COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

Metodología de proyección emocional en la predicción de cotizaciones bursátiles

Autor: Guillermo Fernández Prota

Director: Dr. Antonio García de Garmendia

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

"Metodología de proyección emocional en la predicción de cotizaciones bursátiles"

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2022/23 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Guillermo Fernández Prota Fecha: 04/ 07/ 2023

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

4 Julio 2023

Fdo.: Dr. Antonio García de Garmendia Fecha: 04/ 07/2023

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Metodología de proyección emocional en la predicción de cotizaciones bursátiles

Autor: Guillermo Fernández Prota

Director: Dr. Antonio García de Garmendia

Madrid

METODOLOGÍA DE PROYECCIÓN EMOCIONAL EN LA PREDICCIÓN DE COTIZACIONES BURSÁTILES

Autor: Fernández Prota, Guillermo.
Director: García de Garmendia, Antonio.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este proyecto tiene como objetivo desarrollar una metodología para predecir oportunidades en el mercado bursátil utilizando las emociones generadas por los titulares de las noticias. Se emplea un dataset de titulares categorizados para entrenar un clasificador temático basado en el algoritmo de Naive-Bayes. Para el análisis de sentimientos se ha utilizado la biblioteca Flair que emplea redes neuronales para analizar los sentimientos de los titulares. Se creó un dataset de titulares para diseñar y comparar modelos de regresión, seleccionando el más adecuado para nuestra base de datos. El modelo final es un regresor polinomial que pronostica cambios porcentuales en el valor del IBEX 35 a partir del análisis de sentimientos de los titulares de noticias. Por último, se realiza un estudio de viabilidad económica para evaluar la posible creación de una empresa basada en la metodología desarrollada.

Palabras clave: Análisis de Sentimientos, Regresión Polinomial, Bursátil, Titulares de Noticias

ABSTRACT

This project aims to develop a methodology to predict stock market opportunities using emotions generated by news headlines. A dataset of categorized headlines is used to train a thematic classifier based on the Naive-Bayes algorithm. For sentiment analysis, the Flair library was used which employs neural networks to analyze headline sentiments. A dataset of headlines was created to design and compare regression models, selecting the most appropriate one for our database. The final model is a polynomial regressor that predicts percentage changes in the value of the IBEX 35 from the sentiment analysis of news headlines. Finally, an economic feasibility study is performed to evaluate the possible creation of a company based on the developed methodology.

Keywords: Sentiment Analysis, Polynomial Regression, Stock Market, News' Headlines

Índice de la memoria

1. INTRODUCCIÓN

Los seres humanos somos seres emocionales, nos dejamos llevar por nuestros sentimientos en la gran mayoría de decisiones que tomamos[1], sea de manera consciente o inconsciente. Cuando sentimos emociones positivas, corremos más riesgos a la hora de tomar decisiones y cuando sentimos emociones negativas, solemos ser más cautos[2]. A pesar de que todas los prescriptores recomiendan adoptar decisiones de inversión de una manera estrictamente objetivo, lo cierto es, y este proyecto lo evidencia, las emociones son un factor determinante dentro de la inversión en bolsa. También está probado que los inversores que son capaces de comprender y manejar correctamente sus emociones, suelen obtener una rentabilidad más elevada[3] que la de aquellos que sienten demasiadas emociones e incluso de los que sienten pocas.

Se ha demostrado que hasta los detalles más pequeños pueden influir emocionalmente en las personas y, con ello, en los mercados financieros. Por ejemplo, los resultados de un equipo de futbol influyen las fluctuaciones de la bolsa. Hay estudios que evidencian que la calidad de las inversiones de los hinchas de un determinado equipo de futbol italiano está ligada enormemente a los resultados de su equipo y a los del equipo rival[4]. Incluso algo tan trivial como el tiempo atmosférico puede influir en las fluctuaciones del mercado, de tal manera que en los días nublados o con mal tiempo, el valor de las acciones de la bolsa de Nueva York tiende a ser menor que a los días en los que el tiempo era bueno[5]. Uno de los objetivos de este proyecto es demostrar que las emociones influyen directamente al mercado de valores mediante un análisis teórico y posteriormente un análisis práctico en el que querremos encontrar esa relación.

[1] Jennifer S. Lerner, Ye Li,Piercarlo Valdesolo, and Karim S. Kassam, 2015.
[2] QiweiYang , ShiqinZhou, RuoleiGu, YanWu, 2020.
[3] Lisa Feldman, 2007.
[4] Demir y Rugoni, 2006.
[5] Brian M. Lucey, 2005.

El proyecto "Metodología de proyección emocional en la predicción de cotizaciones bursátiles" tiene como objetivo entender cómo afectan las emociones a las fluctuaciones del mercado. Para ello hemos creado una metodología capaz de detectar los sentimientos que provocan en el lector los titulares de los principales medios de comunicación del país, y predecir cómo esas emociones afectan al mercado, en concreto en el IBEX 35.

Con este objetivo, se ha creado un clasificador de noticias, con Naive-Bayes, a partir del cual se clasificará cada titular de prensa en una de estas ocho posibles temáticas: Macroeconomía, Sostenibilidad, Innovación, Regulaciones, Alianzas, Reputación y Otras. Posteriormente, con la biblioteca Flair de Python, cuya metodología se basa en el uso de redes neuronales, clasificaremos las emociones encontradas en los titulares clasificados. Finalmente, mediante un modelo de regresión polinomial, encontraremos la relación entre las emociones expresadas en la prensa y el rendimiento del IBEX 35 de esas fechas e incluso seremos capaces de predecir el comportamiento de los mercados.

Estructuraremos el trabajo en tres secciones, primero haremos un estudio de la literatura, en segundo lugar, haremos un planteamiento matemático teórico que tendrá como objetivo explicar de forma teórica los métodos que utilizaremos y, por último, explicaremos la implementación práctica del proyecto, donde también haremos un análisis de un caso práctico en el que estudiaremos el desempeño del método en un caso real proyectando el modelo frente al índice IBEX 35.

El estudio de la literatura se dividirá en cuatro bloques, primero se estudiarán los distintos métodos para clasificar textos según su temática, donde destacan los siguientes métodos: Naive Bayes, Artificial Neural Networks, and Decision Trees, Support Vector Machines, K-Nearest Neighbors. En este bloque exploraremos las distintas técnicas que podremos implementar y los pasos a seguir a la hora de crear este tipo de modelos.

En el segundo bloque estudiaremos las distintas técnicas para analizar los sentimientos que se expresan en textos escritos, y las técnicas de preprocesado de texto, POS[6], y NLP[7].

El tercer bloque tiene un enfoque estrictamente teórico, que tiene como objetivo crear un marco conceptual de cómo los seres humanos nos dejamos influenciar por nuestras emociones constantemente y estudiar cómo correlacionan las emociones en un ámbito bursátil.

La finalidad del cuarto y último bloque es estudiar el mercado de valores, haciendo un recorrido por los diversos campos que afectan a las cotizaciones de las acciones, como la información desestructurada, las noticias, las *fake-news* y, finalmente, las emociones.

En el planteamiento matemático explicaremos los pasos a seguir y la teoría de los métodos elegidos. Para el clasificador de Naive-Bayes utilizaremos un *dataset* que recoge miles de titulares de noticias con su temática, lo que, con el método de Naive-Bayes, nos ayudará a predecir la temática más probable al introducir un titular según las palabras que se usan. En el análisis de los sentimientos, decidimos hacer un enfoque binomial, utilizando la biblioteca Flair, a partir de la cual recibiremos la polaridad de los sentimientos expresados en el titular. Con la limitación de que, para poder utilizar este método, hemos de traducir los titulares al inglés primero, por lo que se puede perder algo de precisión. Finalmente, para la parte del modelo de regresión, utilizaremos una regresión polinomial ya que, de inicio, desconocemos la relación que existe entre los sentimientos expresados en las noticias y los cambios de valor del IBEX 35. La razón principal por la que elegimos la regresión polinomial es que nos permite calcular el grado del polinomio que más se ajuste a nuestro dataset e incluso ajustarse a cualquier dataset nuevo, sin necesidad de ajustar el método.

En la implementación práctica explicaremos el uso de la biblioteca de Python scikit-learn para crear el clasificador de Naive-Bayes, la manera de aumentar la precisión en la clasificación de temáticas y, finalmente, mostramos los resultados obtenidos con un set de

[6] Por sus siglas en inglés: *part of speech*
[7] Por sus siglas en inglés: Natural language processing

titulares de ejemplo. También explicaremos los entresijos de la biblioteca Flair, para detectar sentimientos y creamos un nuevo *dataset*, a partir del cual estudiaremos los sentimientos que ser transmiten en la prensa española en relación a los mercados durante un periodo de observación de 14 días.

Finalmente, para el modelo de regresión se hará una comparación de la precisión de distintos modelos, lineal, árbol de decisiones y polinomial, para demostrar que el polinomial es la opción que mejor se ajusta al *dataset* elaborado. Obtuvimos resultados muy satisfactorios con el modelo de regresión polinomial, con una $R^2 = 0.9239$ *y la* $R^{*2} = 0.6702$, este segundo dato nos confirma que el sistema no está sobreajustado.

Por último, testaremos el modelo con un caso práctico, comparando los cambios de valor porcentuales diarios del IBEX 35 con la polaridad y valor asignado mediante un sistema de ponderación de los sentimientos. Una vez desarrollado el método, investigaremos los titulares mediáticos durante un número determinado de días para intentar predecir con un mínimo grado de fiabilidad la relación entre los sentimientos generados por los medios de comunicación, y el comportamiento del mercado.

Como anticipo de las conclusiones, se pudo comprobar que, tal y como se esperaba en el diseño de este proyecto, existe una clara relación entre las emociones mostradas en la prensa y el rendimiento del IBEX 35.

En resumen, para este proyecto se hará un estudio de la literatura, lo que nos servirá para crear un marco conceptual de los estudios y descubrimientos más relevantes en las áreas que trataremos, posteriormente, se explicarán los métodos que utilizaremos de forma teórica y, finalmente, se explicará la implementación práctica junto con un estudio de un caso particular, el rendimiento del IBEX 35.

La finalidad de este proyecto es la de intentar contestar nuestra hipótesis inicial, que las emociones influyen en gran medida en las decisiones que tomamos a la hora de invertir y por ende en los mercados de valores y sus fluctuaciones. También se pretende encontrar la

relación que existe entre los sentimientos/emociones que encontramos en las noticias de los medios de comunicación nacionales y el desempeño del IBEX 35.

2. ESTADO DE LA CUESTIÓN

El objetivo de este capítulo es el de analizar cómo afectan las emociones a las cotizaciones del mercado de valores y, en particular, cómo influyen las mismas en la toma de decisiones de los inversores. Para hacer este análisis dividiremos el estudio en cuatro bloques fundamentales.

El primer bloque tratará sobre las técnicas y características importantes de cómo clasificar texto según su temática. Se analizará la metodología, poniendo especial énfasis en los métodos de preparación del contenido, es decir, la lematización, el preprocesado de las palabras y las técnicas de truncamiento.

En el segundo bloque se estudiarán las técnicas más relevantes para recopilar las emociones que se perciben en un texto. Concretamente, nos interesan especialmente las técnicas de Machine Learning, para poder entrenar nuestro propio código con la ayuda de la biblioteca Flair en Python, que será capaz de devolvernos las emociones predominantes de cualquier texto.

El tercer bloque tiene un enfoque distinto, se trata de un bloque teórico, a partir del cual queremos conseguir un entendimiento profundo acerca de cómo actuamos los seres humanos, y cómo afectan las emociones que sentimos en nuestra toma de decisiones, ya sea de manera consciente o inconsciente.

El cuarto y último bloque, estudiará el mercado de valores. Realizaremos un recorrido por los distintos factores que pueden alterar las fluctuaciones del mercado. Empezaremos estudiando la importancia de la información desestructurada, seguido de la influencia que tienen los medios de comunicación en el mercado y en los inversores, para, finalmente, acabar con cual es la relación entre las emociones que sentimos y cómo estas alteran el mercado.

En resumen, este capítulo tiene como fin ofrecer una visión general de las investigaciones más importantes de los cuatro bloques que se han comentado previamente, lo cual nos servirá como un marco conceptual necesario para comprender la relevancia y originalidad del trabajo. Realizaremos un análisis crítico a partir del cual expresaremos las carencias y elementos a destacar de aquellos artículos cuya relevancia sea mayor.

CLASIFICACIÓN DE UN TEXTO SEGÚN SU TEMÁTICA

La clasificación de textos según su temática es una de las partes fundamentales de este proyecto, el objetivo es analizar el uso de las palabras y la repetición de las mismas para clasificar por temáticas su contenido.

En el contexto actual, resulta especialmente importante la técnica de clasificación de textos según su temática, ya que la cantidad de información disponible en línea es enorme, lo que dificulta la organización y gestión de textos. Para muchas empresas la clasificación de textos se ha convertido en una necesidad para poder gestionar grandes cantidades de información, optimizar su uso, y facilitar la identificación de tendencias y patrones, lo que es clave para un siguiente objetivo que persigue este trabajo que es el patrón de comportamiento que determina la toma de decisiones a partir de las emociones que se generan con dicha información.

Para este proyecto, abordaremos este subproblema mediante el uso de un *dataset* de noticias. El procedimiento consistirá en comparar todas las palabras de un titular con el dataset, después asignar un valor a cada posible temática en función de las palabras que se usen en el texto y, por último, elegir como temática principal aquella cuya puntuación final sea más alta.

Para llevar a cabo este proceso, será clave el uso de diferentes técnicas de procesamiento de lenguaje, como la tokenización, la eliminación de stop-words, y la lematización, con el objetivo de identificar las características que tienen mayor impacto en la clasificación del texto y simplificar el proceso.

Para resumir, este subproblema del trabajo se enfocará en el uso del *dataset* de titulares de noticias y de diversas técnicas de procesamiento del lenguaje con el fin de desarrollar un modelo de clasificación de textos basados en el análisis temático. Es importante para el conjunto del proyecto ya que, con la clasificación por temática, seremos capaces de focalizar la búsqueda de emociones propias de esos contenidos. Por ejemplo, en una noticia sobre deportes lo normal es encontrar emociones como la alegría, euforia o frustración, mientras

que un texto sobre guerras, lo normal es encontrar emociones como miedo, ira o solidaridad. Ser capaces de conocer la temática de un texto antes de analizar las emociones del mismo nos permitirá aumentar la precisión, y ser capaces de focalizar los análisis en las emociones que son determinantes para cada temática.

En el estudio de la clasificación de textos o noticias por temática se han producido grandes avances en los últimos años debido a la revolución tecnológica de las últimas dos décadas que han permitido perfeccionar las técnicas más tradicionales. Durante los últimos años se han realizado varios estudios teóricos [8] sobre las técnicas más apropiadas para realizar la clasificación por temática, para varios idiomas[9], pero principalmente para el inglés. El primer paso suele ser el de recopilar los textos, los métodos varían desde los más complejos como un *web-scrapping*, hasta simplemente copiar y pegar el texto.

El *web-scrapping* consiste en capturar información de manera automática, mediante un código o una herramienta, como por ejemplo mediante una extensión de Google. Esta técnica permite guardar y organizar la información relevante de una página web, supongamos que se trata de un periódico digital, con el *web-scrapping* podrías recopilar las noticias que se publican de manera automática, organizando sus distintas estructuras, como puede ser titular, autor, fecha y hora, sección a la que pertenece e incluso, los programas más complejos serían capaces de dar un resumen de la noticia. Por otro lado, los métodos más sencillos consisten simplemente en copiar y pegar el texto de la noticia en un archivo ".txt", que será después leído por el código.

Una vez obtenida la noticia/texto a clasificar, el segundo paso suele ser también siempre el mismo, limpiar el texto mediante diferentes técnicas de procesamiento de lenguaje. Habrá que convertir cada palabra en un *string*, con el objetivo de eliminar aquellas que no aportan valor, como las *stop-words*, los determinantes o las preposiciones y, posteriormente, llevar

[8] Kaur, G., & Bajaj, K., 2016.
[9] Miao, F., Zhang, P., Jin, L., & Wu, H., 2018.

a cabo la extracción de las raíces de las palabras (i.e. "correr", "corriendo" y "correría" es "corr"). Después se selecciona el tipo de algoritmo de truncamiento que se quiere utilizar. El más popular es M.F. Porter Stemmer, también conocido como algoritmo de Porter. Sin embargo, este algoritmo es solo válido para el idioma inglés, el equivalente en el castellano sería la técnica de la lematización. Todos estos procesos son necesarios, y su finalidad es la de simplificar el trabajo del clasificador y aumentar la precisión del proyecto.

En tercer lugar, hay que llevar a cabo una selección de características que permitan elegir a que factores/palabras le queremos dar más peso o importancia. Para ello, encontramos programas en el mercado como BooleanWeighting, Class Frequency Thresh holding, Term Frequency Inverse Class Frequency, o Information Gain.

En cuarto lugar, hay que clasificar las noticias, para lo que se pueden utilizar algoritmos como Naive Bayes, Artificial Neural Networks, and Decision Trees, Support Vector Machines, K-Nearest Neighbors.

Otros estudios defienden que, a la hora de realizar la clasificación de un texto, es necesario distinguir dos fases principales[10], la fase de hipótesis y la de confirmación. Estos análisis se basan en la búsqueda de *pattern-sets*, que son grupos de palabras que están asociados a un concepto, como, por ejemplo, “conflictos”. El objetivo de estos *pattern-sets* es contar cuántas veces y cuántas de esas palabras aparecen en la noticia. Cuando se supera el umbral predefinido, se dice que se ha encontrado un *pattern-set*. Esta teoría expresa la importancia de la segunda fase, la de comprobación, ya que es muy común, sobre todo en el análisis de noticias de actualidad, el uso de palabras propias de otras temáticas para enfatizar los titulares y dotar de dramatismo al contenido.

Una de las limitaciones de este método es la dificultad que tienen los códigos para entender el contexto de un texto escrito sin información extra, como se aprecia en el siguiente ejemplo: “LENDL DEMONSTRATES GRASS COURT MATURITY LONDON, July 2 -

[10] Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio, 2017.

Czechoslovak top seed Ivan Lendl served warning that he may finally have come of age on grass when he emerged victorious from a pitched battle with one of the finest exponents of the fast court game at Wimbledon today. The U.S. and French Open tennis champion has never won a title on grass but he outlasted American 10th seed Tim Mayotte 6-4 4-6 6-4 3-6 9-7 over three and a half hours to join Boris Becker, Henri Leconte and Slobodan Zivojinovic in Friday's semifinals. The titanic struggle on court one upstaged the centre court clash between seventh seed Leconte and the remarkable Australian Pat Cash, which had been billed as the day's main attraction [...]"

Lo normal para un lector humano es pensar que este texto habla de deporte, más concretamente de tenis. Sin embargo, para un código, no es tan sencillo de distinguir, ya que no se menciona tenis en ningún momento ni se da alusiones a un tema deportivo, por tanto, para el código esta información trata de una temática de una batalla o de guerra, por eso la fase de confirmación es determinante en el proceso. Para corregir este tipo de errores, que son difíciles de interpretar por las máquinas, se crean una serie de reglas que se obtienen de forma empírica y requieren de una alta supervisión humana, lo que encarece el proceso considerablemente. Sin embargo, para textos de mayor extensión este problema es más raro que se dé, consiguiendo una precisión de más del 85%.

En un trabajo cuyo objetivo es la creación de una metodología de proyección emocional con el fin de predecir oportunidades bursátiles es prácticamente obligatorio estudiar las *fake-news.* Estas noticias falsas, las cuales son definidas[11] como aquellas que tienen como objetivo confundir al lector, jugando con sus emociones con un fin de influir en el público por intereses sociales, políticos o, como en el caso de este trabajo, generando decisiones económicas individuales, colectivas o alterando la evolución habitual del mercado de valores.

El proceso para clasificar noticias según su veracidad es muy parecido al de su temática, primero se recolectan los datos, y después se hace un preprocesado de las palabras, se

[11] Nikam, S. S., & Dalvi, R, 2020.

lematiza y se clasifica con diversos algoritmos, según las necesidades y características de los textos a tratar.

Este bloque ha servido para resaltar la importancia de las técnicas NLP[12] cuyo objetivo es permitir que las máquinas y códigos entiendan el lenguaje. Abarca todo el preprocesado del texto, la lematización, tokenización, eliminación de stop-words y las palabras que no añaden información, como determinantes o preposiciones. Gracias a estas técnicas lograremos aumentar considerablemente la precisión de nuestro código.

[12] Procesamiento del lenguaje natural, NLP por sus siglas en inglés.

CLASIFICACIÓN DE EMOCIONES DE UN TEXTO ESCRITO

El análisis y la clasificación de las emociones en textos escritos es un campo de investigación en constante evolución, y de gran importancia para la predicción de comportamientos en base a las percepciones que esos textos generan. Ser capaces de identificar emociones a partir de un lenguaje escrito proporciona una información esencial para la toma de decisiones, especialmente en áreas como la publicidad, el marketing, la política y en las consultoras que investigan el comportamiento de los consumidores. El ámbito en el que se basa este trabajo es el impacto de las emociones de los públicos en la evolución de la cotización en el mercado de valores.

Para el análisis y la clasificación de emociones en textos se han estudiado técnicas basadas en el procesamiento de lenguaje y en el aprendizaje automático. El objetivo es desarrollar un modelo de clasificación de emociones sencillo, que sea capaz de identificar las emociones predominantes generadas por un texto que, junto con la clasificación de textos por temática, sea la base del modelo de predicción de oportunidades bursátiles a través del estudio de las emociones.

Con estos objetivos, se ha estudiado la literatura comparada para encontrar un modelo replicable que cumpla con nuestros requisitos, seguido de un planteamiento matemático y una posterior conexión con los otros bloques del trabajo.

La metodología consiste en clasificar el análisis de las emociones según la valoración que cada palabra de la biblioteca de palabras de Flair tiene asignadas a cada palabra y que permite distinguir la polaridad de los sentimientos de una frase.

El estudio de la clasificación de emociones en textos es un tema relativamente reciente, se lleva analizando desde finales de los años noventa. Sin embargo, en los últimos años, gracias a los avances tecnológicos, principalmente en las ramas de programación e inteligencia artificial, se ha producido un notable progreso sobre todo en relación a los resultados cualitativos.

Casi todos los tratados acerca de este tema parten de una estructura y técnicas similares para preparar el texto a analizar. Se suele empezar por eliminar el ruido[13], es decir, quitar errores de escritura, quitar las *stop-words* (comas, puntos y demás signos de puntuación); lematizar las palabras y hacer técnicas de reconocimiento de nombre, como por ejemplo nombres de personas, ubicaciones o entidades para mejorar la precisión de reconocimiento de emociones y disminuir la cantidad de elementos a analizar, abaratando costes y tiempos; a continuación, se definen las características a buscar, se convierten secciones de texto en vectores de características, con el objetivo de facilitar la clasificación y se observa la presencia y frecuencia de ciertas palabras, POS y el uso de negación.

Después de la preparación del texto se procede a la clasificación, donde los enfoques más clásicos son los de *Machine Learning* y *Lexicón*. Los estudios realizados muestran que tanto las técnicas de *Machine Learning* como las de *Lexicón*, tienen una precisión muy parecida[14], con la ventaja de que *Lexicón* no requiere de un entrenamiento previo.

Actualmente se están desarrollando técnicas basadas en *Machine Learning*[15] entre las que destacan el uso de *Syntax Trees* y *Naive Bayes* [16] donde también se apoyan en técnicas NLP (Natural language processing) y POS (*part of speech*).

Destaca el experimento binomial[17] donde hicieron un análisis que clasificaba críticas de películas en positivas o negativas mediante *Machine-Learning* y *term-counting*. Esta última técnica consiste en contar cuantas veces aparecen los términos en el texto para clasificarlo, para ello es necesario realizar previamente una tokenización/lematización, una construcción del vocabulario, es decir identificar todos los términos que aparecen en el texto y asignarles un índice, después de hacer el recuento de términos, los resultados se muestran

[13] Miranda, C. H., & Guzman, J., 2017
[14] Miranda, C. H., & Guzman, J., 2017
[15] Kennedy, A. and Inkpen, D, 2006
[16] Zou, H., Tang, X., Xie, B., & Liu, B, 2015
[17] Un estudio relevante en cuanto a técnica, pero alejado en cuanto a su temática, es el que llevaron a cabo *Kennedy, A. y Inkpen, D*. (2006).

vectorialmente para que sea más fácil trabajar con ellos. Este método es muy sencillo y, a la vez, muy efectivo para capturar la frecuencia de uso de las palabras.

Otro de los enfoques más exitosos es el de ASNA[18] y ESNA[19] que se basa en la utilización de técnicas NLP junto con la base de datos de *SenseNet*, que asigna un valor numérico a cada línea de texto para posteriormente clasificarla entre 8 tipos de emociones y una emoción neutra. Sin embargo, la mayor limitación respecto a estos artículos es que, al tratarse de una empresa privada, no se conoce con precisión su funcionamiento ni han hecho público el funcionamiento de su código, aunque si sabemos que lo han utilizado para diseñar un buscador, tipo Google, que te muestra noticias y recomendaciones según los gustos y emociones de cada persona.

A pesar de que este proyecto este centrado en la recolección de emociones a partir de textos escritos, se han hecho grandes avances en la clasificación de emociones en discursos[20] o entrevistas. Para ello se utilizan varios algoritmos de aprendizaje automático, como arboles de decisiones o redes neuronales, estas últimas las más precisas, para que, gracias a grabaciones de voz, se pueda aprender a detectar las emociones del hablante.

Este bloque nos permite volver a destacar la importancia de la compresión y uso de técnicas de preprocesado y POS[21]. Es particularmente interesante saber que, tanto técnicas de MLA[22] cómo las de Lexicon, obtienen precisiones parecidas y, además, Lexicon es considerablemente más sencillo de programar.

[18] S. M. Al Masum, M. T. Islam and M. Ishizuka,2005
[19] S. Mostafa Al Masum, H. Prendinger and M. Ishizuka, 2006
[20] Casale, S., Russo, A., Scebba, G., & Serrano, S., 2008
[21] Siglas en inglés: Part-of-Speach
[22] Machine Learning Algorithms

COMO AFECTAN LAS EMOCIONES A NUESTRA TOMA DE DECISIONES

Los humanos somos seres emocionales. Con frecuencia, nuestras decisiones no están basadas en la lógica sino en lo emocional, desde las cuestiones más triviales, hasta las más complejas e importantes. Como veremos más adelante, con carácter general, cuando nos encontramos en un estado emocional positivo, de euforia o felicidad somos más propensos a tomar decisiones más arriesgadas, mientras que cuando sentimos miedo o tristeza, tendemos a ser más cautelosos en la toma de decisiones con el objetivo de reducir riesgos o situaciones que percibamos como peligrosas.

Este capítulo teórico es fundamental para el desarrollo del trabajo, analizaremos cómo las emociones nos afectan a la hora de tomar decisiones, y estudiaremos de forma más profunda los comportamientos típicos de la mayoría de personas ante ciertas emociones. Será la base teórica sobre la que se basarán los siguientes capítulos.

Filósofos de la antigua Grecia, como Aristóteles y Platón ya estudiaron las emociones como parte determinante del comportamiento. Para el primero[23], las emociones eran una parte integral de la experiencia humana, y dedicó gran parte de su tiempo a explorar la relación entre las emociones, el comportamiento humano y la moral. Platón[24], a diferencia de Aristóteles, pensaba que las emociones eran impulsos irracionales que podían interferir en la capacidad de razonar de las personas y, por ello, les alejaba de la búsqueda de la verdad y la sabiduría.

Otro de los grandes autores que estudió este tema fue Sigmund Freud[25], que defendía que las emociones tenían una gran influencia en el comportamiento humano, y que surgían a partir de un conflicto entre las pulsiones internas y las demandas del mundo exterior. Freud también afirmaba que las emociones impactaban de forma integral en todas las facetas del ser humano, en toda su consciencia, es decir desde lo más superficial hasta lo más profundo

[23] Aristóteles, 350 A.C.
[24] Platón, 380 A.C.
[25] Sigmund Freud, 1923

del subconsciente, pero que dependiendo de la parte a la que más impacten, el resultado de las decisiones podía ser diferente.

Estudios más recientes apuntan a que, aunque no seamos conscientes, hasta las emociones menos conectadas con la decisión que estamos tomando, acaban afectando a la misma[26]. Para confirmar esta hipótesis el equipo liderado por Qiwei Yang, investigó el efecto que tenían distintos tipos de emociones accidentales en la toma de decisiones a la hora de realizar una serie de apuestas. A los participantes se les indujo una serie de emociones recordándoles situaciones emocionales de su vida, y se estudiaron los afectos que produjeron a la hora de apostar. Consiguieron demostrar que mientras los participantes sentían miedo, tomaban decisiones menos arriesgadas, tenían una mayor motivación y hacían un mejor uso de sus recursos cognitivos.

Otra de las conclusiones a las que se ha llegado[27] es que las emociones son predecibles, a veces beneficiosas y otras perjudiciales y unas potentes impulsoras de la toma de decisiones. También confirman que la toma de decisiones puede tener la forma de influencias incidentales o integrales, como en el anterior estudio, y que las emociones incidentales pueden producir influencias no deseadas e incluso subconscientes, como afirmaba Freud. Por otro lado, aunque las emociones afectan a nuestra capacidad de toma de decisiones de muchas maneras, los autores descubrieron que los principales cambios vienen en la forma de pensar, en la profundidad de la capacidad de reflexión, o las metas que nos fijamos.

Otra línea de investigación es el estudio que publicó Cambridge de Hans-Rüdiger Pfister y Gisela Böhm en el que se propone una división en cuatro categorías[28] sobre cómo las emociones afectan a las decisiones que tomamos. La primera categoría es la de la información, la cual proporciona indicadores evaluativos que se incorporan a la construcción de preferencias de una persona, informan del (des)agrado de las acciones y consecuencias. En segundo lugar, la función de velocidad, permite tomar decisiones bajo presión, de una

[26] QiweiYang, ShiqinZhou, RuoleiGu, YanWu, 2020.
[27] Jennifer S. Lerner, Ye Li,Piercarlo Valdesolo, and Karim S. Kassam, 2015.
[28] Hans-Rüdiger Pfister & Gisela Böhm, 2023.

manera parecida al pensamiento rápido que propone Kahneman[29]. En tercer lugar, la función de relevancia, dónde el tomador de decisiones pone el foco de atención en aquello que considera de mayor relevancia. Y, por último, la función de compromiso, que se basa en la presión social que sufren las personas a la hora de tomar una decisión.

Respecto a las decisiones sobre temas financieros, se ha descubierto[30] que también están afectadas por las emociones incidentales, aquellas que son ajenas a la decisión que estamos sopesando. Este tipo de emociones "colaterales" acaban impactando en los públicos y afectando a la calidad y precisión de las inversiones.

Para demostrar esto, se indujo a los participantes del estudio una serie de emociones incidentales o secundarias. A la mitad del grupo se les indujo una emoción de enfado y a la otra mitad felicidad. Después se dividió en dos a los participantes del estudio y se les junto en parejas para que realizaran un "juego de ultimátum", es decir, que uno de los participantes le ofreciera al otro cómo dividir $10. Tenía dos opciones, o quedarse él el 75% del dinero y entregar el resto, o al revés. Si el receptor de la oferta no aceptaba ese 25%, ninguno de los dos recibiría dinero. Después se cambiaron las parejas y el receptor inicial se convirtió en el oferente y se repitió el proceso, pero pudiendo hacer la oferta que quisieran.

Se comprobó que aquellos que habían estado sujetos a enfados o emociones negativas acabaron ganando menos dinero que aquellos a los que se les había inducido una emoción positiva. El 40% de las personas felices no aceptaron la oferta injusta mientras que el 73% de los enfadados la rechazaron. Los enfadados en la siguiente ronda fueron más cuidadosos y ofrecieron quedarse con menos dinero que los anteriores (5.8$ de media). Obtuvieron unas conclusiones muy interesantes para nosotros, confirmaron que las personas sujetas a emociones negativas (enfado) fueron más justos a la hora de hacer ofertas, es decir, que los seres humanos somos más cautos y tomamos menos riesgos cuando sufrimos emociones negativas.

[29] Daniel Kahneman, 2011.
[30] Andrade, E. B., & Ariely, D, 2009.

Un estudio[31] cuya importancia es mayúscula para comprobar todas estas hipótesis en el ámbito económico trazó la relación entre el rendimiento de trabajadores de banca de inversión y las emociones que sufrían durante sus horas de trabajo. Descubrieron que existe una relación muy estrecha entre las "corazonadas" y las emociones. El estudio reveló que aquellos *traders* que obtenían mejores resultados, tenían un mayor control de sus emociones y sabían cómo gestionarlas, mientras que aquellos que obtenían menor rentabilidad, gestionaban peor sus emociones, sobre todo ante situaciones desfavorables.

Otra de las características propias de aquellos *traders* con mayor rendimiento es su capacidad para acompañar sus corazonadas con datos e información, y no dejarse llevar solo por su intuición, mientras que los *traders* con peor rendimiento tienden a dejarse llevar por su intuición.

Este tercer bloque nos sirve para conseguir una mejor comprensión de los procesos lógicos y emocionales que sigue el ser humano durante la toma de decisiones. Gracias al estudio de la literatura sobre cómo afectan las emociones a las decisiones, podemos confirmar ciertas hipótesis iniciales que teníamos cómo que cuando los seres humanos sentimos miedo somos más cautelosos, y que mientras estamos felices o eufóricos, corremos más riesgos. Esto tendrá grandes implicaciones en el modelo de predicción de oportunidades bursátiles que plantearemos más adelante, dónde utilizaremos los conocimientos obtenidos en este capítulo como base teórica.

[31] Fenton-O'Creevy, M., Soane, E., Nicholson, N., & Willman, P, 2011.

EFECTO DE LAS EMOCIONES A LAS INVERSIONES BURSÁTILES

La relación entre las emociones, las decisiones que tomamos y, por consiguiente, nuestro rendimiento en el mercado de valores, es un problema muy complejo que se lleva estudiando durante muchos años. Para muchos, la clave a la hora de invertir nuestro dinero es dejar las emociones de lado, e intentar ser lo más objetivo posible con la información que disponemos. Esta es la teoría, pero los humanos somos seres emocionales, gobernados por nuestros impulsos y sentimientos, por lo que la realidad de nuestro comportamiento es muy compleja.

En esta parte del proyecto trabajaremos a partir de los estudios anteriores para crear nuestra metodología de proyección emocional que permita predecir posibles patrones de comportamiento efectivos para detectar oportunidades bursátiles. Antes de esto estudiaremos el efecto que tiene la información desestructurada, las noticias que leemos en periódicos, redes sociales o vemos en la televisión y las emociones en el mundo bursátil y en los inversores.

La información desestructurada es aquella que no sigue unos formatos o estructura predeterminada, a diferencia de la información estructurada, la cual está organizada en tablas o bases de datos. La información desestructurada no tiene un esquema fijo, puede ser, entre otras, en forma de texto libre, imágenes, videos, audios o publicaciones en las redes sociales.

En los últimos años se ha demostrado que este tipo de información puede llegar a influenciar en gran medida las cotizaciones de la bolsa de valores o incluso el del mercado de materias primas[32]. En concreto, el estudio de la actividad en la red social Twitter[33] sobre un determinado tema podría indicarnos la dirección del movimiento del mercado. Se llevó a cabo una recopilación de cerca de 380 mil tweets relacionados con el anuncio de los presupuestos de India en 2015. Tras hacer un análisis de sentimientos de los mensajes se reconoció como emoción predominante la alegría y la satisfacción por los resultados, lo que

[32] Feuerriegel, S., & Neumann, D, 2013.
[33] Khatua, A., & Khatua, A., 2016.

supuso un crecimiento del mercado. Llegaron a la conclusión de que el estudio de las redes sociales puede ser útil para predecir los movimientos del mercado. Plantearon la pregunta de cómo de influyentes son los medios electrónicos en la creación de opiniones de las masas, lo cual se tratará en este proyecto.

En el estudio[34] de los artículos más relevantes sobre la predicción de movimientos en el mercado durante los últimos años (2014-2018), se concluyó que los datos más usados y más precisos para predecir los movimientos de mercado eran los datos técnicos. Sin embargo, también descubrieron que aquellos modelos que utilizaban además datos obtenidos de las redes sociales, más concretamente, los sentimientos expresados en las redes, conseguían obtener mayor precisión. Afirmaron que los modelos más precisos son aquellos que usan Machine Learning Algorithms (MLA), los cuales funcionan mejor que los modelos de deep learning (SVM o ANN).

El 80% de la información perteneciente a una industria esta desestructurada[35] en forma de PowerPoint, audio, video o en redes sociales. Esto, junto con la gran influencia que tienen las noticias de los medios tradicionales, constituye una gran parte de los elementos que influyen en el movimiento de mercado, por lo que los autores recomiendan realizar un estudio híbrido, añadiendo los medios tradicionales a la información desestructurada, lo que permite mejorar la precisión.

Como se ha mencionado anteriormente, los medios tradicionales y las noticias publicadas tienen un gran impacto en las cotizaciones de las empresas. Estos medios tienen la capacidad de influir a cientos de miles de personas cada vez que escriben un artículo y por ello, si quisieran, podrían manipular el mercado de valores a su antojo afectando a nuestros sentimientos o dándonos solo la información que ellos consideran relevante. Evidentemente, los códigos éticos, los editores y los supervisores vigilan que estos comportamientos no se produzcan y, cuando sucede, están fuertemente castigados.

[34] O Bustos, A. Pomares-Quimbaya, 2020.
[35] Rajakumar, M. P., Jegatheesan, R., Chandy, R., & Sampath, T, 2019.

Pero volviendo al impacto de las noticias en las cotizaciones, tenemos un claro ejemplo con la valoración bursátil de las denominadas empresas verdes[36]. En general, estas empresas tienen objetivos a medio/largo plazo, por lo que sus cotizaciones deberían ser más o menos estables. Sin embargo, se ha comprobado que su cotización se ve fuertemente afectada por las noticias que surgen sobre estas empresas. Este cambio no suele ser permanente, dura poco tiempo, lo que les sirve para concluir que, realmente, la valoración de este tipo de compañías es muy sensible a las noticias que surgen a su alrededor.

Este efecto es aún más exagerado en mercados emergentes, como se comprobó en el caso de Uganda[37], donde se descubrió que las noticias positivas tenían un efecto mayor en las cotizaciones que una noticia negativa de la misma magnitud. Para llegar a esta conclusión emplearon el modelo de auto regresión generalizada condicional heteroscedástica exponencial[38] ,cuyo objetivo es predecir la volatilidad y hacer un análisis de riesgo. Este artículo que acabamos de mencionar es especialmente interesante para nosotros, ya que demuestra lo influyente que son las noticias en las fluctuaciones de mercado, aunque tiene la limitación de tratarse de un estudio sobre un mercado emergente y nuestro trabajo se basará en un mercado maduro/desarrollado.

Otro grupo analizó el efecto que tiene las propias comunicaciones de las empresas[39], por ejemplo, mediante ruedas de prensa, presentaciones de resultados o la presentación de planes estratégicos, en la valoración de la empresa en el mercado. Se estudió el rendimiento anómalo de una compañía respecto el resto del mercado con el objetivo de ver que *inputs* son los que hace que la valoración de la empresa crezca. Llegaron a la conclusión de que los comunicados sobre los temas de energía renovable, test de fármacos e investigación médica eran los que más hacían crecer el valor de la empresa.

[36] Justin Robinson, Adrian Glean, Winston Moore, 2018.
[37] Emenike, K. O., & Enock, O. N., 2020.
[38] GARCH, por sus siglas en inglés: Generalized Autoregressive Conditional Heteroskedasticity
[39] S. Feuerriegel, A. Ratku and D. Neumann, 2016.

Volviendo al tema de las emociones, la mayoría de artículos afirman juegan un papel fundamental en la toma de decisiones de inversión, pero algunos estudios indican que no se puede asegurar que los inversores, sobre todo los que tienen más experiencia, respondan emocionalmente ante los cambios de valor[40] en el mercado, a no ser, claro, que tenga una relevancia personal.

Durante los últimos años han surgido muchos modelos cuyo objetivo es predecir las fluctuaciones del mercado con el objetivo de maximizar los beneficios obtenidos al invertir en el mercado de valores. Uno de ellos es un modelo híbrido[41], que tiene en cuenta tanto el efecto de los sentimientos como el precio y fluctuación histórica de las acciones. Para ello utiliza varias técnicas de POS en noticias de medios tradicionales, analizando los sentimientos que un lector encontrará al leer dichas noticias y comparando con las fluctuaciones de valor históricas de la acción pretende predecir el valor que alcanzará a corto plazo, basándose en experiencias pasadas, consiguiendo casi un 85% de precisión.

Otro modelo híbrido se basa en la ingeniería de Kansei junto con *self-organazing maps*[42] cuyo modelo se basa en el estudio de riesgos para minimizar las acciones cuyo riesgo es alto y conseguir un mayor beneficio.

Se consiguieron pruebas empíricas de que las emociones que los inversores al actuar en el mercado, pueden ser beneficiosas[43] siempre y cuando sean capaz de gestionarlas correctamente. Los inversores que sienten una gran pasión por sus inversiones consiguieron mejores resultados en su estudio. Según ella esto se debe a que son capaces de mantener sus emociones fuera de sus decisiones y que estos fueron capaces de entender las emociones que sentían, lo que les permitía ser más objetivos a la hora de tomar decisiones. Resulta interesante el hecho de que contradiga una recomendación habitual en el mundo de las inversiones, la cual afirma que es importante dejar las emociones de lado a la hora de invertir.

[40] Darren Duxbury, Tommy Gärling, Amelie Gamble & Vian Klass, 2020.
[41] Sonam and M. Devaraj, 2020.
[42] Hai V. Pham, Eric W. Cooper, Thang Cao, Katsuari Kamei, 2014.
[43] Lisa Feldman, 2007.

Este estudio promueve la comprensión de las emociones que sentimos, creando una nueva corriente que da importancia a los sentimientos y no los ignora e intenta que desaparezcan.

Estas conclusiones casan perfectamente con otro estudio[44], donde tras analizar los sentimientos y mensajes de *traders* durante varios años observaron que aquellos que sentían pocas emociones y los que sentían muchas tuvieron un rendimiento peor que los *traders* que exhibían una cantidad de emociones moderadas, este estudio lo llevaron a cabo haciendo un análisis de la actividad online de los traders, concretamente sus mensajes, y la calidad de sus *trades.*

Los humanos nos dejamos influir enormemente por nuestras emociones, las pequeñas cosas pueden llegar a marcar grandes diferencias en nuestra capacidad de toma de decisiones[45]. Existe una diferencia cuantitativa en cuanto al precio de las acciones en la bolsa de Nueva York, dependiendo del tiempo[46] que hacía ese día. Los días nublados, en los que el tiempo se consideraría malo, el valor de las acciones tiende a ser menor que los días soleados. Por otro lado, aquellas familias que sean más sociales e interactúen más con otras familias tenderán a invertir más en bolsa, definiendo el componente social como algo muy importante a tener en cuenta. También afirman que los resultados del equipo de rugby influyen enormemente a las inversiones de los seguidores de dicho equipo, en los meses donde los resultados son buenos, los beneficios son mejores que los obtenidos cuando los resultados son malos. A esta misma conclusión llegaron en otro estudio[47], dónde se analizaron las inversiones de los seguidores de un equipo comparándolas con los resultados de su equipo frente a su eterno rival.

Este último bloque nos ha permitido centrarnos en el objetivo del proyecto, que es estudiar cómo afectan las emociones a la forma que tenemos los seres humanos de invertir y a las fluctuaciones del mercado. Conseguimos confirmar la hipótesis inicial que teníamos, de que

[44] Bin Liu, Ramesh Govindan, Brian Uzz, 2016.
[45] John W. Goodell, Satish Kumar, Purnima Rao, Shubhangi Verma, 2022.
[46] Brian M. Lucey, 2005.
[47] Demir y Rugoni, 2006.

las emociones tienen un papel primordial en las inversiones que realizamos, descubriendo que hasta las pequeñas circunstancias del día a día pueden ser diferenciales, cuestiones como que tu equipo gane ese fin de semana o que ese día este nublado, pueden ser sesgos inconscientes determinantes para la toma de decisión de inversión (y para el resultado de las mismas).

Es muy interesante el estudio de Lisa Feldman, a partir del cual hemos descubierto que, a la hora de invertir, lo importante no es el hecho de no sentir emociones, sino el hecho de ser capaz de comprender y trabajar con las emociones que sentimos.

CONCLUSIONES

El estudio y análisis de la literatura correspondiente a los cuatro bloques en los que se dividió el trabajo desde un comienzo ha sido fundamental para identificar las principales tendencias y enfoques metodológicos que han surgido en los distintos campos que hemos tratado.

Gracias a este análisis exhaustivo de la literatura, hemos sido capaces de comprender la importancia de las técnicas de preprocesado de textos, que son multidisciplinares, y se han encontrado en tres de los cuatro bloques, y son fundamentales para el correcto desarrollo de la metodología que utilizaremos en los siguientes capítulos.

También hemos podido aumentar nuestros conocimientos sobre el funcionamiento de la mente humana y los procesos lógicos y emocionales que seguimos las personas durante la toma de decisiones. Poniendo particular interés en las emociones que sentimos al leer noticias o al invertir en bolsa.

Por otro lado, hemos confirmado la importancia de la información desestructurada y la gran influencia que tienen los medios en las fluctuaciones de los mercados. Estas dos afirmaciones serán de especial relevancia a la hora de crear nuestra metodología y las tendremos muy presentes.

En resumen, el estudio de la literatura es fundamental para poder confirmar ciertas suposiciones iniciales que teníamos, para observar los distintos métodos que han surgido a lo largo de los años y las principales soluciones que otros investigadores proponen al problema que nos atañe. En este punto, estamos en disposición de presentar un enfoque teórico propio que aborde los desafíos indicados en la literatura.

3. DE LA NARRATIVA A LAS EMOCIONES Y DE LAS EMOCIONES AL VALOR

El propósito de este capítulo es explicar la formulación matemática y el método que usaremos para realizar la clasificación de un texto según su temática y la clasificación del mismo texto según las emociones que el autor ha plasmado en él. Este capítulo será la base a partir de la cual podremos llevar a cabo el objetivo de este proyecto, que es la creación de una metodología de predicción de oportunidades bursátiles gracias al estudio de las emociones humanas.

Dividimos en dos secciones este capítulo, la primera será el estudio de la clasificación por temática y análisis de sentimientos de un titular y la segunda el modelo de regresión polinomial que explicará cómo afectan las emociones de la primera sección a las fluctuaciones del mercado. Esta división en dos secciones se debe a las similitudes que tienen, en cuanto a técnicas de preparación del texto, la clasificación por temática y emociones. En ambos es necesario llevar a cabo técnicas de preprocesado del texto, NLP y POS, con el objetivo de simplificar la compresión del texto para el código y aumentar su precisión. En esta primera sección del capítulo, se tratará primero el problema de la clasificación por temática y, posteriormente, el de la clasificación de las emociones.

Se ha decidido centrar el análisis en el estudio de los titulares de las noticias, frente a otros textos más largos como pueden ser las propias noticias y/o los editoriales de opinión. El titular ya ofrece una primera valoración de la noticia y, con frecuencia, es lo único que se lee en una lectura rápida, y con lo que el lector[48] se suele "quedar" de una información.

[48] Lagerwerf, L., & Govaert, C. G., 2021.

En la segunda parte del capítulo explicaremos el método que vamos a utilizar para definir el impacto que tienen los titulares de los medios de comunicación analizados en las fluctuaciones del mercado. Nos servirá de base lo mencionado en las dos últimas secciones. El objetivo es diseñar un método que permita predecir oportunidades bursátiles basándonos, fundamentalmente, en los sentimientos que se inducen en los titulares y en la categoría a la que pertenecen dichos titulares.

Realizaremos una recopilación de noticias de los principales medios de comunicación nacionales: El confidencial, El País, La Razón y El Mundo. Se recopilarán veinte noticias al día de cada uno de los periódicos durante un periodo de tiempo significativo, con el objetivo de estudiar su temática, polaridad de las emociones transmitidas y las fluctuaciones del mercado durante ese periodo de tiempo y, a partir de esos datos, determinar mediante un modelo de regresión, el tipo de relación que existe entre esas variables que consideraremos dependientes. Utilizaremos un modelo de regresión polinomial ya que, inicialmente, desconocemos el tipo de relación entre nuestras variables.

Se ha elegido un modelo de regresión polinomial respecto a otros modelos debido a que nos permite crear el modelo, aunque no sepamos en un principio el tipo de relación que existe entre nuestras variables dependientes. Sin embargo, deberemos ser cautos a la hora de no sobreajustar nuestro modelo.

CLASIFICACIÓN SEGÚN SU TEMÁTICA

El objetivo de esta sección es explicar el proceso que usaremos para clasificar los titulares de las noticias según su temática. El primer paso es la extracción del titular de los sitios web de los periódicos más relevantes a nivel nacional. Para ello, utilizaremos una extensión de Google llamada "Web Scrapper"[49], que nos permite obtener un archivo *.xlsx* con los datos deseados. Esta herramienta facilita el acceso a todos los titulares de los principales medios de comunicación online del país. Para intentar garantizar la relevancia de las noticias, utilizaremos las 20 primeras noticias que encontramos en los siguientes medios digitales, El País, La Razón, El Confidencia, y El Mundo. A continuación, se muestra un pequeño esquema de los pasos que vamos a seguir en esta sección:

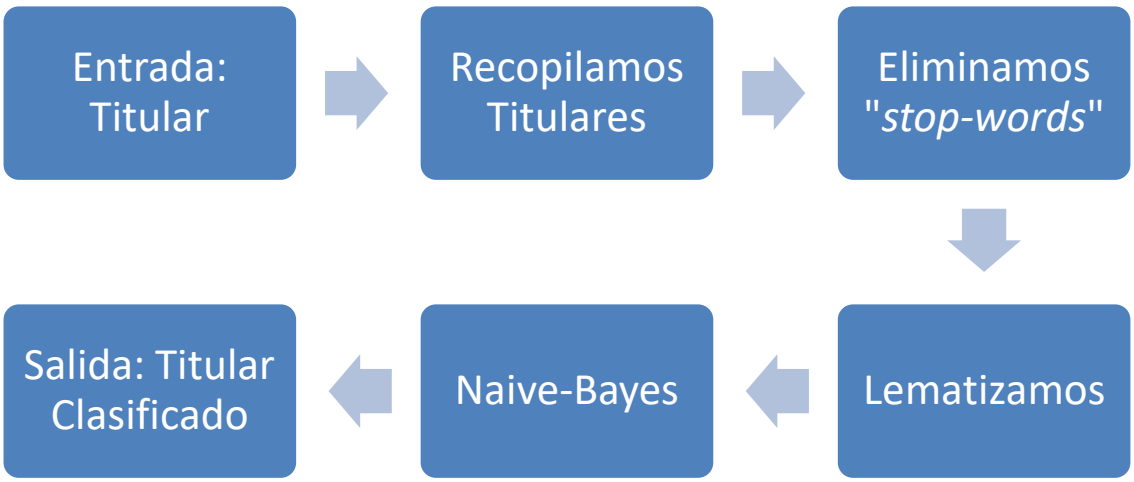

Ilustración 1: Esquema del proceso a seguir en la clasificación de titulares según su temática. Fuente: Elaboración propia, 2023.

[49] Herramienta desarrollada por la empresa Data Miner

Se almacenará el texto de cada titular en un tensor $\boldsymbol{T_{i,j}}$, donde *i* es el periódico al que pertenece la noticia y *j* es el número de noticia de las veinte que recolectamos por periódico. $\boldsymbol{T_{i,j}}$ es un tensor de orden dos, es decir, una matriz, de orden $\boldsymbol{M_{lxl}}$ diagonal, que está formada por *l* elementos, dónde cada elemento de la diagonal es una palabra del titular, de tal forma que λ_1 es la primera palabra y λ_l es la última palabra.

Se definen a continuación los distintos términos de $\boldsymbol{T_{i,j}}$ y $\boldsymbol{M_{lxl}}$:

a. *i*=1 es El confidencial,
b. *i*=2 es El País,
c. *i*=3 es La Razón
d. *i*=4 es El Mundo
e. *j* representa el número del titular al que se refiere, un número del 1 al 20.
f. *l* representa el número de palabras dentro de cada titular $\boldsymbol{T_{i,j}}$.

Por ejemplo, la portada de El Confidencial recoge la siguiente noticia, que usaremos de ejemplo a lo largo de este capítulo:

$$\boldsymbol{T_{1,1}} = "Vox\ no\ garantiza\ investir\ a\ Azcón\ en\ Aragón\ \ tras\ cerrar\ la\ Mesa:$$

$$"Depende\ del\ pacto\ programático""$$

Obteniendo la siguiente matriz: $$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{Vox} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{programático} \end{bmatrix}_{\boldsymbol{lxl}}$$

Cada elemento del tensor $\boldsymbol{T_{i,j}}$ está caracterizado por las distintas λ_k , dónde λ es cada uno de los autovalores del tensor, es decir, cada una de las palabras del tensor y *k* expresa el orden de la palabra. Por ejemplo, para *k*=4 o *k*=16:

$$\boldsymbol{\lambda_4} = "investir" \text{ o } \boldsymbol{\lambda_{16}} = "programático"$$

Una vez expresado cada titular como un tensor distinto, el siguiente paso consiste en eliminar aquellas palabras que no aportan valor, la lista de palabras que eliminaremos se incluye en el Anexo B. Denominaremos $\boldsymbol{L_m}$ a la lista de palabras que no aportan valor, dónde *m* indica la posición de la palabra en la lista, por ejemplo $\boldsymbol{L_2} = "estad"$. Guardamos el resultado de cada autovalor $\boldsymbol{\lambda_k}$ tras haber hecho la eliminación de palabras no relevantes en los autovalores $\boldsymbol{\beta_k}$ del nuevo tensor $\boldsymbol{\dot{T}_{i,j}}$.

$$Sí\ \lambda_k - L_m = 0\ \ \forall m,k \rightarrow Eliminamos\ palabra \rightarrow\ \lambda_k = 0 \qquad [1]$$

$$Sí\ \lambda_k - L_m \neq 0\ \ \forall m,k \rightarrow No\ eliminamos\ palabra \rightarrow\ \beta_k =\ \lambda_k \qquad [2]$$

De tal forma que expresamos el nuevo tensor $\dot{T}$ de la siguiente forma:

$$\boldsymbol{\dot{T}_{i,j}} = \sum \beta_k\ \forall k \qquad [3]$$

En el ejemplo puesto anteriormente, el resultado será el siguiente:

$$\dot{T}_{i=Confidencial,j=1}$$
$$= Vox\ no\ garantiza\ investir\ Azcón\ Aragón\ tras\ cerrar\ Mesa: "Depende\ pacto\ programático"$$

Una vez realizado el segundo paso, procedemos a lematizar las palabras restantes[50]. Para mejorar la precisión del sistema, no lematizaremos aquellas palabras que reconozcamos cómo entidades[51], es decir, lugares, personas, organizaciones y derivados. Denominamos $\boldsymbol{E_n}$ a la lista (*E*) de palabras que están reconocidas como entidades y su posición en dicha lista (*n*).

De esta forma el nuevo tensor se expresará de la siguiente forma:

$$Sí\ \beta_k\ - E_n = 0\ \ \forall n,k \rightarrow No\ se\ lematiza\ la\ palabra \rightarrow\ \Omega_k\ = \beta_k \qquad [4]$$

[50] Utilizamos la biblioteca 'nltk' en Python.
[51] Utilizaremos la biblioteca Flair en Python.

$$Sí\ \beta_k - L_m \neq 0\ \forall m,k \rightarrow Lematizamos\ la\ palabra \rightarrow\ \Omega_k = \beta_k - raíz \qquad [5]$$

$$\ddot{T} = \sum \Omega_k\ \forall k \qquad [6]$$

Donde Ω es la palabra β, pero lematizada.

Volviendo al ejemplo con el que estamos trabajando:

$$\ddot{T}_{i=Confidencial,j=1} = vox\ no\ garantiz\ invest\ azcon\ aragon\ tras\ cerr\ mesa:$$

$$"depend\ pact\ programatico"$$

Una vez finalizado el preprocesado del texto podemos empezar el proceso de clasificar los titulares según temática. Para ello, utilizaremos una base de datos llamada "*Spanish News Classification*"[52], que nos proporciona una colección de noticias, que incluye un enlace a la noticia, el cuerpo de la noticia y la temática de la misma. No es un data set perfecto, ya que no nos proporciona el título de la noticia, por lo que tuvimos que desarrollar un código que abriese los enlaces que proporciona el *dataset* y, obtuviese el título[53] de las noticias[54]. Otra limitación del *dataset* es la cantidad de temáticas que abarca, solo separa las noticias en 8 temáticas distintas, las cuales serán denominadas "*y" Macroeconomics (y=1), Sustainability (y=2), Innovation (y=3), Regulations (y=4), Alliances (y=5), Reputation (y=6) y Other (y=7).*

Una vez el *dataset* está ajustado a nuestras necesidades, usaremos el método de Naive-Bayes[55] para "comparar" las palabras usadas en nuestros titulares con las del *dataset* y poder

[52] El dataset está disponible y es de uso público en la página web *Kaggle*, su autor es Kevin Morgado.
[53] Con la biblioteca de Python BeautifulSoup.
[54] Explicado en más detalle en el apartado de código 2, del anexo A.
[55] Naive (ingenuo en español), ya que se hace la suposición ingenua de independencia condicional entre las características.

"predecir" la temática más probable del titular. Para llevar a cabo el método nos apoyaremos en el teorema probabilístico de Bayes:

$$P(A_y|B) = \frac{P(A_y \cap B)}{P(B)} = \frac{P(A_y) \times P(B|A_y)}{P(A_1) * P(B|A_1) + \cdots + P(A_8) * P(B|A_8)} \quad [7]$$

Donde:

- $P(A_y|B)$ son las probabilidades finales o a posteriores, es decir, la probabilidad de que un determinado titular sea de una categoría (y) u otra.
- $P(A_y)$ es la probabilidad a priori, es decir, la probabilidad de que cada titular pertenezca a una categoría antes de tener en cuenta sus características observadas. Se obtiene dividiendo el número de instancias en cada clase entre el número total de instancias. Nuestro *dataset* tiene 1217 noticias, de las cuales 340 son de macroeconomía, luego $P(A_1) = \frac{N_{A1}}{N_T} = \frac{340}{1218} = 0.279$. Donde N_{A1} es el número total de titulares sobre macroeconomía del dataset y N_T es el número total de titulares del dataset, 1217.
- $P(B|A_y)$ denominadas verosimilitudes o probabilidades condicionales, es decir, las probabilidades de un suceso condicionado por otro.

Utilizando las explicaciones anteriores sobre el teorema de Bayes[56], anteriores, podemos resumir el modelo que utilizaremos en los siguientes pasos:

[56] Rish, I., 2001.

I. Preparación de los datos de entrenamiento: Se adecua el *dataset*, cada instancia tiene ciertas características y una etiqueta de la clase asociada, en nuestro caso se trata del titular de la noticia y la categoría a la que pertenece, respectivamente.

II. Estimación de las probabilidades a priori $P(A_y)$: De la misma manera que en los problemas probabilísticos de Bayes, se calculan las probabilidades a priori $P(A_y)$, es decir, la probabilidad de que cada titular $\boldsymbol{T_{i,j}}$ pertenezca a una categoría A_y antes de tener en cuenta sus características observadas. Se obtiene dividiendo el número de instancias en cada clase N_{A1} entre el número total de instancias N_T. A continuación, se muestra una tabla donde se indica el número de instancias por cada categoría A_y posible del *dataset* y una gráfica de barras de la prioridad a priori $P(A_y)$.

A_y	$P(A_1) = \frac{N_{A1}}{N_T}$	N_A
Macroeconomía	0.2793	340
Sostenibilidad	0.1125	137
Innovación	0.1602	195
Regulaciones	0.1166	142
Alianzas	0.2029	247
Reputación	0.0213	26
Otras	0.1068	130

Total	1	N_T =1217

Tabla 1: Probabilidad a priori y número de instancias de cada clase. Fuente: elaboración propia, 2023.

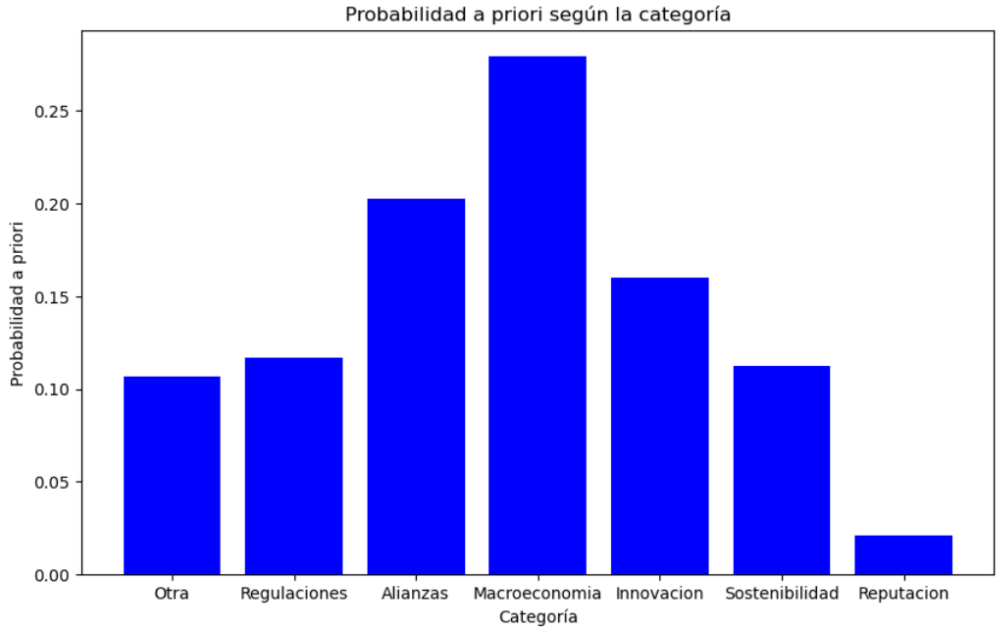

Ilustración 2: Probabilidad a priori. Fuente: elaboración propia, 2023.

III. Estimación de las probabilidades condicionales $P(B|A_y)$: Observando las características de los titulares, se calcula la probabilidad condicional $P(B|A_y)$, se calcula la posibilidad de que una palabra λ especifica aparezca en un título.

IV. Cálculo de la probabilidad posterior $P(A_y|B)$: Se utiliza el teorema de Bayes, para que, dado un título de una noticia $\boldsymbol{T_{i,j}}$, se calcule la probabilidad posterior $P(A_y|B)$ de que pertenezca a una de las ocho categorías posibles y estudiando sus características. Siguiendo el teorema de Bayes, se multiplica[57] la probabilidad

[57] Ecuación 7: $P(A_y|B) = \frac{P(A_y \cap B)}{P(B)} = \frac{P(A_y) \times P(B|A_y)}{P(A_1) * P(B|A_1) + \cdots + P(A_8) * P(B|A_8)}$

a priori de cada clase, por la probabilidad condicional $P(B|A_y)$ correspondiente para cada categoría *y*, posteriormente, se normaliza para obtener una probabilidad total.

V. Clasificación: El titular $\boldsymbol{T_{i,j}}$ se clasifica en la categoría y con la probabilidad posterior $P(A_y|B)$ más alta.

Siguiendo con el ejemplo[58] con el que hemos estado trabajando:

$$\boldsymbol{T_{1,1}} = \text{“}Vox\ no\ garantiza\ investir\ a\ Azcón\ en\ Aragón\ tras\ cerrar\ la\ Mesa\text{:}$$

$$\text{"}Depende\ del\ pacto\ programático\text{"}$$

$$Obtenemos\ y = Macroeconomía$$

En esta sección del capítulo hemos explicado los pasos que se suelen seguir en las técnicas de preprocesado del texto. Posteriormente hemos explicado de forma teórica la técnica clasificación de *Naive Bayes*. Finalmente, con el objetivo de comprobar la eficacia del método, hemos calculado la precisión del clasificador, que ha sido aceptable, sobre todo por sus altos porcentajes en las categorías de macroeconomía, innovación y alianzas, las cuales consideramos más importantes a la hora de influir en la fluctuación del mercado.

[58] Todo este proceso se hace de manera automática gracias a un código que hemos desarrollado, disponible en el Anexo A. Se comprobó la precisión del código mediante la técnica de validación cruzada y se obtuvo una precisión del 75.19%.

CLASIFICACIÓN DE LAS EMOCIONES DE UN TEXTO

El objetivo de esta sección es explicar el método que utilizaremos para clasificar las emociones observadas en un titular. En un principio planteamos la idea de utilizar un diccionario de emociones, que nos diese las emociones más probables de cada palabra para hacer un sumatorio, y elegir como emoción predominante aquella cuyo valor sea mayor. Sin embargo, como nos hemos decidido por utilizar titulares de noticias en vez de textos de mayor longitud, el hecho de evaluar varias emociones en un texto tan corto nos genera desconfianza sobre la precisión que podremos esperar. Por ello, nos hemos decidido a un enfoque binomial, en el que leeremos el titular de una noticia y decidiremos si se aprecian emociones positivas, negativas o neutras, dándole un valor perteneciente al intervalo (0,1) que expresará el grado de confianza de la predicción. A continuación, se muestra un diagrama de los pasos que vamos a seguir en esta sección del capítulo:

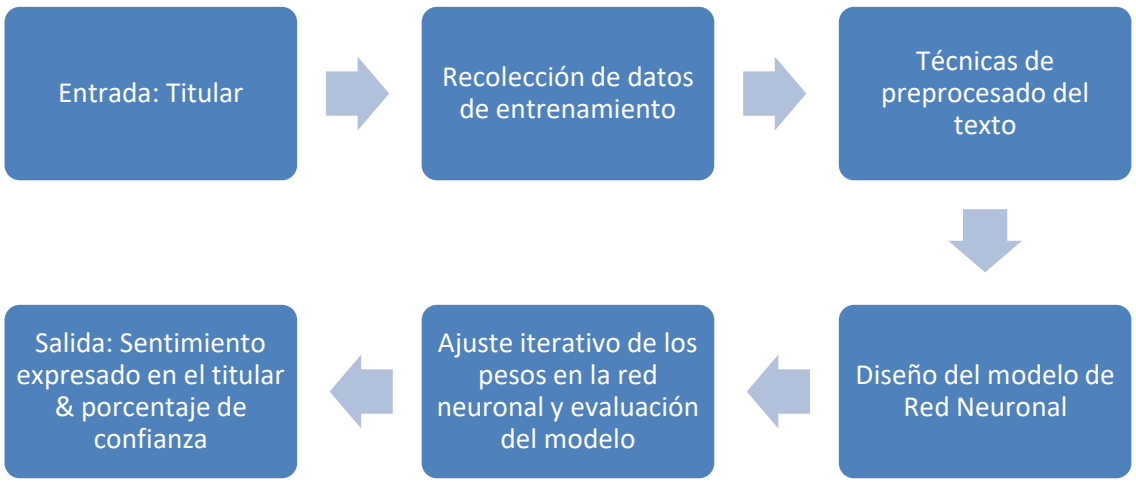

Ilustración 3: Esquema de los pasos a seguir en esta sección. Fuente: Elaboración propia, 2023.

De la misma manera que en la sección anterior, el titular se guarda en forma de texto en un tensor $\boldsymbol{T_{i,j}}$, donde *i* es el periódico al que pertenece la noticia y *j* es el número de noticia de las veinte que recolectamos por periódico.

$\boldsymbol{T_{i,j}}$ es un tensor de orden dos, es decir, una matriz, de orden $\boldsymbol{M_{lxl}}$ diagonal, que está formada por *l* elementos, dónde cada elemento es una palabra del titular, de tal forma que λ_1 es la primera palabra y λ_l es la última palabra.

Se realiza el mismo método de preprocesado, de tal forma que acabamos con un tensor como el de la *ecuación 6*, de la sección anterior*:*

$$\ddot{T}_{i,j} = \sum \Omega_k \; \forall k \qquad [6]$$

Utilizaremos la biblioteca Flair[59], una de las más famosas en cuanto a análisis de sentimientos se refiere. Una limitación a la que nos hemos enfrentado es que la biblioteca Flair solo funciona con textos en inglés o alemán, por lo que cada titular lo hemos traducido al inglés para que funcione correctamente.

Flair utiliza redes neuronales, a continuación, se explicará la metodología a seguir para utilizar redes neuronales[60] poniendo de ejemplo cómo lo hace Flair:

I. Al igual que con Naive-Bayes, el primer paso es la recolección de datos de entrenamiento, que consiste en ejemplos clasificados con etiquetas de positivo, negativo o neutro. Posteriormente, los datos se dividen en *train* y *test*, para comprobar la eficacia del sistema.

II. El texto a analizar ha de ser tratado mediante técnicas de preprocesado del lenguaje, como las vistas en la anterior sección, con el objetivo de convertir el

[59] Herramienta creada por Zalando Research, en noviembre de 2018. Están especializados en las técnicas NLP y su proyecto es de código abierto, por lo que es accesible para todo el mundo.
[60] Galushkin, A. I., 2007.

mismo en una serie de representación numérica para que pueda ser procesado por la red neuronal. Llamaremos a esta transformación Ψ_{RN} cuya entrada es $\ddot{T}_{i,j}$, texto ya lematizado, y su salida es $N_{i,j}$ dónde *N* es la representación numérica del titular.

- $N_{i,j} = \Psi_{RN}(\ddot{T}_{i,j})\ \forall i,j$ [8]

III. Se diseña el modelo de red neuronal a utilizar. Se debe de elegir una arquitectura adecuada para el análisis de sentimiento, en el caso de Flair, se utiliza LSTM[61] o las GRU[62] que son capaces de detectar y reconocer dependencias a lo largo del texto. Estas redes pueden incluir *embedding*, para representar palabras λ, capas recurrentes, para reconocer agrupaciones de palabras que suelen ir juntas y, por último, capas de salidas, para realizar la clasificación de sentimientos[63].

IV. Se recuperan las divisiones de los datos de *train* y *test*, para comprobar la eficacia del sistema, durante el proceso de aprendizaje, se va ajustando iterativamente los pesos de la red neuronal, con el fin de minimizar la función de pérdidas, que mide las diferencias entre las etiquetas predispuestas y las que son predichas por las redes neuronales.

V. Por último, una vez acabado el aprendizaje, se evalúa el modelo, para ello evaluamos su rendimiento, fijándonos en métricas como la precisión, *recall y F1-score*[64]. Si los resultados son satisfactorios ya se puede utilizar el modelo.

[61] Por sus siglas en inglés: Long Short-Term Memory.
[62] Por sus siglas en inglés: Gated Recurrent Unit.
[63] Galushkin, A. I., 2007.
[64] Puntuación que combina la precisión y exhaustividad (recall) de un modelo de clasificación.

De tal forma que la entrada de la red neuronal es la transformada mencionada anteriormente $N_{i,j}$ a la que se aplica el proceso ξ_{RN} el proceso de predicción de sentimientos que lleva a cabo la red neuronal, cuya salida es el sentimiento (Positivo, Negativo o Neutro) y el porcentaje de confianza.

$$Sentimiento = \xi_{RN}(N_{i,j})\ \forall i,j \qquad [9]$$

Recuperando el ejemplo con el que trabajamos la sección anterior, se pueden observar los siguientes resultados al ejecutar nuestro código[65]:

```
Vox no garantiza investir a Azcón en Aragón  tras cerrar la Mesa: "Depende del pacto programático"
Vox does not guarantee to invest Azcón in Aragon after closing the table: "It depends on the programmatic pact"
Sentimiento: NEGATIVE
Porcentaje: 99.98%
```

Ilustración 4: Resultados del análisis de sentimiento. Fuente: Elaboración propia, 2023.

Hemos impreso por pantalla también la traducción al inglés para poder comprobar que se ha hecho una traducción fiel. Se observa que, tras ejecutar el código, la red neuronal de Flair asigna un 99.98% de confianza a que ese titular es negativo.

Se puede considerar más que satisfactorio el método obtenido para clasificar los sentimientos encontrados en un texto. Utilizamos la biblioteca Flair, una de las más reconocidas a nivel mundial, que utiliza el método de redes neuronales, uno de los más utilizados y fiables según la literatura observada[66].

[65] El código está disponible y explicado en el Anexo A.
[66] Severyn, A., & Moschitti, A., 2015.

MODELO DE REGRESIÓN POLINOMIAL

Un modelo de regresión tiene como objetivo identificar la relación existente entre una o más variables independientes con una variable dependiente[67]. En nuestro caso, la variable independiente son los sentimientos expresados $S_{i,j}$[68] que estarán ponderados según la clasificación del titular según su temática $A_{i,j}$[69], siendo la variable S_T los sentimientos totales ponderados de un día *i*. La variable dependiente será V [70], que representa el valor de una determinada acción o conjuntos de acciones[71].

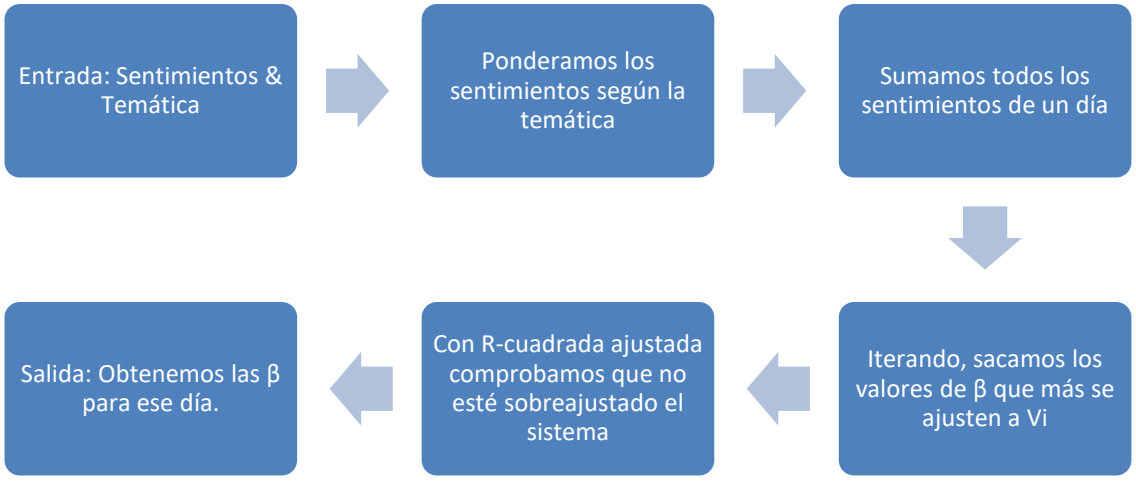

Ilustración 5: Proceso del método de regresión polinomial. Fuente: Elaboración propia, 2023.

[67] Ostertagová, E., 2012.
[68] *i* representa el periódico al que hace referencia y *j* al número del artículo al que se refiere, al igual que en la sección anterior.
[69] Ibid.
[70] Representa el valor de un único día.
[71] Se explicará en el capítulo de implementación del método.

De tal forma que la expresión con la que trabajaremos será la siguiente:

$$V_i = \beta_0 + \beta_1 * S_{T_i} + \beta_2 * {S_T}^2{}_i + \beta_3 * {S_T}^3{}_i + \cdots + \varepsilon \quad para\ i = 1,2,3 \ldots n \qquad [10]$$

El error ε está definido como la diferencia entre los valores reales de la variable dependiente y los predichos por el modelo[72], se asume que está distribuido por una normal con una media cero y una varianza σ^2 definida por el MSE[73].

$$MSE = \sum_{s=1}^{n} \frac{(V_s - \widehat{V_s})}{n-(k+1)} \qquad [11]$$

Dónde k representa el orden máximo del polinomio V, V_s simboliza los resultados observados, $\widehat{V_s}$ los obtenidos por parte de la variable dependiente y n el número de observaciones.

El objetivo de este capítulo será determinar el número de betas (β) a utilizar, es decir, el orden del polinomio, y su valor. Al conseguir eso habremos logrado definir un modelo de regresión polinomial. Se estudiarán los titulares del día juntos con las fluctuaciones del mercado para determinar los valores de β que mejor se ajusten a la realidad.

Primero debemos hacer una serie de suposiciones iniciales, que luego, a medida que vayamos iterando el método, iremos cambiando para que se ajusten más a la realidad. Denominaremos P_y, al valor (P) que se le asignará a cada una de las siete categorías (y), ya que, como es lógico, por lo general, no podrán influir tanto en el mercado de valores las noticias sobre reputación cómo las de macroeconomía. Los valores iniciales son los siguientes[74]:

- $P_1 = 1$

[72] Ostertagová, E., 2012.
[73] Por sus siglas en inglés: Mean Squared Error
[74] Se recuerda que el orden expresado en la sección anterior es: *Macroeconomics (y=1), Sustainability (y=2), Innovation (y=3), Regulations (y=4), Alliances (y=5), Reputation (y=6) y Other (y=7)*

- $P_2 = 0.9$
- $P_3 = 0.9$
- $P_4 = 0.8$
- $P_5 = 0.8$
- $P_6 = 0.5$
- $P_7 = 0.2$

Utilizaremos estos valores P_y para ponderar los sentimientos $S_{i,j}$ de tal forma que los sentimientos ponderados totales de un día S_T, se regirán por la siguiente ecuación:

$$S_T = \sum_{i,j}(S_{i,j} * P_y) \qquad [12]$$

Como el número de noticias será el mismo cada día, asignaremos un valor -1 a las noticias negativas y un valor 1 a las noticias positivas. De tal forma que, tras la ponderación, obtengamos una valoración global del sentimiento predominante ese día.

$$Sí \qquad S_{i,j} \rightarrow Positivo \rightarrow S_{i,j} = +1 \qquad [13]$$

$$Sí \qquad S_{i,j} \rightarrow Negativo \rightarrow S_{i,j} = -1 \qquad [14]$$

Para darle valor a las β utilizaremos el método de comprobación R^2, o R-squared[75] cuya fórmula es la siguiente[76]:

$$R^2 = 1 - \frac{\sum_{s=1}^{n}(V_s - \widehat{V_s})}{\sum_{s=1}^{n}(V_s - \overline{V})} \qquad [15]$$

$\bar{V}$ simboliza la media aritmética de la variable *V*. El valor de R^2 está siempre definido entre cero y uno. Si es mayor que 0.9 se considera una aproximación muy buena, entre (0.8, 0.9)

[75] Coeficiente de determinación o R-cuadrado.
[76] Ostertagová, E., 2012.

se considera buena y entre (0.6,0.8) se considera aceptable. Cuando R^2 es menor que 0.5, eso quiere decir que la regresión explica menos de la mitad de las variaciones de los datos. Consideraremos que los coeficientes de la ecuación (β) serán válidos cuando R^2sea mayor que 0.8.

Para comprobar que nuestra regresión no está sobreajustada usaremos la técnica de R^2 ajustada, R^{*2}:

$$R^{*2} = R^2 - \frac{(1-R^2)\times k}{n-(k+1)} \qquad [16]$$

Cuando R^{*2} es mucho mas pequeña que R^2 se considera que el sistema está sobreajustado, nos interesa una R^{*2} lo más grande posible, aunque siempre será menor que R-squared[77].

Este es el proceso que habría que seguir para encontrar el modelo de regresión polinomial diario. El objetivo es ser capaces de recolectar noticias y crear un polinomio diario durante un periodo lo más largo posible para crear una especie de *dataset*, que almacene dicho polinomio y los sentimientos globales del mercado ese día. El objetivo final sería que, al examinar la prensa del día, ser capaces de predecir los movimientos del mercado basándonos en la predictibilidad del ser humano ante ciertas emociones.

[77] Francisco Borrás, apuntes de Modelos Cuantitativos ICAI-ICADE, 2023.

CONCLUSIONES

Se ha planteado un modelo para clasificar los titulares extraídos de periódicos digitales según su temática y otro modelo para analizar la polaridad de los sentimientos que se plantean en dichos titulares.

Se ha realizado un ejemplo teórico y práctico de cómo extraer la temática de un titular utilizando Naive-Bayes y cómo utilizar una red neuronal para hacer un análisis de sentimientos binomial del titular mencionado anteriormente. Los resultados del ejemplo, combinando ambas metodologías estudiadas es el siguiente:

```
Vox no garantiza investir a Azcón en Aragón  tras cerrar la Mesa: "Depende del pacto programático"
Macroeconomia
Sentimiento: NEGATIVE
Porcentaje: 99.98%
```

Ilustración 5: Combinación de ambos códigos, en los resultados se expresa la temática a la que pertenece la noticia y el sentimiento que transmite con su grado de confianza. Fuente: Elaboración propia, 2023.

El modelo propuesto utiliza técnicas distintas, Naive-Bayes y redes neuronales. Naive-Bayes tiene como ventajas frente a otros métodos su simplicidad, se basa en un modelo probabilístico simple, el teorema de Bayes y es capaz de manejar características irrelevantes o redundantes con facilidad[78]. Por otro lado, las redes neuronales es un método más complejo, cuya principal ventaja es su facilidad para adaptarse a las necesidades específicas del método[79] y al usar la herramienta Flair, conseguimos ese extra de precisión y adaptabilidad, sin tener que programarlo.

En la segunda sección del capítulo hemos explicado el método de regresión polinomial que vamos a seguir. La entrada de este método son los sentimientos $S_{i,j}$, los cuales ponderaremos P_y según la importancia que le damos a la temática del titular $A_{i,j}$. A partir del sumatorio de

[78] Rish, I., 2001.
[79] Severyn, A., & Moschitti, A.,2015.

sentimientos de un día, calcularemos de forma iterativa las β de la ecuación de regresión y comprobaremos la eficacia del método mediante las pruebas de R^2 y R^{*2}.

La principal ventaja que plantea la regresión polinomial frente a la lineal, no lineal, o múltiple, es su adaptabilidad, ya que no sabemos todavía el tipo de relación que van a tener nuestras variables entre sí, por lo que un enfoque que mantenga las opciones abiertas nos beneficia considerablemente.

Otra razón por la que este método es beneficioso es que no nos cierra puertas a convertirlo en una regresión de avance paso a paso, a partir de la cual podremos ir añadiendo más entradas al sistema para que sea más completo. Por lo que sí, en el futuro, queremos retomar el proyecto y añadirle más complejidad podremos hacerlo.

4. DE LA TEORÍA A LA PRÁCTICA

En este capítulo mostraremos de manera práctica el método que hemos desarrollado en la sección anterior[80]. Empezaremos hablando del método para clasificar los titulares según su temática, seguido por el análisis de sentimientos de dichos titulares. Después explicaremos el proceso práctico para implementar la regresión polinomial, y lo compararemos con otros métodos que se han descartado como la regresión lineal o un árbol de regresión. Finalmente, haremos un caso práctico dónde intentaremos predecir las fluctuaciones del IBEX 35 a partir de los sentimientos encontrados en la prensa de ese día y el método de regresión que hemos implementado.

Para el caso práctico hemos creado un *dataset* de 40 noticias diarias que abarcan un periodo de 14 días hábiles de bolsa, del 1 de junio de 2023 al 20 de junio de 2023, ambos inclusive. Se han recolectado las 40 primeras noticias del día del periódico Expansión con el fin de coger solo los titulares más relevantes. El *dataset* contiene dos datos principales, el titular y la fecha a la que corresponde. Se ha escogido este periódico frente a los otros cuatro con los que hemos trabajado a lo largo del proyecto con el fin de ser imparcial.

El objetivo de este *dataset* es el de construir una base de datos a partir de la cual extraigamos la categoría a la que pertenece los titulares y los sentimientos que transmiten. Posteriormente, se calculará el sentimiento global ponderado de cada día, se comparará con las fluctuaciones del mercado, en concreto, con el cambio de valor porcentual del IBEX 35 y se comprobará la eficacia del método diseñado.

[80] Se ha conseguido automatizar todo el proceso mediante códigos, los cuales se incluirán en el Anexo A.

CLASIFICACIÓN POR TEMÁTICA

A partir de Naive Bayes hemos creado un clasificador por temática de los distintos titulares $\boldsymbol{T_{i,j}}$ que nos encontramos en los medios digitales. Como se mencionó en el capítulo anterior, este clasificador se apoya fuertemente en las ecuaciones probabilísticas de Bayes[81]. Sin embargo, lo más importante de este clasificador es, sin lugar a dudas, los datos que utiliza para comparar: el *dataset*.

Hemos utilizado un *dataset* llamado "*Spanish News Classification*"[82] el cual proporcionaba un enlace a la noticia, el cuerpo y la temática de la noticia de varios periódicos. Nos interesaba el titular de la noticia, por lo que tuvimos que desarrollar un código que fuese capaz de abrir el enlace y extraer el titular. De esa manera adaptamos el *dataset* a nuestras necesidades. Se extrajeron 1217 noticias, que pertenecen a 8 temáticas distintas (*y)*, cerca de un 28% eran de macroeconomía (*y=1)*. A continuación, se muestra una gráfica de nubes de palabras dónde se muestran aquellas palabras que se utilizan más a lo largo de los titulares del *dataset*:

Ilustración 3: Palabras más usadas del dataset. Fuente: Elaboración propia, 2023.

[81] La ecuación tiene la siguiente forma: $P(A_y|B) = \frac{P(A_y \cap B)}{P(B)} = \frac{P(A_y) \times P(B|A_y)}{P(A_1) * P(B|A_1) + \cdots + P(A_8) * P(B|A_8)}$

[82] Creado por Kevin Morgado.

También mostramos las palabras más utilizadas de cada categoría:

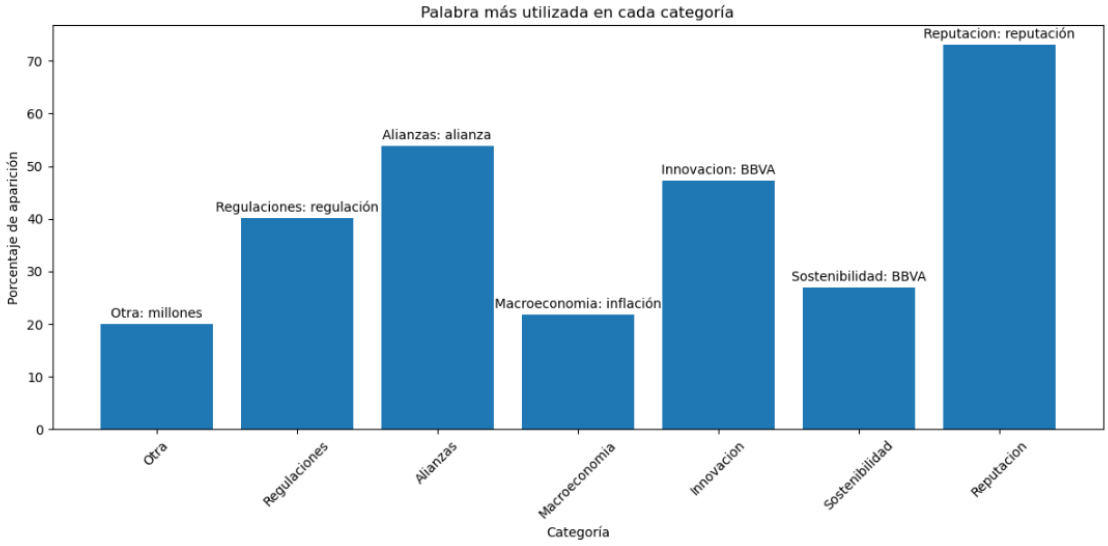

Ilustración 4: Palabras más usadas por categoría. Fuente: Elaboración propia, 2023.

Se puede observar la palabra λ= "BBVA" siendo predominante en la *ilustración 1* y siendo la palabra más usada en las categorías *(y)* de Sostenibilidad e Innovación. Esto muestra la importancia de un buen *dataset*, que contenga información de varios medios y durante un periodo de tiempo suficientemente alto.

Una vez completado el proceso de adaptar el dataset a nuestras necesidades, se entrenó el clasificador con la intención de mejorar su predicción mediante el uso de la biblioteca scikit-learn en Python. El primer paso fue la conversión de los titulares $\boldsymbol{T_{i,j}}$ en una representación numérica de características vectoriales. Se dividen los datos del *dataset* en *train* y *test* y a partir de los datos de *train* se intenta predecir las categorías de los datos del *test*, y se realizan varias iteraciones de esta validación cruzada, lo que ayuda conseguir una estimación más robusta con el clasificador. Finalmente se obtiene una precisión del clasificador Naive Bayes. En nuestro caso hemos conseguido una precisión media del 75.19%. A continuación, se muestra una gráfica de barras dónde se expresa la precisión por categoría del clasificador:

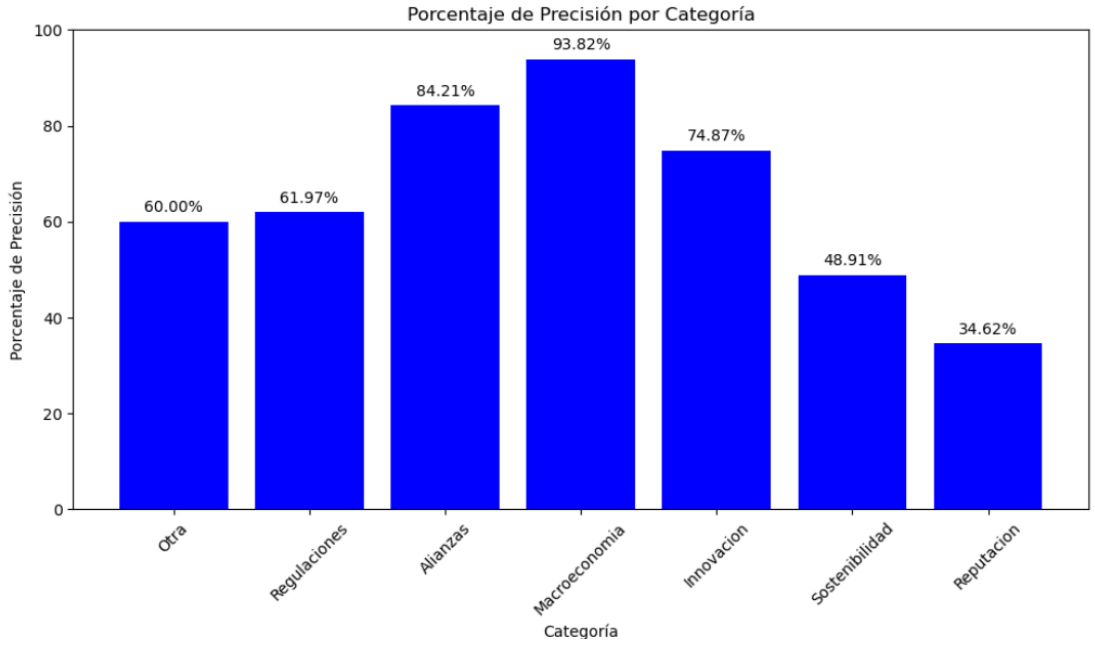

Ilustración 5: Precisión por categorías del clasificador. Fuente: Elaboración propia,2023.

La precisión es menor en categorías cómo Sostenibilidad o Reputación, pero no nos preocupa en exceso debido a que la gran mayoría de noticias que nos encontraremos y las más influyentes en las fluctuaciones del mercado son de macroeconomía, cuya categoría tiene una precisión de acierto muy alta. Se muestra también la matriz de confusión *(ilustración 4)* para entender mejor los fallos del método.

A continuación, se adjunta dos tablas para explicar mejor la matriz de confusión, pero antes se explicará el significado de los términos usados:

- Precisión: Representa la proporción de predicciones positivas correctas, respecto a todas las precisiones positivas por cada categoría
- Recall: También conocido como tasa de verdaderos positivos, es la proporción de instancias positivas correctamente identificadas por el modelo respecto a total de instancias positivas en los datos.
- F-1 Score: Combina precisión y recall en un solo valor.

- Support: número de instancias por cada categoría.

	Precisión	Recall	F1-Score	Support
Alianzas	0.85	0.84	0.85	247
Innovación	0.74	0.75	0.75	195
Macroeconomía	0.66	0.94	0.78	340
Otra	0.71	0.60	0.65	130
Regulación	0.91	0.62	0.74	142
Reputación	1	0.35	0.51	26
Sostenibilidad	0.84	0.49	0.62	137

Tabla 1: Representación de precisión, recall, F1-score y Support para las distintas categorías. Fuente: Elaboración propia, 2023.

	Precisión	Recall	F1-Score	Support
Precisión			0.75	1217
Promedio del macro	0.82	0.65	0.70	1217
Promedio ponderado	0.77	0.75	0.74	127

Tabla 2: Resumen por características. Fuente: Elaboración propia, 2023.

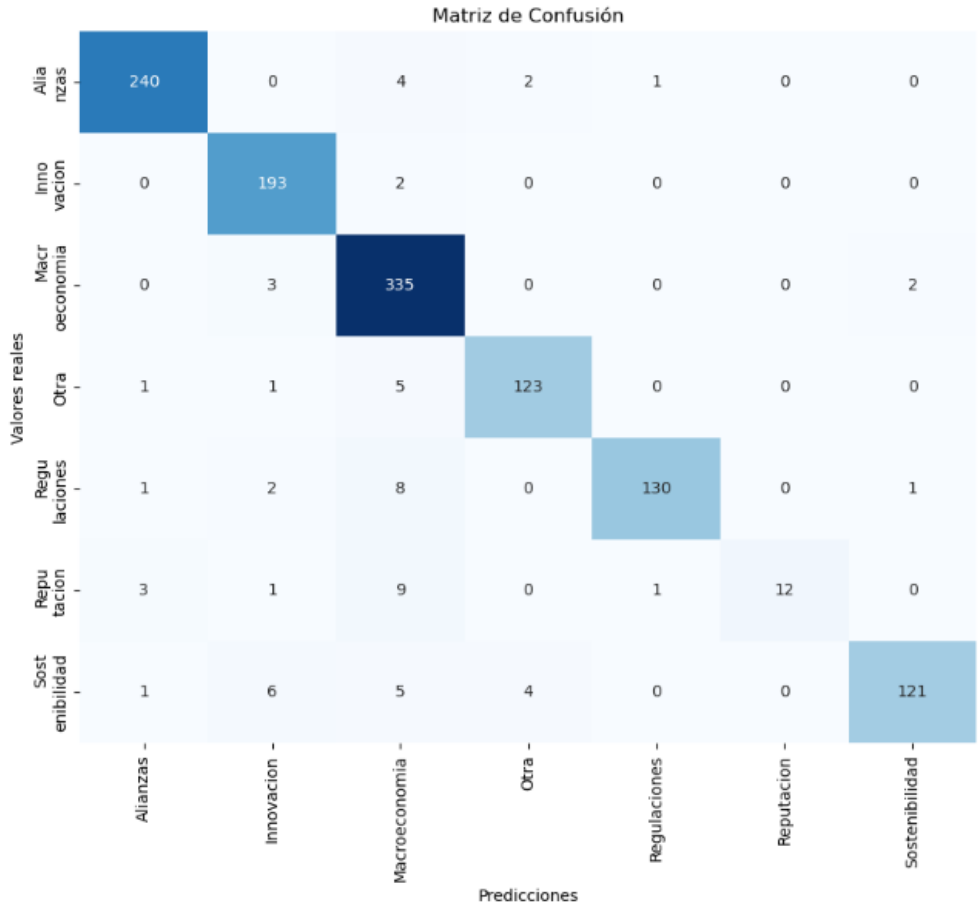

Ilustración 6: Matrix de confusión. Fuente: Elaboración propia, 2023.

De la matriz de confusión y las tablas observamos que el rendimiento del clasificador es aceptable, pero hay algunas categorías como Macroeconomía y Sostenibilidad que tienen una precisión y recall más bajo que otras categorías, lo que nos confirma nuestras sospechas iniciales de que existen ciertas diferencias de rendimiento por categoría.

Estos resultados pueden chocar con los obtenidos en la *Ilustración 3*, pero hay que tener en cuenta que la matriz de confusión muestra información en términos de predicciones correctas e incorrectas para cada categoría mientras que la gráfica de barras de la *Ilustración 3* muestra el porcentaje de precisión relativa por categoría, teniendo en cuenta el tamaño de cada categoría.

Una vez finalizado el proceso de aprendizaje y comprobado las limitaciones de nuestro clasificador, el siguiente paso es ponerlo en práctica para poder utilizarlo en las siguientes etapas del método. Para ello, recolectamos las noticias de distintos periódicos en un único archivo y utilizamos el código desarrollado del clasificador de Naive Bayes para clasificarlas. Obteniendo los resultados que podemos observar en el siguiente ejemplo:

```
Vox no garantiza investir a Azcón en Aragón  tras cerrar la Mesa: "Depende del pacto programático"
Macroeconomia
El PP pide a Feijóo un golpe de autoridad ante el "desgobierno"
Macroeconomia
Objetivo 2023: derogar el bibloquismo
Macroeconomia
El referéndum exigido por los comunes 'revienta' la estrategia de Díaz frente al nacionalismo
Macroeconomia
Sumar propone reducir la jornada laboral por ley a 37.5 horas en 2024 y seguir bajándola a 32 horas
Regulaciones
España crece un 0.6% \nhasta marzo y recupera el PIB anterior a la pandemia cuatro años después
Macroeconomia
Los amigos maduritos del presidente son una fuerza colosal
Otra
Sin garantías: qué se \ndebe cambiar tras la tragedia del Titan
Macroeconomia
¿La moda de atizar al rico? Por qué se ríen de los que iban dentro
Macroeconomia
Los cinco tripulantes murieron tras una "implosión catastrófica"
Macroeconomia
Colapsó sobre sí mismo por la presión: así fue el accidente del submarino
Macroeconomia
Los fallos de las turbinas de Gamesa derrumban un 31% en bolsa a Siemens Energy
Macroeconomia
Iberdrola amenazó con cortar la luz a 2.500 clientes de otra empresa
Alianzas
```

Ilustración 7: Ejemplo de clasificación de noticias. Fuente: Elaboración propia, 2023.

En esta sección del capítulo hemos realizado un recorrido práctico del método que vamos a implementar para realizar la clasificación por temática de los titulares. Se han realizado varias medidas para comprobar la eficacia del sistema, las cuales han sido aceptables. De cara a mejorar el método, la forma más efectiva y dónde hay mayor margen de mejora es la parte del *dataset*, con un *dataset* mejor, es decir, más completo, que abarque más noticias, con temáticas más variadas y en un periodo de recolección más grande, podríamos aumentar considerablemente la precisión del sistema.

ANÁLISIS DE LOS SENTIMIENTOS PERCIBIDOS EN LOS TITULARES

Para implementar el método de análisis binomial de los sentimientos expresados en un titular se utilizará la biblioteca de Python Flair. Esta herramienta es una de las más completas en cuanto a técnicas de NLP y preprocesado del texto se refiere. Como se mencionó en el capítulo anterior, Flair utiliza una metodología de redes neuronales para su funcionamiento.

Se utilizará Flair en este trabajo en dos instancias distintas. La primera de ella es para identificar entidades, como pueden ser localizaciones, entidades gubernamentales, asociaciones o personas. Esto nos ayudará a no tenerlas en cuenta a la hora de lematizar las palabras. Por otro lado, se utilizará para analizar de forma binomial los sentimientos expresados en los titulares. Una de las mayores limitaciones a las que nos enfrentamos es la incompatibilidad de las bibliotecas de sentimientos de Flair con el castellano, solo está disponible para el inglés y el alemán. Para solucionar este problema, se traducirán los titulares al inglés y se procederá a analizarlos.

Para realizar este proceso haremos una recolección de titulares de varios periódicos de un determinado día y mediante un código[83] nos devuelve la polaridad del sentimiento y un porcentaje de confianza de cada titular.

Para la siguiente sección del capítulo, la regresión polinomial, se ha creado un nuevo dataset, con noticias que abarcan un intervalo de 14 días en los que la bolsa está abierta, procedentes del periódico Expansión. Se ha elegido un periódico distinto a los cuatro con los que trabajaremos para la predicción, para tener un estilo de escritura único y distinto del de los otros cuatro. Procederemos en esta sección a estudiar la polaridad de los sentimientos expresados en el *dataset*.

En este nuevo banco de datos tenemos un total de 554 noticas, de las cuales, 225 son positivas y 329 son negativas. Como el dataset recogerá solo 40 noticias por día, los

[83] Proceso completamente automatizado que está disponible en el Anexo-A

sentimientos positivos y los negativos ponderarán el doble que en la predicción ya que para la predicción seleccionaremos 80 noticias.

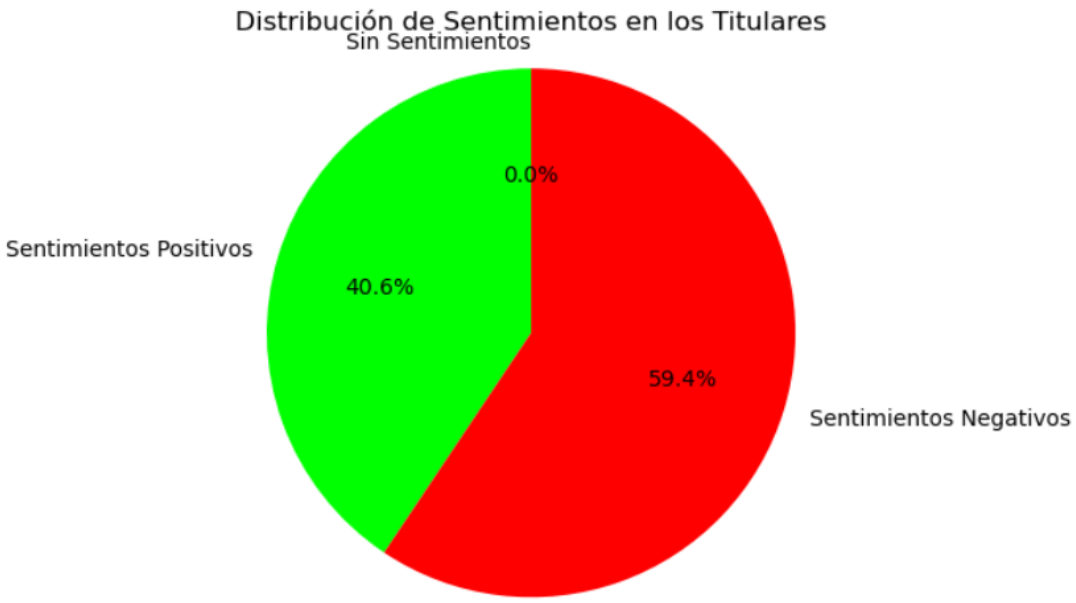

Ilustración 8: Distribución de sentimientos en el dataset. Fuente: Elaboración propia, 2023

Tras aplicar la ponderación y limitación de noticias por día que vimos en el anterior capítulo, los resultados son los siguientes:

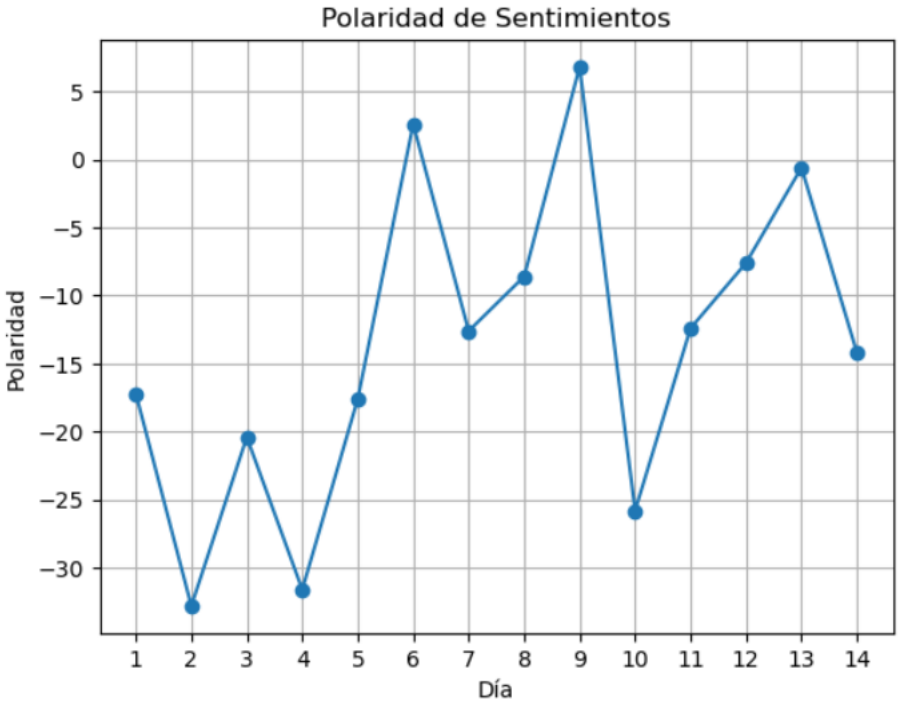

Ilustración 9: Polaridad de sentimientos tras aplicar la ponderación y límite de noticias. Fuente: Elaboración propia, 2023.

Como se puede observar, el periodo estudiado, que corresponde a casi tres semanas, tiene una polaridad prácticamente negativa, 12 de 14 días la polaridad global de las noticias es negativa.

En esta sección hemos visto el proceso práctico que seguiremos para implementar el análisis de la polaridad de los sentimientos de un titular, también hemos discutido el *dataset* que hemos creado para la siguiente sección y los sentimientos que hemos encontrado en él.

REGRESIÓN POLINOMIAL

En esta sección del capítulo discutiremos sobre el enfoque práctico del método para predecir oportunidades bursátiles analizando los sentimientos que encontramos en los titulares de las noticias. El objetivo de esta sección es argumentar por qué es la regresión polinomial la mejor opción para nuestro proyecto, con los datos que tenemos.

Para predecir las fluctuaciones del mercado estudiaremos los cambios porcentuales que sufre el IBEX 35 a lo largo de un día y lo compararemos con los sentimientos encontrados en la prensa para ese mismo día. Es importante destacar que en un principio no somos capaces de conocer el tipo de relación que existe entre los sentimientos y dichos cambios porcentuales, por eso, en una primera instancia enfocamos el problema como una regresión polinomial, que permite adaptar el grado del polinomio a nuestras necesidades.

A lo largo de esta sección trabajaremos con los siguientes datos:

$$S_{T_i} = [17.200, -32.800, -20.4, -31.5999, -17.6, 2.600, -12.6, -8.6001, 6.7999,$$

$$-25.8, -12.4, -7.6, -0.60, -14.2]$$

$$V_i = [1.3, 1.63, -0.3, 0.23, 0.53, -0.23, -0.31, 0.37, -0.11, 1.06, -0.02, 0.68, -0.66, 0.08]$$[84]

Al aplicar una regresión polinomial a los datos de partida obtenemos la siguiente respuesta:

Obtenemos un polinomio de orden nueve con las siguientes β, R^2 y R^{*2}:

$$\beta = [-4.23, 0.398, 27.27, -1.119, -56.508, 0.286, 40.212, 1.253, -6.338, 0]$$

$$R^2 = 0.9239 \ y \ R^{*2} = 0.6702.$$

[84] V_i es el cambio porcentual del IBEX 35 cada día de los que se han extraído noticias, los datos se han obtenido de la página web de Expansión.

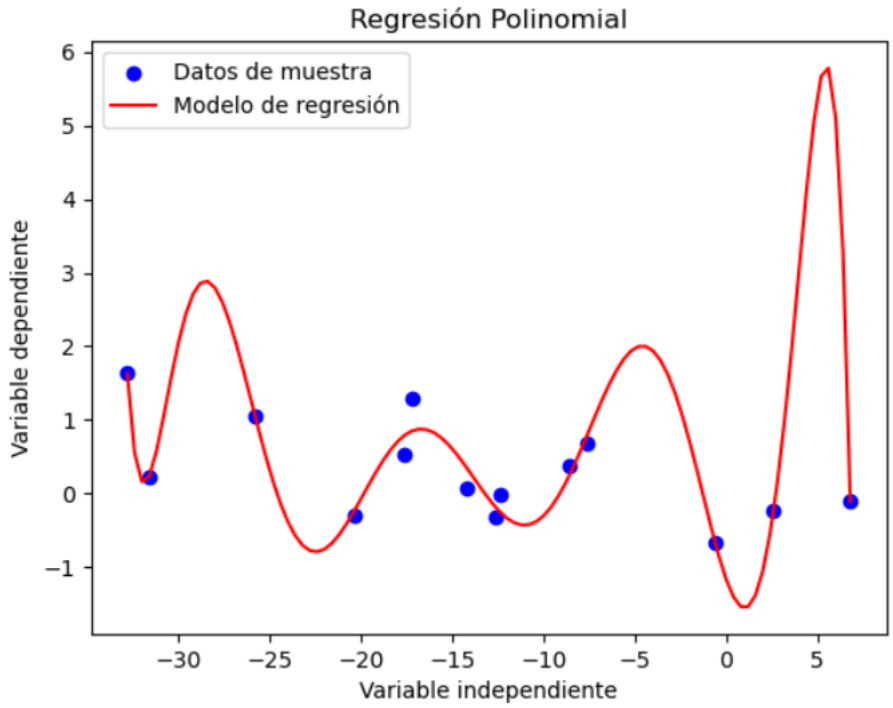

Ilustración 10: Regresión Polinomial. Fuente: Elaboración propia, 2023

La $R^2 = 0.9239$ *y la* $R^{*2} = 0.6702$ son dos valores más que aceptables. El valor de R^{*2} alto nos hace indicar que el sistema no esta sobreajustado, pero, se puede deber a la pequeña cantidad de datos que tenemos. Por otro lado, un valor de $R^2 = 0.9239$ indica que el 92,39% de la varianza de la variable dependiente está explicada por los datos de la variable independiente.

Por otro lado, la regresión lineal, más sencilla que la regresión polinomial, fue el primer modelo propuesto, sin embargo, se ajusta mal a los datos obtenidos, $R^2 = 0.3580$, lo que nos hace descartarla inmediatamente, a continuación, se muestra una gráfica representación de la regresión lineal:

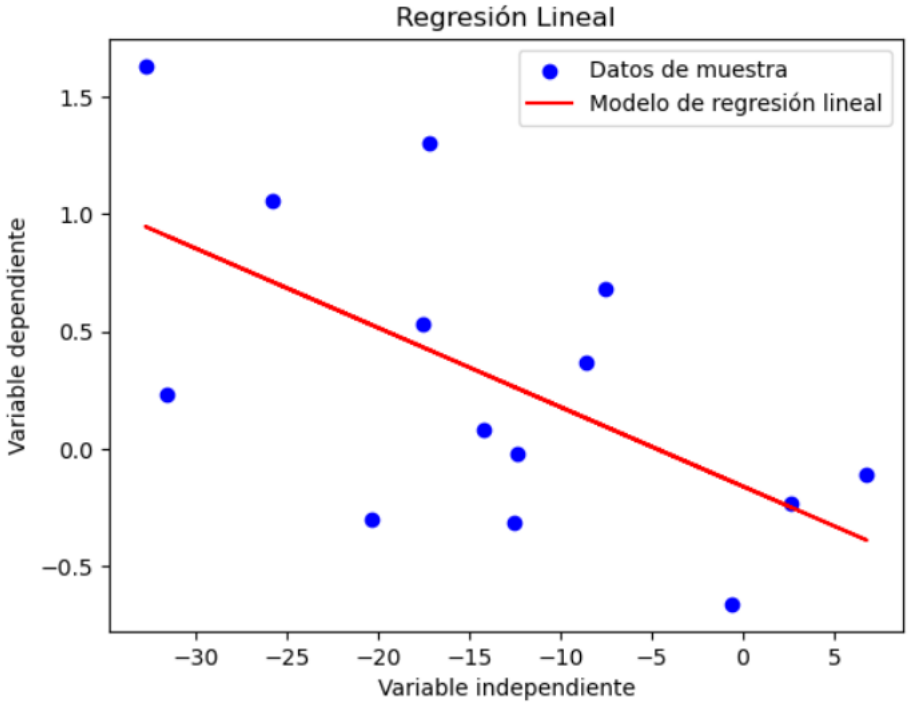

Ilustración 11: Regresión lineal. Fuente: Elaboración propia, 2023.

Finalmente, también se planteó un modelo de regresión por árbol de decisiones, que se ajustaba bastante bien a los datos:

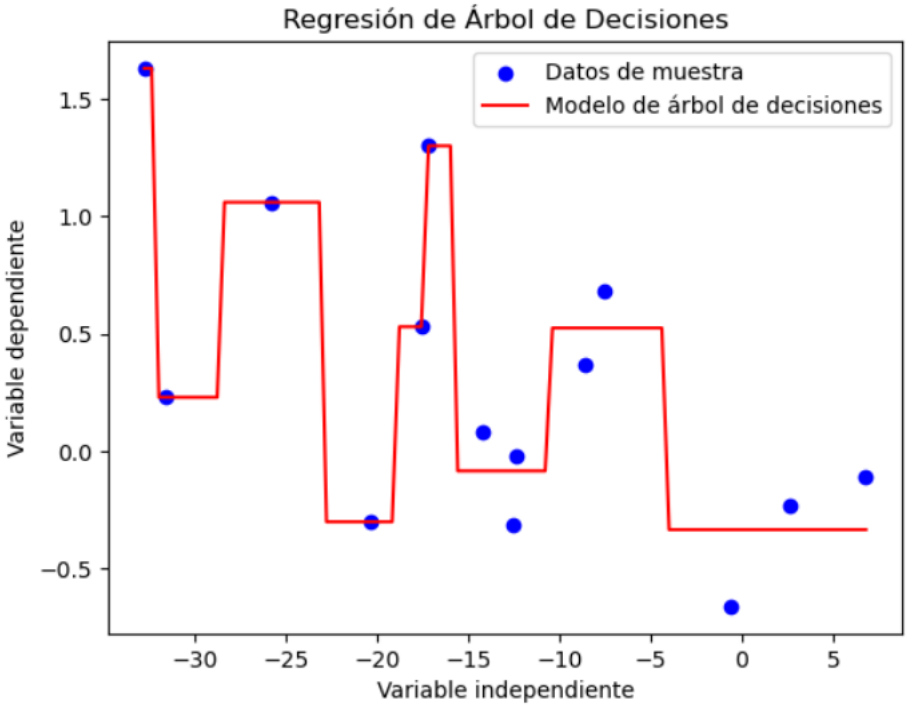

Ilustración 12: Grafica de regresión por árbol de decisiones. Fuente: Elaboración propia, 2023.

Obtuvimos unos resultados bastante buenos, con una $R^2 = 0.9487$ y con un número de ramas "bajo", nueve. Se descartó este enfoque, que podría ser válido perfectamente, debido a la baja cantidad de datos que se disponen en el intervalo de menos treinta y cinco a menos quince, lo que supone que el modelo esté sobreajustado en esa zona.

En esta sección se han evaluado los distintos métodos de regresión que podrían usarse, finalmente decidiendo la regresión polinomial cómo la mejor opción para nuestros datos. En cuanto a la posibilidad de mejora de esta parte del modelo, volvemos a insistir en la importancia de una buena base de datos, por limitaciones de tiempo, solo hemos sido capaces de extraer noticias de 14 días hábiles de bolsa, lo que hace que nuestro modelo no alcance su máximo potencial, lo ideal sería tener una base de datos amplia, que cubra un periodo mucho más largo lo que produciría una información más equilibrada y con menos sesgos de actualidad.

CASO PRÁCTICO: IBEX 35

Se ha elegido realizar un estudio del IBEX 35 ya que es el principal índice bursátil de referencia en la bolsa española formado por las 35 empresas con mayor liquidez y capitalización bursátil del mercado español. Se ha razonado, que, al estudiar los titulares de las noticias de los principales medios de comunicación en España, el IBEX será una forma adecuada de comprobar la eficacia del método propuesto. Se espera que una gran parte de los inversores de empresas del IBEX lean la prensa española y tomen decisiones de inversión en torno a la información que esta ofrece.

Nuestro *dataset* abarca desde el 1 de junio de 2023 hasta el 20 de junio de 2023, 14 días en los que la bolsa estuvo abierta. Durante esos días el IBEX 35 tuvo el siguiente comportamiento:

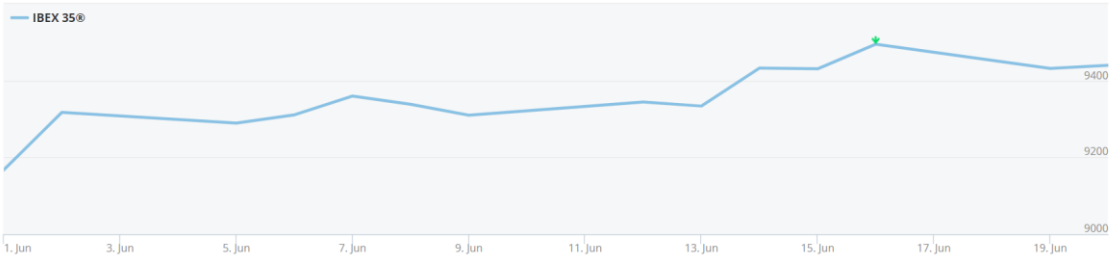

Ilustración 13: Desempeño del IBEX 35 desde el 1 de junio hasta el 20 de junio. Fuente: BME, 2023

Mientras que la gráfica de valor de los sentimientos extraídos del dataset durante ese mismo periodo de tiempo es:

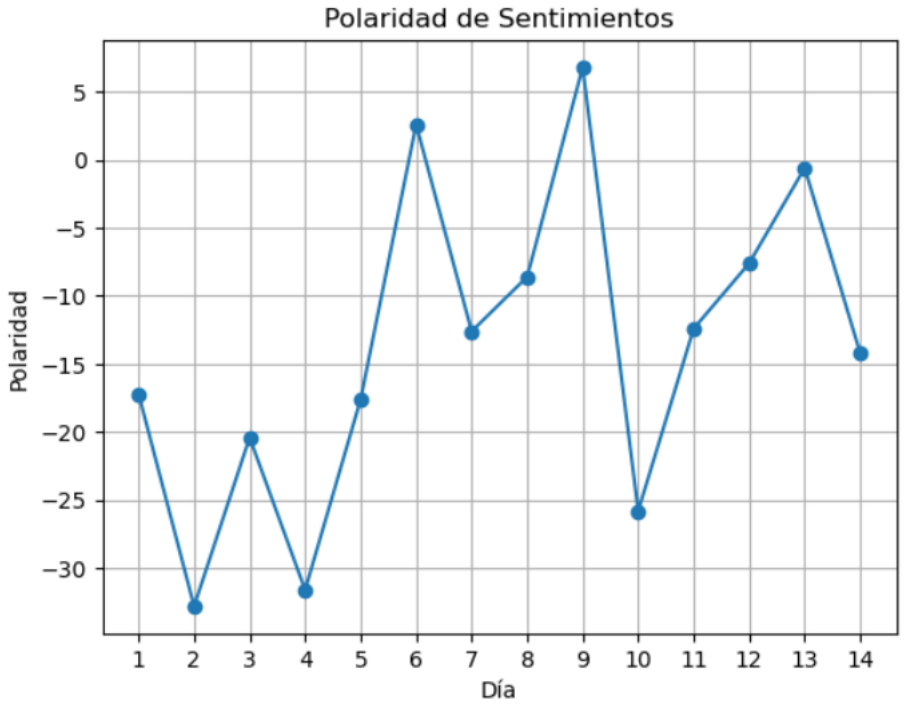

Ilustración 14:Sentimientos encontrados en la prensa desde el 1 de junio hasta el 20 de junio. Fuente: Elaboración propia, 2023.

A simple vista no se aprecia ningún tipo de dependencia entre las dos gráficas, lo que tiene sentido ya que, anteriormente, se ha hallado que su relación es un polinomio de orden nueve. Resulta curioso observar que los días más negativos como el día 2 de la *Ilustración 12,* que coincide con el día 2 de junio, supone un crecimiento de un 1.63% o como el día 4, que coincide con el día 6 de junio, también se aprecia un crecimiento, esta vez del 0.23%. Mientras que los días donde se aprecian sentimientos positivos, como el día 9, que corresponde con el día 13 de junio, bajó un 0.11% o como el día 6 que corresponde con el 8 de junio donde también hubo un bajón del 0.23%.

Parece contraproducente de primeras, pero cómo se observó en el capítulo dos, los seres humanos solemos ser más cautos ante sentimientos negativos como miedo o incertidumbre, lo que podría significar que los inversores tomarán mejores decisiones y por eso el IBEX sube, mientras que, cuando los seres humanos tenemos sentimientos positivos como euforia o alegría, nos dejamos llevar por ellos y es posible que los inversores esos días hayan tomado peores decisiones y por eso el mercado baje. Otra forma de apoyar esta hipótesis es

fijándonos en la *Ilustración 6* y la *Ilustración 12*, en la primera se nos muestra que cerca del 60% de las noticias recolectadas durante ese periodo de tiempo son negativas, y en la segunda, se nos muestra que, en el cómputo global, solo hay dos días en ese periodo de tiempo que tengan un sentimiento global positivo, sin embargo, el IBEX sube 300 puntos durante este periodo de tiempo, de 9100 en el mínimo del día 1 a 9400 en el mínimo del día 20.

El objetivo de este proyecto era el de crear una metodología de proyección emocional para la predicción de oportunidades bursátiles. Por ello, la parte más importante del proyecto es la de poder predecir las fluctuaciones del mercado. Se han recolectado las noticias de tres días, el 23, 27 y 28 de junio de 2023 de los cuatro medios digitales que se han mencionado a lo largo del proyecto. Los resultados de los sentimientos globales de esos días son los siguientes $S_{T_i} = [-16.099, -25, -22.099]$, mientras que los cambios porcentuales de valor del IBEX 35 esos mismos días han sido

$V_{iReal} = [-1.06, 1.28, 0.94]$.

Los valores que se han predicho para esos sentimientos son los siguientes:

$$V_i = [\,0.839209756, 0.316010820, -0.768858110]$$

Como se puede observar los cambios de valor predichos no son exactamente los mismos a los valores reales. A continuación, se muestra una gráfica para que se visualicen mejor las diferencias.

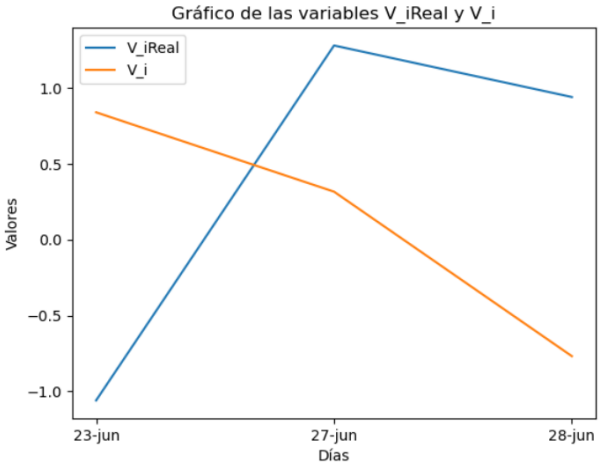

Ilustración 15: Gráfica que muestra los resultados de las predicciones. Fuente: Elaboración propia, 2023.

CONCLUSIONES

Hemos creado una metodología que extrae noticias de los medios de comunicación mediante la herramienta web-scrapper, clasifica esas noticias según su temática, analiza los sentimientos que se expresan en los titulares de dichas noticias, y mediante un modelo de regresión polinomial es capaz de, predecir los movimientos del mercado. Se ha construido un modelo de regresión polinomial bastante bueno, que explica alrededor del 94% de la varianza de los datos y que no estaba sobreajustado, $R^{*2} \approx 0.67$.

A pesar de que las predicciones no son exactas, seguimos creyendo fuertemente que las emociones son un pilar fundamental en las fluctuaciones del mercado, pero hay muchas más circunstancias que hay que tener en cuenta para poder predecir correctamente las fluctuaciones del mercado.

Otro de los motivos por el que el predictor no es del todo preciso es por la base de datos tan limitada en el tiempo, que permite probar el modelo, pero resulta insuficiente para consolidar conclusiones. Lo ideal, al menos, sería tener datos de por lo menos tres o cuatro meses, unos 60 u 80 días de bolsa.

Se proponen dos posibles soluciones para mejorar el método, la primera es invertir dinero en dos buenas bases de datos, una para la clasificación por temática y otra para la regresión polinomial, y la segunda es hacer el estudio para noticias y bolsa en países angloparlantes, las bases de datos son más ricas y completas en inglés, lo que mejoraría considerablemente la precisión del modelo, gracias al fácil acceso a datos.

5. MEMORIA ECONÓMICA

El objetivo de este capítulo es el de estudiar la viabilidad económica del proyecto. Más allá del trabajo académico, hay que analizar el binomio coste-beneficio de un desarrollo real aplicado a la interrelación entre emisiones y decisiones de inversión, y la capacidad del modelo de producir un valor predictivo sobre los mercados.

Se contratará el acceso a un agregador de noticias, como News API, para poder capturar un volumen estimado de 250k correspondientes a un periodo de cinco años, lo que supondrá un coste de 450€. Esto nos permitiría acceder al histórico del agregador y crear un dataset muy completo, con el que podríamos mejorar la precisión del clasificador de Naive-Bayes y el dataset de la regresión polinomial.

Con el fin de mejorar las traducciones que realizamos en el análisis de sentimientos, será necesario contratar la versión *Ultimate* de *DeepL*, 50€ al mes, que nos permitirá mejorar considerablemente el nivel del análisis de sentimientos, evitando que se pierdan matices en la traducción que pueden ser muy valiosos para la generación de emociones.

Respecto al capital humano, se precisará la incorporación temporal de un ingeniero recién graduado que desarrolle el proyecto, con un coste anual estimado de 21.000€[85]. Este ingeniero estará encargado de la creación y desarrollo de los dos nuevos *datasets* y ajustar el método de regresión polinomial para que se ajuste lo máximo posible. Debería completarse esta tarea en un mes, lo que supondría 1750€ para la empresa por sus servicios.

Será necesario contratar a un administrativo, un estudiante de periodismo, que trabajará cuatro horas semanales, 200€ al mes. Su función será la de extraer las noticias de la red y hacer funcionar el código a partir de las noticias obtenidas. Esta tarea es muy sencilla y solo debería llevarle veinte minutos, lo que le permitiría tener 25 minutos para leer las

[85] Salario medio del ingeniero recién egresado

traducciones comprobando que son fieles y leer los titulares de las noticias para ver si se ha asignado bien el sentimiento, es decir, también tendrá la función de encontrar anomalías en el proceso. A su vez, los titulares del día que ha extraído, se incluirán en el dataset, para que siga creciendo.

Para llevar a cabo este proyecto habrá que crear una Sociedad Limitada, que requiere un capital social mínimo de 3.000€ y unos gastos de gestión cercanos a los 300€.

Todo esto asciende el gasto total antes de iniciar el proyecto de:

$$Gasto_{Total\ inicial} = 450€ + 1750€ + 300 = 2500€$$

Mientras que los gastos totales cada mes será de:

$$Gasto_{Total\ mensual} = 50€ + 200€ = 250€$$

Para realizar los cálculos de la viabilidad económica haremos una serie de hipótesis sobre el rendimiento del método, supondremos que nuestro método es capaz de devolvernos un 10% anualmente, un 1% menos que el retorno histórico medio anual del S&P500. Cada mañana se leerán los titulares de ese día y se calculará la predicción de movimiento de mercado, en función de eso se escogerá una posición de corto o largo para ese día, vendiendo la posición a final del día o al alcanzar la fluctuación predicha. Se situarán límites de pérdidas en la mitad de la posición contraria a la predicha, es decir, si el código predice que subirá un 2% el IBEX 35 ese día, se situará el límite de pérdida en una bajada del 1%, con el objetivo de minimizar las pérdidas a la vez que damos cierto margen para que el mercado se mueva.

La inversión inicial será de 20.000€, por lo que, si todos los beneficios que obtenemos, se reinvierten, obtendremos la siguiente ecuación para obtener los ingresos:

$$M = 20.000 \times (1 + 0.10)^n \qquad [1]$$

Donde $\boldsymbol{n}$ significa el número de años que pasan y $\boldsymbol{M}$ la cantidad de dinero que obtenemos gracias al método.

Por otro lado, tenemos que tener en cuenta los gastos de funcionamiento de la empresa y los gastos iniciales, por lo que obtenemos la siguiente ecuación:

$$G = 2.500 + 250 \times 12 \times n \qquad [2]$$

Si a los resultados anuales que obtenemos de la *ecuación 1* le restamos el capital inicial de 20.000€, y juntamos dichos resultados con los resultados de los gastos de la *ecuación 2*, obtendremos el momento en el que el método empieza a dar beneficios, entre el año diez y once. La gráfica siguiente muestra el rendimiento de nuestra inversión y los gastos que tendrá el método a lo largo de los años.

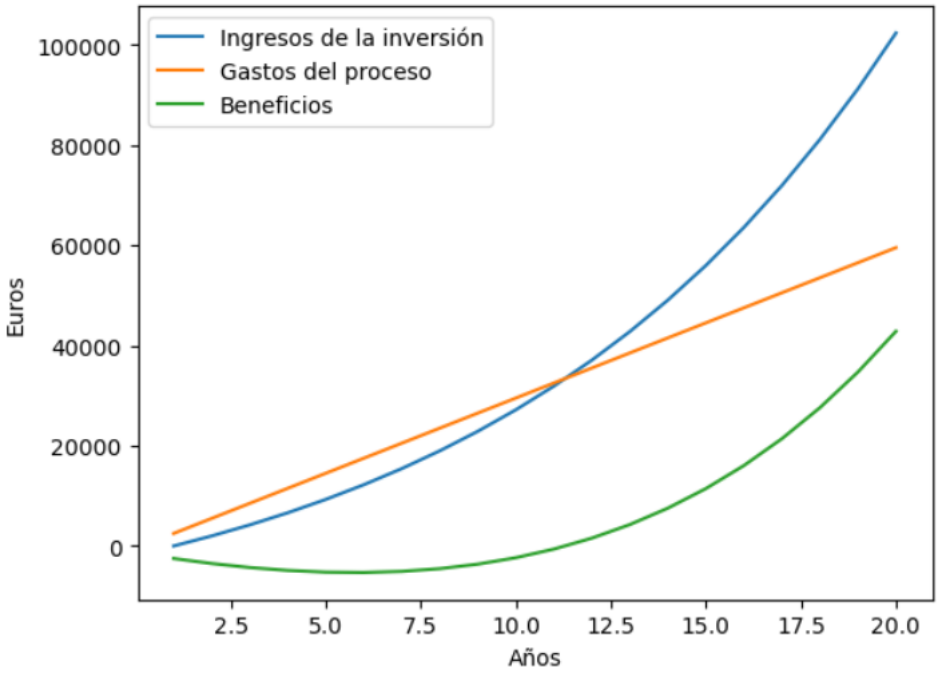

Ilustración 16: Rendimiento de la inversión. Fuente: Elaboración propia, 2023.

Si fuésemos más cautos y predijéramos un retorno menor, de un 7%, con la misma inversión inicial, el método no resultaría rentable, en los veinte años de vida que tendrá el proyecto, no conseguirá recuperar la inversión inicial y cubrir los gastos de funcionamiento, por lo que no tendría sentido implementar el método. Necesitaríamos un poco más de 23 años para empezar a obtener beneficios. A continuación, se muestra el desarrollo del rendimiento de la inversión:

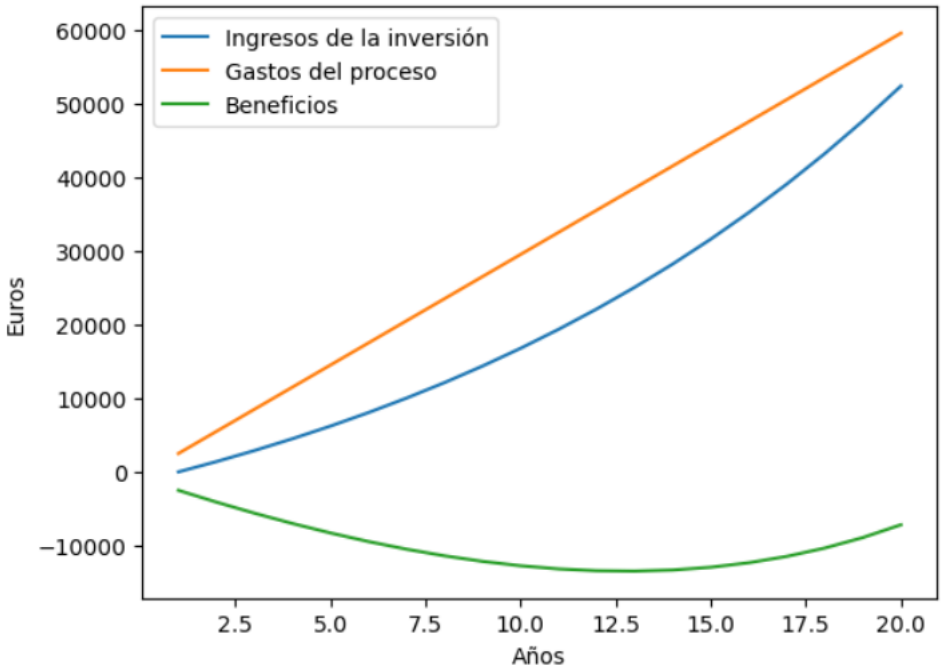

Ilustración 17: Rendimiento de la inversión en el supuesto 2. Fuente: Elaboración propia, 2023.

Una de las grandes limitaciones para la implementación del método es la elevada cantidad de capital inicial que se requiere para obtener una rentabilidad temprana, si no fuésemos capaces de conseguir 20.000 euros de capital, y nos quedásemos en solo 10.000, necesitaríamos un poco más de 22 años para empezar a ser rentables, asumiendo un 10% de retorno.

Si, por el contrario, el capital inicial para invertir no fuese un problema, pero se quisiera conseguir ser rentable en los primeros cinco años, suponiendo un 7% de retorno anual, necesitaríamos un capital inicial de 50.000€ para ser rentables en el periodo de tiempo establecido.

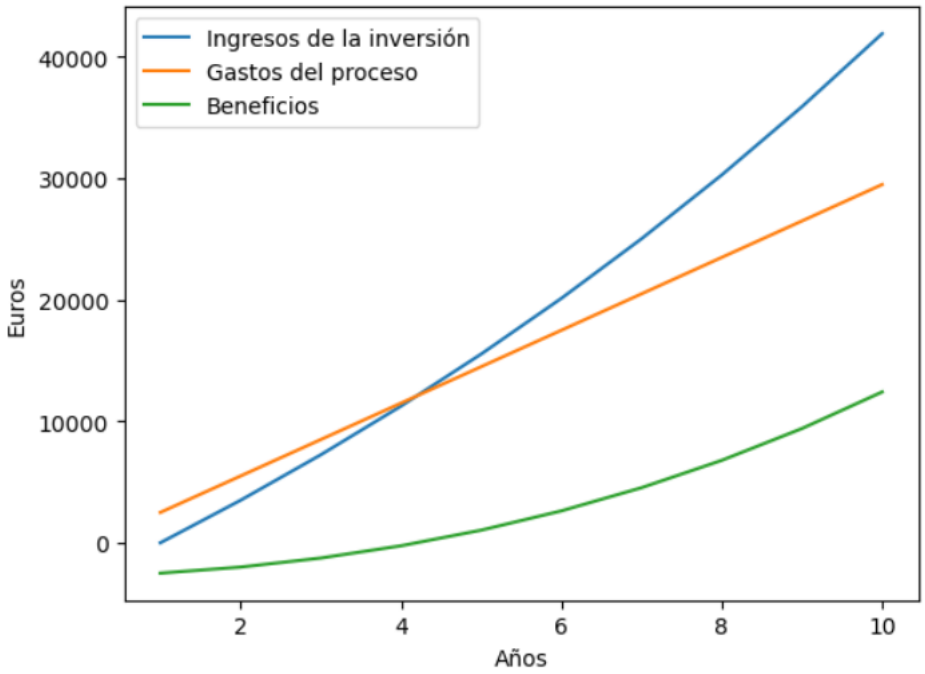

Ilustración 18: Rendimiento de la inversión para el supuesto 4. Fuente: Elaboración propia, 2023.

La mayor limitación a la que se enfrenta el método es la recaudación del capital inicial de inversión, sería necesario obtener un capital mínimo de 20.000 euros, para que, en el mejor de los escenarios, con un retorno del 10%, consiguiéramos ser rentables en un poco más de 10 años, consiguiendo, al cabo de veinte años, un poco más de 40.000 euros de beneficios. La empresa sería más rentable si consiguiésemos recaudar 50.000 euros, lo que nos permitiría empezar a tener beneficios en menos de cinco años, bajar el rendimiento esperado a 7% y aun así conseguir más de 71.000 euros de beneficios al cabo de los veinte años de vida del proyecto.

Esto significa que para iniciar el proyecto necesitaríamos 52.500 euros, sumando al capital para invertir, los gastos iniciales, también necesitaríamos ser capaces de pagar 3000 euros al año para el funcionamiento de la empresa, lo que ascendería el gasto total en 20 años a 112.500€. Todo este dinero tendría que ser desembolsado por la empresa, la cual no habrá obtenido ningún tipo de beneficio, ya que, para aprovechar el interés compuesto, tendremos que reinvertir todos los beneficios obtenidos. Sin embargo, a los veinte años, habremos obtenido un total de 193.500€, de los cuales 143.500€ habrán sido generados gracias al

método, los otros 50.000€ corresponden al capital inicial. Obteniendo así unos beneficios netos, excluida la inversión inicial, de 71.000€.

Para mitigar los gastos anuales de funcionamientos y conseguir recaudar dinero antes de que pasen los 20 años de vida del proyecto, se podría vender el funcionamiento del código a empresas o particulares que quieran utilizar nuestro método, primero haríamos un periodo de prueba de cinco años, para comprobar que se cumplen las predicciones que hemos hecho. Una vez pasado este periodo de prueba, podemos empezar a comercializar el método cobrando entre un 1 y 5% de los beneficios que obtengan los clientes. Esto nos ayudaría a costear los gastos de mantenimiento e incluso obtener unos beneficios extra.

Crear esta empresa sería beneficioso a largo plazo, pero requiere de una fuerte inversión de capital inicial que no es fácil obtener. Habría que tener en cuenta que se pagará un impuesto de sociedades de al menos el 25% y la inflación correspondiente a los veinte años de vida del proyecto. Si los gastos de lanzamiento se cubren, el proyecto sería viable con un *break even* conseguido a partir del décimo año.

6. CONCLUSIONES Y TRABAJOS FUTUROS

En este proyecto hemos creado una metodología de proyección emocional para predecir posibles oportunidades bursátiles. Se ha realizado un estudio multidisciplinar, con nociones de psicología, economía e ingeniería.

Se ha analizado cómo afectan los sentimientos a la toma de decisiones de las personas, se ha confirmado la hipótesis inicial de que los seres humanos nos dejamos influir en gran medida por nuestros sentimientos y que, cuando estamos contentos o eufóricos, solemos correr más riesgos y tomar decisiones menos lógicas, mientras que, cuando estamos tristes o sentimos emociones negativas, como el miedo, solemos ser más cautos y tomar decisiones más razonadas. También se ha estudiado cómo se trasladan estos sentimientos a las inversiones, obteniendo resultados muy similares.

Por otro lado, se ha estudiado cómo afectan a las fluctuaciones del mercado distintos elementos como la información desestructurada, los titulares de las noticias, las fake-news y los sentimientos de las noticias.

En cuanto al método, se ha logrado automatizar todo el proceso[86] mediante códigos de Python, lo que facilita y agiliza la elaboración. Hemos creado un clasificador utilizando el método de Naive Bayes, que permite una precisión del 75,19%. También se ha desarrollado una metodología que, con la ayuda de la biblioteca Flair, utiliza un sistema de redes neuronales para clasificar los sentimientos de los titulares mediante un enfoque binomial. Por último, hemos creado un modelo de regresión polinomial de orden nueve, que ha conseguido explicar cerca de un 93% de la varianza de los datos y con una $R^{*2} = 0.67$, lo

[86] Salvo la extracción de noticias de las páginas webs de los medios digitales, que se hace con una herramienta externa, aunque también se podría hacer desde Python.

que indica que no está sobreajustado. Se ha realizado un estudio sobre la elección de la regresión polinomial frente a otras opciones de regresión, observando los resultados que hubiésemos obtenido, en cuanto a precisión y sobreajuste se refiere, con otros modelos. En la regresión lineal, conseguimos tan solo una $R^2 = 0.3580$, lo que nos obligó a descartarla directamente. En la regresión por árbol de toma de decisiones, se consiguió una $R^2 = 0.9487$, pero la descartamos ya que se detectó un sobreajuste en la zona de sentimiento más negativos. La elección de una regresión polinomial es la adecuada para nuestro método, no solo por sus altos niveles de precisión, sino también por la adaptabilidad del método. Debido a cómo hemos configurado el código, la regresión polinomial ajustará el grado del polinomio a aquel que consiga ajustarse más al *dataset*, por lo que, si se encuentra o se elabora un *dataset* más completo, que cubra un periodo de tiempo más largo, el método se adaptará a los cambios y obtendremos la respuesta que mejor se adapta a nuestros nuevos datos.

En la implementación práctica del proyecto, se ha realizado un estudio del IBEX 35 y sus fluctuaciones. Hemos creado un *dataset* de noticias desde cero, del que se ha analizado la polaridad de los sentimientos globales ponderados de la prensa durante 14 días hábiles y se ha comparado con los cambios porcentuales de valor en el IBEX. Respecto a la hipótesis inicial de que las emociones afectan considerablemente a las fluctuaciones del mercado, se ha descubierto que tienen una relación inversa, de tal manera que en los días en los que los sentimientos eran más negativos, el cambio de valor porcentual era positivo, mientras que los días en los que la prensa expresaba sentimientos positivos, los cambios de valor eran negativos.

Hemos diseñado un modelo de predicción a partir del *dataset* y los datos de las fluctuaciones del IBEX 35 durante el mes de junio de 2023. El resultado del modelo permite extraer predicciones con una cierta validez, pero es evidente que para un resultado más amplio se habría requerido un periodo de observación y captura de datos más amplio. En un futuro, se podría construir un *dataset* más completo, que abarque un periodo de tiempo de al menos seis meses. Con mayor captura de datos el modelo será más fiable a la hora de predecir las fluctuaciones del mercado.

Hemos sido capaces de encontrar una relación de cómo afectan las emociones a las fluctuaciones del mercado, con un método que explica cerca del 93% de la variabilidad porcentual del valor del IBEX 35, pero no es capaz de predecir correctamente valores futuros. Creemos que esto se debe a un *dataset* no suficientemente representativo. Como se ha mencionado anteriormente, el modelo que hemos creado adaptará su forma para parecerse lo máximo posible a la base de datos que usemos, por lo que, si se mejora el *dataset*, mejorará nuestro método.

Se ha estudiado la viabilidad económica de crear una empresa que utilice nuestra metodología para incorporar la predictibilidad a través de las emociones generadas por la actualidad informativa, como un elemento adicional para la toma de decisiones de inversión, los resultados han sido positivos, esperando, en el mejor de los casos, recuperar la inversión inicial en menos de cinco años. El proyecto tiene una vida útil de veinte años y se espera conseguir algo más de 71.000€ de beneficios, suponiendo que obtenemos un 7% de rendimiento anual.

Por otro lado, se podrían añadir más variables, convirtiendo el modelo en una regresión lineal *stepwise forward*, que consiste en ir añadiendo variables de manera incremental para construir un modelo lineal. Se podría añadir variables como el análisis de sentimientos en las redes sociales y noticias corporativas o implementar acontecimientos de importancia global, como la guerra de Ucrania y observar los cambios en el mercado, para intentar predecir las fluctuaciones cuando se repita un evento de magnitud similar.

7. BIBLIOGRAFÍA

[1] Khatua, A., & Khatua, A. (2015). How Stock Market Reacts to Budget Announcement? Through the Lens of Social Media in Indian Context. Management and Labour Studies, 40(3–4), 239–251. https://doi.org/10.1177/0258042X16634568.

[2] Bustos, A. Pomares-Quimbaya, Stock market movement forecast: A Systematic review, Expert Systems with Applications, Volume 156, 2020, 113464, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2020.113464. (https://www.sciencedirect.com/science/article/pii/S0957417420302888)

[3] Rajakumar, M. P., Jegatheesan, R., Chandy, R., & Sampath, T. (2019). Prediction of stock prices using unstructured and semi-structured qualitative data—a neural network approach. International Journal of Intelligent Engineering & System, 12(2), 156-169.

[4] Kennedy, A. and Inkpen, D. (2006), SENTIMENT CLASSIFICATION of MOVIE REVIEWS USING CONTEXTUAL VALENCE SHIFTERS. Computational Intelligence, 22: 110-125. https://doi.org/10.1111/j.1467-8640.2006.00277.x

[5] S. M. Al Masum, M. T. Islam and M. Ishizuka, "ASNA: An Intelligent Agent for Retrieving and Classifying News on the Basis of Emotion-Affinity," 2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06), Sydney, NSW, Australia, 2006, pp. 133-133, doi: 10.1109/CIMCA.2006.51.

[6] S. Mostafa Al Masum, H. Prendinger and M. Ishizuka, "Emotion Sensitive News Agent: An Approach Towards User Centric Emotion Sensing from the News," IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), Fremont, CA, USA, 2007, pp. 614-620, doi: 10.1109/WI.2007.124.

[7] Casale, S., Russo, A., Scebba, G., & Serrano, S. (2008, August). Speech emotion classification using machine learning algorithms. In 2008 IEEE international conference on semantic computing (pp. 158-165). IEEE.

[8] Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio, "A News Story Categorization System", Proceedings of ANLP-88 and the 2nd Conference on Applied Natural Language Processing, Austin, US, 1988, pp. 9-17.

[9] Miao, F., Zhang, P., Jin, L., & Wu, H. (2018, August). Chinese news text classification based on machine learning algorithm. In 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (Vol. 2, pp. 48-51). IEEE.

[10] Kaur, G., & Bajaj, K. (2016). News classification and its techniques: a review. IOSR Journal of Computer Engineering, 18(1), 22-26.

[11] Nikam, S. S., & Dalvi, R. (2020, October). Machine learning algorithm based model for classification of fake news on twitter. In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 1-4). IEEE.

[12] Darren Duxbury, Tommy Gärling, Amelie Gamble & Vian Klass (2020) How emotions influence behavior in financial markets: a conceptual analysis and emotion-based account of buy-sell preferences, The European Journal of Finance, 26:14, 1417-1438, DOI: 10.1080/1351847X.2020.1742758

[13] Sonam and M. Devaraj, "Analyzing News Sentiments and their Impact on Stock Market Trends using POS and TF-IDF based approach," 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 2020, pp. 1-6, doi: 10.1109/IICAIET49801.2020.9257816.

[14] Hai V. Pham, Eric W. Cooper, Thang Cao, Katsuari Kamei, Hybrid Kansei-SOM model using risk management and company assessment for stock trading, Information Sciences, Volume 256, 2014, Pages 8-24, ISSN 0020-0255. https://doi.org/10.1016/j.ins.2011.11.036. (https://www.sciencedirect.com/science/article/pii/S0020025511006219)

[15] Seo, M. G., & Barrett, L. F. (2007). Being emotional during decision making—good or bad? An empirical investigation. Academy of Management Journal, 50(4), 923-940.

[16] Liu, B., Govindan, R., & Uzzi, B. (2016). Do emotions expressed online correlate with actual changes in decision-making?: The case of stock day traders. PloS one, 11(1), e0144945.

[17] Lucey, B. M., & Dowling, M. (2005). The role of feelings in investor decision-making. Journal of economic surveys, 19(2), 211-237.

[18] John W. Goodell, Satish Kumar, Purnima Rao, Shubhangi Verma, Emotions and stock market anomalies: A systematic review, Journal of Behavioral and Experimental Finance, Volume 37, 2022, 100722, ISSN 2214-6350. https://doi.org/10.1016/j.jbef.2022.100722. (https://www.sciencedirect.com/science/article/pii/S2214635022000557)

[19] Daniel Cabrera-Paniagua, Claudio Cubillos, Rosa Vicari, Enrique Urra, Decision-making system for stock exchange market using artificial emotions, Expert Systems with Applications, Volume 42, Issue 20, 2015, pp 7070-7083, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2015.05.004. (https://www.sciencedirect.com/science/article/pii/S0957417415003231)

[20] Demir, E., & Rigoni, U. (2017). You Lose, I Feel Better: Rivalry Between Soccer Teams and the Impact of Schadenfreude on Stock Market. Journal of Sports Economics, 18(1), 58–76. https://doi.org/10.1177/1527002514551801

[21] Qiwei Yang, Shiqin Zhou, Ruolei Gu, Yan Wu, How do different kinds of incidental emotions influence risk decision making?, Biological Psychology, Volume 154, 2020, 107920, ISSN 0301-0511, https://doi.org/10.1016/j.biopsycho.2020.107920. (https://www.sciencedirect.com/science/article/pii/S0301051120300806)
Keywords: Incidental emotion; Outcome evaluation; Emotional experience; Event-related potential (ERP); Feedback-related negativity (FRN); P3

[22] Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. Annual review of psychology, 66, 799-823.

[23] Andrade, E. B., & Ariely, D. (2009). The enduring impact of transient emotions on decision making. Organizational behavior and human decision processes, 109(1), 1-8.

[24] Pfister, H. R., & Böhm, G. (2008). The multiplicity of emotions: A framework of emotional functions in decision making. Judgment and decision making, 3(1), 5-17.

[25] Fenton-O'Creevy, M., Soane, E., Nicholson, N., & Willman, P. (2011). Thinking, feeling and deciding: The influence of emotions on the decision making and performance of traders. Journal of Organizational Behavior, 32(8), 1044-1061.

[26] Justin Robinson, Adrian Glean, Winston Moore, How does news impact on the stock prices of green firms in emerging markets?, Research in International Business and Finance, Volume 45, 2018, Pages 446-453, ISSN 0275-5319. https://doi.org/10.1016/j.ribaf.2017.07.176. (https://www.sciencedirect.com/science/article/pii/S0275531916304846)

[27] S. Feuerriegel, A. Ratku and D. Neumann, "Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation," 2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA, 2016, pp. 1072-1081, doi: 10.1109/HICSS.2016.137.

[28] Feuerriegel, S., & Neumann, D. (2013). News or noise? How news drives commodity prices.

[29] Emenike, K. O., & Enock, O. N. (2020). How Does News Affect Stock Return Volatility in a Frontier Market? Management and Labour Studies, 45(4), 433–443. https://doi.org/10.1177/0258042X20939019

[30] Kogan, S., Moskowitz, T. J., & Niessner, M. (2020). Fake news in financial markets. SSRN.

[31] Miranda, C. H., & Guzman, J. (2017). A review of Sentiment Analysis in Spanish. Tecciencia, 12(22), 35-48.

[32] Zou, H., Tang, X., Xie, B., & Liu, B. (2015, December). Sentiment classification using machine learning techniques with syntax features. In 2015 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 175-179). IEEE.

[33] Platón. (380 A.C.). La República.

[34] Aristóteles. (350 A.C.). Ética a Nicómaco.

[35] Aristóteles. (350 A.C.). Retórica.

[36] Freud, S. (1923). El yo y el ello. Buenos Aires, Argentina: Amorrortu Editores.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Galushkin, A. I. (2007). Neural networks theory. Springer Science & Business Media.

[37] Severyn, A., & Moschitti, A. (2015, August). Twitter sentiment analysis with deep convolutional neural networks. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 959-962).

[38] Lagerwerf, L., & Govaert, C. G. (2021). Raising clickworthiness: Effects of foregrounding news values in online newspaper headlines. News values from an audience perspective, 95-119.

[39] Ostertagová, E. (2012). Modelling using polynomial regression. Procedia Engineering, 48, 500-506.

[40] Francisco Borrás. (2023). Modelos Cuantitativos para la economía y la empresa [Material de clase no publicado]. Universidad Pontificia Comillas, Madrid.

ANEXO A: CÓDIGOS

CÓDIGO 1: PREPROCESADO DEL TEXTO

Este código extrae de un archivo .xlsx los titulares de las noticias extraídas previamente con la extensión de Google. En este caso, extrae las noticias de El Confidencial. Una vez las extrae, imprime el titular por pantalla. Después elimina las palabras que no aportan valor (Anexo-A.1) e imprime otra vez los titulares, pero solo con las palabras importantes. Por último, lematiza las palabras, dejando solo la raíz de éstas. Utilizamos la biblioteca Flair para reconocer entidades, como pueden ser personas, lugares o empresas y esas palabras no las lematiza.

```
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
from flair.models import SequenceTagger
from flair.data import Sentence

# Configuración para guardar las salidas en un archivo de texto UTF-8
output_file = open('PreprocesadoConfidencial2.txt', 'w', encoding='utf-
8')

# Cargar el archivo CSV especificando la codificación UTF-8
df =
pd.read_csv('C:/Users/willy/TFG/WebScrapping/ElConfidencial_Noticias1.csv
', encoding='utf-8')

# Imprimir las 21 primeras filas, solo la primera columna
output_file.write("Las noticias correspondientes al Confidencial
son:\n\n\n")
output_file.write(df.iloc[:21, 0].to_string(index=False) + '\n')
output_file.write('-' * 120)

# Eliminar palabras que no aportan valor
custom_stop_words = ['sí', 'no', 'tampoco', 'nunca', 'siempre', 'quizás',
'tal', 'vez', 'cierto', 'falso', 'verdadero', 'mentira', 'afirmar',
'negar']
stop_words = set(stopwords.words('spanish'))

df.iloc[:, 0] = df.iloc[:, 0].apply(lambda x: ' '.join([word for word in
str(x).split() if word.lower() not in stop_words or word.lower() in
custom_stop_words]))
```

```
# Imprimir después de eliminar palabras que no aportan valor
output_file.write("\n\nDespués de quitar las palabras que no aportan
valor:\n\n\n")
output_file.write(df.iloc[:21, 0].to_string(index=False) + '\n')
output_file.write('-' * 120)

# Lematizar las palabras que no son entidades
tagger = SequenceTagger.load('ner')
stemmer = SnowballStemmer('spanish')

def lemmatize_word(word):
    sentence = Sentence(word)
    tagger.predict(sentence)
    if len(sentence.labels) > 0 and sentence.labels[0].value in ['MISC',
'PER', 'LOC']:
        return word
    else:
        return stemmer.stem(word)

df.iloc[:, 0] = df.iloc[:, 0].apply(lambda x: '
'.join([lemmatize_word(word) for word in str(x).split()]))

# Imprimir después de lematizar las palabras
output_file.write("\n\nDespués de lematizar: \n\n\n")
output_file.write(df.iloc[:21, 0].to_string(index=False) + '\n')
output_file.write('-' * 120)

# Cerrar el archivo de salida
output_file.close()
```

Código 1: Preprocesado de texto

La respuesta del código es la siguiente:

```
Las noticias correspondientes al Confidencial son:


 Vox no garantiza investir a Azcón en Aragón  tras cerrar la Mesa: "Depende del pacto programático"
                                  El PP pide a Feijóo un golpe de autoridad ante el "desgobierno"
                                                        Objetivo 2023: derogar el bibloquismo
      El referéndum exigido por los comunes 'revienta' la estrategia de Díaz frente al nacionalismo
Sumar propone reducir la jornada laboral por ley a 37.5 horas en 2024 y seguir bajándola a 32 horas
    España crece un 0.6% \nhasta marzo y recupera el PIB anterior a la pandemia cuatro años después
                                       Los amigos maduritos del presidente son una fuerza colosal
                                  Sin garantías: qué se \ndebe cambiar tras la tragedia del Titan
                                ¿La moda de atizar al rico? Por qué se ríen de los que iban dentro
                                  Los cinco tripulantes murieron tras una "implosión catastrófica"
                           Colapsó sobre sí mismo por la presión: así fue el accidente del submarino
                     Los fallos de las turbinas de Gamesa derrumban un 31% en bolsa a Siemens Energy
                                 Iberdrola amenazó con cortar la luz a 2.500 clientes de otra empresa
          ¿Por qué España tiene la inflación más baja de la UE? La clave es el origen de la energía
              McKinsey calcula que la inteligencia artificial automatizará la mitad de los trabajos
                  Díaz Ayuso apuesta por un Gobierno de tecnócratas y asume todo el peso ideológico
                                Rabat bloquea los accesos a Melilla. pero abre la mano en Canarias
                       El espejo extremeño en el que se mira Sánchez para no dimitir aunque pierda
                  La JEC defiende la libertad del presidente para pedir el voto antes de la campaña
                          Alerta del Incibe: actualiza tu iPhone y Mac. hay una brecha de seguridad
                         Lo que pasó en Seattle para que la UNED anulara 81 pruebas de selectividad
----------------------------------------------------------------------------------------------------
```

Ilustración 19: Primera respuesta al código, simplemente se imprimen los titulares obtenidos.

```
Después de quitar las palabras que no aportan valor:


Vox no garantiza investir Azcón Aragón tras cerrar Mesa: "Depende pacto programático"
                                          PP pide Feijóo golpe autoridad "desgobierno"
                                                  Objetivo 2023: derogar bibloquismo
            referéndum exigido comunes 'revienta' estrategia Díaz frente nacionalismo
  Sumar propone reducir jornada laboral ley 37.5 horas 2024 seguir bajándola 32 horas
          España crece 0.6% marzo recupera PIB anterior pandemia cuatro años después
                                           amigos maduritos presidente fuerza colosal
                                        garantías: debe cambiar tras tragedia Titan
                                             ¿La moda atizar rico? ríen iban dentro
                                cinco tripulantes murieron tras "implosión catastrófica"
                                       Colapsó sí mismo presión: así accidente submarino
                                fallos turbinas Gamesa derrumban 31% bolsa Siemens Energy
                                      Iberdrola amenazó cortar luz 2.500 clientes empresa
                                      ¿Por España inflación baja UE? clave origen energía
                          McKinsey calcula inteligencia artificial automatizará mitad trabajos
                                 Díaz Ayuso apuesta Gobierno tecnócratas asume peso ideológico
                                        Rabat bloquea accesos Melilla. abre mano Canarias
                              espejo extremeño mira Sánchez no dimitir aunque pierda
                                 JEC defiende libertad presidente pedir voto campaña
                                Alerta Incibe: actualiza iPhone Mac. brecha seguridad
                                    pasó Seattle UNED anulara 81 pruebas selectividad
----------------------------------------------------------------------------------------------------
```

Ilustración 20: Segunda respuesta del código, se eliminan las palabras que no aportan valor respecto a la respuesta anterior.

```
Después de lematizar:


vox no garantiz invest azcon aragon tras cerr Mesa: "depend pact programatico"
                                        pp pid feijo golp autor "desgobierno"
                                                 objet 2023: derog bibloqu
               referendum exig comun 'revienta' estrategi Díaz frent nacional
            Sumar propon reduc jorn laboral ley 37.5 hor 2024 segu baj 32 hor
         españ crec 0.6% marz recuper pib anterior pandemi cuatr años despues
                                       amig madurit president fuerz colosal
                                      garantias: deb cambi tras tragedi tit
                                         ¿la mod atiz rico? rien iban dentr
                                  cinc tripul mur tras "implosion catastrofica"
                                  colaps si mism presion: asi accident submarin
                             fall turbin games derrumb 31% bols siemens energy
                                  iberdrol amenaz cort luz 2.500 client empres
                                  ¿por españ inflacion baja ue? clav orig energ
                    mckinsey calcul inteligent artificial automatiz mit trabaj
                          Díaz Ayuso apuest gobiern tecnocrat asum pes ideolog
                                    Rabat bloque acces Melilla. abre man canari
                               espejo extremeñ mir Sánchez no dimit aunqu pierd
                                    jec defiend libert president ped vot campañ
                                  alert incibe: actualiz iphon mac. brech segur
                                           pas seattl uned anul 81 prueb select
--------------------------------------------------------------------------------
```

Ilustración 21: Respuesta tercera, nos devuelve la respuesta anterior, pero lematizando las palabras.

CÓDIGO 2: AJUSTE DEL DATASET

Este código sirve para extraer los títulos del *dataset* a partir de los enlaces, primero accede al archivo .csv dónde está guardado el *dataset*, una vez dentro, extrae la categoría de la noticia y se mete en el enlace de la noticia, cogiendo el titular de esta. Posteriormente guarda en otro archivo .csv el titular y el tipo de la noticia. El objetivo de hacer esto era conseguir un *dataset* que nos sirviese para comparar las palabras de nuestros titulares, con la de los titulares del *dataset* y poder predecir su temática más probable utilizando Naive Bayes.

```
import csv
import pandas as pd
import requests
from bs4 import BeautifulSoup

# Ruta del archivo CSV de entrada
input_file = 'df_total.csv'

# Ruta del archivo CSV de salida
output_file = 'dataset.csv'

# Función para extraer el título de la noticia utilizando
BeautifulSoup
def extract_title_from_html(html):
    soup = BeautifulSoup(html, 'html.parser')
    title_tag = soup.find('title')
    if title_tag:
        return title_tag.get_text()
    return None

# Leer el archivo CSV de entrada
df = pd.read_csv(input_file)

# Listas para almacenar los títulos y los tipos de noticias
titles = []
categories = []

# Iterar sobre cada fila del DataFrame
for index, row in df.iterrows():
    # Obtener el enlace de la primera columna
    link = row[0]

    # Realizar la solicitud HTTP al enlace
    response = requests.get(link)

    # Extraer el título y el tipo de noticia si la solicitud es
exitosa
    if response.status_code == 200:
        title = extract_title_from_html(response.text)
```

```
        category = row[2]

        titles.append(title)
        categories.append(category)

# Crear un nuevo DataFrame con los resultados
df_output = pd.DataFrame({'Title': titles, 'Category': categories})

# Guardar el DataFrame en un archivo CSV
df_output.to_csv(output_file, index=False, encoding='utf-8')
```

Código 2: Extracción de titulares y preparación del nuevo dataset

CÓDIGO 3: CLASIFICACIÓN DE UN TEXTO SEGÚN SU TEMÁTICA

Este código clasificará los titulares según su temática. A continuación, se muestra el código que hemos utilizado y los resultados que hemos obtenido al utilizar las noticias extraídas de el periódico El Confidencial.

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB

# Cargar el archivo CSV creado anteriormente
df = pd.read_csv('dataset.csv', encoding='utf-8')

# Crear el vectorizador de palabras
vectorizer = CountVectorizer()

# Ajustar y transformar los datos de entrenamiento
X_train_vectors = vectorizer.fit_transform(df['Title'])

# Crear el clasificador Naive Bayes Multinomial
clf = MultinomialNB()

# Entrenar el clasificador
clf.fit(X_train_vectors, df['Category'])

# Función para predecir la categoría de un título dado
def predecir_categoria(titulo):
    titulo_vector = vectorizer.transform([titulo])
    categoria_predicha = clf.predict(titulo_vector)
    return categoria_predicha[0]

# Leer el archivo de texto con los titulares
with open('ConfidencialNoticiasEnteras.txt', 'r', encoding='utf-8') as
file:
    titulares = file.readlines()

# Crear el archivo de texto para guardar los resultados
with open('TextoClasificado.txt', 'w', encoding='utf-8') as file:
    for titular in titulares:
        titular = titular.strip()
        categoria_predicha = predecir_categoria(titular)
        file.write("{}\n{}".format(titular, categoria_predicha))
        file.write("\n")

print("Proceso completado. Los resultados se han guardado en el
archivo 'TextoClasificado.txt'.")
```

Código 3: Clasificación de los titulares según su temática

Los resultados obtenidos son los siguientes:

```
Vox no garantiza investir a Azcón en Aragón  tras cerrar la Mesa: "Depende del pacto programático"
Macroeconomia
El PP pide a Feijóo un golpe de autoridad ante el "desgobierno"
Macroeconomia
Objetivo 2023: derogar el bibloquismo
Macroeconomia
El referéndum exigido por los comunes 'revienta' la estrategia de Díaz frente al nacionalismo
Macroeconomia
Sumar propone reducir la jornada laboral por ley a 37.5 horas en 2024 y seguir bajándola a 32 horas
Regulaciones
España crece un 0.6% \nhasta marzo y recupera el PIB anterior a la pandemia cuatro años después
Macroeconomia
Los amigos maduritos del presidente son una fuerza colosal
Otra
Sin garantías: qué se \ndebe cambiar tras la tragedia del Titan
Macroeconomia
¿La moda de atizar al rico? Por qué se ríen de los que iban dentro
Macroeconomia
Los cinco tripulantes murieron tras una "implosión catastrófica"
Macroeconomia
Colapsó sobre sí mismo por la presión: así fue el accidente del submarino
Macroeconomia
Los fallos de las turbinas de Gamesa derrumban un 31% en bolsa a Siemens Energy
Macroeconomia
Iberdrola amenazó con cortar la luz a 2.500 clientes de otra empresa
Alianzas
¿Por qué España tiene la inflación más baja de la UE? La clave es el origen de la energía
Macroeconomia
```

Ilustración 22: Parte de los resultados tras ejecutar el código 3

CÓDIGO 4: CLASIFICACIÓN BINOMIAL DE EMOCIONES EN UN TEXTO

El siguiente texto se apoya de la biblioteca Flair para crear un mecanismo de predicción de emociones. También utiliza la biblioteca googletrans, para traducir los titulares de las noticias, por lo que se espera una pequeña pérdida de precisión.

```
from flair.models import TextClassifier
from flair.data import Sentence
from googletrans import Translator

# Ruta del archivo de titulares
archivo_titulares = "ConfidencialNoticiasEnteras.txt"

# Ruta del archivo de resultados
archivo_resultados = "Resultados_Emociones.txt"

# Crear el clasificador de sentimientos
classifier = TextClassifier.load("en-sentiment")

# Crear una instancia del traductor
translator = Translator()

# Leer los titulares de noticias
with open(archivo_titulares, 'r', encoding='utf-8') as file:
    titulares = file.readlines()

# Crear el archivo de resultados
with open(archivo_resultados, 'w', encoding='utf-8') as file:
    for titular in titulares:
        titular = titular.strip()

        # Traducir el titular al inglés
        traduccion = translator.translate(titular, dest='en').text

        # Realizar análisis de sentimientos con Flair
        sentence = Sentence(traduccion)
        classifier.predict(sentence)
        label = sentence.labels[0].value
        score = sentence.labels[0].score

        # Escribir el titular, la traducción y el resultado en el archivo
        file.write("{}\n{}\nSentimiento: {}\nPorcentaje:
{:.2f}%\n\n".format(titular, traduccion, label, score * 100))

print("Proceso completado. Los resultados se han guardado en el archivo
'Resultados_Emociones.txt'.")
```

Código 4: Clasificación binomial de las emociones de los titulares extraídos.

Tras ejecutar el código obtenemos los siguientes resultados:

```
Vox no garantiza investir a Azcón en Aragón  tras cerrar la Mesa: "Depende del pacto programático"
Vox does not guarantee to invest Azcón in Aragon after closing the table: "It depends on the programmatic pact"
Sentimiento: NEGATIVE
Porcentaje: 99.98%

El PP pide a Feijóo un golpe de autoridad ante el "desgobierno"
The PP asks Feijóo a blow of authority to the "disgust"
Sentimiento: NEGATIVE
Porcentaje: 99.61%

Objetivo 2023: derogar el bibloquismo
Objective 2023: repeal bibloquism
Sentimiento: NEGATIVE
Porcentaje: 85.53%

El referéndum exigido por los comunes 'revienta' la estrategia de Díaz frente al nacionalismo
The referendum required by the commons 'bursts' Diaz's strategy against nationalism
Sentimiento: POSITIVE
Porcentaje: 74.50%

Sumar propone reducir la jornada laboral por ley a 37.5 horas en 2024 y seguir bajándola a 32 horas
Add proposes to reduce the working day by law to 37.5 hours in 2024 and continue to lower it to 32 hours
Sentimiento: NEGATIVE
Porcentaje: 99.18%

España crece un 0.6% \nhasta marzo y recupera el PIB anterior a la pandemia cuatro años después
Spain grows 0.6%\ nhasta March and recovers the GDP before the pandemic four years later
Sentimiento: POSITIVE
Porcentaje: 75.22%
```

Ilustración 23: Ejemplo de resultados del análisis de sentimientos de los titulares

CÓDIGO 5: COMBINACIÓN DE LOS CÓDIGOS USADOS PARA LA CLASIFICACIÓN POR TEMÁTICA Y PARA EL ANÁLISIS DE SENTIMIENTOS

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from flair.models import TextClassifier
from flair.data import Sentence
from googletrans import Translator

# Cargar el archivo CSV creado anteriormente
df = pd.read_csv('dataset.csv', encoding='utf-8')

# Crear el vectorizador de palabras
vectorizer = CountVectorizer()

# Ajustar y transformar los datos de entrenamiento
X_train_vectors = vectorizer.fit_transform(df['Title'])

# Crear el clasificador Naive Bayes Multinomial
clf = MultinomialNB()

# Entrenar el clasificador
clf.fit(X_train_vectors, df['Category'])

# Función para predecir la categoría de un título dado
def predecir_categoria(titulo):
    titulo_vector = vectorizer.transform([titulo])
    categoria_predicha = clf.predict(titulo_vector)
    return categoria_predicha[0]

# Ruta del archivo de titulares
archivo_titulares = "ConfidencialNoticiasEnteras.txt"

# Ruta del archivo de resultados de temática y sentimientos
archivo_tematica_sentimientos = "Tematica_y_Sentimientos.txt"

# Crear el clasificador de sentimientos
classifier = TextClassifier.load("en-sentiment")

# Crear una instancia del traductor
translator = Translator()

# Leer los titulares de noticias
with open(archivo_titulares, 'r', encoding='utf-8') as file_titulares:
    titulares = file_titulares.readlines()

# Crear el archivo de resultados de temática y sentimientos
```

```
with open(archivo_tematica_sentimientos, 'w', encoding='utf-8') as
file_tematica_sentimientos:
    for titular in titulares:
        titular = titular.strip()

        # Clasificar la temática del titular
        categoria_predicha = predecir_categoria(titular)

        # Traducir el titular al inglés
        traduccion = translator.translate(titular, dest='en').text

        # Realizar análisis de sentimientos con Flair
        sentence = Sentence(traduccion)
        classifier.predict(sentence)
        label = sentence.labels[0].value
        score = sentence.labels[0].score

        # Escribir el titular, temática, sentimiento y porcentaje en
el archivo
        file_tematica_sentimientos.write("{}\n{}\nSentimiento:
{}\nPorcentaje: {:.2f}%\n\n".format(titular, categoria_predicha,
label, score * 100))

print("Proceso completado. Los resultados se han guardado en el
archivo 'Tematica_y_Sentimientos.txt'.")
```

Código 5: Mezcla de los códigos 4 y 3

Los resultados obtenidos son una combinación de los resultados de los dos anteriores códigos, como era de esperar. Después del titular se aprecia la categoría del titular, seguido del análisis de seguimiento binomial y el porcentaje de confianza con el que afirma la polaridad de los sentimientos.

```
Vox no garantiza investir a Azcón en Aragón  tras cerrar la Mesa: "Depende del pacto programático"
Macroeconomia
Sentimiento: NEGATIVE
Porcentaje: 99.98%

El PP pide a Feijóo un golpe de autoridad ante el "desgobierno"
Macroeconomia
Sentimiento: NEGATIVE
Porcentaje: 99.61%

Objetivo 2023: derogar el bibloquismo
Macroeconomia
Sentimiento: NEGATIVE
Porcentaje: 85.53%

El referéndum exigido por los comunes 'revienta' la estrategia de Díaz frente al nacionalismo
Macroeconomia
Sentimiento: POSITIVE
Porcentaje: 74.50%

Sumar propone reducir la jornada laboral por ley a 37.5 horas en 2024 y seguir bajándola a 32 horas
Regulaciones
Sentimiento: NEGATIVE
Porcentaje: 99.18%

España crece un 0.6% \nhasta marzo y recupera el PIB anterior a la pandemia cuatro años después
Macroeconomia
Sentimiento: POSITIVE
Porcentaje: 75.22%

Los amigos maduritos del presidente son una fuerza colosal
Otra
Sentimiento: POSITIVE
Porcentaje: 99.87%
```

Ilustración 24: Resultados del código 5

CÓDIGO 6: CÁLCULO DE PRECISIÓN DEL CLASIFICADOR

El siguiente código sirve para obtener la precisión que tiene el método de Naive-Bayes que hemos diseñado para clasificar los titulares según su temática. Para ello realiza una validación cruzada, que se repite varias veces, dividiendo el *dataset* en dos, *test* y *train* y comprueba la eficacia del método y a su vez mejora la precisión.

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import cross_val_score

# Cargar el archivo CSV creado anteriormente
df = pd.read_csv('dataset.csv', encoding='utf-8')

# Crear el vectorizador de palabras
vectorizer = CountVectorizer()

# Ajustar y transformar los datos de entrenamiento
X_train_vectors = vectorizer.fit_transform(df['Title'])

# Crear el clasificador Naive Bayes Multinomial
clf = MultinomialNB()

# Realizar validación cruzada
scores = cross_val_score(clf, X_train_vectors, df['Category'], cv=5)

# Imprimir la precisión promedio
print("Precisión promedio: {:.2f}%".format(scores.mean() * 100))
```

Código 6: Cálculo de precisión del código.

CÓDIGO 7: CÁLCULO DE LOS SENTIMIENTOS GLOBALES PARA VARÍAS FECHAS

El siguiente código accede a los archivos .csv dónde se almacenan los titulares de las noticias con los sentimientos asociados. Se pondera según su temática y posteriormente te imprime por pantalla el resultado de los sentimientos para cada día, $S_{i,j}$.

```
import csv

# Definir las ponderaciones de cada categoría
ponderaciones = {
    'Macroeconomia': 1,
    'Sostenibilidad': 0.9,
    'Innovación': 0.9,
    'Regulaciones': 0.8,
    'Alianzas': 0.8,
    'Reputación': 0.5,
    'Otros': 0.2
}

# Lista para almacenar los sentimientos globales
X = []

# Cargar y procesar los archivos manualmente
archivos = ['AAAEA1.csv', 'AAAEA2.csv','AAAEA5.csv', 'AAAEA6.csv',
'AAAEA7.csv', 'AAAEA8.csv', 'AAAEA9.csv', 'AAAEA12.csv', 'AAAEA13.csv',
'AAAEA14.csv', 'AAAEA15.csv', 'AAAEA16.csv', 'AAAEA19.csv',
'AAAEA20.csv']  # Agrega los nombres de tus archivos aquí

for archivo_csv in archivos:
    # Inicializar el valor del sentimiento global
    sentimiento_global = 0

    # Leer el archivo CSV y calcular el sentimiento global
    with open(archivo_csv, 'r', encoding='utf-8') as file:
        reader = csv.reader(file)
        next(reader)  # Ignorar la primera fila (cabecera)
        for row in reader:
            categoria = row[1]
            sentimiento = row[2]

            # Obtener la ponderación correspondiente a la categoría
            ponderacion = ponderaciones.get(categoria, 0)

            # Asignar valor al sentimiento según su tipo
            if sentimiento == 'POSITIVE':
                sentimiento_valor = 2
            elif sentimiento == 'NEGATIVE':
                sentimiento_valor = -2
```

```
            else:
                sentimiento_valor = 0

            # Calcular el valor ponderado del sentimiento
            sentimiento_ponderado = sentimiento_valor * ponderacion

            # Sumar el valor ponderado al sentimiento global
            sentimiento_global += sentimiento_ponderado

    # Agregar el sentimiento global a la lista X
    X.append(sentimiento_global)

# Imprimir el vector X con los sentimientos globales
print("Valores de los sentimientos globales:")
print(X)
```

Código 7: Cálculo de sentimientos globales de varios días.

EL resultado del código para el dataset con el que trabajamos es el siguiente:

Valores de los sentimientos globales:

[-17.200000000000003, -32.800000000000004, -20.4, -31.599999999999994, -17.6, 2.600000000000001, -12.6, -8.600000000000001, 6.799999999999999, -25.8, -12.4, -7.6, -0.6000000000000001, -14.2]

CÓDIGO 8: CAMBIOS PORCENTUALES DIARIOS DEL IBEX 35

El siguiente código accede a un archivo .cvs dónde están almacenados los datos correspondientes a los movimientos del IBEX 35 durante el periodo seleccionado. Te pide que le ingreses las fechas entre las que quieres sacar los datos y te imprime por pantalla los cambios de valoración porcentual del IBEX 35, V_i.

```
import pandas as pd

# Leer el archivo CSV con el separador y el separador decimal adecuados
df = pd.read_csv('IBEX27_06_23.csv', sep=';', decimal=',',
parse_dates=['Fecha'], dayfirst=True)

# Solicitar las fechas de inicio y fin al usuario
fecha_inicio = pd.to_datetime(input("Ingrese la fecha de inicio
(dd/mm/aaaa): "), format='%d/%m/%Y')
fecha_fin = pd.to_datetime(input("Ingrese la fecha de fin (dd/mm/aaaa):
"), format='%d/%m/%Y')

# Filtrar los datos entre las fechas especificadas
df_fechas = df[(df['Fecha'] >= fecha_inicio) & (df['Fecha'] <=
fecha_fin)]

# Extraer la columna "Cambio %" y revertir el orden
cambio_porcentaje = df_fechas['Cambio %'][::-1]

# Imprimir los cambios porcentuales de cada día entre las fechas
print("Cambios porcentuales entre", fecha_inicio.date(), "y",
fecha_fin.date(), ":")
print(", ".join(cambio_porcentaje.astype(str)))
```

Código 8: Cálculo de cambios de valor porcentuales en el IBEX 35.

El resultado por pantalla que obtendríamos sería el siguiente:

```
Ingrese la fecha de inicio (dd/mm/aaaa): 01/06/2023
Ingrese la fecha de fin (dd/mm/aaaa): 20/06/2023
Cambios porcentuales entre 2023-06-01 y 2023-06-20:
1.3, 1.63, -0.3, 0.23, 0.53, -0.23, -0.31, 0.37, -0.11, 1.06, -0.02,
0.68, -0.66, 0.08
```

CÓDIGO 9: CÁLCULO DE LA REGRESIÓN POLINOMIAL Y GRÁFICA ASOCIADA

El siguiente código calcula la regresión polinomial con los resultados obtenidos en los dos últimos códigos. Busca el orden de polinomio que mejor se ajuste a los datos que tenemos, mediante varias iteraciones. Te imprime por pantalla el orden del polinomio, R^2, R^{*2} y las betas obtenidas para dicho polinomio. Posteriormente crea la gráfica referida a la regresión polinomial.

```
#regresión polinomial
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

# Datos de ejemplo
X = np.array([-17.200, -32.800, -20.4, -31.5999, -17.6, 2.600, -12.6, -
8.6001, 6.7999, -25.8, -12.4, -7.6, -0.60, -14.2])
y = np.array([1.3, 1.63, -0.3, 0.23, 0.53, -0.23, -0.31, 0.37, -0.11,
1.06, -0.02, 0.68, -0.66, 0.08])

# Normalizar las variables
X_normalized = (X - np.mean(X)) / np.std(X)
y_normalized = (y - np.mean(y)) / np.std(y)

best_degree = None
best_model = None
best_r2 = -np.inf

# Iterar sobre diferentes grados de polinomio
for degree in range(1, 20):
    # Ajustar un modelo de regresión polinomial
    poly_features = PolynomialFeatures(degree=degree)
    X_poly = poly_features.fit_transform(X_normalized.reshape(-1, 1))
    model = LinearRegression()
    model.fit(X_poly, y_normalized)

    # Calcular el coeficiente de determinación (R^2)
    y_pred = model.predict(X_poly)
    r2 = r2_score(y_normalized, y_pred)

    # Actualizar el mejor modelo si R^2 es mayor que 0.8
    if r2 > 0.8:
        best_degree = degree
        best_model = model
        best_r2 = r2
```

```
        break

# Calcular R^2 ajustada
n = len(y_normalized)
p = best_degree + 1  # Número de características incluyendo el término
constante
r2_adj = 1 - (1 - best_r2) * (n - 1) / (n - p - 1)

# Imprimir R^2 y R^2 ajustada
print("R^2:", best_r2)
print("R^2 ajustada:", r2_adj)
print("El orden del polinomio es", best_degree)

# Obtener las betas
betas = best_model.coef_
betas_original = betas / np.std(y_normalized)

# Imprimir las betas
print("Betas:")
for i, beta in enumerate(betas_original):
    print("β{}: {:.3f}".format(i, beta))

# Gráfica del modelo
X_plot = np.linspace(min(X), max(X), 100)
X_plot_normalized = (X_plot - np.mean(X)) / np.std(X)
X_plot_poly = poly_features.transform(X_plot_normalized.reshape(-1, 1))
y_plot_normalized = best_model.predict(X_plot_poly)
y_plot = y_plot_normalized * np.std(y) + np.mean(y)

plt.scatter(X, y, color='blue', label='Datos de muestra')
plt.plot(X_plot, y_plot, color='red', label='Modelo de regresión')
plt.title('Regresión Polinomial')
plt.xlabel('Variable independiente')
plt.ylabel('Variable dependiente')
plt.legend()
plt.show()
```

Código 9: Cálculo de la regresión polinomial y su gráfica.

El código nos imprime por pantalla los siguientes resultados:

```
R^2: 0.9239093421985515
R^2 ajustada: 0.6702738161937231
El orden del polinomio es 9
Betas:
β0: 0.000 β1: -6.338 β2: 1.253 β3: 40.212 β4: 0.286 β5: -56.508
β6: -1.119 β7: 27.270 β8: 0.398 β9: -4.230
```

CÓDIGO 10: CÁLCULO DE LA REGRESIÓN LINEAL Y GRÁFICA ASOCIADA

El siguiente código crea una regresión lineal con los datos obtenidos en los códigos 8 y 9 y posteriormente imprime por pantalla el valor de R^2 y la gráfica asociada a la regresión lineal.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

# Datos de ejemplo
X = np.array([-17.200, -32.800, -20.4, -31.5999, -17.6, 2.600, -12.6, -
8.6001, 6.7999, -25.8, -12.4, -7.6, -0.60, -14.2])
y = np.array([1.3, 1.63, -0.3, 0.23, 0.53, -0.23, -0.31, 0.37, -0.11,
1.06, -0.02, 0.68, -0.66, 0.08])

# Reshape para que X sea una matriz de columna
X_reshaped = X.reshape(-1, 1)

# Crear un objeto de regresión lineal
model = LinearRegression()

# Ajustar el modelo a los datos
model.fit(X_reshaped, y)

# Predecir los valores de y
y_pred = model.predict(X_reshaped)

# Calcular el coeficiente de determinación (R^2)
r2 = r2_score(y, y_pred)

# Imprimir el coeficiente de determinación
print("R^2:", r2)

# Gráfica del modelo
plt.scatter(X, y, color='blue', label='Datos de muestra')
plt.plot(X, y_pred, color='red', label='Modelo de regresión lineal')
plt.title('Regresión Lineal')
plt.xlabel('Variable independiente')
plt.ylabel('Variable dependiente')
plt.legend()
plt.show()
```

Código 10: Cálculo de la regresión lineal y su gráfica.

Este código imprime por pantalla: R^2: 0.3580054415662153.

CÓDIGO 11: CÁLCULO DE LA REGRESIÓN DE ÁRBOL DE TOMA DE DECISIONES Y SU GRÁFICA ASOCIADA.

El siguiente código crea una regresión de árbol de toma de decisiones con los datos obtenidos en los códigos 8 y 9 y posteriormente imprime por pantalla el valor de R^2 y la gráfica asociada a la regresión de árbol de toma de decisiones.

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import r2_score

# Datos de ejemplo
X = np.array([-17.200, -32.800, -20.4, -31.5999, -17.6, 2.600, -12.6,
-8.6001, 6.7999, -25.8, -12.4, -7.6, -0.60, -14.2])
y = np.array([1.3, 1.63, -0.3, 0.23, 0.53, -0.23, -0.31, 0.37, -0.11,
1.06, -0.02, 0.68, -0.66, 0.08])

# Inicializar el número máximo de ramas y el modelo
max_branches = 2
model = DecisionTreeRegressor(max_leaf_nodes=max_branches)

# Entrenar el modelo inicial
model.fit(X.reshape(-1, 1), y)
y_pred = model.predict(X.reshape(-1, 1))
r2 = r2_score(y, y_pred)

# Aumentar gradualmente el número máximo de ramas hasta que R^2 supere
0.8
while r2 < 0.9:
    max_branches += 1
    model = DecisionTreeRegressor(max_leaf_nodes=max_branches)
    model.fit(X.reshape(-1, 1), y)
    y_pred = model.predict(X.reshape(-1, 1))
    r2 = r2_score(y, y_pred)

# Imprimir el valor de R^2 y el número de ramas
print("R^2:", r2)
print("Número de ramas:", max_branches)

# Generar valores de X para la gráfica
X_plot = np.linspace(min(X), max(X), 100).reshape(-1, 1)

# Realizar predicciones para los valores de X de la gráfica
y_plot = model.predict(X_plot)

# Graficar los datos y el modelo ajustado
plt.scatter(X, y, color='blue', label='Datos de muestra')
```

```
plt.plot(X_plot, y_plot, color='red', label='Modelo de árbol de
decisiones')
plt.title('Regresión de Árbol de Decisiones')
plt.xlabel('Variable independiente')
plt.ylabel('Variable dependiente')
plt.legend()
plt.show()

# Realizar una predicción para un número ingresado por el usuario
while True:
    try:
        x_pred = float(input("Ingresa un número para predecir: "))
        break
    except ValueError:
        print("Error: Ingresa un número válido.")

y_pred = model.predict([[x_pred]])
print("La predicción para", x_pred, "es:", y_pred)
```

Código 11: Cálculo de la regresión de árbol de toma de decisiones y su gráfica.

Del código anterior obtenemos la siguiente respuesta:

```
R^2: 0.9487881251724646
Número de ramas: 9
```

CÓDIGO 12: PREDICCIÓN DE VALORES PARA EL MODELO DE REGRESIÓN POLINOMIAL

El siguiente código utilizará el modelo de regresión lineal creado anteriormente para, a partir de los resultados obtenidos, predecir los movimientos del mercado al introducir el valor del sentimiento global S_T del día seleccionado.

```
#Predecir
import numpy as np
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression

# Datos de ejemplo
X = np.array([-17.200, -32.800, -20.4, -31.5999, -17.6, 2.600, -12.6, -
8.6001, 6.7999, -25.8, -12.4, -7.6, -0.60, -14.2])
y = np.array([1.3, 1.63, -0.3, 0.23, 0.53, -0.23, -0.31, 0.37, -0.11,
1.06, -0.02, 0.68, -0.66, 0.08])

# Normalizar las variables
X_normalized = (X - np.mean(X)) / np.std(X)
y_normalized = (y - np.mean(y)) / np.std(y)

# Grado del polinomio
degree = 9

# Ajustar un modelo de regresión polinomial
poly_features = PolynomialFeatures(degree=degree)
X_poly = poly_features.fit_transform(X_normalized.reshape(-1, 1))
model = LinearRegression()
model.fit(X_poly, y_normalized)

# Valor de X para predecir
X_pred = np.array([-16.099,-25,-22.099,] )  # Introduce aquí el valor de
X que deseas predecir

# Normalizar el valor de X
X_pred_normalized = (X_pred - np.mean(X)) / np.std(X)

# Transformar el valor de X a características polinomiales
X_pred_poly = poly_features.transform(X_pred_normalized.reshape(-1, 1))

# Realizar la predicción
y_pred_normalized = model.predict(X_pred_poly)

# Denormalizar la predicción
y_pred = y_pred_normalized * np.std(y) + np.mean(y)

# Imprimir el valor predicho
print("Valor predicho de y:", y_pred)
```

Código 12: Predicción de futuros valores del mercado.

EL código anterior, para los sentimientos introducidos, devuelve los siguientes valores:

```
Valor predicho de y: [ 0.83920976  0.31601082 -0.76885811]
```

ANEXO B: GLOSARIO

LISTA DE PALABRAS QUE SE ELIMINAN EN EL PREPROCESADO

A continuación, se mostrará una serie de palabras que están recogidas dentro de la biblioteca NLTK[87] y que no se considera que aporten valor:

'unos', 'estad', 'está', 'hubieras', 'e', 'estos', 'sentidas', 'vuestro', 'tuvieseis', 'serían', 'nos', 'sentido', 'estarías', 'estoy', 'mío', 'cual', 'tengo', 'pero', 'hubiésemos', 'le', 'tenga', 'estados', 'nuestra', 'tuviera', 'estamos', 'otra', 'sus', 'es', 'estéis', 'para', 'otros', 'habrían', 'ni', 'habrán', 'tuvo', 'tus', 'me', 'ella', 'sería', 'tuviesen', 'entre', 'serás', 'estuviéramos', 'habréis', 'tengas', 'esa', 'teníais', 'tuviese', 'estando', 'estuvieron', 'habida', 'hayamos', 'las', 'suyo', 'quienes', 'tengamos', 'estuvimos', 'tendrán', 'habían', 'tenían', 'teníamos', 'durante', 'habría', 'fueras', 'tienes', 'tanto', 'estuvieseis', 'tendremos', 'no', 'tuve', 'habíamos', 'porque', 'seréis', 'yo', 'tuvierais', 'tengáis', 'ya', 'hemos', 'tendríamos', 'nada', 'mis', 'tu', 'fueses', 'hubieseis', 'estará', 'su', 'habidas', 'ti', 'mías', 'suyas', 'estadas', 'del', 'ha', 'siente', 'teniendo', 'mucho', 'eres', 'ellos', 'también', 'tienen', 'serías', 'sea', 'mí', 'poco', 'estuvieses', 'fuisteis', 'seré', 'nuestros', 'hubierais', 'con', 'estaré', 'esas', 'estuvieran', 'tendrá', 'estarían', 'tuvieron', 'tuyas', 'habéis', 'hayáis', 'que', 'son', 'fuésemos', 'y', 'somos', 'una', 'estuviésemos', 'sois', 'has', 'fue', 'seáis', 'sean', 'habremos', 'estén', 'estuviesen', 'habiendo', 'seríamos', 'tuviste', 'estuviera', 'tendréis', 'otro', 'tenidas', 'contra', 'tuvieses', 'el', 'esto', 'estés', 'tenidos', 'míos', 'fueran', 'fuerais', 'quien', 'tenemos', 'hasta', 'seremos', 'estáis', 'muy', 'qué', 'hubiesen', 'uno', 'tendrían', 'hubo', 'habríamos', 'era', 'seas', 'algunos', 'por', 'tuvieras', 'suyos', 'habías', 'algo', 'vuestra', 'esté', 'habido', 'estábamos', 'hubimos', 'han', 'antes', 'mía', 'habidos', 'estaba', 'vosotras', 'hay', 'habríais', 'estuvieras', 'hubisteis', 'fuese', 'fuera', 'tuyos', 'estar', 'seríais', 'eras', 'hayan', 'todos', 'he', 'tenías', 'estaríais', 'estaban', 'están', 'tendrías', 'sentidos', 'fueron', 'haya', 'ante', 'de', 'hubiste', 'estabas', 'habrás', 'les', 'a', 'hubieron', 'hubiese', 'sentida', 'soy', 'tenía', 'tuvisteis', 'cuando', 'hubiéramos', 'mi',

[87] Por sus siglas en ingles Natural Language Toolkit.

'tendría', 'estuvisteis', 'serán', 'tendrás', 'estaríamos', 'estas', 'estaréis', 'hube', 'ese', 'nosotras', 'estarás', 'muchos', 'tendríais', 'estuviese', 'hayas', 'sentid', 'tuyo', 'este', 'estemos', 'habrías', 'te', 'vuestros', 'fui', 'tuvieran', 'fueseis', 'él', 'habrá', 'tened', 'nuestras', 'se', 'estuvo', 'sobre', 'un', 'en', 'estás', 'eran', 'estarán', 'fuimos', 'estuve', 'fuéramos', 'al', 'tendré', 'será', 'os', 'estuviste', 'los', 'más', 'eso', 'tuviésemos', 'todo', 'otras', 'tú', 'desde', 'tenida', 'estado', 'esos', 'estada', 'habíais', 'vuestras', 'estaremos', 'estabais', 'tenido', 'hubiera', 'seamos', 'o', 'tengan', 'hubieses', 'éramos', 'tuviéramos', 'tuya', 'suya', 'tuvimos', 'esta', 'como', 'hubieran', 'erais', 'habré', 'estuvierais', 'nosotros', 'nuestro', 'fuiste', 'había', 'vosotros', 'donde', 'sintiendo', 'tenéis', 'lo', 'tiene', 'ellas', 'fuesen', 'la', 'algunas', 'sin'.

GLOSARIO DE TÉRMINOS EN INGLÉS

Término en español	**Original, en inglés**	**Mini-definición**
Base de datos	Dataset	Lugar donde se almacenan datos
Ingenuo	Naive (Bayes)	Que el método hace una suposición simplificada
Redes neuronales artificiales	Artificial Neural Networks	Modelo computacional que emula las conexiones del cerebro humano
Árbol de decisiones	Decision Trees	Modelos de aprendizaje automático que emula las conexiones de las ramas de un árbol
Máquinas de soporte vectorial	Support Vector Machines	Algoritmo de aprendizaje automático cuyo objetivo es la separación de datos en sus clases óptimas
K-Vecinos más Cercanos	K-Nearest Neighbors	Método de clasificación
Partes Del Discurso	Part Of Speech	Categorización de las palabras según su categoría gramatical

Procesador del lenguaje natural	Natural language processing	Estudia las interacciones entre las computadoras y el lenguaje humano
Noticias falsas	Fake-news	Aquellas noticias que tienen como objetivo la desinformación.
Aprendizaje Automático	Machine Learning	Rama de la inteligencia artificial que se enfoca a crear algoritmos y modelos
Palabras vacías	Stop-words	Aquellas palabras que no aportan información a una oración.
Rascado de la web	Web-scrapping	Proceso a partir el cual se obtiene de manera automática la información de una web.
Cadena	String	Secuencia de caracteres
Ponderación booleana	Boolean Weighting	Técnica que asigna 1 o 0 en función de si un término está presente o no
Umbral de frecuencia de clase	Class Frequency Thresholding	Umbral a partir del cual se considera que un término pertenece a una categoría

Frecuencia de término inversa de frecuencia de clase	Term Frequency Inverse Class Frequency	Técnica para ponderar la importancia de los términos de un documento
Ganancia de información	Information Gain	Medida propia de los árboles de decisión que indica la importancia de un término
Conjunto de patrones	pattern-sets	Conjunto de reglas o patrones que sigue un código de aprendizaje automático
Recuento de términos	term-counting	Técnica que consiste en contar cuantas veces se repite un término en un texto
Inversores	Traders	Personas que invierten su dinero en un activo
Datos de entrada	Inputs	Elemento que se introduce como entrada en un código o sistema
Mapas autoorganizados	Self-Organazing Maps	Algoritmos que permiten mostrar datos de alta dimensionalidad

Entrenamiento	Train	División del dataset a partir de la cual se entrena el código
Test	Test	División del dataset a partir de la cual se testea el código
Técnica de incrustación	Embedding	Técnica a partir de la cual se asigna una representación vectorial a palabras
Sensibilidad	Recall	Porcentaje de casos positivos que el sistema pudo identificar correctamente respecto al total
Selección paso a paso hacia adelante	Stepwise forward	Técnica de la regresión que sirve para añadir entradas de una a una para comprobar su relevancia en el sistema

Tabla 1: Glosario de términos en inglés, su traducción al español y definición.