182
8	$\langle 15 \rangle$	*all*-{W}	0,629	0,973	0,644	0,707	0,954	0,727	$\pm$ 0,049	0,341
9	$\langle 15 \rangle$	*all*-{Sp}	0,748	0,979	0,759	0,721	0,946	0,735	$\pm$ 0,018	—

Tabla B.3: Desempeño promedio del modelo RNN propuesto usando la mejor arquitectura (una capa Bi-LSTM con quince unidades ocultas) para nueve diferentes conjuntos de atributos de entrada. Por cada conjunto de atributos se realizaron cinco repeticiones y se reporta el promedio de esas cinco repeticiones para cada una de las métricas usadas. Las métricas usadas son sensibilidad (sens), especificidad (spec) y F1-score (F1), todas reportadas sobre el conjunto de entrenamiento y el de validación. Para el conjunto de validación también se reporta el intervalo de confianza con un nivel de significancia del 95 % para la métrica F1-score en la validación. También se reporta para cada modelo el p-valor de un t-test entre el F1-score en la validación de ese modelo contra el mejor modelo de la tabla. Un p-valor pequeño provee evidencia para rechazar la hipótesis de que los F1-scores de ambos modelos son iguales, indicando que el mejor modelo es estadísticamente mejor.

En estos experimentos, se examinaron nueve conjuntos de atributos de entrada. Se usó un modelo con todos los atributos como base para la comparación y ocho más que eliminan uno o dos atributos del total. Se eliminaron T y P juntos porque son semánticamente el mismo atributo (son la versión simplificada y detallada del mismo atributo). Por último, se eliminaron B y S juntos para medir el impacto de los *embeddings* contextuales basados en *BERT*. Los *embeddings* contextuales basados en *BERT* mostraron el impacto más significativo. Los resultados muestran que los atributos D, E, W y Sp tienen un impacto insignificante en el desempeño. Se puede concluir de estos resultados que en presencia de los otros atributos estos cuatro podrían ser eliminados sin afectar al desempeño. Por otro lado, los atributos T y P tienen un impacto pequeño, pero estadísticamente significativo cuando son eliminados.

detalle la inclusión o exclusión de la información gramatical (T, P y D) durante los experimentos en el conjunto de test.

Las filas 7 a 9 de la tabla B.3, correspondiente a los modelos 7 a 9, muestran los resultados del modelo que tiene todos los atributos menos los *embeddings* de entidades (E), los resultados del modelo con todos los atributos menos los *embeddings* de palabra *Word2vec* (W) y los resultados del modelo con todos los atributos menos los *embeddings* contextuales *spaCy* (Sp), respectivamente. El modelo 9 es el modelo con el mejor desempeño. Sin embargo, el p-valor alto muestra que no hay diferencia estadísticamente significativa entre este modelo y los otros dos. Más aún, no hay diferencia estadísticamente significativa entre el mejor modelo y el modelo con todos los atributos (modelo 1). Estos resultados indican que la eliminación de cualquiera de estos tres atributos (E, W o Sp) tiene un impacto insignificante en el desempeño en términos de F1-score en la validación. Se analiza en profundidad el efecto de incluir o excluir estos atributos en el conjunto de test. Sin embargo, a partir de estos resultados, se tiene evidencia que sugiere que en presencia de los otros atributos estos tres atributos pueden ser eliminados del modelo sin afectar el desempeño.

Bibliografía

[AdC03] Acid, S., and de Campos, L. M. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research 18*, 1 (May 2003), 445–490.

[AGZ+18] Abualsaud, M., Ghelani, N., Zhang, H., Smucker, M. D., Cormack, G. V., and Grossman, M. R. A system for efficient high-recall retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (New York, NY, USA, 2018), SIGIR '18, Association for Computing Machinery, p. 1317–1320.

[AH19] Alsmadi, I., and Hoon, G. K. Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications 31*, 8 (Aug 2019), 3819–3831.

[Ahe16] Ahelegbey, D. F. The econometrics of bayesian graphical models: a review with financial application. *Journal of Network Theory in Finance 2*, 2 (2016), 1–33.

[Ahn06] Ahn, D. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events* (USA, 2006), ARTE '06, Association for Computational Linguistics, p. 1–8.

[AIR96] Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*, 434 (1996), 444–455.

[AOGD+16] Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., and Gomes, J. B. A rule dynamics approach to event detection in twitter

with its application to sports and politics. *Expert Systems with Applications 55* (2016), 351–360.

[APL98] Allan, J., Papka, R., and Lavrenko, V. On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1998), SIGIR '98, Association for Computing Machinery, p. 37–45.

[AX20] Alshaher, H., and Xu, J. A new term weight scheme and ensemble technique for authorship identification. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis* (New York, NY, USA, 2020), ICCDA 2020, Association for Computing Machinery, p. 123–130.

[BBS09] Barnett, L., Barrett, A. B., and Seth, A. K. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett. 103* (Dec 2009), 238701.

[BCFS19] Balashankar, A., Chakraborty, S., Fraiberger, S., and Subramanian, L. Identifying predictive causal factors from news streams. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, November 2019), Association for Computational Linguistics, pp. 2338–2348.

[BDL+15] Bronstein, O., Dagan, I., Li, Q., Ji, H., and Frank, A. Seed-based event trigger labeling: How far can event descriptions get us? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 372–376.

[BGJM17] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics 5* (06 2017), 135–146.

[BHBL11] Bizer, C., Heath, T., and Berners-Lee, T. *Linked Data: The Story so Far*. Semantic Services, Interoperability and Web Applications: Emerging Concepts. IGI Global, Hershey, PA, USA, 2011, pp. 205–227.

[Bor18] Boroş, E. *Neural Methods for Event Extraction*. Theses, Université Paris Saclay (COmUE), September 2018.

[Car90] Card, D. The impact of the mariel boatlift on the miami labor market. *ILR Review 43*, 2 (1990), 245–257.

[CG09] Cecchini, Rocío L.Lorenzetti, C. M., and G., A. Evolving disjunctive and conjunctive topical queries based on multi-objective optimization criteria. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* (2009).

[CK93] Card, D., and Krueger, A. B. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania. Working Paper 4509, National Bureau of Economic Research, October 1993.

[CNL03] Chieu, H. L., Ng, H. T., and Lee, Y. K. Closing the gap: Learning-based information extraction rivaling knowledge-engineering methods. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (Sapporo, Japan, July 2003), Association for Computational Linguistics, pp. 216–223.

[CSG+08] Chiquet, J., Smith, A., Grasseau, G., Matias, C., and Ambroise, C. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics 25*, 3 (12 2008), 417–418.

[Cun21] Cunningham, S. *Causal Inference: The Mixtape*. Yale University Press, 2021.

[CXL+15] Chen, Y., Xu, L., Liu, K., Zeng, D., and Zhao, J. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 167–176.

[CY13] Cooper, G. F., and Yoo, C. Causal discovery from a mixture of experimental and observational data. *arXiv preprint arXiv:1301.6686* (2013).

[CZLZ16] Chen, K., Zhang, Z., Long, J., and Zhang, H. Turning from tf-idf to tf-igm for term weighting in text classification. *Expert Systems with Applications 66* (2016), 245–260.

[CZSL18] Chen, X., Zhou, X., Sellis, T., and Li, X. Social event detection with retweeting behavior correlation. *Expert Systems with Applications 114* (2018), 516–523.

[DCLT18] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[DGB+10] Deisy, C., Gowri, M., Baskar, S., Kalaiarasi, S., and Ramraj, N. A novel term weighting scheme midf for text categorization. *Journal of Engineering Science and Technology 5*, 1 (2010), 94–107.

[DH50] DOLL, R., and HILL, A. B. Smoking and carcinoma of the lung; preliminary report. *British medical journal 2*, 4682 (Sep 1950), 739–748. 14772469[pmid].

[DHZ17] Duan, S., He, R., and Zhao, W. Exploiting document level information to improve event detection via recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Taipei, Taiwan, November 2017), Asian Federation of Natural Language Processing, pp. 352–361.

[DLY14] Deng, Z.-H., Luo, K.-H., and Yu, H.-L. A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications 41*, 7 (2014), 3506–3513.

[DMP+04] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*

(LREC'04) (Lisbon, Portugal, May 2004), European Language Resources Association (ELRA).

[DMPS15] Domeniconi, G., Moro, G., Pasolini, R., and Sartori, C. A study on term weighting for text categorization: A novel supervised variant of tf.idf. In *Proceedings of 4th International Conference on Data Management Technologies and Applications* (Setubal, PRT, 2015), DATA 2015, SCITEPRESS - Science and Technology Publications, Lda, p. 26–37.

[DS04] Debole, F., and Sebastiani, F. Supervised term weighting for automated text categorization. In *Text Mining and its Applications* (Berlin, Heidelberg, 2004), S. Sirmakessis, Ed., Springer Berlin Heidelberg, pp. 81–97.

[DSDN18] Dasgupta, T., Saha, R., Dey, L., and Naskar, A. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 306–316.

[DTY+02] Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Zhang, M., Wu, X.-B., and Yang, M. A linear text classification algorithm based on category relevance factors. In *Digital Libraries: People, Knowledge, and Technology* (Berlin, Heidelberg, 2002), E.-P. Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs, and T. Costantino, Eds., Springer Berlin Heidelberg, pp. 88–98.

[DU19a] Dogan, T., and Uysal, A. K. Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications 130* (2019), 45–59.

[DU19b] Dogan, T., and Uysal, A. K. On term frequency factor in supervised term weighting schemes for text classification. *Arabian Journal for Science and Engineering 44*, 11 (Nov 2019), 9545–9560.

[EM07] Eaton, D., and Murphy, K. Exact bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (San Juan, Puerto Rico, 21–24

Mar 2007), M. Meila and X. Shen, Eds., vol. 2 of *Proceedings of Machine Learning Research*, PMLR, pp. 107–114.

[ES19] Eckroth, J., and Schoen, E. A genetic algorithm for finding a small and diverse set of recent news stories on a given subject: How we generate aaai's ai-alert. *Proceedings of the AAAI Conference on Artificial Intelligence 33*, 01 (Jul. 2019), 9357–9364.

[FFSG20] Fan, B., Fan, W., Smith, C., and Garner, H. Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing & Management 57*, 1 (2020), 102131.

[FGM05] Finkel, J. R., Grenager, T., and Manning, C. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (Ann Arbor, Michigan, June 2005), Association for Computational Linguistics, pp. 363–370.

[FHK+20] Fischbach, J., Hauptmann, B., Konwitschny, L., Spies, D., and Vogelsang, A. Towards causality extraction from requirements. In *2020 IEEE 28th International Requirements Engineering Conference (RE)* (2020), pp. 388–393.

[FLSZ18] Feng, G., Li, S., Sun, T., and Zhang, B. A probabilistic model derived term weighting scheme for text classification. *Pattern Recognition Letters 110* (2018), 23–29.

[FQL18] Feng, X., Qin, B., and Liu, T. A language-independent neural network for event detection. *Science China Information Sciences 61*, 9 (Aug 2018), 092106.

[Fre98] Freitag, D. Information extraction from html: Application of a general machine learning approach. In *AAAI/IAAI* (1998), pp. 517–523.

[FS16] Fattah, M., and Sohrab, M. Combined term weighting scheme using ffnn, ga, mr, sum, & average for text classification. *International Journal of Scientific and Engineering Research 7*, 8 (2016), 2031–2040.

[Gar97] Garcia, D. Coatis, an nlp system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management* (Berlin, Heidelberg, 1997), E. Plaza and R. Benjamins, Eds., Springer Berlin Heidelberg, pp. 347–352.

[GCR16] Grossman, M. R., Cormack, G. V., and Roegiest, A. Trec 2016 total recall track overview. In *TREC* (2016).

[GM02] Girju, R., and Moldovan, D. I. Text mining for causal relations. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference* (2002), AAAI Press, p. 360–364.

[Gra69] Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*, 3 (1969), 424–438.

[GSS00] Galavotti, L., Sebastiani, F., and Simi, M. Experiments on the use of feature selection and negative evidence in automated text categorization. In *Research and Advanced Technology for Digital Libraries* (Berlin, Heidelberg, 2000), J. Borbinha and T. Baker, Eds., Springer Berlin Heidelberg, pp. 59–68.

[HA10] Hirata, Y., and Aihara, K. Identifying hidden common causes from bivariate time series: A method using recurrence plots. *Phys. Rev. E 81* (Jan 2010), 016203.

[HAT+92] Hobbs, J. R., Appelt, D., Tyson, M., Bear, J., and Israel, D. SRI international: Description of the FASTUS system used for MUC-4. In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992* (1992).

[HCJ+16] Huang, R., Cases, I., Jurafsky, D., Condoravdi, C., and Riloff, E. Distinguishing past, on-going, and future events: The EventStatus corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas, November 2016), Association for Computational Linguistics, pp. 44–54.

[HDPM18] Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models:. *Journal of Causal Inference 6*, 2 (2018), 20170016.

[HJCV17] Huang, L., Ji, H., Cho, K., and Voss, C. R. Zero-shot transfer learning for event extraction. *arXiv preprint arXiv:1707.01066* (2017).

[HLSP17] Harnack, D., Laminski, E., Schünemann, M., and Pawelzik, K. R. Topological causality in dynamical systems. *Phys. Rev. Lett. 119* (Sep 2017), 098301.

[Hma20] Hmamouche, Y. Nlints: An r package for causality detection in time series. *The R Journal 12* (06 2020), 21–.

[HR11] Huang, R., and Riloff, E. Peeling back the layers: Detecting event role fillers in secondary contexts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 2011), Association for Computational Linguistics, pp. 1137–1147.

[HZM+11] Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., and Zhu, Q. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 2011), Association for Computational Linguistics, pp. 1127–1136.

[JG08] Ji, H., and Grishman, R. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT* (Columbus, Ohio, June 2008), Association for Computational Linguistics, pp. 254–262.

[JLH18] Jacobs, G., Lefever, E., and Hoste, V. Economic event detection in company-specific news text. In *Proceedings of the First Workshop on Economics and Natural Language Processing* (Melbourne, Australia, July 2018), Association for Computational Linguistics, pp. 1–10.

[JY16] Jagannatha, A. N., and Yu, H. Bidirectional rnn for medical event detection in electronic health records. *Proceedings of the conference. As-*

sociation for Computational Linguistics. North American Chapter. Meeting 2016 (Jun 2016), 473–482. 27885364[pmid].

[KBR91] Kaplan, R. M., and Berry-Rogghe, G. Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition 3*, 3 (1991), 317–337.

[KCN00] Khoo, C. S. G., Chan, S., and Niu, Y. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (USA, 2000), ACL '00, Association for Computational Linguistics, p. 336–343.

[KF09] Koller, D., and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[KJRI91] Krupka, G., Jacobs, P., Rau, L., and Iwanska, L. GE: Description of the NLToolset system as used for MUC-3. In *Third Message Uunderstanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991* (1991).

[KS05] Kipper-Schuler, K. *VerbNet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and Information Science Department, Universiy of Pennsylvania, Philadelphia, PA, 2005.

[kSM12] kwang Song, S., and Myaeng, S. H. A novel term weighting scheme based on discrimination power obtained from past retrieval results. *Information Processing & Management 48*, 5 (2012), 919–930. Large-Scale and Distributed Systems for Information Retrieval.

[KSv+14] Kemmeren, P., Sameith, K., van de Pasch, L., Benschop, J., Lenstra, T., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C., van Heesch, S., Kashani, M., Ampatziadis-Michailidis, G., Brok, M., Brabers, N., Miles, A., Bouwmeester, D., van Hooff, S., van Bakel, H., Sluiters, E., Bakker, L., Snel, B., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M., and Holstege, F. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell 157*, 3 (2014), 740–752.

[LCJ03] Lee, C.-S., Chen, Y.-J., and Jian, Z.-W. Ontology-based fuzzy event extraction agent for chinese e-news summarization. *Expert Systems with Applications 25*, 3 (2003), 431–447.

[LCLZ18] Liu, J., Chen, Y., Liu, K., and Zhao, J. Event detection via gated multilingual attention mechanism. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[LG10] Liao, S., and Grishman, R. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Uppsala, Sweden, July 2010), Association for Computational Linguistics, pp. 789–797.

[Lin57] Lind, J. *A Treatise on the Scurvy*. A. Millar, 1757.

[LJH13] Li, Q., Ji, H., and Huang, L. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Sofia, Bulgaria, August 2013), Association for Computational Linguistics, pp. 73–82.

[LK02] Leopold, E., and Kindermann, J. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning 46*, 1 (Jan 2002), 423–444.

[LLH18] Liu, X., Luo, Z., and Huang, H. Jointly multiple events extraction via attention-based graph information aggregation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, October-November 2018), Association for Computational Linguistics, pp. 1247–1256.

[Llo82] Lloyd, S. Least squares quantization in pcm. *IEEE Transactions on Information Theory 28*, 2 (1982), 129–137.

[LLS09] Liu, Y., Loh, H. T., and Sun, A. Imbalanced text classification: A term weighting approach. *Expert Systems with Applications 36*, 1 (2009), 690–701.

[LLZR21] Li, Z., Li, Q., Zou, X., and Ren, J. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing 423* (2021), 207–219.

[LMG11] Largeron, C., Moulin, C., and Géry, M. Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM Symposium on Applied Computing* (New York, NY, USA, 2011), SAC '11, Association for Computing Machinery, p. 924–928.

[LML+16] Lorenzetti, C., Maguitman, A., Leake, D., Menczer, F., and Reichherzer, T. Mining for topics to suggest knowledge model extensions. *ACM Trans. Knowl. Discov. Data 11*, 2 (December 2016).

[LMR14] Leake, D., Maguitman, A., and Reichherzer, T. Experience-based support for human-centered knowledge modeling. *Knowledge-Based Systems 68* (2014), 77–87. Enhancing Experience Reuse and Learning.

[LS04] Liu, H., and Singh, P. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal 22*, 4 (Oct 2004), 211–226.

[LSLL09] Liu, D.-R., Shih, M.-J., Liau, C.-J., and Lai, C.-H. Mining the change of event trends for decision support in environmental scanning. *Expert Systems with Applications 36*, 2, Part 1 (2009), 972–984.

[LSLT05] Lan, M., Sung, S.-Y., Low, H.-B., and Tan, C.-L. A comparative study on term weighting schemes for text categorization. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (2005), vol. 1, pp. 546–551 vol. 1.

[LTSL09] Lan, M., Tan, C. L., Su, J., and Lu, Y. Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 4 (2009), 721–735.

[MAC14] Ma, H., Aihara, K., and Chen, L. Detecting causality from nonlinear dynamics with short-term time series. *Scientific Reports 4*, 1 (Dec 2014), 7464.

[Mai19] Maisonnave, M. Detección de textos similares a través de una técnica de agrupamiento basada en densidad. In *Communication at the XV Dr. Antonio Monteiro Congress, Bahía Blanca, Argentina* (2019).

[MCCD13] Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[MDT+20a] Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A., and Milios, E. Event detection dataset. Mendeley Data, V1 - http://dx.doi.org/10.17632/7d54rvzxkr.1, 2020.

[MDT+20b] Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A. G., and Milios, E. E. Assessing causality structures learned from digital text media. In *Proceedings of the ACM Symposium on Document Engineering 2020* (New York, NY, USA, 2020), DocEng '20, Association for Computing Machinery.

[MDT+21] Maisonnave, M., Delbianco, F., Tohmé, F., Maguitman, A., and Milios, E. Detecting ongoing events using contextual word and sentence embeddings. *arXiv preprint arXiv:2007.01379* (2021).

[MDTM18] Maisonnave, M., Delbianco, F., Tohmé, F. A., and Maguitman, A. G. A supervised term-weighting method and its application to variable extraction from digital media. In *XIX Simposio Argentino de Inteligencia Artificial (ASAI)-JAIIO 47 (CABA, 2018)* (2018), p. 40–53.

[MDTM19a] Maisonnave, M., Delbianco, F., Tohmé, F., and Maguitman, A. Economic relevant news from the guardian. Mendeley Data, V3 - http://dx.doi.org/10.17632/yt8j2f3hpp.3, 2019.

[MDTM19b] Maisonnave, M., Delbianco, F., Tohmé, F. A., and Maguitman, A. G. A flexible supervised term-weighting technique and its application to variable extraction and information retrieval. *Inteligencia Artificial 22*, 63 (Feb. 2019), 61–80.

[MDTM21] Maisonnave, M., Delbianco, F., Tohmé, F., and Maguitman, A. Assessing the behavior and performance of a supervised term-weighting tech-

nique for topic-based retrieval. *Information Processing & Management 58*, 3 (2021), 102483.

[Mil95] Miller, G. A. Wordnet: A lexical database for english. *Commun. ACM 38*, 11 (November 1995), 39–41.

[MLRM04] Maguitman, A., Leake, D., Reichherzer, T., and Menczer, F. Dynamic extraction topic descriptors and discriminators: Towards automatic context-based topic search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management* (New York, NY, USA, 2004), CIKM '04, Association for Computing Machinery, p. 463–472.

[MSC+13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26* (10 2013), 3111–3119.

[NCG16] Nguyen, T. H., Cho, K., and Grishman, R. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (San Diego, California, June 2016), Association for Computational Linguistics, pp. 300–309.

[NFCG16] Nguyen, T. H., Fu, L., Cho, K., and Grishman, R. A two-stage approach for extending event detection to new types via neural networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP* (Berlin, Germany, August 2016), Association for Computational Linguistics, pp. 158–165.

[NG15] Nguyen, T. H., and Grishman, R. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 365–371.

[NG16] Nguyen, T. H., and Grishman, R. Modeling skip-grams for event detection with convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Austin, Texas, November 2016), Association for Computational Linguistics, pp. 886–891.

[NG18] Nguyen, T., and Grishman, R. Graph convolutional networks with argument-aware pooling for event detection. In *Thirty-second AAAI conference on artificial intelligence* (2018).

[NMB17] Nicholson, W., Matteson, D., and Bien, J. Bigvar: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:1702.07094* (2017).

[Nov90] Novak, J. D. Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching 27*, 10 (1990), 937–949.

[ONSH20] Ombadi, M., Nguyen, P., Sorooshian, S., and Hsu, K.-l. Evaluation of methods for causal discovery in hydrometeorological systems. *Water Resources Research 56*, 7 (2020), e2020WR027251. e2020WR027251 2020WR027251.

[PB15] Peters, J., and Bühlmann, P. Structural intervention distance for evaluating causal graphs. *Neural Computation 27*, 3 (2015), 771–799.

[PBM16] Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 78*, 5 (2016), 947–1012.

[PCI+03] Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering 3* (2003), 28–34.

[Pea85] Pearl, J. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA* (1985), pp. 15–17.

[Pea09] Pearl, J. *Causality*, 2 ed. Cambridge University Press, 2009.

[PFM18] Pinho, C., Franco, M., and Mendes, L. Web portals as tools to support information management in higher education institutions: A systematic literature review. *International Journal of Information Management 41* (2018), 80–92.

[PJS17] Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference : Foundations and Learning Algorithms*. The MIT Press, 2017.

[PK07] Pechsiri, C., and Kawtrakul, A. Mining causality for explanation knowledge from text. *Journal of Computer Science and Technology 22*, 6 (2007), 877.

[PM18] Pearl, J., and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*, 1st ed. Basic Books, Inc., USA, 2018.

[PNI+18] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[PR09] Patwardhan, S., and Riloff, E. A unified model of phrasal and sentential evidence for information extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, August 2009), Association for Computational Linguistics, pp. 151–160.

[PSM14] Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha, Qatar, October 2014), Association for Computational Linguistics, pp. 1532–1543.

[QWQ11] Quan, X., Wenyin, L., and Qiu, B. Term weighting schemes for question categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 5 (2011), 1009–1021.

[RBB+19] Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J. Inferring causation from time

series in earth system sciences. *Nature Communications 10*, 1 (Jun 2019), 2553.

[RCGC15] Roegiest, A., Cormack, G. V., Grossman, M. R., and Clarke, C. Trec 2015 total recall track overview. *Proc. TREC-2015* (2015).

[RDM12] Radinsky, K., Davidovich, S., and Markovitch, S. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web* (New York, NY, USA, 2012), WWW '12, Association for Computing Machinery, p. 909–918.

[Rij79] Rijsbergen, C. J. V. *Information Retrieval*, 2nd ed. Butterworth-Heinemann, USA, 1979.

[Ril96a] Riloff, E. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2* (1996), AAAI'96, AAAI Press, p. 1044–1049.

[Ril96b] Riloff, E. An empirical study of automated dictionary construction for information extraction in three domains. *Artificial Intelligence 85*, 1 (1996), 101–134.

[RJ76] Robertson, S. E., and Jones, K. S. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 3 (1976), 129–146.

[RNK+19] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances 5*, 11 (2019).

[Rob04] Robertson, S. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation 60*, 5 (Jan 2004), 503–520.

[San08] Sandhaus, E. The new york times annotated corpus LDC2008T19. *Linguistic Data Consortium, Philadelphia 6*, 12 (2008), e26752.

[SB88] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management 24*, 5 (1988), 513–523.

[SBMM19] Samant, S. S., Bhanu Murthy, N. L., and Malapati, A. Improving term weighting schemes for short text classification in vector space model. *IEEE Access 7* (2019), 166578–166592.

[Sch00] Schreiber, T. Measuring information transfer. *Phys. Rev. Lett. 85* (Jul 2000), 461–464.

[Scu10] Sculley, D. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, Association for Computing Machinery, p. 1177–1178.

[SG91] Spirtes, P., and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review 9*, 1 (1991), 62–72.

[SGSH00] Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

[SHHK06] Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research 7*, 72 (2006), 2003–2030.

[Sim80] Sims, C. A. Macroeconomics and reality. *Econometrica 48*, 1 (1980), 1–48.

[SIS+11] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *J. Mach. Learn. Res. 12*, null (July 2011), 1225–1248.

[SKW07] Suchanek, F. M., Kasneci, G., and Weikum, G. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web* (New York, NY, USA, 2007), WWW '07, Association for Computing Machinery, p. 697–706.

[SLB+21] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE 109*, 5 (2021), 612–634.

[SMY+12] Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. Detecting causality in complex ecosystems. *Science 338*, 6106 (2012), 496–500.

[Sno56] Snow, J. On the mode of communication of cholera. *Edinburgh medical journal 1*, 7 (Jan 1856), 668–670. 29647347[pmid].

[SPP+05] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*, 5721 (2005), 523–529.

[SQCS18] Sha, L., Qian, F., Chang, B., and Sui, Z. Jointly extracting event triggers and arguments by dependency-bridge rnn and tensor-based argument interaction. *Proceedings of the AAAI Conference on Artificial Intelligence 32*, 1 (Apr. 2018).

[SSG+98] Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research 33*, 1 (1998), 65–117. PMID: 26771754.

[STA06] Surdeanu, M., Turmo, J., and Ageno, A. A hybrid approach for the acquisition of information extraction patterns. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)* (2006).

[TC60] Thistlethwaite, D. L., and Campbell, D. T. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology 51*, 6 (1960), 309.

[Tho53] Thorndike, R. L. Who belongs in the family? *Psychometrika 18*, 4 (Dec 1953), 267–276.

[Tib96] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[TLL20] Tang, Z., Li, W., and Li, Y. An improved term weighting scheme for text classification. *Concurrency and Computation: Practice and Experience 32*, 9 (2020), e5604. e5604 CPE-19-0287.R1.

[TM94] Tokunaga, T., and Makoto, I. Text categorization based on weighted inverse document frequency. In *Special Interest Groups and Information Process Society of Japan (SIG-IPSJ* (1994), pp. 33–39.

[TWC+20] Tong, M., Wang, S., Cao, Y., Xu, B., Li, J., Hou, L., and Chua, T.-S. Image enhanced event detection in news articles. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 05 (Apr. 2020), 9040–9047.

[Var14] Varian, H. R. Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*, 2 (May 2014), 3–28.

[VH99] Voorhees, E., and Harman, D. Overview of the eight text retrieval conference. In *Proc. TREC-8, the 8th text retrieval conference* (1999).

[vHP81] van Rijsbergen, C., Harper, D., and Porter, M. The selection of good search terms. *Information Processing & Management 17*, 2 (1981), 77–91.

[VSHK16] Verberne, S., Sappelli, M., Hiemstra, D., and Kraaij, W. Evaluation and analysis of term scoring methods for term extraction. *Information Retrieval Journal 19*, 5 (Oct 2016), 510–545.

[WBL+14] Wu, S., Bondugula, S., Luisier, F., Zhuang, X., and Natarajan, P. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2014).

[WGG17] Wu, H., Gu, X., and Gu, Y. Balancing between over-weighting and under-weighting in supervised term weighting. *Information Processing & Management 53*, 2 (2017), 547–557.

[WL21] Weng, J., and Lee, B.-S. Event detection in twitter. *Proceedings of the International AAAI Conference on Web and Social Media 5*, 1 (Aug. 2021), 401–408.

[WSMM06] Walker, C., Strassel, S., Medero, J., and Maeda, K. ACE 2005 multilingual training corpus LDC2006T06. *Linguistic Data Consortium, Philadelphia 57* (2006).

[WZ13] Wang, D., and Zhang, H. Inverse-category-frequency based supervised term weighting schemes for text categorization. *Journal of Information Science and Engineering 29*, 2 (2013), 209–225. cited By 28.

[YGTH00] Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. Automatic acquisition of domain knowledge for information extraction. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics* (2000).

[ZJS19] Zhang, T., Ji, H., and Sil, A. Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. *Data Intelligence 1*, 2 (04 2019), 99–120.

[ZLZ+16] Zhao, S., Liu, T., Zhao, S., Chen, Y., and Nie, J.-Y. Event causality extraction based on connectives analysis. *Neurocomputing 173* (2016), 1943–1950.

[ZWM+17] Zhao, S., Wang, Q., Massung, S., Qin, B., Liu, T., Wang, B., and Zhai, C. Constructing and embedding abstract event causality networks from text snippets. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2017), WSDM '17, Association for Computing Machinery, p. 335–344.