FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES
ICADE

Trabajo Fin de Grado

ANÁLISIS DE TEXTO EN FINANZAS

Autor: Ana Berjón Valles
Director: Carlos Bellón Núñez-Mera

MADRID
Diciembre 2021

Este trabajo fue escrito con LaTeX y compilado en TeXmaker usando la distribución TeX-2013. Las familias de fuentes usadas son Bitstream Charter, Utopia, Bookman, and Computer Modern. A menos que se indique lo contrario, todas las figuras fueron creadas por el autor usando Microsoft Visio®, Adobe Illustrator®, y MATLAB®.
Agradecimientos a Jaime Boal, autor de la plantilla.

A Pablo,
porque este trabajo también es tuyo

Resumen/Abstract

ANÁLISIS DE TEXTOS FINANCIEROS

Autor: Ana Berjón Valles
Director: Bellón Núñez-Mera, Carlos
Entidad colaboradora: ICADE – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Los avances logrados en el siglo XXI a nivel tecnológico han afectado drásticamente a todos los aspectos de la vida, incluídas las finanzas. En este contexto de digitalización, se desarrolla el presente proyecto que busca conseguir un sistema de inteligencia artificial que mediante técnicas de NLP sea capaz de detectar el sentimiento con el que se redactan las noticias e identificar si los mercados actúan en consecuencia empleando el modelo de los tres factores de Fama y French. El sistema propuesto logra ambos objetivos, pero no consigue demostrar su correlación.

Palabras clave: NLP, Event Studies, Noticias financieras, Modelo de los 3 factores.

FINANCIAL TEXT ANALYSIS

Author: Berjón Valles, Ana
Supervisor: Bellón Núñez-Mera, Carlos
Collaborating Entity: ICADE – Universidad Pontificia Comillas

ABSTRACT

The technological advances of the 21st century have drastically affected all aspects of life, including finance. In this context of digitalization, this project aims to develop an artificial intelligence system that through NLP techniques is able to detect the sentiment with which the news is written and identify whether the markets act accordingly using the 3-Factor model of Fama and French. The proposed system achieves both objectives, despite not being able to demonstrate its correlation.

Keywords: NLP, Event Studies, Financial News, 3-Factor model

Índice general

Índice de figuras

Índice de tablas

1

Introducción

Este primer capítulo presenta las razones fundamentales por las que se ha decidido llevar a cabo este proyecto, su principal objetivo y las herramientas empleadas en su ejecución. Además señala la organización del documento para facilitarle su lectura.

La expansión tecnológica resultó dramática a partir de los años 2000, cuando se popularizó que en cada casa hubiera al menos un dispositivo. Las nuevas capacidades de estos instrumentos, su incremento exponencial de potencia, memoria y procesamiento, llevaron a la aplicación de nuevas técnicas a todos los aspectos de la vida. Y las finanzas no serían menos.

Los mercados bursátiles pronto se digitalizaron, y no tardaron en aparecer algoritmos encargados de predecir los devenires de los mercados, automatizando la compraventa de acciones y buscando obtener la máxima rentabilidad en prácticamente cualquier situación. Ademas, aprovechando todas estas capacidades que tiene actualmente la tecnología, surgen nuevas técnicas: *Machine Learning, Deep Learning*... que consiguen obtener resultados muy precisos para diversas situaciones.

1.1. Motivación

En este contexto de digitalización, la opinión de un medio de comunicación como el *Financial Times* alcanza 26 millones de lectores mensuales , es decir, sus artículos llegan a una cantidad de gente que en el siglo anterior resultaría inconcebible. En este sentido, no solo es importante qué cuentan, sino también cómo lo cuentan. De esta forma, el modo de redacción de un determinado hecho puede resultar decisivo en entornos financieros, donde la confianza es crucial para el devenir de los mercados.

El presente estudio nace para tratar de determinar la influencia derivada de estas diferencias en la narración de una noticia, tratando de relacionar el sentimiento con el que se relata un suceso en un lugar y momento concreto con el desempeño de los mercados en ese día.

1.2. Objetivos

El objetivo principal de este proyecto es analizar si existe correlación entre el sentimiento con el que se redactan las noticias financieras y la evolución de su precio en los mercados. Para ello se ha tomado un dataset de noticias financieras obtenidas de distintas fuentes (*Financial Times, The Wall Street Journal*) y se analiza la intención tras la noticia, clasificándola en positivo, negativo o neutro. Este resultado será comparado con la evolución del precio de la acción aquel día en el mercado indicado para tratar de determinar si realmente está correlacionado o no.

1.3. Herramientas

Este proyecto consiste en un desarrollo software. Para llevarlo a cabo, se ha empleado el lenguaje de programación *Python*, sobre el que se ha generado el modelo de NLP. Para ello se han utilizado dos entornos distintos: *Jupyter Notebook* y *Visual Studio Code*.
Por otra parte, se ha llevado a cabo un control de versiones en un repositorio Git, sobre el que se ha ido edificando la solución.

1.4. Organización del documento

El documento se estructura en ocho capítulos, siendo el presente el primero de ellos, que introduce el problema del proyecto. El siguiente contiene la revisión de la literatura existente sobre temas de *machine learning* en materia de procesamiento de lenguaje, y los *event studies*. A continuación se detalla el desarrollo del proyecto y se finaliza con tres capítulos que incluyen resultados, conclusiones y áreas de mejora. Finalmente se recoge la bibliografía con las referencias citadas a lo largo del documento.

2

Estado del arte

Este capítulo encuadra el proyecto en el contexto de las tecnologías ya desarrolladas. Se indican algoritmos empleados en la elaboración de este sistema así como posibles alternativas realizando una revisión de la literatura publicada por diversos investigadores hasta la fecha. La información parte de los puntos más generales hacia la especificación de cada materia.

Este capítulo se encuentra dividido en dos partes fundamentales: la primera de ellas estudiará temas más próximos a modelos de tratamiento del lenguaje en programación y la segunda tratará temas financieros. De esta forma, comentará las diversas tecnologías que participan en estos sistemas, realizando un análisis de distintas técnicas de NLP (*Natural Language Processing*) y describiendo las metodologías requeridas para determinar la existencia de la correlación entre el sentimiento y los precios.

2.1. NLP

Natural Language Processing se puede definir como un subconjunto de técnicas de Inteligencia Artificial que trata de reducir la brecha de comunicación entre humanos y ordenadores. Para ello, se busca que los ordenadores sean capaces de entender, interpretar, utilizar y procesar el lenguaje humano aplicando diversas técnicas.

Los orígenes de esta tecnología se encuentran en los años 1940, durante la Segunda Guerra Mundial. La idea inicial era mucho más simple que lo que hoy se conoce como NLP. Se conocía como *Machine Translation* (MT) e intentaba conseguir traducción de inglés a ruso (y viceversa) empleando ordenadores. Con el tiempo, este concepto inicial fue volviéndose más ambicioso, persiguiendo la comunicación hombre-máquina. En cualquier caso, a día de hoy, la traducción simultánea es una de las aplicaciones más comunes de estos algoritmos y se consiguen resultados muy fiables.

La información dentro de un texto, una frase, o cualquier otra forma de lenguaje se considera no estructurada, ya que los datos no se están almacenados siguiendo ningún tipo de estructura, como la que podría encontrarse al realizar mediciones de elementos cuantitativos o recuperar

las respuestas de un cuestionario. Por ello, antes de realizar cualquier tipo de estudio habrá que extraerla. En este sentido, es de vital importancia ser capaz de identificar y recoger aquellos elementos que realmente puedan aportar resultados válidos.

Dependiendo de la finalidad del algoritmo que se esté desarrollando, los procesos de NLP a implementar serán muy diversos. Dentro de este amplio campo hay infiinidad de procesos, que se pueden agrupar en distintas categorías. En función de la amplitud de la clasificación, se pueden encontrar entre cinco y siete grupos fundamentales. En esta ocasión se englobarán en cinco:

- **Análisis léxico y morfológico**: aglutinan el análisis de las palabras como tal basándose en su estructura, independientemente de su contexto en la frase. Es uno de los grupos más amplios, y bien podría dividirse en dos: análisis léxico y análisis morfológico.
- **Análisis sintáctico**: la sintaxis pretende comprender el significado de las oraciones, buscando entender la estructura de los distintos tipos de frases. Este tipo de estudio determina cómo el orden de las palabras afecta a su significado y así decidir cómo se relacionan entre sí.
- **Integración discursiva**: supone una abstracción superior, tratando de enlazar palabras concretas, generalmente pronombres, con elementos de oraciones anteriores o siguientes. Esta combinación de formas gramaticales y significado es lo que en general permite obtener textos hilados, generando matices en el lenguaje.
- **Análisis semántico**: la semántica es la parte de la lingüística que estudia el significado de las expresiones. Estas funciones buscan transferir el significado de palabras individuales a conjuntos de palabras.
- **Análisis pragmático**: trata descubrir la intencionalidad de una oración considerando el propósito comunicativo y social. Se intenta abstraer el significado literal de la frase para comprender el significado real que se quiere transmitir. Para este análisis es necesario mayor contexto para la frase, generalmente un diálogo.

2.1.1. Modelos

En el Capítulo 1 se ha destacado la importancia del *Machine Learning* como técnica innovadora y disruptiva, pero no se ha explicado en qué consiste ni por qué es tan relevante. En esta sección se darán unas breves nociones sobre el tema y se discuten en mayor profundidad diversos modelos generados mediante esta técnica.

Machine Learning se describe como la técnica que proporciona a los ordenadores la capacidad de aprender sin ser programados para ello de forma explícita. Es una aplicación de la inteligencia artificial que hace que una máquina sea capaz de aprender de experiencias anteriores para realizar predicciones futuras. En este contexto, las experiencias son simplemente datos introducidas con anterioridad. Se distinguen tres tipos fundamentales de aprendizaje: supervisado, no supervisado y aprendizaje por refuerzo.

El aprendizaje supervisado procesa los datos junto con sus características, etiquetas y objetivos. Estas etiquetas son las que van consiguiendo que el modelo obtenga una correlación entre los resultados y las características. Dos de las tareas más comunes dentro de este tipo de aprendizaje son la clasificación y la regresión. Empleando clasificación, la máquina busca predecir valores discretos, es decir, debe encontrar la categoría más probable para los nuevos

ejemplos. Respecto a la regresión, el objetivo es encontrar el valor de una variable continua. Estos modelos son, generalmente, los que resultan más efectivos ya que son entrenados con conjuntos de datos más detallados, enriqueciendo el modelo haciéndolo más sensible a matices.

En caso de disponer de un conjunto de datos sin clasificar, el tipo de aprendizaje sería no-supervisado. En estas ocasiones no hay etiquetas u objetivos para cada uno de los ejemplos y por ello una de las soluciones más utilizadas es el *clustering*, donde se intenta agrupar muestras similares.

El aprendizaje por refuerzo aplica al buscar modelos muy orientados a una finalidad concreta, donde deben aprender a maximizar una dimensión o lograr algún objetivo muy complejo. En estos casos se propicia el aprendizaje dándole una retroalimentación sencilla al sistema, conocida como señal de refuerzo (*reinforcement signal*).

En general, un modelo se puede encontrar con dos problemas fundamentales: *over-fitting* y *under-fitting*, aunque no de manera simultánea. *Under-fitting* se produce cuando el modelo tiene pocos datos o pocas variables y características y por lo tanto no puede aprender correctamente, lo que se traduce un un gran sesgo. Por el contrario, el sobre-aprendizaje u *over-fitting* se da cuando el modelo emplea funciones muy complejas y pierde la capacidad de generalizar ante nuevos modelos. Esto supone una alta varianza entre sus resultados.

2.1.1.1. Pre-procesamiento

Para llegar a comprender el significado concreto de un texto, se pueden llevar a cabo muchos y muy diversos análisis siendo las cinco mayores agrupaciones de estos estudios los descritos en la introducción de esta sección. Sin embargo, cada uno de estos grupos engloba gran cantidad de pruebas distintas. En todos los casos es preciso llevar a cabo un preprocesamiento del texto donde se engloban distintas tareas propias de estos análisis. En esta sección, se hablará de algunas de las más comunes y que, en ocasiones, se pueden emplear para distintos tipos de estudios.

Uno de los elementos fundamentales del preprocesamiento de textos es la ***tokenización***. Es un procedimiento base en este tipo de estudios y no se entiende un análisis NLP sin él. Su importancia es tal, que se emplea tanto en métodos NLP tradicionales como en *transformers*[1] y en la mayoría de ocasiones, es el punto de partida.

La *tokeninzación* consiste en separar una porción de texto en entidades más pequeñas llamadas *tokens*. Existen tres tipos fundamentales: palabras, caracteres o sub-palabras (partes de palabras). Cada uno de los distintos tipos de *tokenización* en función del tipo de *token* tienen ventajas y desventajas.

Se emplea para obtener el vocabulario del corpus, es decir, el conjunto de los diferentes *tokens* de un texto. Este vocabulario se puede formar utilizado todos los *tokens* o solamente los que aparecen con mayor frecuencia para después aplicar otras técnicas más complejas que permiten extraer conocimiento. Entre estas técnicas se encuentran la "cuenta vectorizada" (*"Count Vectorizer"*) o el TF-IDF.

En segundo lugar, el ***stemming***. Este procedimiento es muy similar a la *lemmatización*, pero esta se discutirá más adelante. El *stemming* es una técnica que busca reducir las palabras eliminando lo que suelen ser sufijos o prefijos. De esta forma, se queda con la raíz o lexema

[1] Los transformes se explicarán con más detenimiento en la Sección 2.1.2

del vocablo, que es invariable en todas las palabras de la misma familia. Esta raíz, también denominada *stem*, no tiene por qué ser una palabra con significado completo, sencillamente el punto de partida para obtener nuevas palabras. Con este procedimiento, se pueden producir dos tipos de errores fundamentales: *over stemming* y *under stemming*.

Over stemming se produce cuando se eliminan más letras de las que sería necesario, y esto puede provocar que dos palabras que realmente no están relacionadas se confundan. Un ejemplo de esto podría darse con las palabras "universidad" y "universo", en el caso de que el modelo llegara a reducir ambas a "univer". Esto daría a entender, erróneamente, que ambas tienen el mismo significado. Por el contrario, *under stemming* provoca que dos palabras que contienen el mismo significado se asocien a *stems* diferentes.

Porter propuso en 1980 un algoritmo para esta tarea (Porter, 1980). Se basaba en eliminar los distintos sufijos asumiendo que eran una combinación de sufijos más sencillos. A día de hoy, continúa siendo uno de los principales métodos, ya que destaca por su simplicidad y rapidez, pero existen otros que también son muy utilizados. Algún otro algoritmo había sido ya propuesto anteriormente, como es el caso del de Lovins y desde entonces se han elaborado muchos otros algoritmos, como por ejemplo el de Krovetz, pero todos ellos presentan sus ventajas y desventajas.

Otro elemento importante, y muy parecido al anterior, es la ***lemmatización***. En esta ocasión también se procede a eliminar las partes de la palabra que no aportan significado. Sin embargo, mientras que en el caso anterior el resultado podía ser un conjunto de palabras sin significado propio, en este caso se obtiene la palabra primitiva desde la que se obtienen otras palabras derivadas.

La *lemmatización* requiere un mayor número de recursos el que *stemming* para poder ser ejecutado correctamente. Además, en muchos casos, los resultados no difieren en gran medida. Aplicar ambos simultáneamente carece de sentido y será preciso evaluar el objetivo final del modelo para determinar cuál de ellos es conveniente aplicar.

Por último, se trata el tema de la eliminación de las ***stop words***. Este término se refiere a aquellas palabras que no aportan significado o que no proporcionan información de cara a la aplicación de determinadas técnicas de análisis. Entre estas palabras se encuentran algunas de las más frecuentes en un idioma, siendo algunos ejemplos claros del idioma inglés las palabras *the*, *is*, o *and*. De hecho, en esta categoría se suelen englobar conjunciones, preposiciones, adverbios y artículos.

A la hora de eliminar estas palabras, se lleva a cabo un proceso de tokenización por palabras. A continuación, se toma una lista de *stop words* y se va comparando de forma secuencial cada palabra de la lista con las del texto. En caso de que coincidan, se elimina la palabra. Existen infinidad de listas ya diseñadas, pero se pueden construir a partir de un texto o conjunto de ellos. A la hora de crear estas listas se suelen tomar las palabras más frecuentes, las menos frecuentes y aquellas con un bajo IDF (*Inverse Document Frequecy*).

Para aplicar esta última técnica se requiere un conjunto de documentos distintos y se busca conocer aquellos términos que aparecen en un menor número de ellos. El valor de IDF se obtiene de la siguiente forma: $IDF(ti) = log(\frac{N}{M})$, siendo N es el número total de documentos y M el número de documentos en el que dicho término aparece.

2.1.1.2. Named Entity Recognition and Classification

Esta técnica forma parte del pre-procesado de la información no estructurada y es una de las técnicas más comunes y utilizadas en las soluciones NLP, casi con independencia de la aplicación final a la que se vaya a destinar el modelo.

Esta técnica consiste en reconocer las *Named Entities*[2] (en castellano: *Entidades Denominadas*) que son palabras con propiedades similares a otras pertenecientes a un grupo concreto. De esta forma se definen los *rigid designator* (*designador rígido*) que son miembros dentro de una clase semántica cuya información varía dependiendo de los puntos de interés.

Estos conceptos surgieron en la conferencia MUC-6 (*Message Understanding Conferences*) cuyo objetivo principal pasaba por lograr identificar ENAMEX (persona, localización, organización) y NUMEX (tiempo, moneda y expresiones en porcentaje) a partir de información no-estructurada recogida en informes de empresas y mensajes militares (Grishman & Sundheim, 1996).

En los 25 años transcurridos desde aquella conferencia el campo ha ido experimentando numerosos avances, entre los que cabe destacar la aplicación de las técnicas a otros idiomas y la utilización de distintas aproximaciones para resolver estos problemas, basadas en *Machine Learning*.

Uno de los factores fundamentales a la hora de aplicar esta técnica de extracción de información es el lenguaje sobre el que se aplica. Para empezar, la mayor parte de la información disponible se encuentra en inglés, seguido de cerca por otras lenguas europeas. En estos idiomas, las mayúsculas generalmente son un claro indicador de que el sistema se encuentra ante una *named entity*. Sin embargo, se han desarrollado algoritmos para gran variedad de lenguas, desde los idiomas del sudeste asiático hasta ruso o vietnamita (Goyal, Gupta, & Kumar, 1996).

Pero ese no es el único factor que afecta al NER. Los distintos tipos de generos literarios y de textos presentan gran variedad de *named entities*, con distintas propiedades y características. Por ello no serán iguales las entidades en un texto sobre medicina, que en un libro de fantasía para niños.

Como se ha indicado al comienzo de la sección, la mayoría de los modelos actuales de NERC se apoyan en modelos de *Machine Learning*. En función de las características del set de datos disponible, se tenderá a utilizar un tipo de modelo u otro. Los modelos supervisados requieren un detallado conjunto de datos para facilitar el reconocimiento de patrones y *clusters*.

En caso de no disponer de un set de datos tan descriptivo, se deberá elegir un aprendizaje semi-supervisado, que opta por combinar corpus etiquetados y no etiquetados para elaborar hipótesis. En este contexto, el método más conocido es *bootstrapping* (Bhagavatula, 2012). También se puede disponer de aprendizaje no supervisado, que se suele basar en *clustering* elaborado a partir de reglas estadísticas para extraer estas entidades. Otra opción al basarse en aprendizaje no supervisado consiste en buscar reglas de asociación a través del texto, pero en general los resultados son peores que empleando otros tipos de aprendizaje.

Hay gran cantidad de clasificadores en *named entity*. A continuación se mencionan algunos de los más básicos (Goya et al., 2018)

[2] A lo largo del documento se emplearan los nombres en inglés de diversos elementos ya que es la forma más común de denominarlos.

- Naive Bayes: clasificador sencillo basado en probabilidad. Este clasificador sigue el teorema de bayes, asumiendo que los atributos empleados para la clasificación son condicionalmente independientes.
- CRF: *Conditional Random Field* modelo probabilístico que se emplea habitualmente para etiquetar. Es un clasificador continuo que funciona con características dependientes.
- SVM: *Support Vector Machine* clasificador linear no probabilístico que tiende a alcanzar altos niveles de precisión.
- HMM: *Hidden Markov Model* modelo estadístico que asigna una secuencia objetivo a cada secuencia de palabras.
- MaxEnt *Maximum Entropy* clasificador probabilístico exponencial que se usa para resolver problemas complejos.

2.1.1.3. Análisis de sentimiento

El análisis de sentimiento es un campo de estudio que surgió a principios de los años 2000 con el trabajo de Pang, Lee y Vaithyanathan (2002). El objetivo es clasificar los documentos desde un punto de vista diferente, ignorando, entreo otros, el tema del texto. Se busca conocer si el sentimiento que transmite un texto es positivo o negativo. Para ello se analizaron críticas de películas y se aplicaron algoritmos de *machine learning* basados en estadísticos clásicos. Los resultados de este primer estudio no fueron especialmente halagüeños, pero con el paso del tiempo, estos sistemas han ido desarrollando nuevas y mejoradas técnicas.

Actualmente hay dos aproximaciones principales: el análisis de sentimiento basado en reglas y el automático. Sus enfoques son muy distintos entre sí, aunque el segundo de ellos presenta generalmente resultados más eficientes. Esto se debe, como veremos a continuación con un poco más de detalle, a que la estrategia basada en reglas carece de flexibilidad y precisión.

El análisis de sentimiento automático es el más eficiente de los dos ya que realmente profundiza en el texto para devolver resultados. Para ello, se emplean algoritmos de clasificación de aprendizaje supervisado. Entre estos algoritmos de clasificación destacan la regresión linear, el clasificador bayesiano ingenuo (*Naive Bayes*), SVM (*Support Vector Machines*) o las redes neuronales convolucionales.

Por el contrario, la aproximación basada en reglas precisa de un preprocesado NLP como el descrito en la Sección 2.1.1.1. Una vez realizado este paso, se procede a crear una lista de palabras, también llamada bolsa de palabras (*Bag of Words - BOW*) y se compara con la información de una librería ya creada, que tiene asignados unos valores positivos o negativos para cada uno de los términos. La mayoría de estas librerías, también llamadas “diccionarios de sentimiento” están entrenados en inglés, aunque cada vez es más común encontrarlos en más idiomas.

Dado que un mismo término empleado en diferentes contextos puede tener distintas connotaciones, se han creado librerías especializadas en los distintos campos del saber. De esta forma, siempre que se use una librería apropiada, el análisis de sentimiento basado en reglas puede obtener resultados similares a los del modelo automático, pero con un procesamiento muy inferior.

En este sentido, el mundo de las finanzas no es distinto a muchos otros y en el año 2011 Loughran y Mcdonald (2011) hicieron un estudio para mejorar el estudio del sentimiento

en textos de esta índole. En el artículo se recogen críticas a la librería *Harvard IV-4*,[3] ya que cataloga como negativas palabras que, en argot financiero no tienen por qué serlo. Algunas de estas palabras son: *tax, cost, capital, board, liability* o *depreciation*.

Este diccionario financiero, conocido como LM (*Loughran and McDonald*) es muy extenso: contiene 354 palabras positivas y 2392 negativas. Respecto a otros diccionarios previos que versaban sobre este tema, presenta dos ventajas principales: es muy completo, y se creó ya teniendo como objetivo el análisis de las comunicaciones financieras. Ha sido utilizado en multitud de estudios para medir el tono de artículos y columnas, algunos de ellos llegando a confirmar lo que se pretende conseguir en este proyecto (Loughran & Mcdonald, 2016).

2.1.2. Transformers

En su entrenamiento emplean *Transformers*, que es una arquitectura de redes neuronales especializada en resolver tareas de conversión de una secuencia a una secuencia nueva (*Sec2Seq*). Esta arquitectura se propuso en 2017, en el paper "*Attention Is All You Need*" (Vaswani et al., 2017). Esta arquitectura solventa problemas de memoria de las redes neuronales recurrentes, mejoran el rendimiento y permiten entrenar más rápido con cuerpos más pesados.

Esta estructura está compuesta por un *Encoder* y un *Decoder*. El primero procesa la secuencia de entrada y obtiene un contexto, que luego utilizará el segundo para generar la de salida. Entrando un poco más en el detalle de su funcionamiento, se mantiene la información de la posición y lleva a cabo una codificación basada en una función sinusoidal que garantiza que habrá un valor único para cada paso (posición de la palabra), las palabras se separan tantos pasos como distanciadas estén en el texto y es determinístico.

Sin embargo, el algoritmo principal de esta arquitectura es la *Auto-atención* (*Self-Attention*) que es el que permite conocer con qué otra palabra de la secuencia está relacionada la palabra que se está procesando en un instante concreto de tiempo. Dicho de otra forma, vendría a resolver el problema de identificar la palabra dentro de la oración a la que hace referencia un pronombre. Para ello se realiza un producto escalar entre el vector de la palabra en ese punto y las demás de la secuencia. Se pondera entre el tamaño total de la cadena para evitar errores al operar sobre secuencias muy grandes, normalizando así el valor de salida. Finalmente se multiplica el valor obtenido (entre 0 y 1 tras la normalización) para descartar las palabras no significativas.

2.1.3. Grandes modelos

A lo largo de esta sección, se han ido comentando algunos algoritmos a implementar para obtener modelos adecuados. Se ha introducido la dificultad de su entrenamiento y la necesidad de tener conjuntos de datos fiables y bien estructurados. Esta complejidad, sumada al alto consumo de recursos que supone mantener grandes modelos hace que en general, se opte por generar sistemas pequeños y manejables destinados a solucionar problemas concretos.

Sin embargo, algunas multinacionales dedicadas a desarrollo de software sí generan este tipo de modelos, que después rentabilizan vendiendo a empresas menores la posibilidad de usarlos a través de APIs. Emplean gran cantidad de recursos y tiempo para llevar a cabo su entrenamiento y son muy completos en cuanto a las funcionalidades que son capaces de desempeñar. Estos modelos se entrenan con información disponible en Internet. No necesariamente conjuntos de

[3] Un diccionario de sentimiento general (no especializado en ningún argot concreto) de los más utilizados.

datos como tal, sino todas las páginas disponibles de manera gratuita. Esto les proporciona un entrenamiento exhaustivo a partir del cual se pueden conseguir resultados increíbles.

En esta sección, se comentan algunos de los modelos más importantes: GPT-3, deBERTa y sus variantes, y Copilot, y todos ellos emplean esta arquitectura de red neuronal descrita en la Sección 2.1.2.

2.1.3.1. BERT

BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019) es un modelo presentado en 2018 por Google que marcó tendencia. Su modelo permitió analizar el contexto antes y después de cada palabra, concepto que fue disruptivo. Este modelo consigue generar buenos resultados en once tareas diversas relacionadas con el campo del NLP, entre las que destaca la posibilidad de responder preguntas, NER y otros elementos de comprensión. Sus mayores logros incluyen mejorar en más de un 7 % los modelos anteriores y llegando a superar en algunos parámetros la eficiencia humana.

Desde 2018, se ha ido presentando varios modelos que perfeccionaban en distintos aspectos al original: RoBERTa (*Robustly Optimized BERT Pretraining Approach*) (Liu et al., 2019), ALBERT (*A Lite BERT*) (Lan et al., 2019) o StructBERT (Wang et al., 2019), siendo el último modelo presentado deBERTa (*Decoding-enhanced BERT with Distangled Attention*) (He, Liu, Gao & Chen, 2020) . No todos estos modelos han sido elaborados por Google, son mejoras sobre su modelo inicial.

El modelo más reciente, DeBERTa (He et al., 2020) fue elaborado por Microsoft y presenta dos mejoras principales relacionadas con la atención (algoritmo descrito en la Sección 2.1.2) y el decodificador de salida.

2.1.3.2. GPT-3

GPT es una serie de modelos de procesamiento del lenguaje entrenados y comercializados por la empresa OpenAI. El modelo GPT-2 (Radford et al., 2018) supuso una revolución cuando se estrenó en el año 2019 ya que fue el mayor sistema de NLP entrenado hasta la fecha: un *transformer* que considera 1.5 miles de millones de parámetros. Para preparar este modelo se utilizó el texto de 45 millones de páginas web. GPT-2 es capaz de generar textos coherentes y cumplir satisfactoriamente gran cantidad de tareas de índole similar.

Sin embargo, la auténtica revolución llegó un año más tarde, en el momento en el que OpenAI introdujo su nuevo modelo: GPT-3 (Brown et al., 2020). Este *transformer* considera 175 miles de millones de parámetros, lo que supone un incremento de más del 1000 %. Ha sido preparado de forma genérica, es decir, sin buscar eficiencia en una tarea concreta y para ello se han utilizado todo tipo de textos, sin discernir. Emplea la misma arquitectura y el mismo modelo que GPT-2, aunque aplica distintos patrones de atención en las diferentes capas de la red.

El modelo se ha evaluado en 24 tareas diferentes de NLP y en algunas de ellas ha llegado a superar la eficiencia de modelos entrenados específicamente para realizar esa actividad. Destaca la elaboración de artículos de noticias, donde los textos elaborados por este modelo son indistinguibles de los que podría haber realizado un ser humano.

2.1.3.3. Copilot

Finalmente, en esta sección se hablará de *Copilot*, que es el último gran hito del campo del NLP, anunciado el pasado 29 de junio de 2021. Es un modelo elaborado por GitHub en colaboración con OpenAI y es uno de los grandes modelos más especializados que existen.

Su aplicación está muy centrada en un único campo concreto: la elaboración de código en distintos lenguajes de programación. Aunque GPT-3 era capaz de crear código, no había sido especialmente entrenado para ello. Por ello ha surgido esta nueva versión, cuyo aprendizaje se ha realizado con todo tipo de repositorios públicos de GitHub. Su utilización es muy sencilla: el programador describe lo que necesita que realice su función y el modelo automáticamente genera código que satisface esas necesidades.

Este modelo sigue en fase de pruebas, y tiene una precisión del 43 % al primer intento que sube al 57 % cuando se le dan diez intentos. Aunque los números no son muy elevados, son cifras prometedoras, ya que el modelo continúa entrenándose. Por ello, cuando la eficiencia alcance valores más significativos, puede llegar a suponer toda una revolución en el mundo de la programación, ya que no se requerirán conocimientos específicos para desempeñar dichas tareas.

2.2. Event Study

Los mercados bursátiles se caracterizan por la constante apreciación y depreciación de distintos valores financieros. Sin embargo, detrás de estas fluctuaciones se suelen encontrar muchos y muy diversos motivos que determinan el valor final. Este conglomerado de factores dificultan enormemente la evaluación del evento que se busca analizar.

Por ello, en esta sección se analizan distintas aproximaciones para demostrar cómo ha influido un suceso concreto en la valoración del activo financiero estudiado. Estas técnicas reciben el nombre de *Event Study* y se pueden utilizar tanto para explicar el efecto de un suceso, como para intentar predecir cómo reaccionará un activo ante una situación concreta. Estos métodos se basan en la aplicación de técnicas estadísiticas para obtener los resultados buscados.

Para poder aplicar este análisis, independientemente del modelo a seguir, el procedimiento es el siguiente:

1. El primer paso es identificar el periodo de interés sobre el que tiene efecto el evento a estudiar. La duración de este periodo variará en función de la naturaleza del suceso, pero generalmente comenzará el día de la publicación de la información y generalmente abarcará unos pocos días, aunque se puede limitar al propio día de la comunicación.

2. Se debe definir una ventana de tiempo previa al evento. A lo largo de este periodo, se estudiarán los retornos obtenidos por la empresa y se obtendrán los estadísticos a utilizar, generalmente media y varianza. Esto permitirá fijar una línea de referencia a partir de la cual se podrá determinar si el retorno obtenido ha sido anómalo. Esta ventana de tiempo debe tener una duración superior al periodo de interés.

3. Es crucial poder determinar las empresas que serán estudiadas en el análisis. Para tener en cuenta sus precios, deberán pertenecer al grupo de interés de las empresas potencialmente afectadas por el evento.

Desde una primera perspectiva general, los modelos se pueden dividir en dos categorías con respecto a su enfoque, agrupándose en técnicas estadísticas y económicas. En la primera categoría, los modelos se basan en asumir supuestos estadísticos respecto al retorno de los activos. El segundo grupo de *event staudies* se apoya en suposiciones sobre el comportamiento de los inversores y no meramente en estadísticos, aunque siguen siendo necesarios para poder ponerlos en práctica (Mckinlay, 1997).

Respecto a los estadísticos utilizados para poder aplicar estas metodologías, existen diversas aproximaciones, que se desglosarán con mayor detalle en las siguientes subsecciones: retorno de media constante, modelo del mercado, modelo de los tres factores y modelo de los cinco factores.

2.2.1. Modelo del retorno constante

Este modelo es el más sencillo de los analizados en esta sección. Fue definido por Brown y Warner en 1985. Se basa en la presunción de que la media de los retornos de una empresa es constante. De esta forma, solo será necesario comparar el retorno de los días anteriores con los del período de interés que permita obtener determinar si realmente la respuesta del mercado ha sido anómala (Brown & Warner, 1985).

A pesar de su sencillez, esta aplicación suele tener resultados similares a otros más complicados, con la ventaja que supone un modelo más simple en términos de tiempo y recursos de computación. Esto se debe a que, aunque se empleen modelos más sofisticados, la varianza de rendimientos anómalos no suele reducirse en gran medida. Además, aunque se suele utilizar sobre retornos nominales, también se puede aplicar sobre rendimientos reales (Corrado, 2011).

2.2.2. Modelo de mercado

Uno de los mayores problemas que tiene el modelo anterior es que asume que el mercado, en conjunto, se mantiene constante. Esta afirmación es rotundamente falsa, ya que los ciclos económicos determinan que el mercado tiende a subir durante un ciclo de crecimiento y, en general, bajan durante uno de recesión. Por todo esto, el segundo modelo estudiado es ligeramente más complicado que el anterior, ya que no se limita a contemplar los retornos de las empresas analizadas, sino que será necesario asociar las empresas al rendimiento del mercado en el que se encuentra.

Para poder asociar las empresas al mercado se empleará la beta del CAPM (*Capital Asset Pricing Model*) que asocia el riesgo sistémico y los retornos esperados de los activos. Es uno de los métodos más utilizados en finanzas para medir el riesgo que corren los inversores. La fórmula fundamental es:

$$R_i = R_f + \beta \cdot (ER_m - R_f) \tag{2.1}$$

En esta ecuación se identifica:

- CAPM o ER_i es el retorno esperado de la inversión.
- R_f más conocido por su nombre en inglés: *risk free rate* y es la tasa de rendimiento teórica de una inversión con riesgo cero, es decir, el interés que un inversor esperaría de una inversión sin ningún tipo de riesgo durante un periodo concreto.

- β: beta de la inversión, es un parámetro que determina cuánto riesgo tiene la empresa en relación con el mercado. De esta forma, una beta mayor que uno implica que la inversión tiene un riesgo superior al del mercado, y si es menor, implica que es más segura.
- $(ER_m - R_f)$: prima de riesgo del mercado, es el retorno esperado por el mercado por encima de la tasa de riesgo cero. El valor del retorno de mercado se obtiene a partir de los grandes índices de mercado, como el Standard & Poor 500 (S&P 500).

En este aspecto, el parámetro fundamental es la beta, que indica el nivel de riesgo de una empresa en relación al mercado. Para poder calcularla, se emplea la Ecuación 2.2.2. Donde R_i son los retornos de la empresa, R_M los retornos del mercado y $r_{(i,M)}$ la correlación entre ambos retornos.

$$\beta = \frac{Varianza\, R_i}{Varianza\, R_M} \cdot r_{(i,M)} \tag{2.2}$$

A partir de este valor, se puede crear el modelo, que explica los retornos a través de la Ecuación 2.2.2. En ella, ϵ introduce ruido blanco gaussiano a la ecuación de la recta que representa la línea regresión de los retornos de la empresa.

$$R_{it} = \alpha_i + \beta_i\, R_{Mt} + \epsilon_{it} \tag{2.3}$$

2.2.3. Modelo de los tres factores

En 1993, Fama y French (1996) publicaron una primera versión de este modelo. En su investigación, descubrieron que las acciones de valor[4] suelen tener un mejor rendimiento que las acciones de crecimiento[5] e igual sucede con las acciones de pequeña capitalización frente a las de grande. De esta forma, el modelo intenta explicar el resultado diferencial de una cartera (sin considerar la parte referida al tipo sin riesgo) en función de tres parámetros principales:

- **Diferencial de rentabilidad en el mercado**: Retorno del mercado excluyendo la tasa libre de riesgo $(R_M - R_f)$.
- **Tamaño de las acciones**: diferencia entre los retornos de carteras formados por acciones pequeñas otros de acciones de mayor tamaño. El parámetro que aparece en la ecuación es la prima de tamaño, y se designa como SMB (*Small Minus Big*).
- **Valor contable de mercado**: la diferencia entre el retorno de acciones con un *High book to market* y aquellas con un *low book to market*. Este factor se denomina prima de valor, y en la Ecuación 2.2.3 viene representado como HML (*High Minus Low*).

De esta forma, la ecuación final de este modelo quedaría representado en la Ecuación 2.2.3. En esta ecuación, los factores b_i, s_i y h_i son las pendientes de las curvas de regresión sobre una serie temporal.

[4]Acciones de valor: aquellas que se comercian a un precio relativamente bajo considerando sus dividendos, ganancias o ventas.

[5]Acciones de crecimiento: cualquier tipo de valor de una empresa que se espera que crezca a mayor ritmo que el mercado. Generalmente estas acciones no reparten dividendos, ya que las ganancias tienden a reinvertirse.

$$R_i - R_f = \alpha_i + b_i(R_M - R_f) + s_i SMB + h_i HML + \epsilon_i \tag{2.4}$$

Empleando este modelo, Fama y French llevaron a cabo pruebas con diversos tipos de portfolios diferentes, y fueron capaces de explicar hasta un 95 % del retorno de las carteras debidamente diversificadas, siendo el 5 % restante causado por el riesgo no sistemático. En su estudio, se resaltó que los inversores deberían ser capaces de librarse de la volatilidad extra y los rendimientos insuficientes que ocurren periódicamente.

El modelo de los tres factores ha sido modificado y ampliado en gran cantidad de ocasiones para incluir distintos factores. Sin embargo, en 2015, fueron los propios Fama y French quienes expandieron su modelo de los tres factores para incluir alguno más. De esta forma, crearon el **modelo de los cinco factores** (Fama & French, 2015). En esta nueva aproximación, agregaron dos nuevos factores a los tres anteriormente discutidos:

- **Rentabilidad**: según este factor, aquellas empresas que obtienen mayores ganancias futuras tienen un mejor rendimiento en el mercado de valores.
- **Inversión**: relaciona la inversión interna y los rendimientos, sugiriendo que empresas que dirigen sus beneficios hacia grandes proyectos de crecimiento son más susceptibles de experimentar pérdidas en los mercados.

3

Datos

En este apartado se procede a describir los conjuntos de datos con los que se trabajará en este proyecto. También se procederá a describir su obtención y algunas de sus características generales.

Como se ha comentado en la Capítulo 2, es fundamental tener un buen conjunto de datos de cara a elaborar un modelo de inteligencia artificial, especialmente si va a emplear aprendizaje profundo. Aquí puede residir la diferencia entre un buen modelo que obtenga resultados válidos y generalizables o la aparición de sobre-aprendizajes. En esta ocasión, se necesitarán dos tipos de datos diferentes, uno para poder elaborar el modelo de NLP y otro que contenga la información de los precios diarios de las empresas estudiadas, incluyendo la información sobre el índice de mercado y lo necesario para aplicar el modelo de estudio de eventos.

El primer conjunto de datos, utilizado para crear el modelo de inteligencia artificial, consta de diversas noticias del mundo de las finanzas obtenidas de dos periódicos distintos: *Financial times* y *The Wall Street Journal*. El dataset contiene todo el cuerpo de la noticia, la fecha y el periódico del que se ha obtenido. Todas las columnas del conjunto de datos se encuentran recogidas en la Tabla 3.1. Sin embargo, es preciso destacar que las columnas referidas al sentimiento y a la empresa fueron inmediatamente descartadas, ya que su fiabilidad no era demasiado elevada. Consta de un total de 4681 noticias.

Columna	Interpretación
Date	Fecha en la que se publicó la noticia
Media	Medio que publicó la noticia. Tiene dos valores posibles: *ft* y *wsj*
TextBody	Cuerpo de la noticia.
Company	Empresa más mencionada en la noticia.
Pos	Saldo de sentimiento positivo.
Neu	Saldo de sentimiento neutro.
Neg	Saldo de sentimiento negativo.
Compound	Valor final.

Tabla 3.1. Columnas del conjunto de datos de noticias.

La obtención y limpieza de este conjunto de datos no se realizó durante la elaboración de este proyecto, sino que los datos ya se encontraban almacenados. Para obtener esta información, se emplearon técnicas de *web scrapping*, por lo que luego fue necesario llevar a cabo una limpieza de los datos, eliminando las codificaciones erróneas de caracteres especiales y suprimiendo cualquier vestigio de código *HTML* que pudiera quedar.

Respecto a los datos en sí, narran diversos sucesos ocurridos en este periodo de tiempo a distintas empresas farmacéuticas. Entre estas empresas se encuentran algunas tan conocidas como Pfizer, Johnson & Johnson, AstraZeneca, Teva o GSK. Para poder llegar a identificarlas, se empleará un nuevo conjunto de datos que cuenta con la información mostrada en la Tabla 3.2 para cada una de las empresas presentes en bolsa en Estados Unidos.

Columna	Interpretación
Exchange	Mercado bursátil en el que opera.
Ticker	Identificador de la empresa.
Country	País del mercado.
Name	Nombre de la empresa.

Tabla 3.2. Campos del conjunto de datos responsable de la identificación de las empresas.

Es muy importante notar que el conjunto de datos de noticias no está etiquetado, es decir, no vienen dados unos sentimientos de referencia a partir de los cuales poder crear un modelo propio. Esto condicionará en gran medida el sistema, ya que no se podrá emplear un modelo generado específicamente para estos datos, como se explicará en el Capítulo 4.

Tras obtener el sentimiento, se debe detectar si existe una correlación entre estos valores y la evolución del mercado. Para ello se aplicarán técnicas de *Event Studies* (Capítulo 5). De cara a llevar este análisis a cabo, se necesitará un segundo conjunto de datos que contenga los precios de esa acción en el período de tiempo concreto.

Columna	Interpretación
PermCo	Número de identificación asignado por el CRSP (*Center for Research in Security Prices*) que identifica la empresa de manera permanente y unívoca a lo largo de todo su periodo de operación
PemNo	Número de identificación asignado por el CRSP a una empresa y que identifica las acciones en función dela empresa emisora y el tipo
Date	Fecha en la que se obtuvieron estos valores
PRC	Precio no ponderado de las acciones en el día indicado
Vol	Volumen de acciones intercambiadas en el día señalado. Volumen de operaciones.
CfaCpr	Factor de ajuste de los precios.
vwretd	*Value Weighted Return Daily*. Retorno ponderados de las acciones en el día señalado.
sprtm	Retorno del índice S&P500 en la fecha marcada.
Price	Columna añadida a posteriori que engloba el precio ajustado de las acciones por *splits*. Para obtener dicho valor se realizó la operación $^{PRC}/_{CFACPR}$

Tabla 3.3. Columnas del conjunto de datos en el que se encuentran los datos de precios.

Este segundo set estará formado por las columnas mostradas en la Este conjunto de datos tampoco se ha obtenido durante la realización del proyecto y fue proporcionado por el director. Este conjunto de datos, cuenta con columnas descritas en la Tabla 3.3.

Como se explicará en la Capítulo 5, finalmente se va a emplear el modelo de los tres factores. Para ello, será necesaria la utilización de un nuevo conjunto de datos con todos los parámetros necesarios para aplicar este modelo. A pesar de haber sido descritos ya en la Sección 2.2.3, se recogen los nombres de las columnas en la Tabla 3.4.

Columna	Interpretación
Date	Fecha de validez de esos parámetros. Recogida en formato: yyyymmdd.
$R_M - R_F$	Diferencial de rentabilidad de mercado.
SMB	Tamaño de las acciones. Rentabilidad de las pequeñas frente a las grandes.
HLM	Valor contable de mercado. Comparativa basada en el nivel de capitalización de las acciones.
R_F	Tasa de retorno libre de riesgo.

Tabla 3.4. Columnas del conjunto de datos necesario para el modelo de los tres factores.

4

Procesamiento de texto

Este capítulo detalla los pasos realizados en la realización del procesamiento de texto. Se incluye una descripción de los métodos y las herramientas empleadas. También se repasa brevemente los resultados obtenidos.

La primera pieza de este proyecto es el análisis de las noticias, para lo que se emplearán los textos contenidos en el conjunto de datos descrito en la Tabla 3.1. Este análisis busca conseguir dos hitos principales:

- Reconocer las empresas de las que se habla en la noticia, pudiendo identificar si hay más de una.
- Determinar el sentimiento de la noticia hacia cada una de las empresas señaladas.

A partir de esta información, se obtendrá un nuevo conjunto de datos que contendrá una fila por cada empresa mencionada en cada noticia, acompañado por su sentimiento positivo, negativo, subjetividad y polaridad[1].

El primer paso consiste en realizar un análisis de los datos, más allá de la mera descripción de las columnas. El conjunto de datos contiene información de noticias en un amplio periodo de tiempo: la noticia más antigua data del 8 de enero de 2008, mientras que la más reciente es del 21 de diciembre de 2017. Los textos recogidos tratan sobre empresas farmacéuticas, sin embargo, cabe destacar que la última noticia es previa a la pandemia de la Covid-19, por lo que no habrá referencia alguna ni a las vacunas ni a este periodo en general.

En una primera iteración, se decidió descartar el sentimiento contenido en el conjunto de datos, pero no las empresas destacadas. Para obtener esta columna, se había realizado una cuenta del número de veces que se mencionaban las empresas en las noticias, un método poco fiable. Al analizar sus valores se encontró que solo aparecían tres empresas distintas a lo largo de las 4681 noticias: Pfizer, GSK y Exxon. Esto terminó de confirmar que el método no era viable.

[1] Estos valores se explican en la Sección 4.3

Para conseguir obtener unos resultados correctos, se debía definir la estrategia a seguir. De acuerdo con lo explicado en la Sección 2.1, el primer paso consiste en decidir qué tipo de aprendizaje utilizar. Lo óptimo en cuanto a desempeño sería conseguir entrenar un modelo, preparado específicamente para noticias financieras. Sin embargo, para ello sería necesario tener un conjunto de datos etiquetado, con información real sobre el sentimiento con el que cuenta cada uno o, por lo menos, una etiqueta cualitativa acerca de sobre quién habla y qué desprende la noticia.

Esto podría llegar a realizarse: hay herramientas para poder clasificar y a partir de ahí empezar a trabajar, pero este etiquetado tendría que ser manual. Por ello se ha considerado que el coste en horas de trabajo de esta tarea es muy superior a su valor para esta primera aproximación. En consecuencia, debido a la dificultad de entrenar un modelo en estas condiciones, se decidió optar por herramientas pre-entrenadas.

El lenguaje de programación que se emplearía sería Python, ya que es el que cuenta con mayor número de herramientas de aprendizaje automático. En el campo del NLP existen infinidad de librerías que están ya preparadas y han sido entrenadas con una gran cantidad de información. Surgen dos herramientas principales: NLTK y SpaCy, de las cuales se realizará una comparativa en la Sección 4.1.

Respecto a los objetivos a conseguir con este desarrollo (identificación de empresas y cálculo de sentimiento) se han empleado distintos enfoques, que se recogen respectivamente en la Sección 4.2 y Sección 4.3.

4.1. Comparativa de herramientas

Puesto que se va a realizar el desarrollo a partir de herramientas pre-entrenadas, ha sido necesario llevar a cabo una comparativa de las distintas opciones disponibles. Para ello, se ha realizado una búsqueda en Internet, y se ha terminado decidiendo que la situación óptima sería emplear una combinación de librerías. Entre las más destacadas en este campo se encuentran NLTK, TextBlob o spaCy.

NLTK (*Natural Language ToolKit*) es una librería fundamental, capaz de realizar tareas como *stemming*, clasificación, análisis semántico o tokenización. Se suele emplear en entornos educativos y su funcionamiento se asemeja al de una caja de herramientas de algoritmos. En cuanto a su utilización, sus algoritmos suelen tomar como valor de entrada un *string* de información y devuelven un *string* procesado. Sin embargo, cabe destacar que a pesar de su versatilidad, es bastante lenta a la hora de ejecutarse y no es especialmente intuitiva al comenzar a utilizarla.

TextBlob es otra librería de código libre que proporciona distintas herramientas para llevar a cabo multitud de tareas relacionadas con el procesamiento de texto. Se caracteriza por ofrecer una interfaz simple para principiantes. Destaca por su sencillez y por la facilidad para crear prototipos con ella. No obstante, está basada en NLTK, por lo que mantiene su problema principal: la lentitud en el procesamiento.

Scikit-learn es una librería que proporciona gran variedad de algoritmos para la creación de modelos. Ofrece multitud de opciones a la hora de afrontar diversos problemas de clasificación. Se caracteriza por ser muy intuitiva y está muy bien documentada. Sin embargo, su procesamiento no está basado en redes neuronales, lo que hace que sea idónea únicamente para tareas sencillas de tokenización o creación de bolsas de palabras.

La última librería que merece la pena destacar es **spaCy**. Es una librería relativamente nueva que se diseñó para poder integrar procesamiento de texto en entornos de producción. Por ello, es mucho más accesible, eficiente y rápida que cualquiera de las mencionadas anteriormente. A diferencia de las estas, está diseñada orientada a objetos y proporciona un servicio que consigue resolver tareas específicas. Ofrece los mejores algoritmos para cada tarea y devuelve objetos que dejan claro en todo momento qué información contienen y cómo utilizarla, lo que no ocurre con el resto. Su documentación es intuitiva y tiene gran variedad de herramientas disponibles. Su mayor inconveniente, es la gran utilización de memoria que requiere, pero es muy eficiente debido a que está escrita en Cython[2]. Además proporciona soporte para más de 50 idiomas distintos.

Tras valorar todas estas condiciones, se descartó en primer lugar scikit-learn por su simplicidad y los problemas que podría traer al emplearla con algoritmos complejos y la diferencia entre NLTK y TextBlob se reduce a la manera de interactuar con la librería, por lo que se decidió considerarlas un pack.

La clara ganadora de esta comparativa es spaCy, por su fácil utilización, su buena documentación y la versatilidad de herramientas que ofrece. Por ello, se determinó que la base del desarrollo se realizaría empleando esta librería y que se dejaría NLTK como segundo recurso por si fuera necesario implementar algún algoritmo concreto que no se pudiera conseguir con spaCy.

El funcionamiento de spaCy se basa en la creación de un *pipeline*, una tubería que contiene todos los algoritmos que se van a utilizar especificando el orden. Para ejecutarlo, se "introduce" el texto en esta tubería y el resultado será un objeto que contiene todo el procesamiento sobre el párrafo analizado. Una vez realizado, solo habrá que estudiar los resultados obtenidos.

4.2. Reconocimiento de entidades

El primer objetivo de este desarrollo es llegar a conocer las empresas de las que habla la noticia. A priori, esta tarea puede parecer muy sencilla, pero entraña varios problemas que se fueron descubriendo a medida que se fueron probando distintas soluciones.

En primer lugar, un único texto puede contener información de más de una entidad, lo que hace que sea necesario reconocer qué dice de cada una de ellas puesto que las ideas que transmite pueden ser opuestas. Esto es muy frecuente en el caso de realizarse algún tipo de comparativa o contraposición de situaciones. Por otra parte, también es común que se hable de varios tipos de empresas o incluso otro tipo de entidad. De esta forma, será preciso identificar aquellas organizaciones que realmente interesan.

Para llevar a cabo estas tareas, se realizó un proceso iterativo, implementando distintas soluciones y comprobando los resultados con uno o dos ejemplos, de forma que se pudiera comprobar si realmente funcionaba de la manera esperada.[3] A lo largo de esta sección, se demuestra el funcionamiento de los distintos modelos empleando la primera noticia del conjunto

[2]Cython se programa como una combinacióne entre C y Python, donde se pueden declarar tipos como C y se pueden utilizar funciones desarrolladas en este lenguaje. También cuenta con un compilador que proporciona lo mejor de ambos lenguajes.

[3]Esta memoria recoge únicamente la implementación final, omitiendo intentos intermedios. De esta forma, se pretende facilitar la comprensión del lector evitando posibles confusiones con soluciones descartadas.

de datos, publicada en el *Financial Times* el 22 de julio de 2008, cuyo titular es: "A bigger dose: Why generic drug producers are bulking up". [4]

Para afrontar este problema se optó por utilizar la librería spaCy, por los motivos mencionados en la Sección 4.1. Para ello, el *pipeline* incluía los siguientes algoritmos, representados también en la Figura 4.1:

- Tokenizar: segmentación del texto en tokens, del tamaño de palabras/caracteres.
- Tagger: asignación de las etiquetas POS (*Part Of Speech*, que indican la funcionalidad gramatical de la palabra estudiada. Incluye información como la palabra anterior, la posterior o si la primera letra está en mayúsculas. Con esta información, termina determinando el tipo de palabra que es: sustantivo, adjetivo, verbo, adverbio...
- Parser: señala la dependencia del token estudiado con los que le rodean.
- NER: reconocimiento de entidades presentes en el texto.
- Lemmatizador: obtención de los lemmas de cada palabra.
- Textcat: agrega etiquetas categorizando el documento.

Estos son los elementos que se aplican al texto, pero se hace todo de golpe tras una única instrucción. Como se indicó en el Capítulo 2, el orden no suele importar a la hora de aplicar estos algoritmos, por lo que a partir de este momento se hablará de los resultados obtenidos independientemente del orden en el que se hayan considerado.

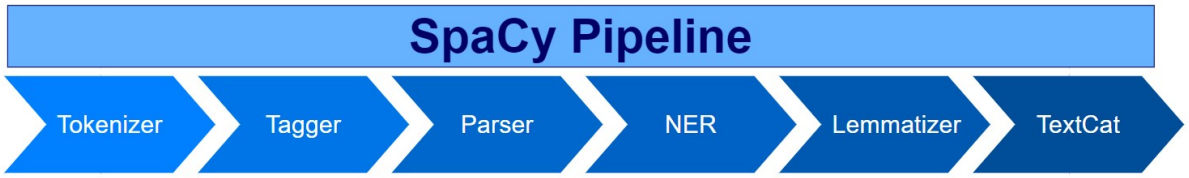


Figura 4.1. Diagrama de los componentes del pipeline de spaCy.

El primer paso consiste en tokenizar el cuerpo de la noticia. La empleada como ejemplo cuenta con un total de 1951 tokens, agrupados en un total de 68 frases diferentes. Desde este momento, la mayoría de los ejemplos se realizarán con una frase aleatoria del texto, concretamente la número 20 (Figura 4.2).

```
A watershed came last month, when Pfizer, the US-based company that is the world's largest pharmaceutical group,
was forced to reach a settlement over extensive legal challenges to Lipitor, a blood-thinning drug that generates
nearly $13bn a year in sales.
```

Figura 4.2. Frase de ejemplo sobre la que se muestra la información recuperada.

Mediante la aplicación de este *pipeline* se consiguió obtener información acerca de los tipos de palabras que forman el texto, a nivel semántico y sintáctico. A nivel semántico, se obtiene la clasificación de la palabra en función de su naturaleza (sustantivo, adjetivo, verbo, adverbio, preposición...). La Figura 4.3 muestra esta información.

[4]Puede visitar el artículo completo o visitar la siguiente URL: https://www.ft.com/content/72671542-5815-11dd-b02f-000077b07658

Word	Semantic Function	Lemma
A	DET	a
watershed	NOUN	watershed
came	VERB	come
last	ADJ	last
month	NOUN	month
,	PUNCT	,
when	ADV	when
Pfizer	PROPN	Pfizer
,	PUNCT	,
the	DET	the
US	PROPN	US
-	PUNCT	-
based	VERB	base
company	NOUN	company
that	DET	that
is	VERB	be
the	DET	the
world	NOUN	world
's	PART	's
largest	ADJ	large
pharmaceutical	NOUN	pharmaceutical
group	NOUN	group
,	PUNCT	,
was	AUX	be
forced	VERB	force
to	PART	to

Word	Semantic Function	Lemma
reach	VERB	reach
a	DET	a
settlement	NOUN	settlement
over	ADP	over
extensive	ADJ	extensive
legal	ADJ	legal
challenges	NOUN	challenge
to	ADP	to
Lipitor	PROPN	Lipitor
,	PUNCT	,
a	DET	a
blood	NOUN	blood
-	PUNCT	-
thinning	VERB	thin
drug	NOUN	drug
that	DET	that
generates	VERB	generate
nearly	ADV	nearly
$	SYM	$
13bn	NOUN	13bn
a	DET	a
year	NOUN	year
in	ADP	in
sales	NOUN	sale
.	PUNCT	.

Figura 4.3. Muestra de cada palabra de la frase ejemplo con su funcionalidad semántica y la raíz de la palabra.

Utilizando esta información y aplicando nuevos algoritmos, el modelo también realiza un reconocimiento de todas las entidades del texto, agrupándolas en distintas categorías. La Figura 4.4 muestra todo el artículo con las entidades reconocidas señaladas en él. Los organismos reconocidos se agrupan en las siguientes categorías[5]:

- Person: personas, incluyendo aquellas que pudieran ser ficticias.
- GPE: entidad geopolítica.
- Norp: nacionalidades o grupos políticos o religiosos.
- ORG: empresas, agencias, instituciones... Organizaciones en general.
- Product: objetos de distintos tipos, diferencia entre productos o servicios.
- Date, Time: Date se emplea para periodos de tiempo mientras que time hace referencia a medidas inferiores al día.
- Cardinal, Ordinal, Money: números en distintos formatos o con distintas unidades.
- LOC: localización geográfica (montañas, lagos, continentes...).

[5]La librería es capaz de reconocer alguna entidad más, como FAC (infraestructuras), WORK_OF_ART (títulos de libros, canciones...) o LAW (documentos relacionados con las leyes) pero son menos comunes en el contexto tratado.

Figura 4.4. Reconocimiento de todas las entidades del texto categorizadas según su tipo.

Una vez identificadas todas las entidades, se debe reconocer aquellas que realmente son industrias farmacéuticas. Inicialmente se consideró la opción de buscar aquellas organizaciones más comúnmente mencionadas en el texto, pero tras unas pequeñas pruebas iniciales se terminó por descartar. En la noticia de ejemplo, las más mencionadas son: "US" (10 menciones), "Barr" (4 menciones), "India" (3 menciones) o Europa (3 menciones), por lo que, claramente, este método no devolvía los resultados esperados.

Para ello se comenzó por estudiar en varios artículos en qué categorías se suele incluir estas empresas, y se ha determinado que la mayoría de ellas se encuentran identificadas como Person u Org. En la Figura 4.4 se observan ambos casos: como **Person** se reconocen Pfizer o Teva, como **ORG** está AstraZeneca y como **Product** se encuentra Novartis. En otros artículos, otras empresas aparecen como **GPE**, pero el porcentaje es muy inferior.

Sobre esta selección de entidades de tipo **Person**, **ORG** y **Product** se ha comparado la lista de empresas farmacéuticas presentes en bolsa[6] y aquellos nombres coincidentes se han tomado como entidades válidas. Para evitar problemas derivados de abreviaturas en la denominación de las empresas[7] se ha llevado a cabo una doble comparativa para reducir la aparición de falsos negativos.

Al lector podría surgirle la duda sobre por qué realizar todo este análisis de entidades para terminar utilizando fuerza bruta en el reconocimiento. Es decir, por qué no realizar directamente la comparativa de la lista de empresas sobre la totalidad del texto. Hay dos razones principales para esto. Por una parte, llevar a cabo ese ejercicio supondría una cantidad enorme de procesamiento, ya que, solo considerando un primer artículo, habría que comparar las casi 2000 palabras que lo forman con las 835 posibles empresas que hay.
Por otra parte, hay que tener en cuenta que las entidades no siempre están formadas por una única palabra, igual que los nombres de las empresas (Johnson & Johnson). Esta situación podría derivar en falsos positivos, duplicando los resultados y adulterando así las conclusiones.

Aplicando esta metodología, en el artículo de ejemplo se encuentran reconocidas las entidades mostradas en la Figura 4.5. En la imagen se muestra en primer lugar la raíz de la entidad identificada en el texto y a continuación los nombres por los que se identifica a la empresa en función del dataset de empresas farmacéuticas.

```
Pfizer
['Pfizer Inc.', 'Pfizer Ltd']
Pfizer
['Pfizer Inc.', 'Pfizer Ltd']
AstraZeneca
['AstraZeneca Plc']
Johnson & Johnson ’s
['Johnson & Johnson']
Sanofi - Aventis
['Sanofi', 'Sanofi', 'Sanofi', 'Sanofi', 'Sanofi']
Novartis
['Novartis AG', 'Novartis AG', 'Novartis AG']
Novartis
['Novartis AG', 'Novartis AG', 'Novartis AG']
Teva
['Teva Pharmaceutical Finance Netherlands III BV', 'Teva Pharmaceutical Industries Ltd', 'Teva Pharmaceutical Industries Ltd.', 'Teva Pha
rmaceutical Industries Ltd.', 'Teva Pharmaceutical Industries, 7.00% Mandatory Convertible Preferred Shares']
```

Figura 4.5. Resultado de las empresas farmacéuticas identificadas en el primer artículo.

Este proceso se ha automatizado para ejecutarlo sobre todas las noticias, obteniendo un nuevo conjunto de datos donde se reconocen las empresas de cada noticia para poder utilizarlas más adelante en el cálculo de sentimiento. Inicialmente se pretende llevar a cabo todo el proceso seguido, pero se toma este punto como valor intermedio.

[6]Este conjunto de datos es el descrito en la Tabla 3.2.
[7]Por ejemplo, el nombre oficial de Pfizer es "Pfizer Inc.", pero es poco común que se utilice completo

4.3. Cálculo de sentimiento

Para poder calcular el sentimiento de las entidades reconocidas, se empleará la información recogida por el *pipeline* representado en la Figura 4.1. Además de toda la información mostrada en la sección anterior, al procesar el texto también se consigue una división sintáctica, clasificando las palabras en cuanto a su funcionalidad sintáctica.

La Figura 4.6 recoge la frase de ejemplo empleada en la sección anterior (Figura 4.2) reconociendo qué función desempeña cada palabra. Se pueden encontrar adverbios modales, sujetos nominales, aposiciones... A un estudiante de lengua castellana le pueden resultar extrañas algunas de estas funcionalidades, pero es necesario resaltar que el análisis sintáctico no es igual en todos los idiomas. Se mantienen algunos elementos fundamentales, pero la clasificación presenta variaciones.

Word	Function
A	det
watershed	nsubj
came	ROOT
last	amod
month	npadvmod
,	punct
when	advmod
Pfizer	nsubjpass
,	punct
the	det
US	npadvmod
-	punct
based	amod
company	appos
that	nsubj
is	relcl
the	det
world	poss
's	case
largest	amod
pharmaceutical	compound
group	attr
,	punct
was	auxpass
forced	advcl
to	aux

Word	Function
reach	xcomp
a	det
settlement	dobj
over	prep
extensive	amod
legal	amod
challenges	pobj
to	dative
Lipitor	pobj
,	punct
a	det
blood	npadvmod
-	punct
thinning	amod
drug	appos
that	nsubj
generates	relcl
nearly	advmod
$	nmod
13bn	dobj
a	det
year	npadvmod
in	prep
sales	pobj
.	punct

Figura 4.6. Muestra de cada palabra de la frase ejemplo con su funcionalidad sintáctica.

Utilizando toda esta información, se puede llegar a realizar un análisis sintáctico completo de la oración, estableciendo las dependencias de unas palabras con otras. Así se pueden reconocer los distintos sintagmas de la oración. Este análisis queda representado en la Figura 4.7. En ella se muestra la oración completa, indicando el tipo de palabra que es y el término al que complementa.

Se puede observar que la palabra que no complementa a ninguna es el primer verbo: "came", del que depende el verbo de la primera subordinada: "forced". De esta forma se van desglosando los distintos sintagmas y su complemento original.

Esta información es útil para determinar qué sintagmas afectan a las entidades reconocidas, evitando cualquier tipo de confusión posible en caso de que aparezca más de una empresa en cada oración.

De cara a estudiar el sentimiento relacionado con una empresa, se ha creado una *Bag of Words* (BOW) que contuviera las palabras presentes en la oración en la que se encuentra la entidad. Inicialmente se consideró la posibilidad de tomar únicamente aquellos sintagmas que influyeran directamente en la palabra de interés, pero terminaba siendo complicado determinar en qué momento comenzar a descartar parte de la frase.

Es decir, la BOW de cada empresa estará formada por aquellas palabras presentes en la oración en la que se mencione excluyendo las *stop words*[8], ya que no aportan sentimiento a la oración e implicará mayor tiempo de procesamiento. Cabe destacar, que en caso de que una empresa se mencione en más de una ocasión, aparecerán dos BOW independientes, cuyo sentimiento se fusionará más adelante.

Una vez preparada esta bolsa de palabras, es necesario analizar la forma de llevar a cabo este análisis de sentimiento. Como se mencionó en la Sección 2.1.1.3, se debe seleccionar con mucho cuidado la librería a utilizar, ya que cada campo de estudio tiene un argot y unas connotaciones concretas para diferentes términos. El diccionario destacado en el mundo de las finanzas es el LM.

Para poder aplicarlo, se utilizó el módulo *pysentiment2*[9]. Esta librería permite aplicar varios diccionarios del campo de las finanzas, como puede ser el *Harvard-IV* o el LM, que será el utilizado en este caso. El módulo cuenta con un tokenizador propio en el que se puede introducir el texto y este resultado se introduce en el evaluador de sentimiento. Sin embargo, para aprovechar todo lo creado anteriormente, se introducirá directamente la BOW creada a raíz de la frase.

Esto devolverá el sentimiento de la bolsa insertada, incluyendo los siguientes parámetros: **pos** que representa el valor positivo, **neg**, puntuación negativa del "texto" considerado, **polarity**, que representa la polaridad del texto, indicando la fuerza de los tokens positivos frente a los negativos del total de los tokens no-neutros. Se define en la Ecuación 4.1. Finalmente, **subjectivity**, que indica la subjetividad del texto, ponderando cuántos tokens tienen sentimiento frente a aquellos neutros. Su valor se calcula con la Ecuación 4.2.

$$Polarity = \frac{N_{pos} - N_{neg}}{N_{pos} + N_{neg}} \tag{4.1}$$

[8] Sección 2.1.1.1

[9] Para más información sobre la librería, pulse aquí o siga el siguiente link: https://pypi.org/project/pysentiment2/

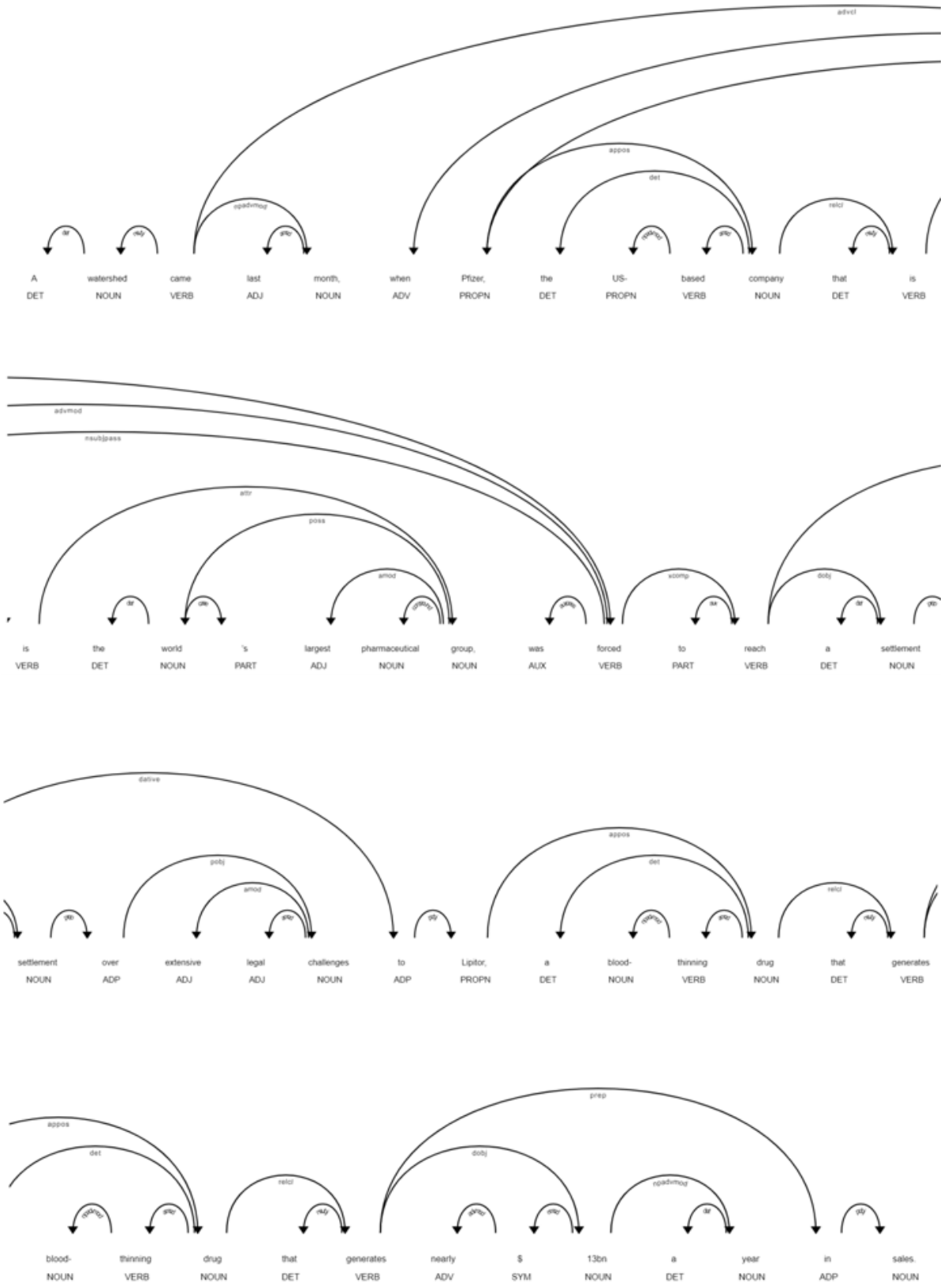


Figura 4.7. Análisis sintáctico de la oración de ejemplo.

$$Subjectivity = \frac{N_{pos} + N_{neg}}{N} \tag{4.2}$$

se estaba creando una BOW para cada mención a una empresa, lo que devuelve un sentimiento por cada mención a la misma. Para solucionar esto, tras realizar el análisis de sentimiento se han combinado ambos resultados. Los valores pos y neg se han sumado, mientras que en los campos de polaridad y subjetividad se ha llevado a cabo una media aritmética. Toda esta información se recopilará para cada una de las noticias, y se almacenará en un documento en formato *json*, que contenga estos cuatro parámetros de sentimiento para cada una de las empresas mencionadas a lo largo de la noticia.

4.4. Comparativa

El siguiente paso ha sido llevar a cabo una comparativa entre los sentimientos presentados por ambos periódicos para una misma empresa en un mismo día. Se han distinguido cuatro categorías:

- Igual sentimiento, positivo: ambos periódicos coinciden en dar un sentimiento positivo.
- Igual sentimiento, negativo: ambos diarios presentan un sentimiento negativo.
- Distinto sentimiento: uno de ellos presenta un sentimiento positivo y el otro negativo.
- Solo informa uno de los dos diarios.

Para ello se ha iterado sobre los resultados obtenidos por ambos periódicos, almacenando en un nuevo conjunto de datos las fechas en las que tienen lugar cada una de las situaciones anteriores. Los resultados de esta segunda parte se describen en la Sección 6.3.

5

Estudios de eventos

El quinto capítulo engloba todo lo relativo al estudio de los mercados bursátiles en la fecha señalada. Se comenta el estudio del efecto causado por cada noticia en los precios de las acciones de las empresas implicadas. Para ello, se aplica la metodología de *event studies* en los valores de mercado.

El segundo gran bloque de este proyecto es reconocer si los retornos obtenidos por las empresas fueron los esperados a pesar de las noticias publicadas en aquel día. Para determinarlo, se aplica la metodología de los *event studies*, descrita en la Sección 2.2. De los modelos destacados en esta sección se ha optado por implementar el de la Sección 2.2.3, el modelo de los tres factores de Fama y French (1996) ya que actualmente es el estándar en finanzas.

A la hora de implementar un modelo de este estilo, se deben considerar dos períodos fundamentales: la ventana y la influencia. La ventana estudia los retornos obtenidos a lo largo de un determinado período de tiempo para poder establecer la normalidad en los retornos, es decir, una línea base que permita determinar lo excepcional de lo observado en el día indicado.

El período de influencia aparece debido a que un gran evento tiene mucha repercusión y su influencia se puede mostrar durante varios días o semanas. Por ello, se hace necesario determinar una franja temporal en la que poder estudiar estos efectos.

En el caso estudiado, se ha fijado la ventana inicial en un año financiero, y la de influencia de un día. Para el tamaño de la ventana se ha valorado utilizar un período de entre uno y dos años. Finalmente, para reducir al máximo el número de interferencias entre eventos, se decidió disminuir el tamaño de esta franja al mínimo[1]. Respecto a la influencia, en esta ocasión se ha optado por dejarlo en un día. Esta decisión se debe a que en la mayoría de los casos, se estudian noticias sin gran importancia y sus efectos se ven limitados al día de publicación.

Para poder aplicar el modelo de los tres factores, se ha recopilado la información necesaria. Se utilizarán dos conjuntos de datos ya descritos en la Capítulo 3. En primer lugar, será necesaria la información del desempeño de las empresas a estudiar, conjunto de datos descrito en la

[1]El tema de las interferencias se comentará en la Capítulo 8.

Tabla 3.3, que cuenta con los precios diarios de las acciones. Los retornos se calcularán siguiendo la Ecuación 5.1, y será la información empleada para sustituir las variables R_i. Este factor de ajuste de precios recoge aquellos elementos que hacen variar el número de acciones a lo largo del tiempo, permitiendo así la comparación de los precios.

El segundo dataset necesario es el de la Tabla 3.4 que recoge los valores de todos los parámetros necesarios para aplicar este modelo: SMB, HML, R_F y $R_M - R_F$ para cada una de las fechas del período en el que se enmarcan las noticias.

$$AdjPrice = \frac{PRC}{CfaCpr} \rightarrow Return = \frac{AdjPrice_t}{AdjPrice_{t-1}} - 1 \tag{5.1}$$

Trayendo de nuevo la ecuación de este modelo (Ecuación 5.2), los únicos elementos que quedarían por calcular son los valores representados en la ecuación con la forma x_i y que representan el efecto de ese parámetro concreto sobre la empresa.

$$R_i - R_f = \alpha_i + b_i \cdot (R_M - R_f) + s_i \cdot SMB + h_i \cdot HML + \epsilon_i \tag{5.2}$$

Para calcularlos se ha llevado a cabo una regresión linear multivariante tomando los datos de un año financiero de los parámetro SMB, HML, $R_M - R_f$ y los retornos de la empresa hasta la fecha. El único valor restante, ϵ_i, se omite ya que es un ruido gaussiano destinado a ajustar el modelo. Esta operación se ha llevado a cabo con la librería *statsmodels* de Python, empleando la función OLS que utiliza el método de los mínimos cuadrados ordianrios.

Se sustituyen los valores en la ecuación principal (Ecuación 5.2), dejando como incógnita el valor R_i y el resultado será el valor esperado de los retornos de la empresa para el día señalado: $E(R_i)$. Se sustituirá en la ecuación con los mismos multiplicadores para todos los días del periodo de la ventana, de esta forma que se pueda comprobar si el resultado es constante.

Con estos valores ya determinados, se procede a compararlos con el valor real ocurrido en el día. La diferencia entre ambos (Ecuación 5.3) devuelve el exceso de retorno (*excess return*) o el retorno anormal (*abnormal return* - AR).

$$R_i - E(R_i) = AR \tag{5.3}$$

En este punto, se procede a comprobar que la media de los AR obtenidos antes del evento es igual a la obtenida con posterioridad. Para ello se plantea el contraste de hipótesis representado en la Ecuación 5.4.

$$\begin{aligned} H_0 &: \frac{1}{M}\sum_{j=1}^{M} AR_{t-j} = \frac{1}{N}\sum_{j=1}^{N} AR_{t+j} \\ H_1 &: \frac{1}{M}\sum_{j=1}^{M} AR_{t-j} \neq \frac{1}{N}\sum_{j=1}^{N} AR_{t+j} \end{aligned} \tag{5.4}$$

En esta ecuación la hipótesis nula (H_0) será que la media en el AR antes y después del evento son iguales y la alternativa (H_1) afirmará que son distintos. En este caso, M es el número de días considerados antes del evento y N la cantidad días posterior.

Como se ha indicado con anterioridad, en este estudio se lleva a cabo una gran simplificación, puesto que el período de influencia se reduce al día del evento. Por eso, en esta ocasión N sería 1, por lo que el contraste de hipótesis se simplifica notablemente. Su final se muestra en la Ecuación 5.5.

$$\begin{aligned} H_0 : \frac{1}{M}\sum_{j=1}^{M} AR_{t-j} = AR_t \\ H_1 : \frac{1}{M}\sum_{j=1}^{M} AR_{t-j} \neq AR_t \end{aligned} \tag{5.5}$$

Para poder resolver este contraste, se calculará la varianza de los AR previos al evento. Para ello se empleará la Ecuación 5.6, donde L_1 es el primer día de la ventana previa y L_2 es el último, es decir, un día antes del evento.

$$\sigma^2 = \frac{1}{L_2 - L_1 - 1}\sum_{j=L1}^{L2} (AR_{i,j} - E(R_i)_j)^2 \tag{5.6}$$

Con esta información, ya se puede calcular el valor del estadístico empleando el retorno real de la empresa en el día del evento y la varianza de los días de la ventana. La ecuación empleada para ello está representada en la Ecuación 5.7.

$$Z = \frac{X - \mu}{\sigma} = \frac{R_i - \overline{ER_i}}{\sigma} \tag{5.7}$$

De acuerdo con el estudio de MacKinlay (1997) resultado se comparará con los valores de una distribución normal de media 0 y desviación σ_i, calculada como se muestra en la Ecuación 5.8.

$$\sigma_i^2 = (L_2 - L_1 + 1) * \sigma^2 \tag{5.8}$$

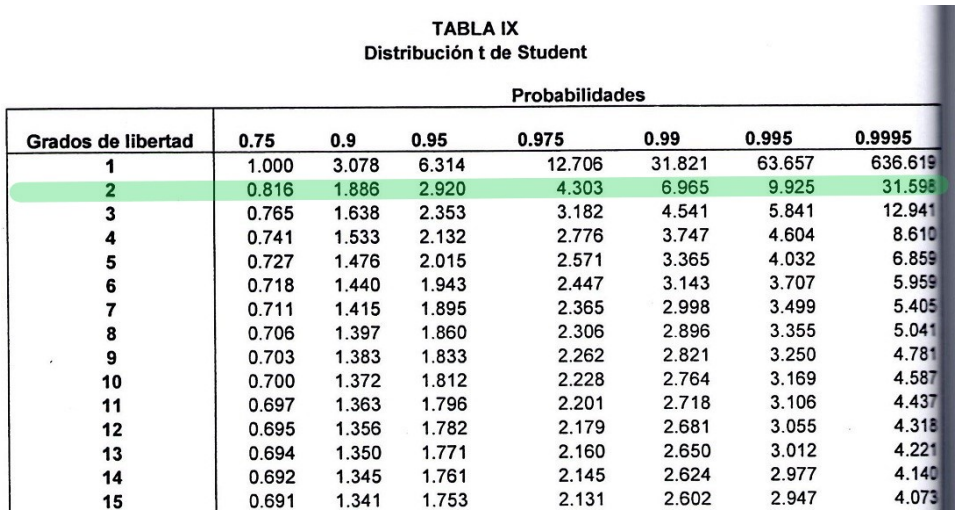


TABLA IX
Distribución t de Student

	Probabilidades						
Grados de libertad	0.75	0.9	0.95	0.975	0.99	0.995	0.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	6.859
6	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	5.405
8	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	4.073

Figura 5.1. Tablas de valores de la distribución t de Student.

Sin embargo, se debe tener en cuenta que para calcular este valor se están estimando dos parámetros: la media y la varianza, por lo que de acuerdo con Corrado (2011), deberá seguir una distribución t de Student de 2 grados de libertad, cuyos valores se muestran en la Figura 5.1. Para automatizar esta comprobación de valores, se ha empeleado el módulo *scipy.stats* de Python, cuya función CDF devuelve este valor.

Con esta comparativa se obtendrá el grado de significación de este resultado, al que se puede rechazar o no rechazar la hipótesis nula y por ello afirmar que el retorno obtenido se sale de lo que se podría considerar normal.

Esta implementación no considera los problemas derivados del *clustering*, es decir, la posibilidad de que se solapen los distintos eventos afectando a la ventana de estudio previa (Corrado, 2011). Se han decidido omitir los efectos de estas situaciones puesto que queda fuera del alcance de este proyecto.

5.1. Segunda parte

En esta sección se pretende conseguir clasificar los eventos en función de si la reacción de los periódicos es igual o distinta. Para ello se obtendrán tres eventos distintos que engloban todos los anteriores, tomando los resultados de la Sección 6.3:

- Mismo sentimiento en ambos periódicos con tendencia positiva.
- Igual sentimiento en los dos con polaridad negativa.
- Distinto sentimiento.

En estos grupos se clasifican las fechas en las que publicaron ambos periódicos para la misma empresa. Para estos momentos concretos, se utilizará el modelo diseñado en la sección anterior y las Ecuación 5.9 y Ecuación 5.10 para obtener el $\overline{CAR}$ y su varianza. En estas ecuaciones, N representa el número de eventos ocurridos en ese grupo.

$$\overline{CAR} = \frac{1}{N}\sum_{i=1}^{N}\overline{AR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=L_1}^{L_2} AR_j \tag{5.9}$$

$$var(\overline{CAR}) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma_i^2 \tag{5.10}$$

Utilizando estos dos resultados, se obtendrá un nuevo estadístico, que de nuevo seguirá una distribución t de Student de dos grados de libertad. Este nuevo valor θ se obtendrá a través de la ecuación Ecuación 5.11.

$$\theta = \frac{\overline{CAR}}{var(\overline{CAR})} \tag{5.11}$$

6
Resultados

En este capítulo se detallan los resultados obtenidos a lo largo de los desarrollos principales de este proyecto, mostrando los puntos más significativos del análisis a través de distintas gráficas.

Tras realizar todo el desarrollo descrito en el Capítulo 4 y el Capítulo 5 solo falta analizar la información obtenida. En este capítulo no se estudia la eficiencia del desarrollo, puesto que se han utilizado herramientas pre-entrenadas. Se examinan los resultados devueltos, intentando señalar aquellos puntos más relevantes, comparando distintos parámetros calculados.

En primer lugar se analizan los resultados del procesamiento de texto, a continuación los de los *Event studies* y finalmente se busca una combinación de ambos.

6.1. Análisis de sentimiento

Esta subsección analiza aquellos resultados que se pueden obtener a partir del análisis de sentimiento de las noticias, comentando la evolución en el tiempo de los distintos valores obtenidos. Cabe destacar que en el caso de los gráficos bicolores, el azul representará las noticias publicadas por el *Financial Times* y las naranjas por el *The Wall Street Journal*. Como se comentó en la Sección 4.3, el modelo diseñado devuelve cuatro parámetros fundamentales: positivo, negativo, subjetividad y polaridad.

Para analizar los resultados obtenidos, se ha optado por realizar una serie de gráficos, utilizando Tableau, un *software* de visualización de datos. El primer gráfico a estudiar es el de la Figura 6.1.

Lo primero que llama la atención es el pico de menciones que aparece el 2 de mayo de 2014, fecha en la que AstraZeneca rechazó de manera definitiva la propuesta de Pfizer para unirse y crear la mayor farmacéutica del mundo[1].

[1]Link a la noticia: pinche aquí o siga esta URL: https://www.elcomercio.es/economia/empresas/201405/19/britanica-astrazeneca-rechaza-oferta-20140519105907-rc.html

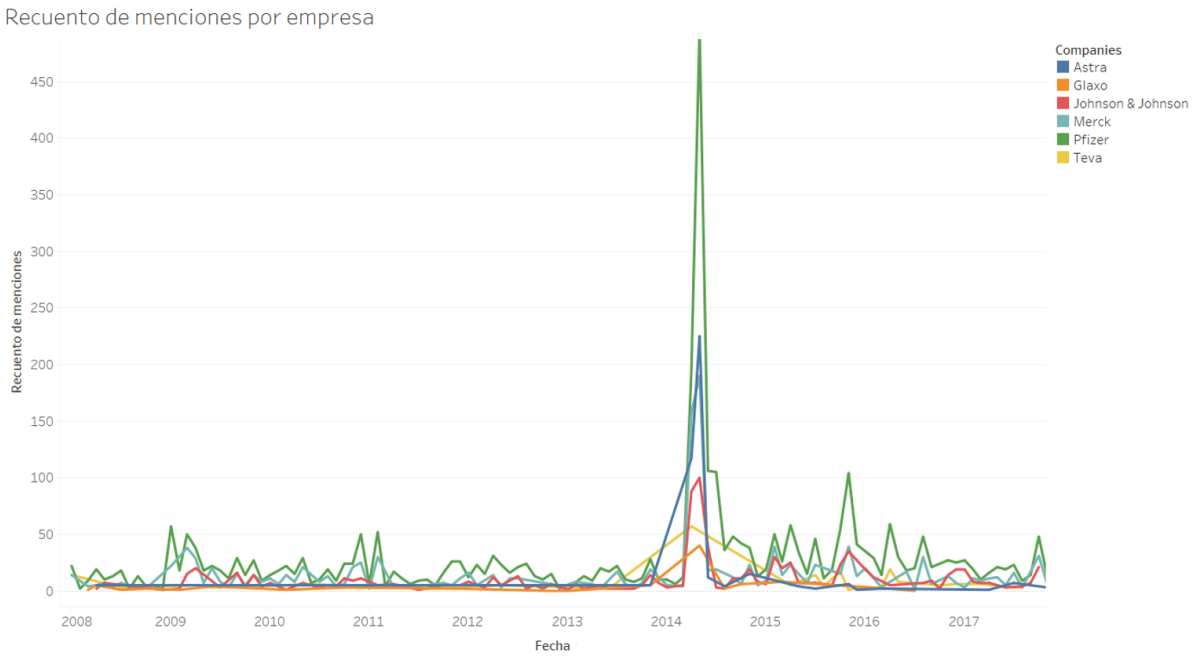


Figura 6.1. Número de menciones de las distintas empresas a lo largo del eje cronológico.

En general, a lo largo del tiempo queda claro que la empresa con mayor número de menciones es Pfizer, seguida de AstraZeneca. También cabe destacar que en general, el número de referencias tampoco suele ser muy elevado, a excepción de algunos casos concretos. De hecho, hay días en los que no hay mención a estas empresas.

En la Figura 6.2 se muestra la polaridad presentada por las distintas noticias, coloreadas en función de la empresa. La polaridad es un factor que compara el tipo de sentimiento (positivo o negativo) frente al total de tokens con sentimiento, sin considerar aquellos neutros (Ecuación 4.1). Varía entre -1 y 1, donde -1 representa el nivel máximo de sentimiento negativo y el 1 nivel máximo de sentimiento positivo. El nivel central, 0, indica que la redacción es neutral, de forma que las facetas positivas mencionadas están al mismo nivel que las negativas.

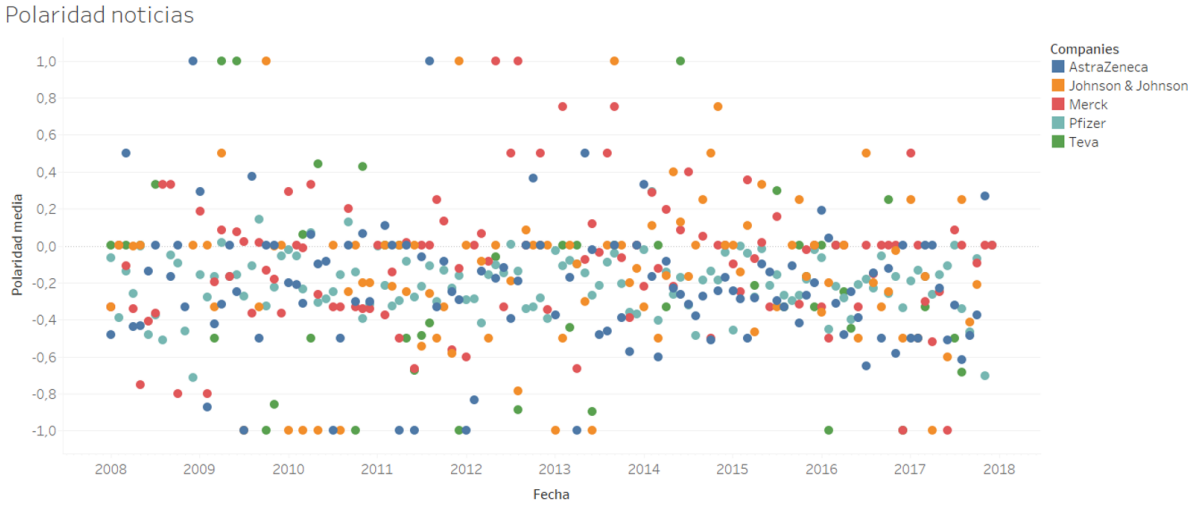
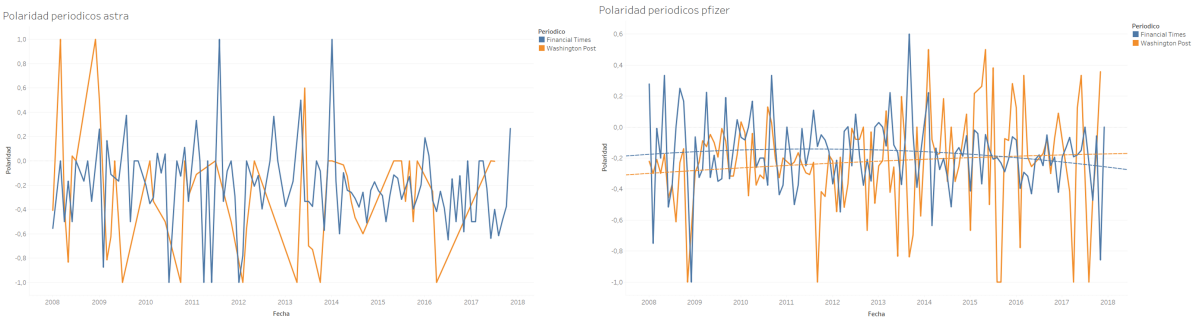


Figura 6.2. Polaridad presentada por las menciones clasificadas por empresa.

Este gráfico demuestra que, salvo en contadas excepciones, la mayoría de las noticias se encuentran entre $-0{,}5$ y $0{,}5$, lo que indica que la polaridad no suele ser muy elevada. La empresa que más aparece entre las máximas polaridades es "Johnson & Johnson", especialmente en el sentimiento negativo.

Sin embargo, al haber tanta información de diferentes fuentes, no es posible analizar nada en gran detalle. Por este motivo se ha decidido llevar a cabo un pequeño desglose de esta información. Se han realizado dos gráficos extra, uno con cada una de las empresas más mencionadas: Pfizer y AstraZeneca.

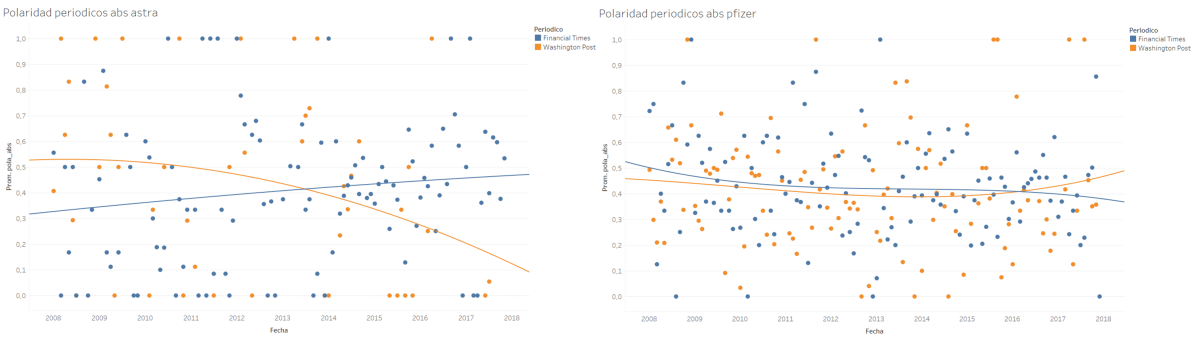
Primero se representó separando estos valores de los del resto de las empresas para extraer algún tipo de conclusión (Figura 6.3). Sin embargo, al tener ambos sentimientos no es posible encontrar un patrón de comportamiento que permita entrever si la polaridad de las noticias sigue una tendencia concreta.



(a) **(b)**

Figura 6.3. Valor absoluto de la polaridad y líneas de tendencia para las empresas (a) AstraZeneca y (b) Pfizer.

Por ello, se realizó la Figura 6.4, que muestra el valor absoluto de los valores de polaridad obtenidos por AstraZeneca y Pfizer respectivametne. En el caso de Pfizer se puede observar cómo ambos periódicos han mantenido una postura similar, muy centrada en una posición intermedia estable a lo largo de estos 10 años estudiados. Sin embargo, a finales de 2017 se empieza a observar un cambio en esta tendencia, donde el " The Wall Street Journal" parece aumentar su sesgo y el "Financial Times" disminuirlo.



(a) **(b)**

Figura 6.4. Valor absoluto de la polaridad y líneas de tendencia para las empresas (a) AstraZeneca y (b) Pfizer.

Respecto a AstraZeneca, la situación es completamente distinta. En el caso del "Financial Times" la polaridad presenta un aumento sostenido en el tiempo, mientras que el "The Wall Street Journal" muestra una reducción drástica. Sin embargo, esta gráfica puede llevar a engaño, ya que realmente hacia los últimos años representados apenas hay menciones de esta empresa en el diario estadounidense y eso acentúa la caída. En cualquier caso es cierto que las pocas noticias que aparecen, presentan polaridades cercanas a cero.

El siguiente parámetro estudiado es la subjetividad, que varía entre cero y uno. Este factor analiza la cantidad de palabras que presentan sentimiento frente a aquellas que no lo tienen (Ecuación 4.2). Si el valor del parámetro es uno, implica que todas las palabras consideradas tienen sentimiento, mientras que un valor cero, indicaría que ninguna lo tiene.

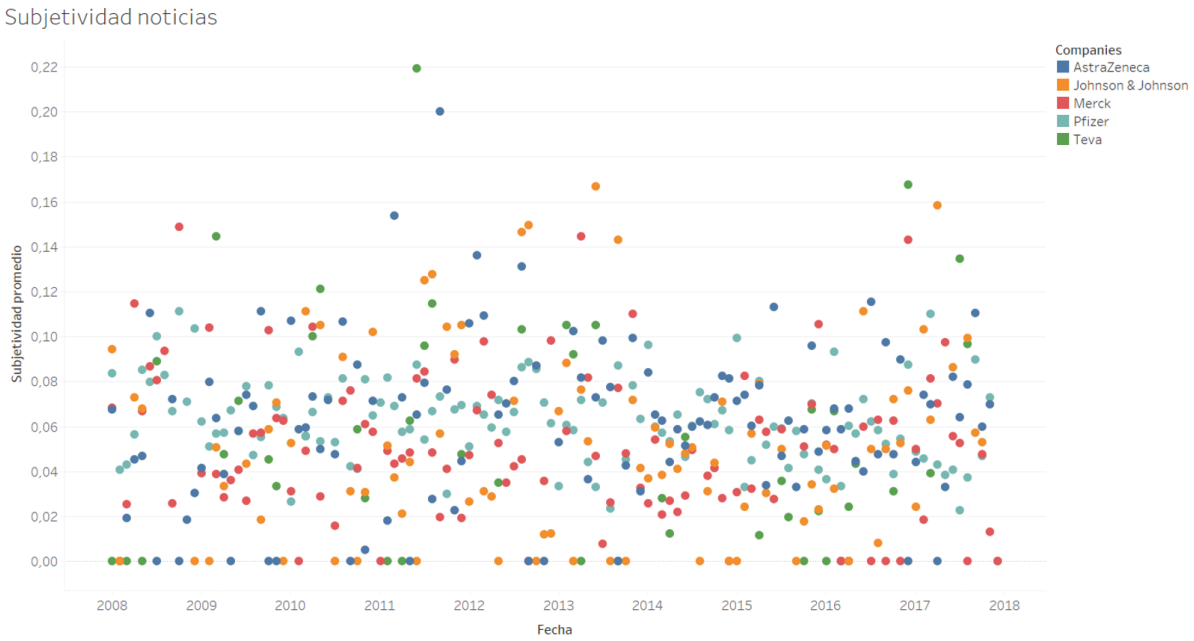
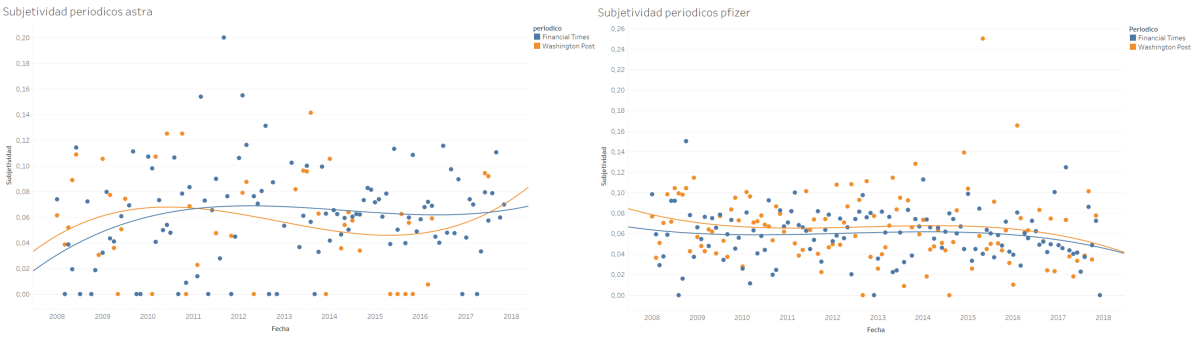


Figura 6.5. Subjetividad presentada por las menciones clasificadas por empresa.

La Figura 6.5 muestra el valor de la subjetividad obtenida por las noticias clasificado por empresas. En primer lugar llama la atención que el valor más elevado es inferior a $0{,}25$, y la mayor parte de los valores se encuentran por debajo de $0{,}15$. Con estos datos se puede afirmar que la subjetividad de estos textos es muy baja, en la mayoría de los casos despreciable.



(a) **(b)**

Figura 6.6. Subjetividad y sus líneas de tendencia para las empresas (a) AstraZeneca y (b) Pfizer.

De nuevo, con la intención de llevar a cabo un desglose más pormenorizado, se han separado los valores recibidos por Pfizer y AstraZeneca. La Figura 6.6 muestra la subjetividad de las noticias para AstraZeneca y Pfizer respectivamente, separando las menciones en función del periódico que las realizara.

Se puede apreciar que los valores presentados por los distintos periódicos no es constante en tiempo, y a su vez, que no sigue la misma tendencia al enfrentarse a distintas empresas. Mientras que al referirse a Pfizer este valor lleva 10 años descendiendo, en el caso de AstraZeneca está en ascenso. En cualquier caso, la diferencia entre estos valores no es relevante, puesto que sus cotas se encuentran próximas a cero.

El último gráfico, Figura 6.7 presenta las puntuaciones positivas y negativas obtenidas por cada noticia. En ambos casos se observa que la mayoría de las noticias no tienen unas valoraciones muy elevadas. Sin embargo, cabe destacar que las máximas negativas son más altas que las máximas negativas.

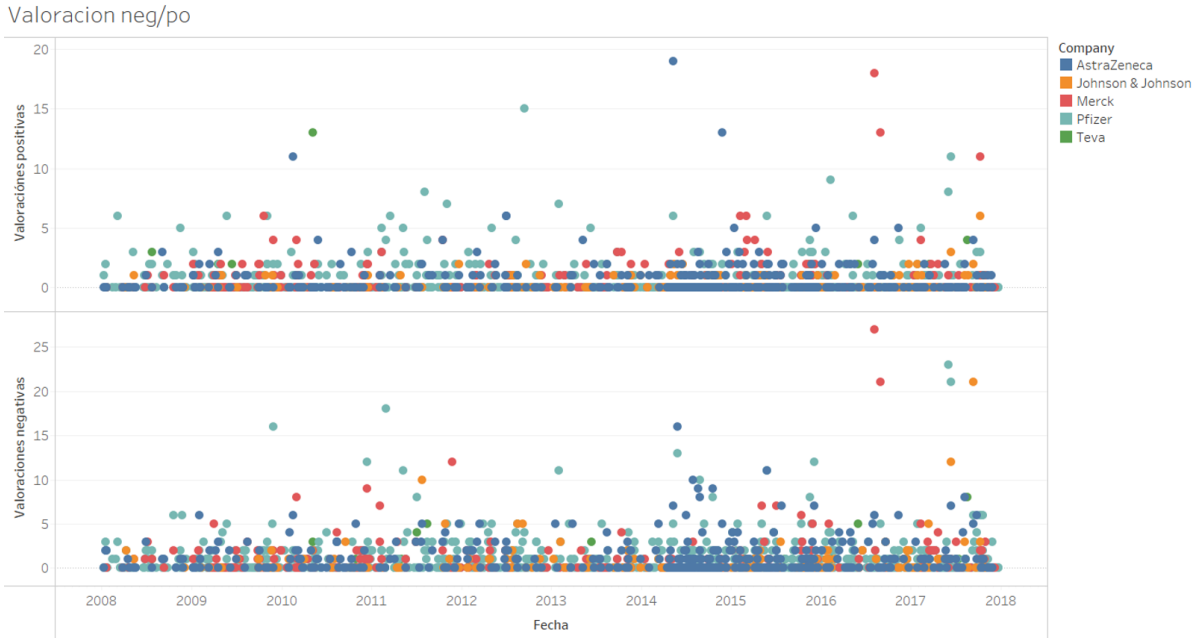


Figura 6.7. Valoraciones positivas y negativas de las menciones clasificadas por empresa.

Observando detenidamente el gráfico, da la impresión de tener más puntuaciones negativas que positivas. Puesto que el número de menciones es constante, se infiere que hay más noticias con una puntuación cero para el sentimiento positivo que para el negativo. Es decir, resulta más sencillo encontrar sentimiento negativo que positivo en las noticias estudiadas.

En resumen, a lo largo de esta subsección se ha descubierto que la narración de las noticias es muy objetiva, es decir, de todas las palabras que hacen referencia a una empresa, solo un pequeño porcentaje de ellas aporta sentimiento. Tras eliminar aquellas palabras neutras, se contempla que no existe una tendencia común en cuanto a la polaridad, intercalándose aquellas noticias fuertemente positivas con las negativas y neutras de manera indiscriminada.

6.2. Event Studies

En esta sección se han utilizado únicamente aquellos días en los que las empresas cuyos datos se encuentran en el dataset de precios aparecen en las noticias. Por ello, de las más de 4000 noticias iniciales, se utilizarán 2306. De estas se deben despreciar 94, aquellas acontecidas a lo largo de 2008, ya que no se dispone de información suficiente para poder calcular los parámetros del modelo de Fama y French.

El resultado de aplicar el modelo de los tres factores devuelve que en 374 ocasiones, el retorno está fuera de lo que cabría esperar, concretamente, en 319 ocasiones se podría rechazar la hipótesis nula a un 10 % de significación, en 53 a un 5 % y en ningún caso a un 1 %. Esto debería indicar que, en la mayoría de los casos, las noticias habrían resultado decisivas.

No obstante, al comparar estos resultados con los valores de sentimiento, en muchas de estas ocasiones no se corresponden: algunas de las noticias publicadas en días con retornos anormales no presentan gran cantidad de sentimiento o incluso ninguno. Esto confirmaría que no existe relación entre la publicación de noticias y su desempeño bursátil.

6.3. Combinación

A pesar de todos los resultados comentados en las secciones anteriores, la comparativa de estos datos resulta difusa, por ello se ha llevado a cabo un último análisis, el descrito como segunda parte en los Capítulo 4 y Capítulo 5 donde se compara el sentimiento calculado para Pfizer[2] de ambos periódicos. Las fechas de las noticias se clasifican en función de si ambos muestran el mismo sentimiento y es negativo, si es igual y positivo o si difieren.

Los resultados muestran que ambos escriben en el mismo día sobre la misma empresa en únicamente 299 ocasiones. De estas, solo en 112 presentaron el mismo sentimiento (12 positivo y 66 en negativo) y en 187 difirieron. La información está desglosada en la Tabla 6.1.

Acuerdo/Desacuerdo	Total	Sentimiento	Casos
Acuerdo	112	Positivo	12
		Negativo	66
		Ambos sin polaridad	34
Desacuerdo	187	Uno sin polaridad	130
		Con polaridad	57

Tabla 6.1. Resultados de la comparativa entre periódicos.

La Tabla 6.2 hace un desglose mayor de la información para aquellos días en los que ambos periódicos difieren. Se puede observar que es mucho más probable encontrar noticias en las que el *Financial Times* presente un sentimiento negativo (78 negativo, 27 positivo, un 75 % de los casos en los que presenta sentimiento) mientras que para el *Wall Stret Journal* el porcentaje está más equilibrado (74 positivo, 60 negativo, es positivo un 55 % de las veces).

[2]Se utiliza solo la empresa Pfizer ya que es la que cuenta con mayor número de noticias para la parte del estudio de significación, en el resto de casos se emplean todas.

Polaridad	Condiciones	Total
Ambos con polaridad	FT Positivo | WSJ Negativo	20
	FT Negativo | WSJ Positivo	37
Uno con polaridad	FT Positivo | WSJ Cero	7
	FT Negativo | WSJ Cero	41
	FT Cero | WSJ Positivo	37
	FT Cero | WSJ Negativo	45

Tabla 6.2. Resultados de la comparativa de polaridades en los casos de diferencia de opinión.

Tomando los grupos descritos anteriormente, se ha realizado el estadístico para el acumulado de cada uno de ellos siguiendo los pasos de la Sección 5.1, pero desgraciadamente, ninguno de los valores obtenidos permitió rechazar la hipótesis cero, puesto que el p-valor se encontraba en las tres ocasiones entre $0{,}4$ y $0{,}7$. El gráfico de la Figura 6.8 muestra los retornos acumulados para todos los eventos de cada grupo, donde cero representa en todos los casos el día del evento. En caso de obtener valores significativos, se debería apreciar un salto en el valor del CAR.

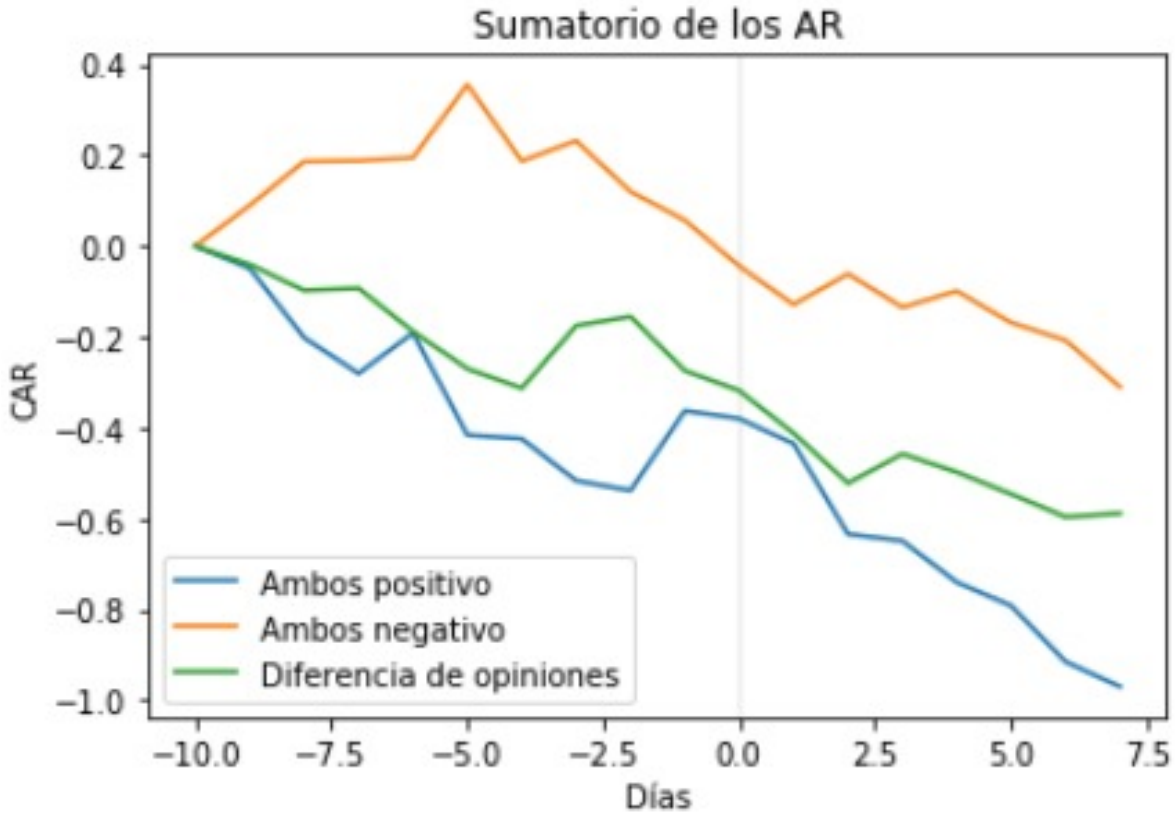


Figura 6.8. Representación del sumatorio de los retornos acumulados para cada grupo estudiado.

Para obtener el gráfico, se han tomado los valores desde el décimo día antes del evento y se han acumulado a partir de ese punto, mostrando la evolución general de todas las noticias englobadas en el grupo señalado.

En los tres casos se observa un pequeño punto de inflexión en la trayectoria de la pendiente, pero no es un cambio drástico. Sin embargo, resulta llamativo que la tendencia comienza días antes cuando ambos periódicos están de acuerdo. Esto se puede deber al hecho de que el

mercado realmente ya conoce la noticia o por lo menos se anticipa a ella, pero tampoco obtiene un resultado significativo.

A pesar de estos resultados, se debe resaltar de nuevo la enorme simplificación de este proyecto, pero también que existen gran cantidad de fuentes de información, por lo que no todo se puede reducir a estos dos periódicos. Además, en la mayoría de casos, pero especialmente con Pfizer y AstraZeneca, se suceden las noticias/eventos en días casi consecutivos. Esto provoca que las ventanas de estudio previas se vean alteradas constantemente y no sean uniformes, perjudicando considerablemente la eficiencia del modelo.

7
Conclusiones

El presente capítulo expone las conclusiones finales del proyecto extraídas a partir de los resultados obtenidos de las diferentes pruebas realizadas.

Durante la elaboración de este proyecto se ha desarrollado un programa en Python capaz de procesar noticias financieras, reconocer las empresas farmacéuticas mencionadas en él e identificar el sentimiento con el que se habla sobre ellas desde un punto de vista económico. A partir de ahí, se calcula un estudio de eventos siguiendo el modelo de los Tres Factores de Fama y French. Para ello, se toman los valores del modelo (SMB, HML, $R_M - R_F$) y los retornos de la empresa para la que se calcula y se obtiene un porcentaje de significación al que indica si el retorno está dentro de lo esperado. La arquitectura general del proyecto se muestra en la Figura 7.1.

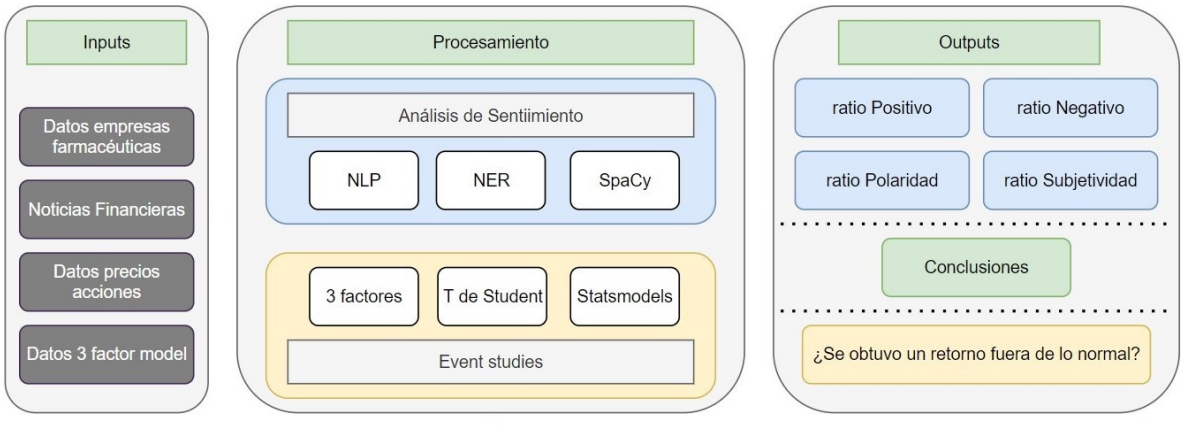


Figura 7.1. Esquema de la arquitectura del sistema.

Respecto a los objetivos del proyecto, señalados en la Sección 1.2, se ha conseguido el principal: analizar la correlación entre el sentimiento de las noticias y la evolución del precio en los mercados. Sin embargo, los resultados no han sido los esperados, ya que se puede confirmar que no existe una correlación directa: el acumulado de los eventos no es significativo a ningún nivel.

Sin embargo, sí se han observado algunas tendencias. Entre ellas, que es más probable que ambos periódicos estudiados difieran en sus opiniones, ya que sucede en un $62{,}5\,\%$ de los casos. Por otra parte, comparando los días en los que ambos publican sobre la misma empresa expresando sentimiento, el *Financial Times* presenta sentimiento negativo en un 75 % de las ocasiones mientras que el *Wall Street Journal* es mucho más optimista: tiene un sentimiento positivo en un 55 % de las ocasiones. Esto abre la puerta a la realización estudios de diversa índole para determinar los motivos tras esta situación.

Como se ha comentado con anterioridad, el desempeño en los mercados no presenta diferencia en el día del evento a ningún nivel de significación a pesar de las noticias publicadas. Por este motivo se puede confirmar que los mercados funcionan bien en estas situaciones, siendo inalterables ante las noticias publicadas en prensa.

Tomando un punto de vista más global, este proyecto es una simplificación de todo lo que se puede llegar a conseguir aplicando las tecnologías correctas. Por ello, el modelo se puede mejorar considerablemente, como se detalla en el Capítulo 8 y tal vez aumentando su precisión se alcance alguna solución que confirme la premisa inicial: la existencia de correlación entre el sentimiento de las empresas mencionadas en las noticias y su desempeño.

En cualquier caso cabe destacar que en los últimos años, concretamente desde el inicio de la COVID-19, las empresas farmacéuticas han cobrado una importancia sin precedentes, lo que se ha traducido en un auge de popularidad. Probablemente, esto haya supuesto modificaciones en el tratamiento que sufren en prensa y sería interesante poder llevar a cabo esta comparativa.

8

Futuros desarrollos

El último capítulo recoge una serie de mejoras que podrían realizarse en el futuro partiendo del elaborado en el presente proyecto. Se relatan las posibles mejoras y se proponen diversas maneras de llevarlas a cabo

Este proyecto es solo el principio de un desarrollo mucho más ambicioso, por lo que los puntos de mejora son amplios y variados. Se han identificado varias propuestas para incrementar el nivel de acierto del desarrollo realizado, incorporando nuevos algoritmos y procedimientos que aseguren su fiabilidad.

Una primera mejora consistiría en basar la creación del BOW en los resultados obtenidos del NER. Para ello será necesario estudiar en un plano lingüístico qué palabras son las que más fuerza tienen sobre un sintagma concreto en la oración. A partir de ahí, se definirán los elementos que pasarán a formar parte de la BOW atendiendo a esta influencia.

Siguiendo la misma línea, se propone incluir algoritmos para reconocer si en una frase consecutiva se está haciendo referencia a algún elemento de la anterior. De esta forma, en una primera oración se puede mencionar una empresa y, en la siguiente, continuar hablando sobre ella haciendo referencia a la misma con la utilización de un pronombre. Para ello, se deberían utilizar algoritmos de *Co-Reference Resolution*. SpaCy tiene un módulo destinado a este tipo de tareas denominado: *neuralcoref*.

Otro posible desarrollo sería la mejora del reconocimiento de empresas. Como se describió en la Sección 4.2, tras llevar a cabo el NER, se contrastan las entidades con una lista de empresas de la industria. El doble reconocimiento sirve para evitar problemas en caso de que no se utilice el nombre completo de la empresa, es decir, cuando no se escriben todas las palabras o se elimina la sigla final. No obstante, no llega a reconocer aquellos casos en los que se utiliza una abreviatura del nombre, y esto es muy frecuente en algunos casos. Por ejemplo, a *Johnson & Johnson* se le suele hacer referencia como *J& J*, o para *GlaxoSmithKline*, poner *GSK* o *Glaxo*.

La última propuesta para el terreno del análisis de sentimiento implicaría el cambio total del proyecto. Supondría incluir el etiquetado de los datos para poder entrenar un modelo que

devuelva resultados fiables y adaptados a las necesidades del estudio. Esto generaría un modelo reutilizable a lo largo del tiempo que permitiría estudiar tendencias en el tiempo.

En el campo de los *event studies*, se podría utilizar el modelo de los cinco factores, pero los resultados no deberían diferir en gran medida. Sin embargo, sí hay una idea importante que no se ha tenido en cuenta a la hora de llevar a cabo este trabajo: el *clustering*.

Para realizar el estudio, siempre se tiene en cuenta la ventana de análisis previa al evento. Sin embargo, en ocasiones pueden ocurrir otros eventos durante este período que contaminan los retornos obtenidos previamente fijando como "normal" unos valores que realmente no lo son. Existen metodologías recursivas para minimizar el impacto de esta situación que se podrían aplicar (Corrado, 2011).

9

Referencias

Bhagavatula, M., Gsk, S., & Varma, V. (2012). Named Entity Recognition an Aid to Improve Multilingual Entity Filling in Language-Independent Approach. Proceedings of the First Workshop on Information and Knowledge Management for Developing Region, 3–10. Presented at the Maui, Hawaii, USA. doi:10.1145/2389776.2389779

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (5 2020). Language Models are Few-Shot Learners. Opgehaal van http://arxiv.org/abs/2005.14165

Brown, S. J., & Warner, J. B. (1985). The Case of Event Studies* (Vol 14, bll 3–31). North-Holland USING DAILY STOCK RETURNS.

Corrado, C. J. (2011). Event studies: A methodology review. Accounting & Finance, 51, 207–234.

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Opgehaal van https://github.com/tensorflow/tensor2tensor

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. Journal of Finance, 51, 55–84. doi:10.1111/j.1540-6261.1996.tb05202.x

Fama, E., F., & French, K., F. (2015). A five-factor asset pricing model. Journal of Financial Economics, 116, 1–22.

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent Named Entity Recognition and Classification tec hniques: A systematic review. Computer Science Review, 29, 21–43. doi:10.1016/j.cosrev.2018.06.001

Grishman, R., & Sundheim, B. (8 1996). Message Understanding Conference- 6: A Brief History. 466–471. doi:10.3115/992628.992709

He, P., Liu, X., Gao, J., & Chen, W. (6 2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. Opgehaal van http://arxiv.org/abs/2006.03654

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (9 2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. Opgehaal van http://arxiv.org/abs/1909.11942

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (7 2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Opgehaal van http://arxiv.org/abs/1907.11692

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. Journal of Finance, 66, 35–65. doi:10.1111/j.1540-6261.2010.01625.x

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 54, 1187–1230. doi:10.1111/1475-679X.12123

Mackinlay, A. C. (1997). Event studies in economics and finance. Journal of Economic Literature, 35, 13–39.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques (bll 79–86). Opgehaal van http://reviews.imdb.com/Reviews/

Porter, M. (1980). An algorithm for suffix stripping. doi:10.1108/00330330610681286

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language Models are Unsupervised Multitask Learners. Opgehaal van https://github.com/codelucas/newspaper

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (6 2017). Attention Is All You Need. Opgehaal van http://arxiv.org/abs/1706.03762

Wang, W., Bi, B., Yan, M., Wu, C., Bao, Z., Xia, J., ... Si, L. (8 2019). StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. Opgehaal van http://arxiv.org/abs/1908.04577