Comillas Pontifical University

Faculty of Economics and Business Administration, ICADE

The effect of financial news on stock prices: insights from NLP techniques

Author: Fernando Santana García

Director: Carlos Bellón Núñez-Mera

Madrid | June 2023

ABSTRACT AND KEYWORDS

Abstract

This dissertation conducts different analyses of financial news's effect on the price dynamics of traded companies. The project focuses on oil & gas companies related news from two different newspapers (the *Financial Times* and the *Wall Street Journal*).

These analyses include the classification of news based on their content and sentiment to analyze newspaper bias and perform event studies of the evolution of their abnormal return over and above the one determined by the Fama-French 3 factor model (Fama & French, 1992)

Different Natural Language Processing (NLP) techniques have been employed to carry out all these analyses, highlighting the use of fine-tuned Large Language Models (LLM), like BERT. These models usually employ a transformer architecture based on the self-attention concept. This project emphasizes the utility of fine-tuning pre-trained models.

Keywords

Sentiment analysis, Natural Language Processing, NLP, transformers, Large Language Models, LLMs, news classification, event studies, FinBERT, fine-tuning, 3-factor model.

RESUMEN Y PALABRAS CLAVES

Resumen

En este proyecto se realizan distintos análisis sobre el efecto de las noticias financieras en la dinámica de precios de las empresas cotizadas en bolsa. Se centra en noticias de los periódicos Financial Times y Wall Street Journal sobre el sector de gas y petróleo.

Estos análisis incluyen la clasificación de las noticias en función de su contenido y sentimiento para analizar el sesgo de los periódicos y posteriormente realizar estudios de eventos de la evolución de su rentabilidad anormal por encima de la determinada por el modelo de 3 factores de Fama-French (Fama & French, 1992)

Para llevar a cabo todos estos análisis se han empleado diferentes técnicas de Procesamiento del Lenguaje Natural (NLP), destacando el fine-tuning de Grandes Modelos de Lenguaje (LLMs), como BERT. Estos modelos suelen emplear una arquitectura de *transformers* basada en el concepto de autoatención. Este proyecto hace hincapié en la utilidad de afinar modelos preentrenados.

Palabras clave

Análisis de sentimiento, Procesamiento del Lenguaje Natural, NLP, *transformers*, Grandes Modelos del Lenguaje (LLMs), estudio de eventos, clasificación de noticias, *FinBERT*, *fine-tuning*, modelo de los 3 factores.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

All tables in this document are self-made based on data frames or results

List of Acronyms

AR	Abnormal Return
AI	Artificial Intelligence
ANN	Artificial Neural Network
BERT	Bidirectional Encoder Representations from Transformers
CAPM	Capital Asset Pricing Model
CBOW	Continuous Baf of Words
DL	Deep Learning
EIA	Energy Information Administration
ELMo	Embeddings from Large Models
EU	European Union
FT	Financial Times
FTSE	Financial Times Stock Exchange
GloVe	Global Vectors
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HML	High Minus Low
IDE	Integrated Development Environment
LLaMA	Large Language Model Meta AI
LLM	Long Language Model
LSTM	Long Short-Term Memory
ML	Machine Learning

NLP	Natural Language Process
NER	Named Entity Recognition
OLS	Ordinary Least Squares
OPEC	Organization of the Petroleum Exporting Countries
PaLM	Pretraining with Adversarial Learning Method
POS	Part of Speech
RNN	Recurrent Neural Network
SMB	Small Minus Big
SG	Skip-Gram
TENER	Transformer Encoder Named Entity Recognition
UK	United Kingdom
US	United States
WSJ	Wall Street Journal

Section 1. ABOUT THE PROJECT

This first chapter will try to establish the project motivation, explaining the reasons for developing it and its objectives (section 1.1), the methodology followed as well as the tools employed to achieve the desired results (section 1.2), and finally, a brief description of the structure and organization of this written document (section 1.3).

1.1 GOALS AND MOTIVATION OF THE PROJECT

Is the same economic news reported similarly neutrally by different newspapers? What if these newspapers are from other countries and can favor or harm the interests of national companies?

Therefore, studying social behavior through a powerful influence tool like newspapers and for a world-spanning industry like Oil & Gas may allow us to extract conclusions about national-bias in finance.

Two primary goals for the project:

- Check whether local newspapers treat news about domestic and foreign companies differently. To do this, sentiment and other analyses will be carried out about a company in this sector (Exxon) in its country of origin (US) and in another country with competing companies (UK).

- Check if the sentiment in news published about companies affects their stock prices. To do so, the first step is to identify if the information discussed is related only to share price commentary already reflected on the stock price or company/market analysis that can impact future market reactions. After classification and sentiment analysis, event studies will explore the importance of the news on market movements.

1.2 METHODOLOGY AND TOOLS TO BE EMPLOYED

The project aims to analyze companies based on published news about them. The primary programming language used has been Python, where different techniques and models that will be explained later were used (*e.g.*, transformers, BERT model). Many of these models are big, heavy models that require a high computational cost that cannot be supported by a regular laptop with an IDE (Integrated Development Environment) such as *PyCharm* or *Visual Studio Code*. Instead, Google Colab has been used to overcome this problem as it offers more powerful resources such as GPU (Graphics Processing Unit) completely free. To take advantage of Google's hardware, we have also used Colab PRO, accessing high-end GPUs (V100 and A100 from Nvidia), more RAM to train big models, and longer runtimes with background execution.

Google Drive was used to store the data as it seemingly integrates with Google Colab and can hold much larger datasets than GitHub, which is especially useful for us treating extensive news article data.

1.3 STRUCTURE OF THE DISSERTATION

This present dissertation, which gathers the work elaborated as the final project for the B.Sc. in Business Analytics, is organized into six different sections:

- The current section, Section 1. , offers a preliminary overview of the conducted research, encompassing the contextual background and the driving forces behind the investigation.
- The main objective of Section 2. is to familiarize the reader with the oil and gas industry, which will be analyzed throughout the project.
- Section 3. will overview the existing literature related to the research topic, including NLP techniques and sentiment analysis.
- Section 4. will describe the data used in the project. It will specify the news and market data sources as well as some preprocessing.
- In Section 5. we will classify financial news into two groups depending on its content: News with and without informational content. It will discuss the process and model used and the results obtained.
- After classifying news in the previous section, Section 6. aims to analyze them. Sentiment analysis will be conducted on the news from the same company in different newspapers and various companies in the same newspaper.
- Section 7. deeps into the event studies, i.e., how sentiments and events affect stock companies' prices.

- Section 8. deals with the conclusions from the project and will streamline the possible future work to improve and continue this project.
- Finally, Section 9. References the papers and data cited throughout the dissertation.

Section 2. OIL AND GAS INDUSTRY CONTEXT

The oil and gas industry is one of the most important sectors across many industries. They are critical energy market players among the most important global energy suppliers. Today, oil is the primary energy source worldwide, which can be proved in facts like the consumption of petroleum in the United States in 2021, which amounted to 19.78 million barrels per day, according to the EIA (US Energy Information Administration)[1].

Throughout this section, we will dive deeper into this industry, explaining how it works (section 2.1), briefly mentioning its historical background (section 2.2), and naming some of its largest companies (section 2.3).

2.1 UNDERSTANDING THE SECTOR

In order to have an overview of this vital sector of the economy, it is essential to have an idea of how this sector works in general terms. It could be said that this industry can be divided into three main pillars:

- Exploration and production. This is the very first step which is known as *Upstream*. It involves studying and examining the earth's subsurface to find

[1] Figures showing the importance of oil in the worldwide economy according to EIA.

oil and gas reserves from which to extract raw materials. Oil and gas producers/operators usually conduct this step by trying to locate those reserves and remove the desired material from them.

- Transportation. This step, also called *Midstream,* is a bridge between the Upstream and Downstream steps. It consists of two main activities: gas and oil transportation through different ways (*e.g.*, subsea pipelines, rails, trucks, or ships) and storage in tanks and terminals. Companies involved in this step are known as infrastructure companies and oversee the efficient and safe movement of oil and gas across regions worldwide.
- Refining and marketing of the final products. This last step of the process is known as *Downstream*. In this step, activities related to refining and processing the crude oil into various products such as diesel, gasoline, or petrochemicals and marketing and distributing the refined products to end consumers (for instance, gas stations) are carried out.

These three steps of the process are graphically summarized in Figure 2.1:

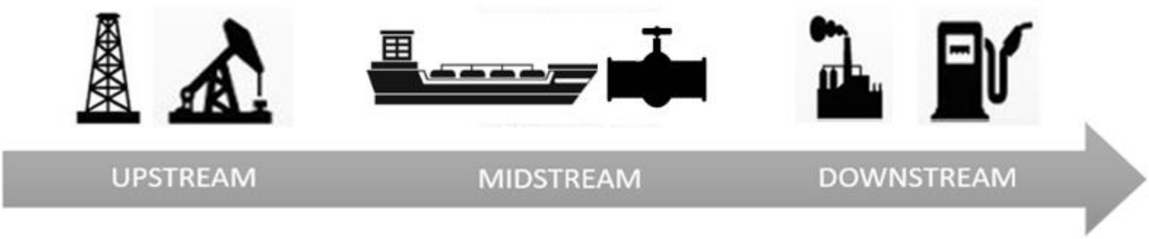


Figure 2.1: Oil and gas industry process. Source: self-made

2.2 MODERN HISTORICAL EVOLUTION

It can be stated that the modern history of the oil and gas industry started around 1847, thanks to the Scottish chemist James Young (Yergin, 1991). He noted a natural petroleum leakage from which he could distill two oil forms: a light and thin variant appropriate for use in lamps and a denser version well-suited for lubrication purposes.

From then on, research and investigations proliferated with progressive and advanced discoveries at the end of the 19th century and the early 20th century, motivated by the advent of the internal combustion engines and the proliferation of automobiles. These advances expanded oil exploration geographically as all countries around the globe wanted to have this 'new gold.'

In 1960, the OPEC (Organization of the Petroleum Exporting Countries) was created, initially formed by Iran, Iraq, Kuwait, Venezuela, and Saudi Arabia. These countries gained control over their resources and were able to affect the global oil market, highlighting their geopolitical significance and dependency during the 1970s oil crisis.

In the past years, there has been a growing awareness and concern over the impact of fossil on citizens and governments. This new mentality leads to a movement towards renewable energy sources or natural gas, creating new challenges and opportunities for innovation and transformation in the oil and gas industry. However, oil is still essential in our day-to-day lives as many activities require it.

Today, OPEC countries (15 members) have significant power, accounting for approximately 44% of global oil production and 81.5% of the world's oil reserves.[2]

2.3 LEADING COMPANIES IN THE SECTOR

Most of the leading companies in this industry participate in the three steps of the process aforementioned in section 2.1. That implies that the same company is in charge of finding and extracting oil and gas from the subsurface, transporting it, and refining it to sell different products to diverse clients. One of the clearest examples in our country, Spain, is the multinational *Repsol*.

The largest oil corporations have their headquarters primarily in countries including the United States, Saudi Arabia, China, the United Kingdom, and France, with most of these entities trading their shares on US stock exchanges. However, *Saudi Aramco* stands as a notable exception. As the world's preeminent integrated oil and gas company, it is listed on the Saudi Stock Exchange, known as Tadawul.

Other important companies are:

- *ExxonMobil Corporation* engages in the exploration, production, trading, transportation, and sale of oil and natural gas across six continents. ExxonMobil merchandises fuels, lubricants, and chemicals under four

[2] OPEC data web page.

renowned brands: Esso, Exxon, Mobil, and ExxonMobil itself. This is the main company chosen to center the analysis conducted in the project.

Figure 2.2: Exxon Mobil Corporation logo. Source: Company website

- *BP* (British Petroleum): UK-based global energy company.
- *Shell* (Royal Dutch Shell): British and Dutch integrated energy company.
- *Chevron*: US-based multinational energy corporation.
- *Total* SA: France-based major international oil and gas company.
- *Gazprom*: Russia-based energy company specializing in natural gas.
- *PetroChina Company Limited* holds the leading position as China's foremost producer and distributor of oil and gas. It accounts for roughly 50% of China's domestic oil volume and approximately 60% of its gas production.
- China Petroleum & Chemical Corp: producer and distributor of a variety of petrochemical and petroleum products

It is important to highlight that Saudi Aramco, Gazrpom PetroChina, and China Petroleum & Chemical Corp are majority state-owned companies by their respective states. On the other hand, BP, Shell, Chevron, Exxon, and Total SA are all owned by a variety of shareholders, including private and institutional investors.

Section 3. STATE OF THE ART OF NLP

In the current section, we conduct an exhaustive literature review of the main methodological issues to address in the project. As a brief reminder, the project's objectives revolve around classifying and analyzing financial news. Hence, this literature review focuses mainly on NLP (Natural Language Process) techniques for analyzing and understanding the human language used in such information in an automated way.

Although the concept of the Natural Language Process (NLP) may sound new and vanguardist, its origin dates back to the 1940s, in the context of the Second World War (Johri et al., 2021). The goal at that time was easy: translating with computers from one language to another (English and Russian). Since then, this field has not stopped advancing until today, when good communication between men and machines is being achieved.

Therefore, it can be said that the fundamental objective of NLP is to provide a computer with the ability to understand and interpret the content of texts and documents, having multiple uses, such as the classification or summarization of texts, topic modeling, or translation itself.

Throughout this section, three major blocks will be analyzed. First and briefly, the more theoretical part concerns the preprocesses and techniques used in the NLP field to understand and comprehend texts (section 3.1). Secondly, in section 3.2, the main architectures used over the years will be reviewed, particularly transformers (the most used architecture nowadays). Finally, a quick review will be made of LLMs (Large Language Models) in section 3.3,which have been very popular recently.

3.1 PREPROCESSING CONCEPTS OF NLP

When a computer processes a text to try to understand it, several steps and actions can be carried out to facilitate the complex comprehension process. The most commonly used are detailed below.

Tokenization is one of the essential steps; it could be said it is the starting point when carrying out NLP tasks. It consists of deconstructing unstructured data and natural language text into distinct portions of information that can be viewed as separate elements called tokens. By doing this, complex texts and language structures can be easily analyzed, which is why many architectures carry out this process, from more traditional models to the most advanced ones. These small entities can be done at different levels, most commonly the word or the sentence level.

When talking about the word tokenization, different levels can also be implemented. This concept is called n-gram, a tuple of *n* tokens where n represents an integer. For instance, it can be talked about unigrams or single words, bigrams which are pairs of words, trigrams, and so on. This approach is commonly used when trying to combine words that frequently occur together and whose collective meaning differs from the interpretation of each word (for instance, 'New York' does not imply a new York).

By carrying out the tokenization of a text, the text corpus is obtained, which is the collection of all tokens available in the text. Once the corpus is identified, the analysis to extract knowledge and conclusions could begin.

However, other steps can also be taken to improve further analysis. For instance, another vital step commonly used once obtained the tokens is the word normalization. They are two different ways of achieving this aim: *lemmatization* and *stemming*.

The first one, lemmatization, implies reducing inflated forms of words to their root form or base (known as the *lemma* of a word). To do it, the morphological aspect of the words is considered, grouping, for example, all the different forms of a verb under the same root. This process depends on the language used as different rules to get inflated forms of words are used in other languages (*e.g.*, genre and number). The second one, stemming, consists of obtaining the stems of words by removing the possible suffixes or prefixes used in each word. Unlike lemmatization, when stemming is applied, nonsense words can be generated, and the word's context is not considered (POS, Part of Speech). For instance, both techniques would reduce the form 'ran' to 'run'; however, stemming would reduce the word 'better' to 'bet,' whereas lemmatization would not change it.

When applying stemming to words, two main errors can be found: over-stemming (when two words with different stems are stemmed to the same root, that is, a false positive) and under-stemming (when two words that should be stemmed to the same root are not, that is, a false negative). An example of the first mistake is 'universe' and 'university,' where both words are stemmed to 'univer' and an example of the second error is 'data' and 'datum,' which could be reduced to 'dat' and 'datu' being that incorrect as both come from 'dat.'

Different algorithms to avoid these problems have been developed through the years. (Jivani, 2011) proposes a comparison among the main ones, differentiating three distinct groups: truncating, statistical, inflectional and derivational methods. All different ways explained and compared in the article to carrying out stemming have advantages and disadvantages, with slight variances in their methodology.

Nevertheless, it is worth highlighting one of them, the Porter algorithm presented in (Porter, 1980), which serves as the basis for all subsequent developed stemming algorithms. Porter's technique has shown excellent and fast performance, and it is based on a simple idea: eliminating the different suffixes assuming they are a combination of simpler suffixes. Improved versions have also been presented based on the original algorithm, like in (Willett, 2006), where the original algorithm introduced in 1980 is compared to modern versions with minor modifications, highlighting its importance.

Both stemming and lemmatization are wildly used in sentiment analysis as it helps to standardize words, reducing the existing redundancy and helping to identify better features. Articles in the literature comparing both techniques show clearly that the election of each method depends on the particular case of application (Khyani & Siddharta, 2021).

Another important step, after tokenization and lemmatization/stemming, is *eliminating the stop words* to reduce the volume of data analyzed. The stop words are the words that do not contribute semantic meaning to sentences, and consequently, they do not have value when studying a text. They are usually the most common and repeated words of a language and include articles (*e.g., a* or *the*), prepositions (*e.g., under* or *on*), and conjunctions (*e.g.*, *and* or *for*).

Finally, the last step is the Named Entity Recognition (NER). It is a widespread and complex process aiming to locate named entities in the text and categorize them into categories, such as person names, organizations, locations, medical codes, time expressions, quantities, percentages, etc. Identifying Named Entities depends not only on the language used in the text but also on the literary genre and the language field. Therefore, it is a complex process that requires powerful models for good results. Nowadays, deep learning approaches, particularly those based on transformer architectures (explained in section 3.2.3) like BERT (Bidirectional Encoder Representations from Transformers), have achieved state-of-the-art performance in NER tasks. An example is (Yan et al., 2019), where TENER is introduced, a Named Entity Recognition architecture that employs a modified Transformer Encoder for character-level and word-level feature modeling. By integrating attention mechanisms that are aware of the direction and relative distance, along with un-scaled attention, it is demonstrated that the Transformer-style encoder is equally effective for NER as other NLP tasks.

3.2 MAIN ARCHITECTURES REVIEW

The following is a journey through the evolution of NLP models, from the most traditional and simple ones to those most used nowadays, allowing one to carry out surprising tasks and produce impressive outcomes.

3.2.1 Word embedding

This concept dates back to the late 1980s and early 1990s and originates in the distributional hypothesis used in linguistics. This hypothesis posits that words usually in the same context tend to have similar meanings. With this in mind, word embedding could be explained as a specific type of word depiction that ensures words of similar meanings share comparable representations. They exist as a variety of vector space models, where vocabulary words are converted into vectors composed of real numbers. The most common methods for generating word embeddings are described below.

The first one is *Word2Vec*. Google developed this technique in 2013, which uses external neural networks to learn word associations from a large text corpus. It has two main architectures, the Continuous Bag of Words (CBOW) and the Skip-Gram (SG), which differ in how to generate the word embedding: CBOW predicts a target word given its context, while SG predicts the context words given a target word. Precisely, the article (Al-saqq & Awajan, 2019) presents a compilation of many practical studies comparing the results of both named architectures when applying them to sentiment analysis. The conclusion extracted from the article is that both methods show outstanding results. However, SK is better for infrequent words, whereas CBOW is faster and works well with frequent words.

Later, in 2014, a new approach that tried to combine the advantages of both CBOW and SG was presented in the article (Pennington et al., 2014) by a group of researchers from the American University of Stanford. Its name is GloVe (Global Vectors for Word Representation), and it is based on the co-occurrence statistics of words in a corpus, aiming to capture the semantic and syntactic relationships between terms. Therefore, it mixes ANN (Artificial Neural Networks) with a statistical approach.

A couple of years later, Facebook's AI Research team presented (Bojanowski et al., 2016) an enhanced version of the Word2Vec technique by introducing subword information, allowing it to handle out-of-vocabulary words and capture word compositionality. This version is called *FastText,* representing words as bags of characters n-grams.

A different approach was also developed in 2018 to overcome all these models. Unlike traditional word embedding techniques, ELMo (Embeddings from Language Models) looks at the entire sentence before assigning each word to its embedding, allowing for contextual information. That results in different embeddings for homonymous words, *i.e.*, words written in the same way but with different meanings.

3.2.2 Sequence to Sequence models

While progress continued to be made in word embedding techniques, another type of model called Sequence-to-sequence (*Seq2Seq*) was also presented in 2014 by Google (Sutskever et al., 2014). These models showed remarkable and improved results compared to previous approaches in tasks that involved sequence generation, like translation or text summarization, as they use Recurrent Neural Networks (RNN), especially Long Short-Term Memory Networks (LSTM), to handle sequential data.

These models typically consist of two parts: an encoder and a decoder. The first part, the encoder, processes the input sequence (a sentence, for instance), converting it to a fixed-length meaningful vector representation called the 'context vector.' The second part, the decoder, takes that representation built by the encoder and generates the output sequence.

These models have suffered many slight modifications depending on the activity target. For sentiment analysis, the topic of this dissertation, many articles exist in the literature. One of them is (Pham et al., 2018), where when applying Seq2Seq models adapted for sentiment analysis, the f1 score increased by 12 compared to the baseline model for a given dataset.

3.2.3 Attention and transformers

Despite the excellent performance of some Seq2Seq models, more advanced techniques have been developed the current state of the art of NLP techniques has been reached with transformers.

The attention mechanism, a component within a neural network, determines the relevance of source parts to the decoder at each step, eliminating the need for the encoder to compress the whole source into a single vector. Hence attention mechanisms improve Seq2Seq models by enabling dynamic attention to different portions of the input during decoding, enhancing context awareness and accuracy.

Transformers, a more complex and evolved concept, was introduced in 2017 in the famous article 'Attention is all you need' (Vaswani et al., 2017). They are only based on attention, not recurrence or convolutions in the encoder and decoder parts as in previous architectures. This structure type that only uses self-attention mechanisms tends to outperform as it can capture long-range dependencies and parallelize computations.

An excellent example of the performance of the different techniques explained over the same dataset can be found in (Habimana et al., 2020). Specifically, the various deep learning methods are evaluated when carrying out sentiment analysis, with excellent results for three models: BERT (this model will be explained in section 3.3), sentiment-specific word embedding models, and cognition-based transformers models.

3.3 LARGE LANGUAGE MODELS

The explained transformers are the base for many Large Language Models (LLM). These are powerful models capable of understanding and generating human-like text as they have been pre-trained on massive amounts of text data. They utilize the self-attention mechanism of transformers to learn contextual representations of words and sentences. That is, while transformers serve as the foundational architecture for capturing contextual dependencies in language, LLM harness this architecture extensively to achieve highly advanced language processing capabilities.

It is not easy to develop these large models due to the massive amount of data and resources they need to be built and maintained. However, some multinationals (*e.g.,* OpenAI, Google, and Meta) are working in this direction, as the results can be incredible and unique. Some of the most recent and known LLM will be discussed below:

3.3.1 BERT

The first acceptable LLM presented was BERT (Bidirectional Encoder Representations from Transformers) by Google Research (Devlin et al., 2018). The revolutionary concept introduced in this model is related to the word 'bidirectional.' This model acquires contextual word representations by incorporating information from both left and right contexts, enabling it to grasp the complex connections between terms and comprehend their contextual significance.

Additionally, BERT has played a significant role in inspiring the creation of several BERT-based models like RoBERTa (Robustly Optimized BERT Pretraining Approach), ALBERT (A Lite BERT), and deBERTa (Decoding-enhanced BERT with Distangled Attention). These models have built upon the foundation of BERT, enhancing its capabilities and addressing specific limitations to push the boundaries of language understanding even further.

3.3.2 GPT

GPT (Generative Pre-trained Transformer) is a family of large-scale language models developed by the American company OpenAI.

The first one was GPT-2 (Radford et al., 2019). It was a revolution as it was the most extensive NLP system trained to date, being a transformer with more than 1.5 thousand million parameters. It was necessary to use more than 45 million web pages to train it.

However, the great revolution came with the presentation of GPT-3 a year later (Brown et al., 2020). This new transformer considers around 175 thousand million parameters, which involves an increment of approximately 1,000%. It showcases substantial improvements over GPT-2 by scaling up the model size and the amount of training data.

GPT-3 harnesses the power of a large-scale language model to generate coherent and contextually relevant text across a wide range of natural language processing tasks, demonstrating the remarkable capabilities of language generation at a scale never seen before.

Furthermore, recently less than a year ago, Open AI launched the famous GPT-4 (OpenAI, 2023). It is also a large-scale, multimodal model with a great novelty: it can accept both image and text inputs and produce text outputs.

3.3.3 Other popular LLMs

BERT was the first popularized LLM, and GPT is the most famous worldwide by non-AI experts. Nevertheless, other companies are also developing their LLMs with good results, triggering great competition in this field of AI, where each newly released model can do more tasks more efficiently and with better results than earlier versions. This race will last many years in the nearest future.

Among those other famous LLMs, it can be mentioned PaLM2 (Pretraining with Adversarial Learning Method), presented in May 2023 by Google, to compete against GPT4 from OpenAI. This model has 340 thousand million parameters. Another important model is LLaMa (Large Language Model Meta AI), released by Meta in February 2023.

3.3.4 FINE TUNING

The concept of fine-tuning an LLM is fundamental and spread nowadays. It involves an extra round of training on a particular dataset or task, applying it to a pre-trained model. This process aims to hone and enhance the model's efficiency for that specific function or context. It capitalizes on the broad language comprehension gained during the initial training but adapts it to suit a particular use case.

For instance, a company can fine-tune an LLM with its own data to have a chatbot focused on the company. An example is JPMorgan, which is developing 'IndexGPT,' an AI stock adviser to replace traditional financial advisors.

Section 4. DATA USED

This section explains the data used throughout the project and some preprocessing to adapt it to this use case. Therefore, the section is divided into two main parts: section 4.1 deals with the data related to the financial news that will be analyzed, and section 4.2 refers to stock and market data, including historical stock market prices.

4.1 NEWS DATA

The analyzed news comes from the Financial Times (section 4.1.1) and the Wall Street Journal (section 4.1.2).

4.1.1 FINANCIAL TIMES ARTICLES

For the project, there is available access to news from the Financial Times from 2008-2017. The dataset from raw data is created, so some preprocessing is needed:

1. <u>Data Extraction</u>: each article, stored as a JSON file in a specified directory, is read. It is parsed for significant mentions of the target organization, in our case, major oil & gas companies: Chevron Corp, Royal Dutch Shell PLC, Gazprom OAO, Total SA, Exxon Mobil Corp, and BP PLC.

2. Data Transformation: each article with a significant mention is transformed into a structured format (rows of a DataFrame), where each column represents a specific attribute for each article.
3. Data Loading: the final DataFrame is exported to a CSV file.

This process is looped for each year from 2008 to 2017. Finally, the data from all these years is consolidated into a single DataFrame. This results in a decade's worth of data in a uniform, structured format, ready for further analysis, as shown in the table below:

Attribute name	Attribute description
id	A unique identifier for each news article in the dataset.
standfirst	A brief summary that provides an overview of the article's content. Missing for most of the articles.
firstPublishedDate	Date when the article was first made available to the public.
webUrl	The URL or web address of the article.
majorMentionedOrgs	The main organizations or companies that are prominently mentioned in the article.
title	The headline or title of the news article.
bodyXML	The main body of the article.

Table 4-1: DataFrame structure of articles' data from the Financial Times.

Further preprocessing is done to analyze later news data detailed in each paragraph instead of each article. Therefore, another dataset is created by separating this one, making a new row for each paragraph of each article. Since the body is in XML format, we can easily separate paragraphs by looking at *<p> </p>*.

4.1.2 WALL STREET JOURNAL ARTICLES

The data is already processed for the Wall Street Journal; no further transformations are needed initially. The format of the DataFrame is the following:

Attribute name	Attribute description
publishedDate	Date when the article was first made available to the public.
Company	The company discussed in the article.
Title	The headline or title of the news article.
textBody	The main body of the article.

Table 4-2: DataFrame structure of articles' data from the Wall Street Journal.

The only processing is again separating the articles by paragraphs, in this case, looking at the *\n'* in the body of the article, which represents line breaks.

4.2 MARKET DATA

4.2.1 STOCK DATA

The Yahoo finance library for Python is used to get stock performance data. The following dataset is obtained by defining the ticker symbol for the desired stock, 'XOM' in this case for Exxon, and the period of interest, 2007-2018.

The last column is added by comparing the close price of each day with the previous day's Adjusted Close price.

Attribute name	Attribute description
Date	Only dates when the stock market is open
Open	Opening price of the stock for the given trading day.
High	Highest price reached by the stock during the trading day.
Low	Lowest price reached by the stock during the trading day.
Close	Closing price of the stock for the given trading day.
Adj Close	The adjusted closing price considers factors such as dividends, stock splits, and other corporate actions.
Volume	Total number of shares traded for the given trading day.
Return	Percentage change in the stock price from the previous trading day, representing the daily return of the stock.

Table 4-3: DataFrame structure of stock data.

4.2.2 US MONTHLY FAMA-FRENCH 3

This dataset provides monthly returns for various factors needed for the Fama-French three-factor model, which will be discussed in the following sections. This model, based on the seminal paper by Eugene Fama and Kenneth French (Fama & French, The cross-section of expected stock returns, 1992), is widely used in finance to explain the evolution of expected returns of stocks based on their exposure to three economy-wide risk factors and the factors' risk premia: the equity market risk premium (Mkt-Rf), the risk premium associated with the differential evolution of small and big companies (SMB), and the so-called 'style-investing' risk premium (HML). The dataset includes several columns, each representing a specific factor. Although there have been significant additions to this model over the years, and the literature currently uses four and five-factor models (e.g., Fama & French, 2016) that include, for example, momentum (Carhart, 1997), (Jegadeesh & Titman, 1993), the original three-factor model is still a relevant baseline.

Attribute name	Attribute description
Date	Only dates when the stock market is open.
Constant	Intercept used in the asset pricing model.
Mkt-RF	Market Risk-Free: Excess return of the market portfolio (the difference between the market return and the risk-free rate).
SMB	Small Minus Big: accounts for the size factor, which measures the performance of small-cap stocks relative to large-cap stocks.
HML	High Minus Low: accounts for the value factor, which measures the performance of value stocks relative to growth stocks.

Table 4-4: DataFrame structure of the data needed for the Fama/French model. Sou

Section 5. NEWS CLASSIFICATION

After explaining the data used and how it was processed, this section 5 describes the first central part of the project: classifying financial news into two groups depending on their type of content. This section will first explain in more detail the goal of this classification (section 5.1), the process carried out to classify them (section 5.2), and finally, the results obtained for two of the newspapers analyzed (section 5.3).

5.1 GOAL

As stated in section 1.1, one of the project's main goals is to check if news published about companies affects the stock market, *i.e.,* to see if the company's stock price fluctuates in any way (positively or negatively) due to the news published about it.

To do so, only news with informational content will be considered as they add information about the company's background or possible future events. The objective is to analyze news where its effect is not already reflected in the stock price. This type of news will be treated as type 1.

The other type of news, without informational content, talks about past events whose impact has already affected the stock price and will not help to study future stock market movements based on published news. For example, an article describing an increase in a company's stock price that took place over the previous days has no informational content for investors over what is already reflected in the historical share price series. This news is type 0.

The goal is to compare the effect of news of type 1 against all the news (0 and 1).

5.2 METHODOLOGY

This section will explain the methodology of fine-tuning (optimizing) a pre-trained BERT model for binary text classification employing a dataset containing text samples labeled either 0 or 1. Both concepts of BERT (LLM using a transformer architecture) and fine-tuning (retraining a model on a particular dataset) have been explained carefully in sections 3.3.1 and 3.3.4, respectively.

After explaining the data preprocess (Section 4.), the data's starting point is a large dataset containing a paragraph of financial news from a particular newspaper on each row. Accessing that data must be in a compatible format with the BERT model extracted from the Hugging Face web page[3]. The transformers library from the same web page was used to achieve this, as it provides high-level APIs for dealing with these heavy transformers-based models. Furthermore, the library mentioned above includes a built-in tokenizer to convert text into tokens suitable for input to the chosen model.

[3] BERT model used

The process of tokenization of the text involved several steps:

- Breaking down the text into words, subwords, or characters (tokens).
- Mapping the tokens to their Ids in the vocabulary of the BERT model.
- Creating unique tokens needed by the model.
- Padding and truncating the tokens to a fixed length of 256.
- Creating attention masks to differentiate the actual tokens from the padding.

Then, as in all correct models of Machine Learning, the data was split into two sets: training and validation, and each one was loaded into PyTorch DataLoader for a more efficient data loading during training.

Once the data is processed, in order to improve the model, it was fine-tuned for this specific task. That is why another student, Laura Carbajo, manually labeled a set of 1,000 paragraphs randomly taken (1 or 0). The results were that 72.8% of the paragraphs were type 1 (informational content), and 27.2% were type 0.

The chosen model is the base uncased BERT, which has 110 million parameters and was trained in English. The model includes a final layer for sequence classification that was additionally trained on this particular task of financial news classification. During training, the model learned how to modify the weights of the pre-established BERT model to lessen the loss on the categorization, thereby fine-tuning the model. The AdamW optimizer provided by the transformers library was used, allowing the optimizer to weight decay fix.

In order to enhance generalization and avoid overfitting, dropout and early stopping were employed, halting the training process if there was no improvement in validation loss for three successive epochs, with a patience of 6 epochs. The model state with the optimal validation loss was preserved for making predictions. The training involved multiple cycles over the training data or epochs. For every data batch, the loss was calculated by comparing the model's predictions with actual labels, which were then fed into weight updates via backpropagation. After each epoch, the model's performance was assessed using the validation set.

Once the model was ready after the fine-tuning, it was time to predict the labels on new sentences of the original dataset. After passing the preprocessed text through the model, the output was converted into probabilities by applying the softmax function, selecting the highest probability label as the prediction.

In conclusion, the explained process presents a detailed walkthrough of fine-tuning a BERT model for text classification, from data preprocessing to model training and prediction.

5.3 RESULTS

The described process has been done through two newspapers from two countries: the British Financial Times and the American Wall Street Journal. In both cases, the news analyzed was about the mentioned company Exxon. Firstly, we will comment on the results obtained when training the model (section 5.3.1). Then the predictions obtained for each of the newspapers analyzed, Financial Times and Wall Street Journal.

5.3.1 TRAINING THE MODEL

Before getting the predictions, the model had to be trained following the earlier process discussed. To do so, news from different companies and newspapers was chosen. The following table (Table 5-1) summarizes some metrics when training:

Epoch	1	2	3	4	5	6	7	8	9	10
Avg training loss	0.59	0.43	0.36	0.31	0.30	0.30	0.30	0.29	0.30	0.30
Training accuracy	0.71	0.82	0.85	0.87	0.88	0.88	0.88	0.89	0.88	0.88
Validation Loss	0.45	0.39	0.38	0.37	0.37	0.37	0.37	0.37	0.37	0.37
Validation Accuracy	0.83	0.83	0.85	0.86	0.86	0.86	0.86	0.86	0.86	0.86

Table 5-1: Execution results when training the model.

As seen in the table, after each epoch, the loss decreases, and the accuracy increases until stabilization for both the training and the validation sets. The model stops training after ten epochs as there were no improvements in the validation loss for five epochs. The training loss can be seen graphically in Figure 5.1:

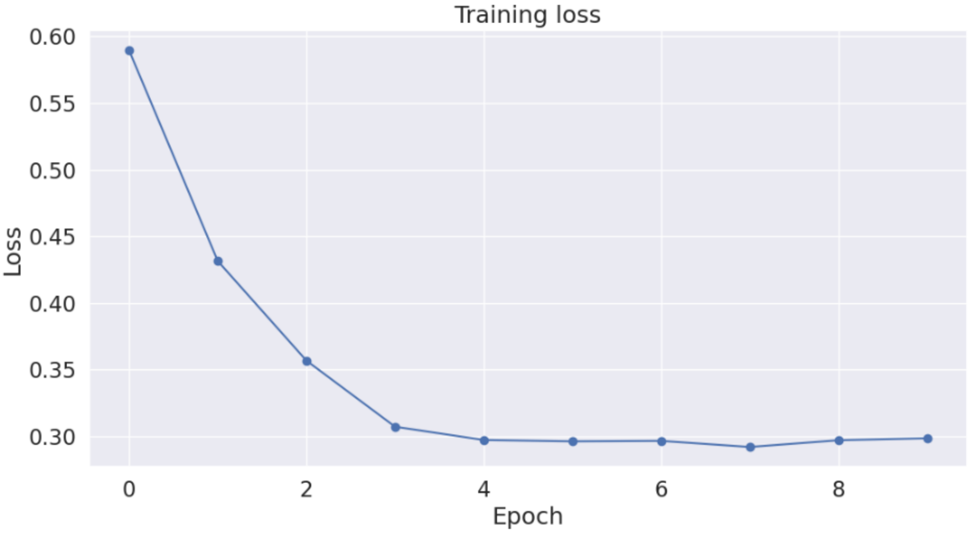


Figure 5.1: Graphical representation of the model training loss. Source: self-made

The obtained values for the validation loss and the accuracy are outstanding. This can also be seen when looking at other metrics, where the results are excellent, indicating an outstanding performance of the model:

ROC AUC	0.9061
Accuracy	0.8636
Precision	0.8814
Recall	0.9123
F1 score	0.8966

Table 5-2: Performance Metrics

5.3.2 PREDICTIONS

After the classification, a new column on the dataset is created: *Classification_label*, which has a value of 1 if it is a paragraph with informational content and 0 if not.

Predictions are focused on news from the company Exxon in the Financial Times and Wall Street Journal, having the following results.

The total number of rows of financial sentences talking about Exxon in the Financial Times news is 11,482. Of them, 65.4% were classified as 1 (informational content), and 34.6% as 0. The volume of information about Exxon in the Wall Street V is higher. Of its 19,052 financial sentences, 72.9% were classified as 1 and 27.1% as 0 (no informational content).

As a reminder, the values obtained in the set of 1,000 sentences labeled manually were 69.3% for class 1 and 30.7% for class 0. These values are in the middle of both newspapers, concluding that obtained values with the fine-tuning BERT model are acceptable. These analyses can be seen graphically in Figure 5.2:

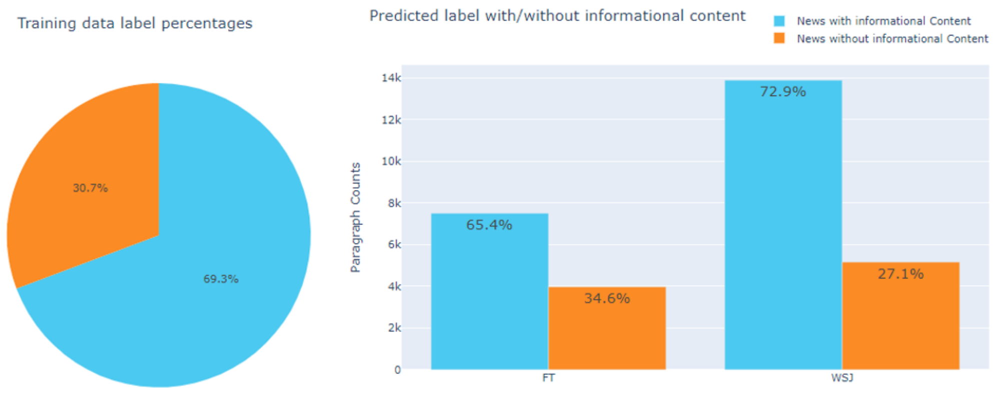


Figure 5.2: News classification results for Exxon. Source: self-made

Section 6. NEWS SENTIMENT ANALYSIS

After classifying news into two groups depending on their informational content, this section will revolve around the sentiment analysis of the classified news. One of the objectives was to see if the fluctuations in a company's stock price are related to the news published through sentiment analysis.

The section is divided into different parts. The first one, section 6.1, tries to find the best pre-trained model to carry out the sentiment analysis by comparing the performance of 5 models on a public and labeled dataset. Section 6.2 will study the sentiment of a particular company, such as Exxon, among the two newspapers studied, whereas section 6.3 is compared the sentiment of different companies in the same newspaper (Financial Times).

6.1 COMPARISON OF DIFFERENT SENTIMENT MODELS

Before carrying out the sentiment analysis on Exxon news, it is vital to determine which already trained model works better. Five different models are compared on a dataset[4] with around 4,800 financial sentences with a label associated with a sentiment. There are three categories: neutral, positive, and negative, with the

[4] Labeled dataset from Kaggle

following proportions: 59% of the sentences are neutral, 28% are positive, and 13% are associated with negative sentiment.

The models compared on this dataset are:

- Textblob uses a pre-trained sentiment analyzer that relies on a lexicon of pre-labeled words, and each is assigned a polarity (positive, neutral, negative) and subjectivity (objective, subjective) score.
- Vader: it also uses a pre-labeled lexicon of words, each assigned with a polarity score as textblob, but this model accounts for factors like intensifiers, negations, and the context of sentences to determine the overall sentiment of a text.
- BERT for sentiment product reviews: this model is the most used BERT for sentiment classification. As explained earlier, it uses a transformer architecture to identify the sentiments. However, it is important to note that it has been trained with product reviews, not anything related to finance.
- FinBERT: FinBERT is a specialized version of the BERT model, which is explicitly fine-tuned for sentiment analysis in the economic and financial context.
- FinBERT Tone: this model is another FinBERT trained with another dataset different from the previous one FinBert. The idea is to see if the training dataset for the financial part affects the result.

The results obtained are shown in Table 6-1:

Model	Accuracy	Precision	Recall	Time (min)
Textblob	0.491	0.516	0.487	1
Vader	0.543	0.605	0.532	1
BERT sentiment	0.31	0.49	0.26	12
FinBERT Tone	0.792	0.794	0.789	12
FinBERT	0.889	0.899	0.878	12

Table 6-1: Results of the comparison of sentiment models.

As seen in the results, the model that outperforms when doing sentiment analysis of financial headlines and sentences is the FinBERT model, a BERT model (transformer) that has been fine-tuned for the financial context. Two main conclusions can be extracted when looking at the results:

- BERT models take longer to run as they are heavy (transformers) models.
- The importance of fine-tuning a model for specific and concrete tasks. The *BERT sentiment for product reviews* model is the most used BERT for sentiment analysis. However, since it is not adapted to our specific task, which has very different language and expressions compared to the financial and economic context, the performance is much worse even if being a complex model.

The metrics mentioned above in the table for each sentiment and the confusion matrix were calculated for all the models, but it is only shown for the best model: FinBERT.

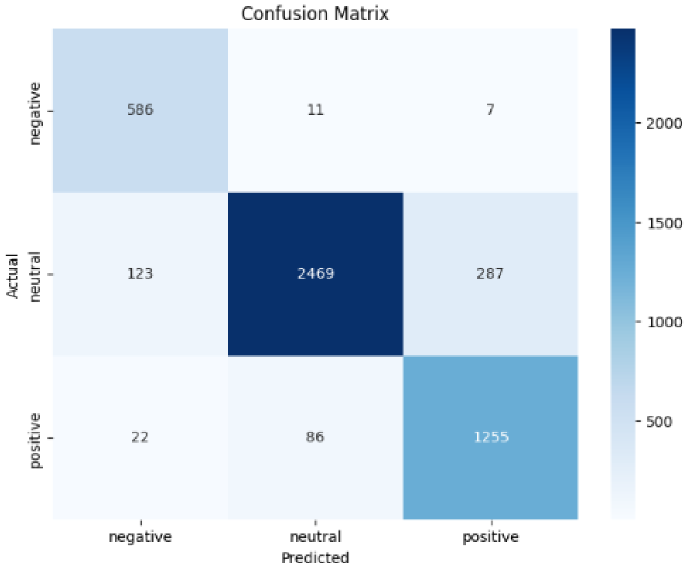


Figure 6.1: Confusion matrix for the FinBERT model. Source: self-made

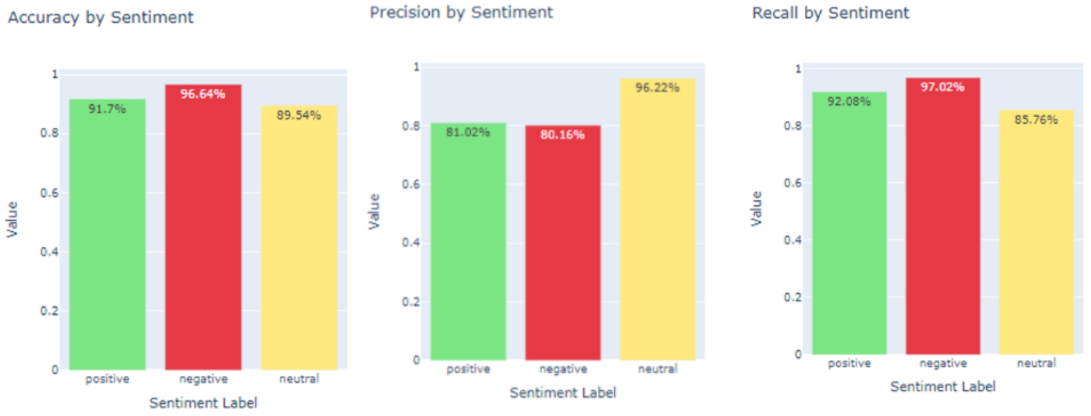


Figure 6.2: Accuracy, precision, recall for each setiment using the FinBERT model. Source: self-made

Both figures above demonstrate the FinBERT model's good results when classifying sentiments. By looking at the recall, it can be stated that the negative emotion is the one the model identifies easily and with fewer errors.

6.2 SENTIMENT COMPARISON OF EXXON NEWS IN THE FINANCIAL TIMES AND THE WALL STREET JOURNAL

This section will conduct a sentiment analysis in Exxon news articles from the Financial Times (FT) and the Wall Street Journal(WSJ) for two scenarios: all news and the filtered paragraphs classified as '1: with informational content' in Section 5.

The decided model to predict sentiment is FinBERT due to its good results in previous sections. Therefore, a function that runs FinBERT across all the rows in the FT and WSJ classified datasets is created.

After predicting the sentiment, the dataset has the following columns:

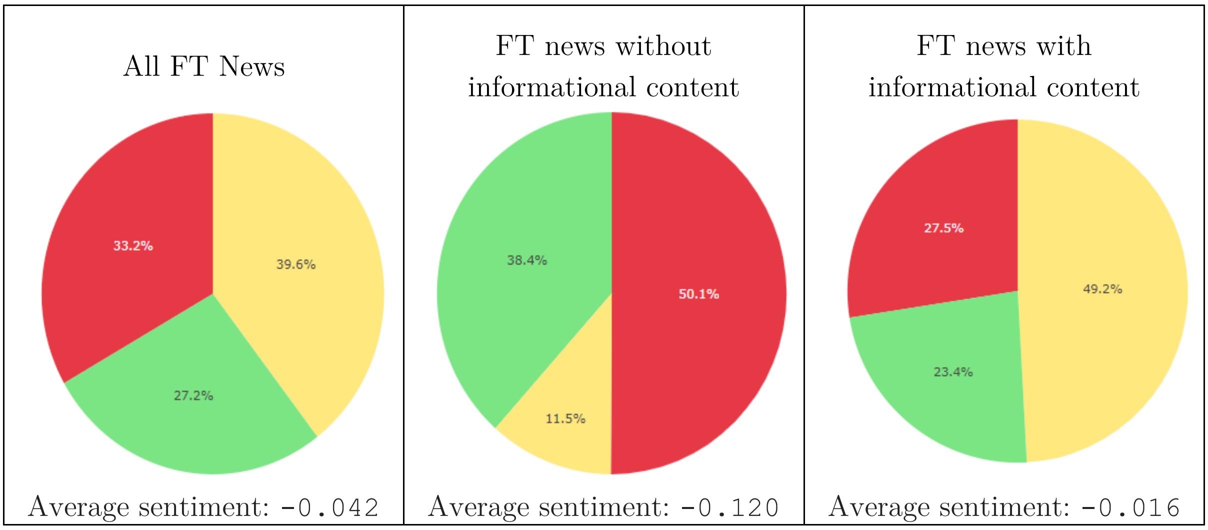
Attribute name	Attribute description
Article_id	A unique identifier for each news article in the dataset.
Paragraph number	Date when the article was first made available to the public.
Title	The headline or title of the news article.
Text	The main body of the article.
Classification_label	1: paragraphs with informational content, 0: without.
Prob_positive	Softmax probability of that paragraph being positive.
Prob_negative	Softmax probability of that paragraph being negative.
Prob_neutral	Softmax probability of that paragraph being neutral.
Total_sentiment	Number between -1 (negatives) and 1 (positives).
Major_sentiment	Categorical (positive, neutral, negative).

Table 6-2: DataFrame structure after predicting the sentiment.

Once all the news has been classified with its sentiment is time to visualize and compare the results obtained. In the graph below, three different things are compared for each newspaper: all the Exxon news (first column on the left), Exxon news that does not have informational content (middle column), and finally, Exxon news that does have informational content (right column).

Each case represents the percentage of news corresponding to each sentiment: green for positive, yellow for neutral, and red for negative sentiment. Furthermore, at the bottom of each pie chart, the total average sentiment of each case is shown.

On the pie charts, we observe approximately evenly weighted slices for All news, while a large portion of negative and also a big portion of positive paragraphs are found in news without informational content, and finally, mainly neutral paragraphs are observed in news with informational content. All average sentiments are slightly negative as it can be predicted by the bigger negative slices in all pie charts, mainly in the middle ones, which are majorly negative.



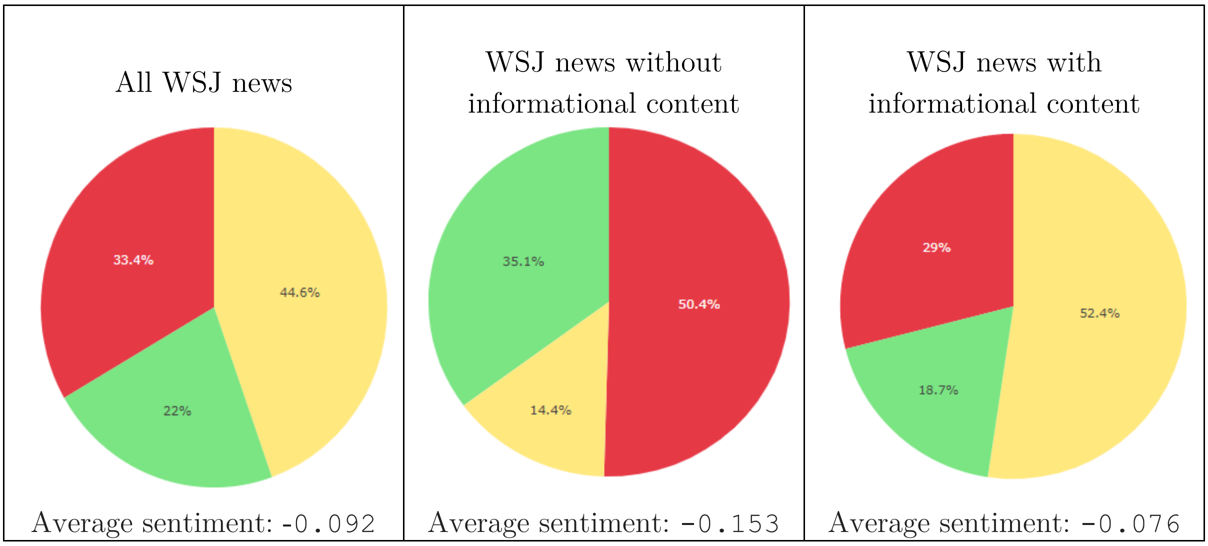


Figure 6.3: Sentiment comparison of Exxon news in FT and WST. Source: self-made

Moreover, the distribution of the total sentiment for each case shown above can be seen below. This analysis is essential and complementary as having a few news with high or strong feelings is not the same as having a lot of news with little or slight emotions.

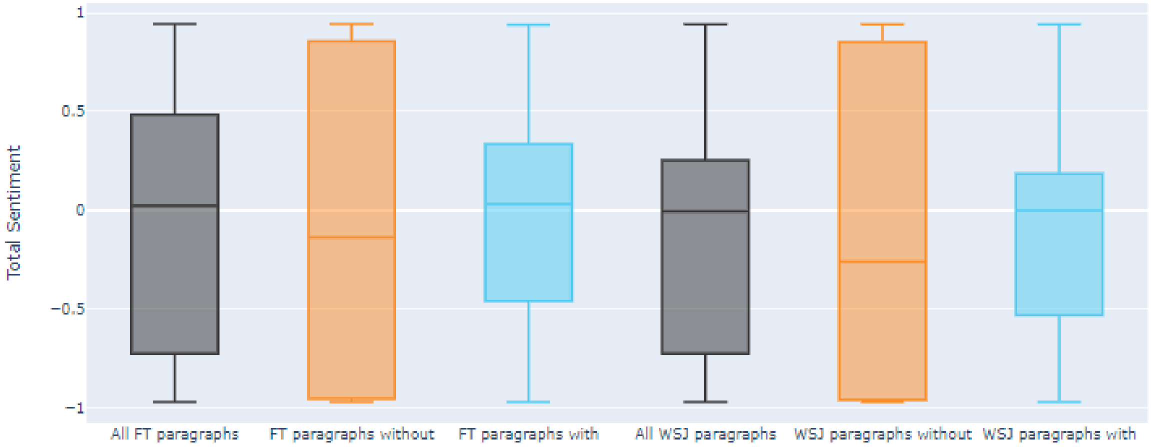


Figure 6.4: Total sentiment distribution of Exxon news paragraphs in FT & WST. Source: self-made

Some conclusions can be extracted from the previous graphs:

- News without informational content (orange boxplots) tends to be more extreme as they are talking about past actions already reflected in the stock market, so the financial sentiment detected by FinBERT is stronger. This is because descriptions of past movements in the stock price tend to be more clearly negative or positive ("stock rises, falls") than deeper analyses.
- As opposed to that, news with informational content is usually more neutral and does not show as much emotion or sentiment. This can be because the language tends to be more prudent when analyzing fundamentals in a company or discussing future predictions.
- Finally, the average sentiment of Exxon news paragraphs is slightly more positive in the Financial Times (British) than in the Wall Street Journal (US), which could be surprising since Exxon is an American company.

  To see if sentiment is significantly different, we compare the means of the two newspapers, which can be done using an independent two-sample t-test if total_sentiments are normally distributed, and the variances of the two populations are equal. We decide to start with a visual inspection:

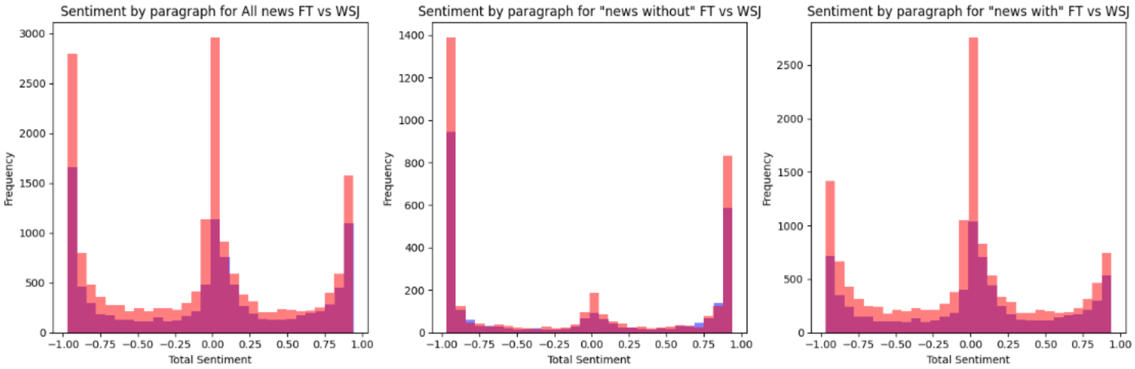


Figure 6.5: Histograms of Sentiment for paragrahs in FT vs WSJ news. Source: self-made

Visually, it is clear that the distributions do not follow a normal distribution in any of the cases. However, we decide to carry out some hypothesis testing to further support our visual conclusion:

We check for normality using the Shapiro-Wilk (Shapiro & Wilk, 1965) test and for equality of variances using Levene's test (Levene, 1960). For the first test, the null hypothesis states that the data is drawn from a normal distribution; the second test's hypothesis states that all input samples are from populations with equal variances. Therefore, low p-values will lead to rejecting the null hypotheses and assuming our data is not normal or has equal variances, while high p-values would mean normality and equal variance.

	Shapiro-Wilk Test		Levene's test	
	Shapiro stat	p-value	Levene stat	p-value
-FT All -WSJ all	0.912 0.918	0 0	66.60	$3.5\ x10^{-16}$
-FT without vs -WSJ without	0.786 0.787	0 0	2.62	0.105
-FT with vs -WSJ with	0.936 0.935	0 0	24.13	$9.0\ x10^{-7}$

Table 6-3: Shapiro-Wilk Test & Levene's Tests Results for sentiment distribution in paragraphs

From the table, we can observe how low p-values indicate rejecting both null hypotheses, indicaing that our data is not normally distributed, and variances are not equal.

Since we cannot assume normality, we proceed with the Mann-Whitney U test (Mann & Whitney, 1947). This is a non-parametric test that does not assume a normal distribution. It basically compares whether the distribution of the two groups is statistically different. The null hypothesis states that for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X.

	Mann-Whitney U test	
	statistic	p-value
FT vs WSJ all	$9.56\,x10^{7}$	$6.1\,x10^{-17}$
FT vs WSJ without informational content	$4.88\,x10^{6}$	0.022
FT vs WSJ with informational content	$5.80\,x10^{7}$	$2.88\,x10^{-25}$

Table 6-4: Mann-Whitney U test Results for FT vs WSJ sentiment distribution in paragraphs

Extremely low p-value (less than 0.05), indicate rejectibg the null hypothesis and conclude that there's a statistically significant difference between the two distributions. This indicates that WSJ and FT sentiment distributions regarding news to the paragraph level is statistically significantly different.

6.2.1 AGGREGATING PARAGRAPHS BY ARTICLE AND DATE

This sentiment analysis is carried out by diving deep into the detail at the news paragraph level. We will now first group by news article and calculate the sentiment for each article as a whole. This way, we avoid that long articles have a bigger influence than short articles.

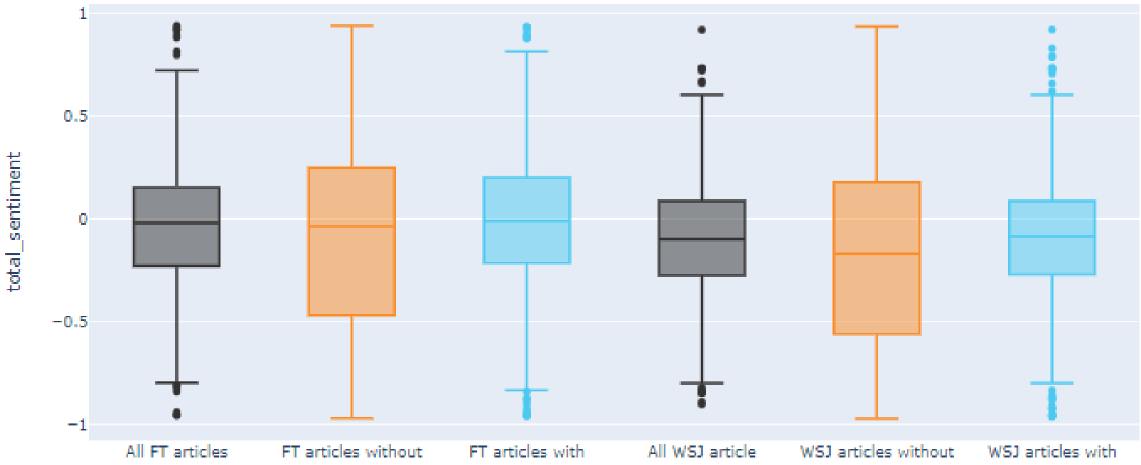


Figure 6.6: Total sentiment distribution of Exxon news articles in FT & WST. Source: self-made

When consolidating to the news article level, we still see the news articles without informational content be more extreme (with a broader distribution). However, the distribution of articles classified as "with informational content" is now practically identical to the distribution of all news.

In order to prepare the data for the next section (section 6.3), we will further group the data by date further and analyze the distribution in the following boxplot.

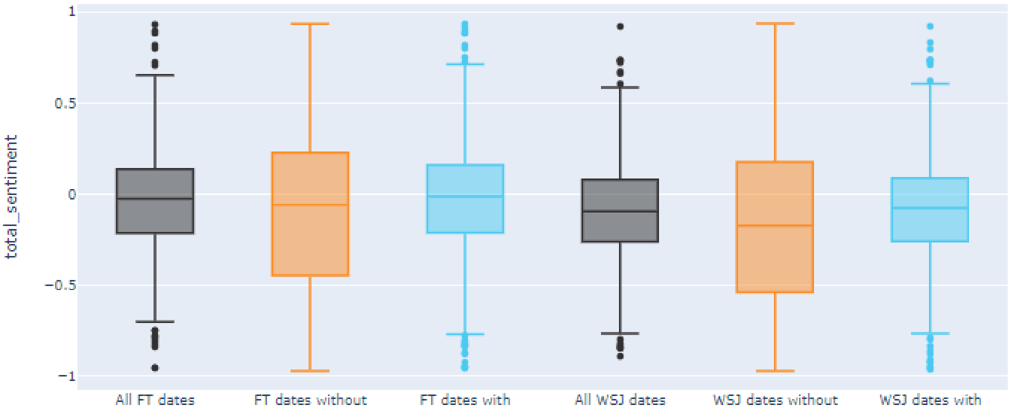


Figure 6.7: Total sentiment distribution of Exxon news Dates in FT & WST. Source: self-made

Now we see no major change drom the previous visualization, All News and News with informational content distributions are very similar. This suggests that previous analyses by different authors that did not classify news into the two groups we are classifying before sentiment analysis obtained equally valid conclusions. Another possibility could be that we more labeled data and/or a better classification model.

We are now intrigued to check the normaility of distributions after aggregating news by Date, so we will plot several histograms as well as performing the test for normality (Shapiro & Wilk, 1965) and for variance equality (Levene, 1960).

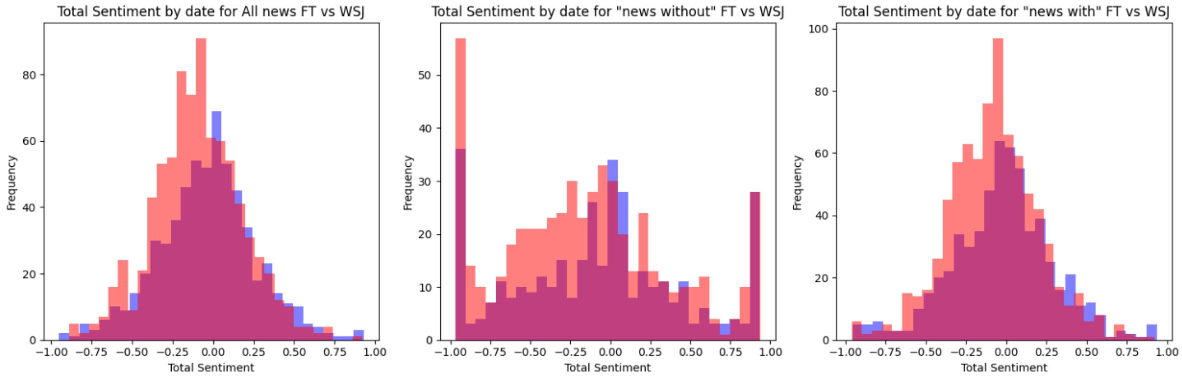


Figure 6.8: Histograms of Sentiment by Date in FT vs WSJ news. Source: self-made

	Shapiro-Wilk Test		Levene's test	
	Shapiro stat	p-value	Levene stat	p-value
-FT All -WSJ all	0.994 0.997	0.005 0.176	5.82	0.015
-FT without vs -WSJ without	0.961 0.962	4.5 $x10^{-9}$ 1.5 $x10^{-11}$	0.164	0.68
-FT with vs - WSJ with	0.990 0.993	1.4 $x10^{-4}$ 2.0 $x10^{-4}$	6.833	0.009

Table 6-5: Shapiro-Wilk Test & Levene's Tests Results for sentiment distribution by Date

From the table of results, looking at the low p values (for those <0.05) we would generally reject the null hypothesis and conclude that distributions are not normal, and variances are not equal. However, the Shapiro-Wilk test can be sensitive to large sample sizes like this one, and even tiny departures from normality can be detected. So, we might reject the null hypothesis for the Shapiro test (distributions not normal), even when a normal distribution would be a fine approximation.

So, when looking at the histograms we do see that the distributions are now much more normal shaped and could be approximated to a normal distribution. Therefore, for a final comparison of the means we are going to carry out a Welch's t-test (Welch, 1947) which is an adaptation of the Student's t-test that does not assume that the two distributions have the same variance. We will also perform the non-parametric Mann-Whitney U test (Mann & Whitney, 1947).

FT vs WSJ	Welch's t test		Mann-Whitney U test	
	statistic	p-value	statistic	p-value
All news	4.18	3.14 $x10^{-5}$	4.22 $x10^{5}$	5.1 $x10^{-6}$
Without info	2.28	0.022	1.44 $x10^{5}$	0.011
With info	4.55	5.68 $x10^{-6}$	4.05 $x10^{5}$	3.33 $x10^{-7}$

Table 6-6: Welch's t test & Mann-Whitney U test Results for FT vs WSJ sentiment dist by Date

Since all p-values are < 0.05 we can reject the null hypothesis in both Welch's t-test (both distributions have equal means) and Mann-Whitney U test (both distributions are equal and there is no difference between the medians), concluding that distributions in a per date level are different in the FT against the WSJ. This could be because these newspapers treat similar news with different sentiments, might publish similar news on different dates, or might publish different types of news.

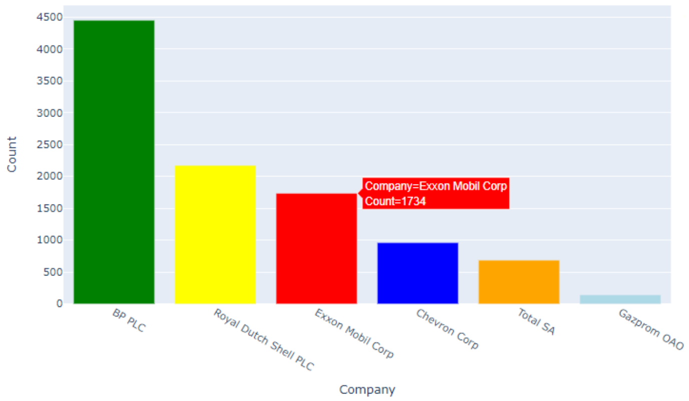
It is also interesting that among news classified as “without” informational content, the distributions are more similar as seen from the lower statistics and higher p-values.

After comparing sentiment distributions of the FT and the WSJ for the company Exxon, we conclude that news for the same company is treated with different sentiment both at the paragraph and date level.

6.3 SENTIMENT COMPARISON OF DIFFERENT OIL & GAS COMPANIES IN FINANCIAL TIMES

In this final section, instead of comparing the same company across two various newspapers, it is chosen just one of them, the Financial Times, and many companies in the oil & gas sector will be compared based on the sentiment of the news published on it. More concretely, the sentiment of each article's title. This is because there are too many articles when considering several companies, and the computing power required for analyzing the complete articles in this section would be too high.

There are a total of 6 companies that will be compared; two of them are British (BP and Shell), two American (Exxon and Chevron), and other two from the European continent (Total SA from France and Gazprom from Russia). The total number of available articles from 2008 and 2017 of these six companies is 10,140, with the following distribution shown in Figure 6.9:



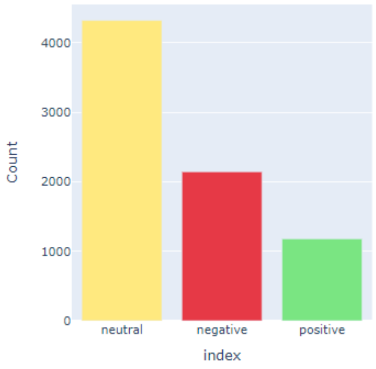


Figure 6.9: Distribution of the number of articles for each company analyzed in the Financial Times.

Figure 6.10: Distribution of the major sentiment for all the 6 companies. Sources: self-made

As discussed earlier, the sentiment value is between [-1,1], being -1 negative, 0 neutral, and 1 positive. When calculating the average sentiment value across the six companies, the result is -0.0931, which is close to neutral, slightly more negative. This can be seen graphically in Figure 6.10, with the major sentiment distribution for the more than 10 thousand articles analyzed. The majority class is neutral, followed by negative and positive articles.

The same analysis can be conducted yearly. The following figures (Figure 6.11 and Figure 6.12) show the total number of articles by each sentiment throughout the years and the yearly distribution of that total sentiment. The previous section commented on the reason for studying this double perspective.

In all the following graphs, the same color legend is applied (yellow for neutral, red for negative, and green for positive).

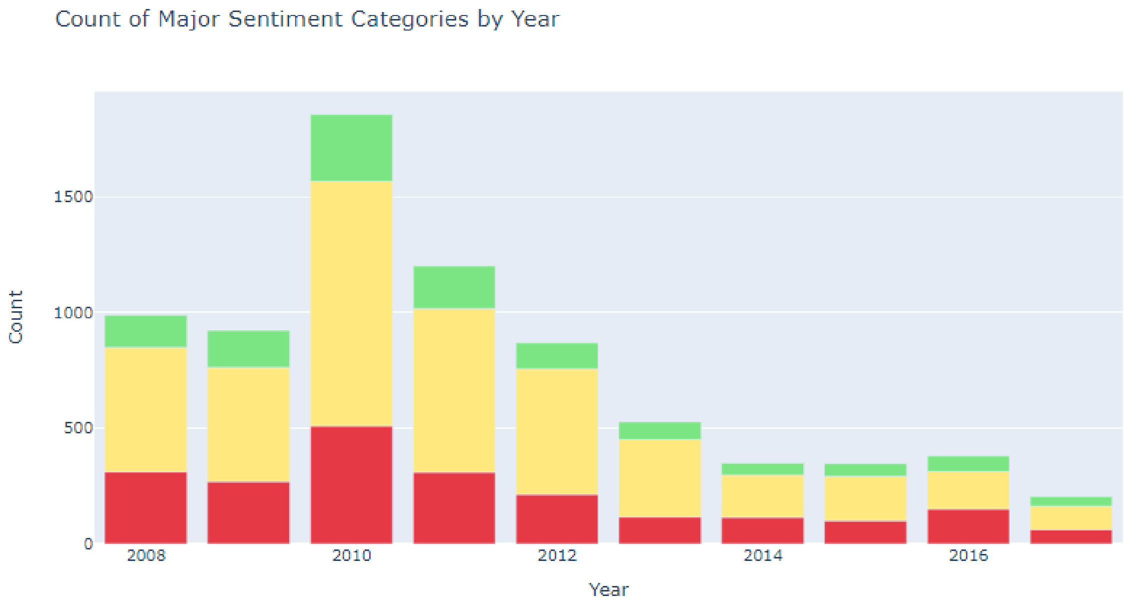


Figure 6.11: Yearly distribution of the major sentiment for all the 6 companies. Source: self-made

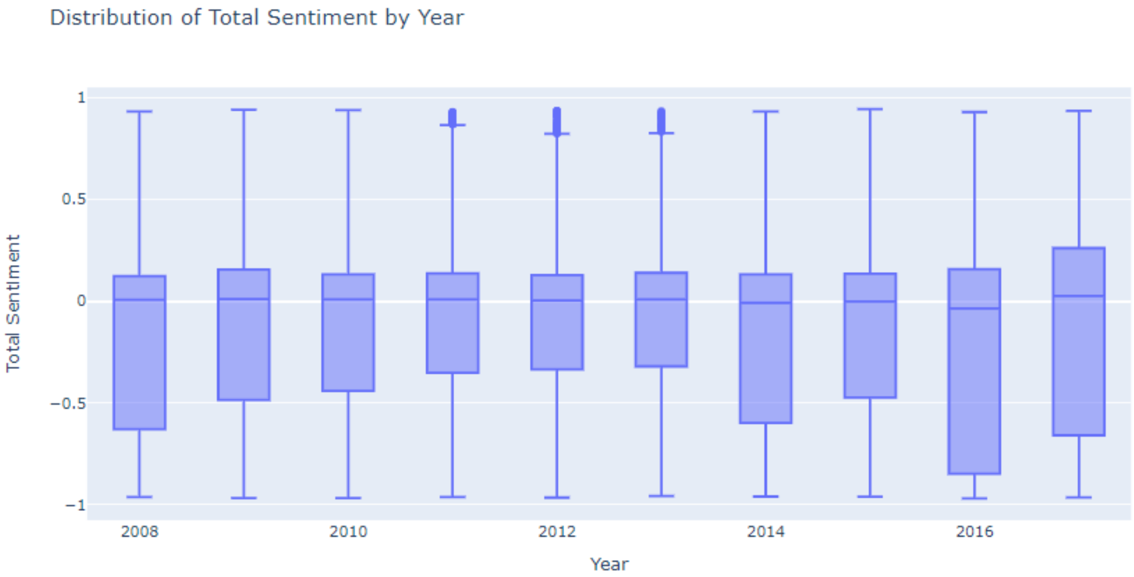


Figure 6.12: Yearly distribution of the major sentiment for all the 6 companies. Source: self-made

From these graphs, it can be stated that in 2010 and 2011, there was a boom in financial news. However, since 2014, the Financial Times experimented with an apparent decrease in financial news published on oil & gas. This can be due to different factors, including overcoming the financial crisis, people's tiredness of financial events, and the change in the focus to other types of news like politics with the Brexit referendum in the UK in 2016 or the presidential elections in the US where Donald Trump won.

That is about the volume of News. Regarding their sentiment, it can be said that the average sentiment is close to 0, neutral. However, the amount of strong negative news is much higher every year than the positive one, evident from the downward skew of the boxplots. From 2008 to 2013, that negative sentiment decreased as the financial crisis was overcoming, but in 2014 started again to fluctuate negatively, reaching its peak in 2016, overlapping with the Brexit referendum.

This same analysis can be done by looking at the month instead of the years. The results can be seen in Figure 6.13. The main conclusions are that June and July are the most active months of the year regarding financial news, and November and December are the least. In June, many companies end the second quarter, governments end their fiscal year too, and the driving season starts in the US. Moreover, some international economic summits, including G20, are held in this season. On the other hand, November and December are traditionally months to review the year's performance, doing balances and not doing that much activity. Moreover, in Western countries, are important holidays like Thanksgiving or Christmas.

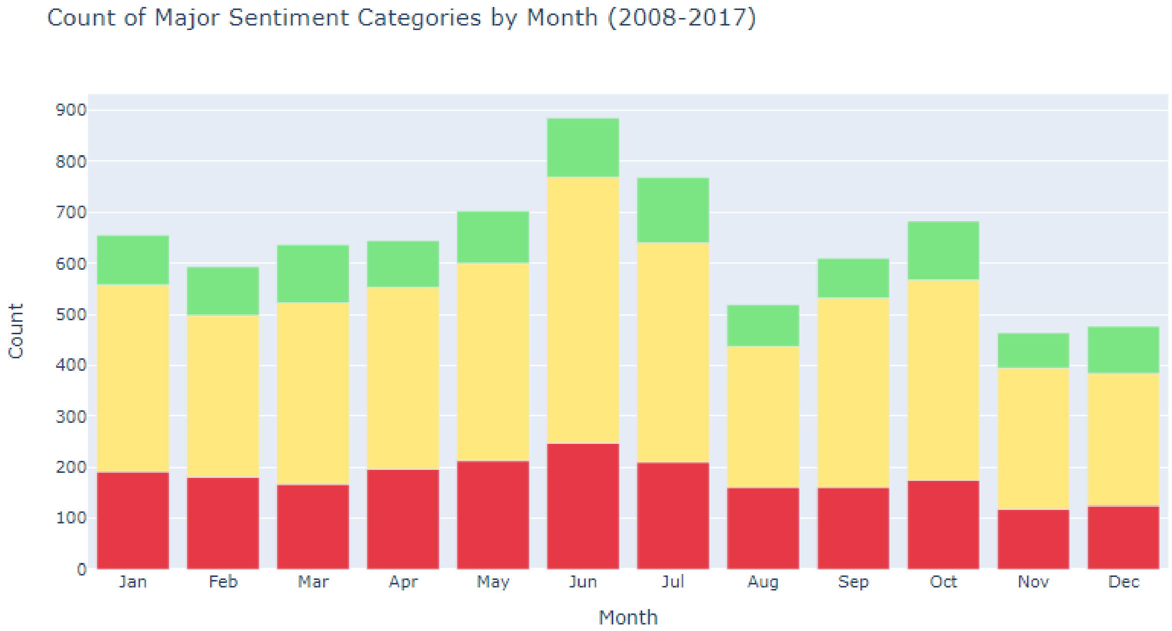


Figure 6.13: Monthly distribution of the major sentiment for all the 6 companies. Source: self-made

After studying all the news together, we analyze each company individually rather than the aggregation of all of them. To do so we create a separate dataframe of new for each company.

The same visualizations are shown again but applied to each company individually. Firstly, in Figure 6.14, a pie chart with the percentage of the news's main sentiment for each company was studied with the total average sentiment.

Secondly, in Figure 6.15, the total sentiment distribution among the companies is shown.

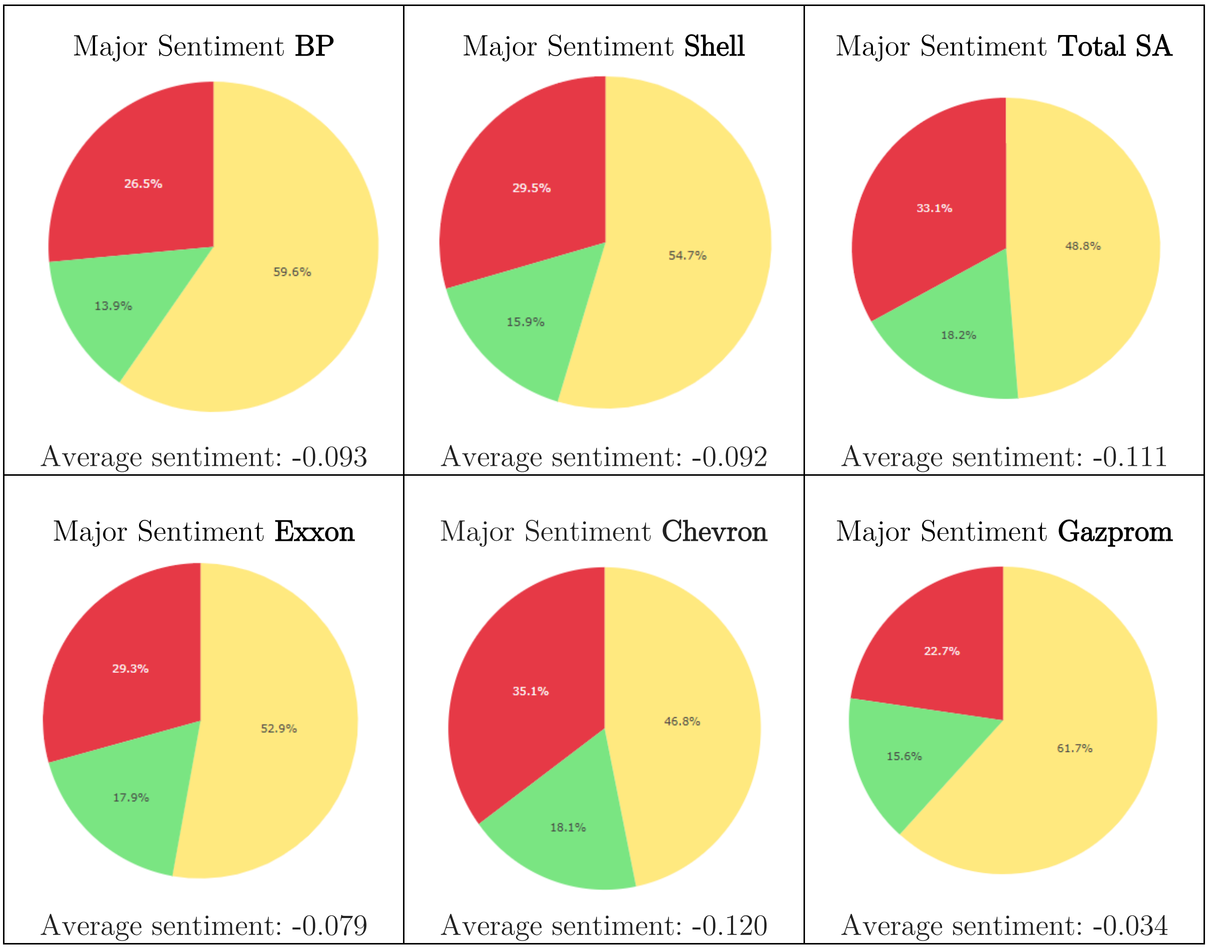


Figure 6.14: Distribution of the major sentiment for each company individually. Source: self-made

The results of the general average sentiment of each company are aligned with the known results; all of them are slightly negative without significant differences. Chevron has the worst feeling, followed by Total SA and Exxon (2 of them are American). On the other hand, the Russian Gazprom is the most neutral, which can be due to the low number of articles available. The two British companies have similar behavior, being more than 1 out of 2 articles neutral. All companies have similar results, meaning the newspaper follows a rigorous methodology when discussing financial news.

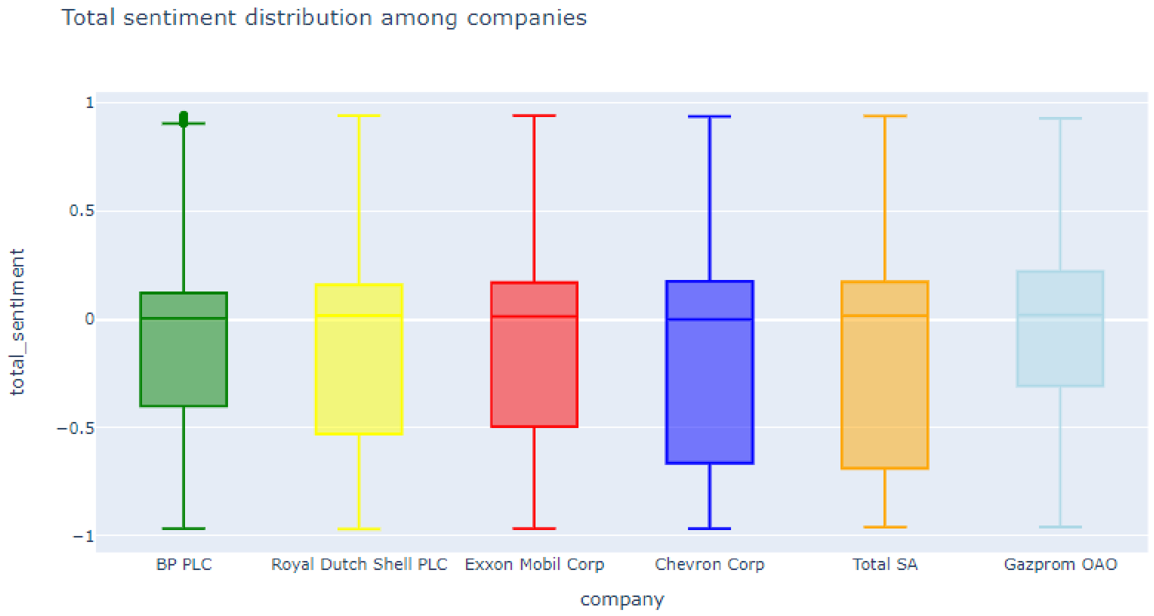


Figure 6.15: Distribution of the total sentiment for each company individually. Source: self-made

When looking at the total sentiment distribution, similar results can be extracted. There is much less positive news than negative. However, the overall sentiment tends to be 0, a neutral value. It stands out Total SA and Chevron, with high negative values.

Finally, in Figure 6.16, Word Clouds are shown for each company's articles containing the most relevant words. They all share common words related to the sector (oil, gas) and have information about their countries, continents, and even the stock market where they are registered (FTSE in London for Shell and BP, Wall Street for Exxon, and Chevron).

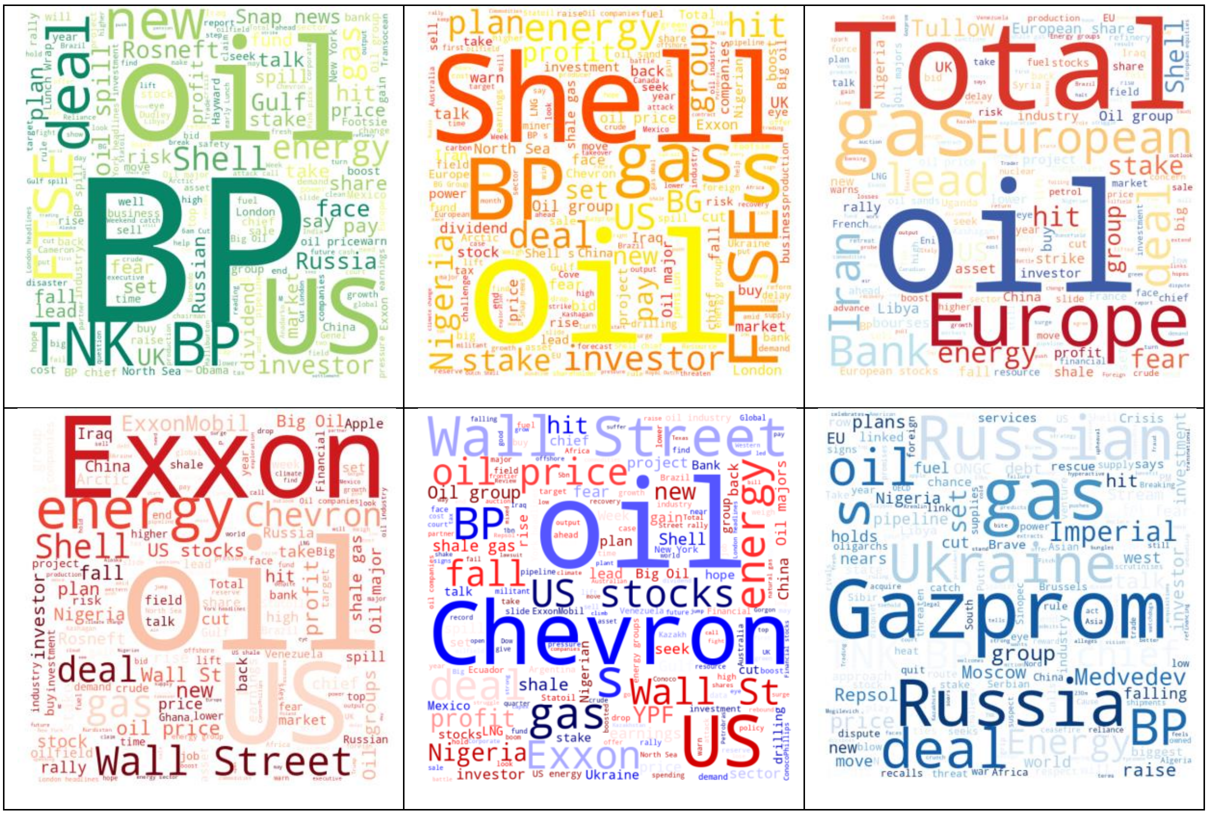


Figure 6.16: Word Cloud for each company. Source: self-made

Section 7. ADDING MARKET DATA

This section deals with event studies, a type of analysis used in finance and economics to evaluate a particular event's impact on a firm's value. It is divided into two significant parts: the first one, section 7.1, tries to explain a little bit theoretically this concept of event studies and the different ways and manners it can be carried out, whereas section 7.3 deals with the development and implementation of the chosen model. Finally, section 7.3 discusses the results of applying that model to the case of the Exxon company in the two analyzed newspapers (Wall Street Journal and Financial Times).

7.1 THEORETICAL CONCEPTS

When discussing event studies, the goal is to examine how the market responds to significant occurrences such as mergers and acquisitions, earnings declarations, macroeconomic updates, and other news or events that could affect a company's worth. Therefore, the core concept involves recognizing a shift in market expectations concerning a company's prospective earnings or risks resulting from the information disclosed by the event. This shift is typically quantified by the abnormal return, which is the discrepancy between the actual return during the event period and the anticipated return per a given model (for example, the market model).

Therefore, this section analyzes different approaches to demonstrate how a specific event has influenced the valuation of the studied financial asset. These techniques can explain an event's effect and predict how an investment will react to a particular situation. These methods are based on applying statistical techniques to obtain the desired results. The standard procedure (MacKinlay, 1997) to use any of them is described below:

1. Identify the period of interest over which the event to be studied has an effect. It is typically the publication day of the new and maybe one or two more days after that.
2. Define a time window before the event, longer than the period of interest. Throughout this period, the returns obtained by the company will be studied, allowing for comparison and deciding whether or not the return received has been anomalous.
3. Define the company to study based on the analyzed news to see the correlation. In this particular case, the company is Exxon.

These models can be divided into two main groups based on their point of view: statistical (statistical assumptions regarding the return on assets) or economic (assumptions about investor behavior). Some of the first ones will be described below, focusing on the one chosen for this project:

- The constant return model, also known as the mean return model, originated as a simplistic assumption in event studies that the expected return of a stock is stable over time, i.e., it neither reacts to market trends nor specific events.

In practice, this model is used as a benchmark to calculate abnormal returns during an event, which are then analyzed to see if a particular event has caused a significant change in the stock's returns.

- Market model. This model tries to overcome the main drawback of the previous model, which is that it is based on a false assumption as the market moves up and down based on economic cycles. This approach introduces an important factor called beta from the CAPM (Capital Asset Pricing Model) (Sharpe, 1964) (Lintner, 1965). It is used to associate companies with the market as it associates systemic risk and expected asset returns.
- Three-factor model. This is a well-known asset pricing model developed by Eugene Fama and Kenneth French in the 1990s (Fama & French, 1992). It extends the Capital Asset Pricing Model (CAPM), which includes only market risk, by adding two additional factors: the size of firms (small minus big market capitalization, or SMB) and the book-to-market ratio (high minus low, or HML). These two additional factors aim to explain the variations in stock returns better. The formula that describes this behavior can be seen below:

$$R_i - R_f = \alpha_i + b_i \cdot \left(R_M - R_f\right) + s_i \cdot SMB + h_i \cdot HML + \epsilon_i$$

Where $\left(R_M - R_f\right)$ represents the market return, excluding the risk-free rate and the factors b_i, s_i and h_i are the slopes of the regression curves in a time series.

This model has undergone many variations and improvements, most notably the one carried out in 2016 by the creators of this model, including two new factors, thus giving rise to the birth of the five-factor model (Fama & French, 2016). These two new factors included are profitability (those companies that obtain higher future earnings have a better performance in the stock market) and investment (relates inward investment and returns, suggesting that companies that direct their profits to large growth projects are more susceptible to experience market losses).

7.2 IMPLEMENTING AND APPLYING THE CHOSEN MODEL

This section will first describe the data needed and used to conduct this analysis (section 7.2.1) and later explain how the chosen three-factor model can be applied in a real example (section 7.2.2).

7.2.1 DATA NEEDED

In order to implement the three-factor model, different sets of data are needed:

- A dataset containing a company's stock prices with the date and the daily return.
- For each day, a dataset containing the three variables explained earlier (risk-free market rates, SMB, and HML).
- The third and last dataset includes the dates of events, that is, dates with extreme news sentiment associated (positive or negative) and its value.

The period studied is from 2008 to 2017, having chosen the window time of a year (time gap to establish and calculate average returns) and the influence time of a single day (time to see the effects on stock price). Furthermore, it must be said that the value taken into account will be the stock's close price.

7.2.2 METHODOLOGY

This section will explain how to implement the chosen model to conduct event studies, following the method described in (MacKinlay, 1997). Let's briefly remind the equation of the model:

$$R_i - R_f = \alpha_i + b_i \cdot \left(R_M - R_f\right) + s_i \cdot SMB + h_i \cdot HML + \epsilon_i$$

The first step is to calculate the actual daily return (R_i) based on the stock's adjusted close price, which can be positive (if the prince went up compared to the previous day) or negative. This can be done following the equation below:

$$R_i = \frac{Close\ price_i}{Close\ price_{i-1}} - 1$$

As the values R_f, R_M, SMB, and HML are already present in a dataset and R_i has just been calculated, the only factor to be calculated in the equation mentioned above are the other terms containing subindex *i*, i.e., α_i, b_i, s_i and h_i. To calculate them, a multivariate linear regression is calculated by taking the available data of a financial year. It has been calculated by applying the Ordinary Least Squares regression (OLS) in Python.

By substituting the values into the general model equation and leaving R_i as the unknown, we obtain the company's expected daily return for a given day, i.e., E(R_i) and will be substituted into the equation with the same multipliers for all days of the window period to check if the result is constant.

The following step is, with that expected value calculated, compare it with the actual return obtained, getting the Abnormal Return (AR) by following the equation described below:

$$AR = R_i - E(R_i)$$

The critical point now is to create a hypothesis contrast to see whether the mean AR before the event equals the mean AR obtained after the event. In this equation, the null hypothesis (H0) will be that the mean in the AR before and after the event are equal, and the alternative (H1) will state that they are different. In this case, M is the number of days considered before the event, and N is the number of days after the event.

$$H_0: \frac{1}{M} \cdot \sum_{j=1}^{M} AM_{t-j} = \frac{1}{N} \cdot \sum_{j=1}^{N} AM_{t+j}$$

$$H_1: \frac{1}{M} \cdot \sum_{j=1}^{M} AM_{t-j} \neq \frac{1}{N} \cdot \sum_{j=1}^{N} AM_{t+j}$$

As stated earlier, the influence time after the event is just one day, i.e., N = 1, allowing for simplification of the equations above. To solve the contrast, it must be calculated the AR variance of the past events by applying the following equation, where L_1 is the first day of the previous window, and L_2 is the last day, i.e., the previous day of the event:

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

Once that value is obtained, the statistic's value can be computed using the actual abnormal return of a day minus the abnormal return of the previous day divided by the abnormal return of the last year:

$$Z = \frac{X - \mu}{\sigma} = \frac{AR_t - AR_{t-1}}{\sigma(AR)}$$

The result will be compared with a normal distribution of mean 0 and standard deviation σ_i, calculated as follows:

$$\sigma_i = (L_2 - L_1 + 1) \cdot \sigma$$

However, it is important to note that two parameters are being estimated in this calculation: the mean and the variance. Consequently, a t-student distribution with 2 degrees of freedom should be used. The significance value is obtained with that comparison, allowing us to accept or reject the null hypothesis and, therefore, to state whether the received return is average.

An important note to remember is that this way of studying each event individually has the problem of not considering different events affecting the stock price simultaneously. This is left for a more in-depth project on this topic.

With this methodology, we will be able to evaluate the stock price reaction to news (information content) and no-news (Chan, 2003).

7.3 RESULTS

After explaining how to apply the model in a real example, it is time to see and discuss the results of using the Exxon company's articles in two newspapers: WSJ (Wall Street Journal) and FT (Financial Times).

7.3.1 OUTLIER DETECTION

The first step is to transform the dataset to apply this study. Instead of having news paragraphs on each row, each row now represents a day with a sentiment value extracted from grouping all the news for the same day and calculating the average of their sentiments (as described at the end of section 6.2).

After that, a process to detect outliers is carried out. This procedure identifies and highlights news articles with remarkably high or low sentiment scores. These 'outliers' could have a more significant influence on stock prices and thus deserve additional scrutiny.

To standardize the sentiment scores of the articles, the statistical z-score is used. This measure is beneficial in spotting outliers as it calculates the number of standard deviations a data point deviates from the mean. Thus, by utilizing z-scores, we can evaluate the degree to which the sentiment of a particular article diverges from the overall sentiment distribution.

We define outliers as those observations with a z-score that surpass a specific threshold=2 standard deviation, either in a positive or negative direction. This threshold signifies a boundary beyond which sentiment scores are exceptionally high or low.

Results are shown on Table 7-1: Results for outlier days in terms of sentiment (threshold = 2 std).

	Days	Non outlier days	Outliers	(%) Positive & Negative Outliers
FT All	387	348 (90%)	39 (10%)	43% +
				57% −
FT with informational	366	319 (87%)	47 (13%)	44% +
				56% −
WSJ All	481	432 (92%)	49 (8%)	18% +
				82% −
WSJ without informational	482	419 (87%)	63 (13%)	21% +
				79% −

Table 7-1: Results for outlier days in terms of sentiment (threshold = 2 std).

As we see Table 7-1: Results for outlier days in terms of sentiment (threshold = 2 std). about 10% of dates across the different datasets are found to be outliers. The percentage of outliers is slightly higher for news with informational content. This might surprise us since when we were doing the analysis at a paragraph level; we saw that non-informational news (paragraphs) was more extreme than informational. However, here we have aggregated paragraphs into articles and articles into dates. So, when aggregating the mean sentiment for paragraphs into articles, extreme positives might cancel out extreme negatives for non-informational news. Overall, we do not see a significant difference in outlier detection between all News and News with informational content when analyzing sentiment in dates.

7.3.2 EVENT STUDY

For this section we filter dates after 2008 (as the window size is one year) and before 2018 (since we do not have complete news for this year). To perform an event study we select, as explained above, dates with extreme sentiment (outliers) in the news and compare the abnormal return on the day before the event with the day after. We also repeat the process with non-outlier dates.

We are going to separate the results into two tables: one with positive sentiment dates and one with negative sentiment dates. The first three columns show total days, outlier and non-outlier days with percentage over total days. The next show results referring to the event study: total significantly important events (percentage of total days), out of those how many where outliers and how many non-outliers (percentage of significant outlier/non-outlier events out of outlier/non-outlier days).

Positive Dates

	Total days	Outlier days	Non outlier days	Total significant events	Significant events for outlier days	Significant events non outlier days
FT All	151	15 (10%)	136 (90%)	55 (33%)	5 (33%)	50 (37%)
FT with informational	149	18 (12%)	131 (88%)	58 (39%)	5 (27%)	53 (37.5%)
WSJ All	94	25 (27%)	69 (73%)	30 (32%)	11 (44%)	19 (27%)
WSJ without informational	110	32 (29%)	78 (71%)	30 (27%)	11 (33%)	19 (24%)

Table 7-2: Results of event analysis for Positive Dates.

Negative Dates

	Total days	Outlier days	Non outlier days	Total significant events	Significant events for outlier days	Significant events non outlier days
FT All	188	16 (9%)	172 (91%)	68 (36%)	4 (25%)	64 (37%)
FT with informational	178	20 (11%)	158 (89%)	60 (33.7%)	4 (20%)	56 (35%)
WSJ All	306	16 (5%)	290 (95%)	98 (32%)	2 (12.5%)	96 (33%)
WSJ without informational	292	19 (7%)	273 (93%)	100 (34%)	6 (31%)	94 (34%)

Table 7-3: Results of event analysis for Negative Dates.

About 1/3 of the total days is found to be a significantly important event, which suggests that news do have a significant impact on the stock market.

However, we do not see a significant difference in this percentage among sentiment outlier dates and non-outlier dates. For some categories, such as positive sentiment in the WSJ, we do find a higher percentage of significant events for outliers (44%, 33% vs 27%, 24%) but then on the same newspaper for negative sentiment we see few significant events (12.4%, 31% vs 33%, 34%).

This indicates that the sentiment results we obtained apparently do not have a big impact on detecting significantly important events. A possibility for improvement could be using different sentiment analysis techniques that, for example, can detect when sentiment is directly related to the company and when it is not. Another possibility is that other factors in the news, such as ‘topic’ have a very significant impact too and we are not considering them.

Section 8. CONCLUSIONS & FUTURE WORK

After finishing this project and writing this dissertation, some important conclusions can be extracted when talking about news classification and sentiment analysis applying NLP techniques and their relationship with event studies in the stock market:

- After classifying news paragraphs into those with informational content (company or market analysis not yet reflected in the stock price) and paragraphs without informational content (commentary about past stock movements), we see how the former has a much less extreme and more neutral sentiment distribution than the latter. This is because commentary about how previous days' share price rises or falls outputs a much stronger sentiment than fundamental analysis of a company, the market, or predictions about the future, where language is more prudent and neutral.
- Continuing with our analysis when aggregating paragraphs into whole news articles and then grouping by date, where we observed how the previous effect disappeared completely. At the date level of detail, the distribution of filtered informational news does not vary significantly from the distribution of all news. This suggests that previous analyses by other authors that did not dive deep into analyzing news by paragraph may extract equally valid conclusions since the effect disappears when aggregating by date to study the effect on the stock market.

- When comparing sentiment distributions between the Financial Times (FT) and The Wall Street Journal (WSJ) with respect to the same company: Exxon, our findings indicate significant differences in sentiment portrayal, both at the paragraph level and at the level of each published date. The observed disparity could be attributed to the newspapers expressing similar news with different sentiment intensity, publishing analogous news on different dates, or even focusing on different facets of news altogether. This conclusion highlights the importance of considering multiple sources in news analytics to gain a comprehensive understanding of sentiment trends.
- Finally, when carrying out the event studies, we find out that news does have a significant impact on the stock market, but the extreme sentiment is not necessarily related. This suggests that a different approach or technique might be necessary to fully detect the desired sentiment, which would be the one related to the company and not necessarily everything mentioned in the article. Another possibility would be focusing on other news factors, such as the topic, which could significantly impact the marke's reaction to the news.

These are some valuable conclusions that open several new questions and future lines of work in order to dive deeper into the different analyses. Some of them include:

- Improve the fine-tuning aspect of BERT for classification using more training data or a better model. The advantage of BERT being open source is the availability of already pre-trained models, so if this classification topic interests the community, different developers can collaborate to create the model. Another possibility to improve the classification is using a larger, better model such as GP4. However, this is not open source, and usage requires paying for API access.

- Different classifications of news might provide different insights. An example would be topic identification and classification, which can be a significant driver of market reaction to news.
- Testing different models or fine-tuning one to detect specifically the sentiment related only to the main company of the article, not every aspect of it. This could be done by retraining an advanced pre-trained sentiment model like FinBert with a manually labeled dataset of directly related and non-related paragraphs.
- A more precise and advanced market model could be used for event studies—for instance, the five-factor model.
- We could vary the event window to encapsulate potential delayed effects of the event on abnormal returns during event studies. Expanding the event window may capture more prolonged impacts of the event on financial returns.

Section 9. REFERENCES

Al-saqq, S., & Awajan, A. (2019). *The use of Word2Vec model in Sentiment Analysis: A Survey.*

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, 5.*

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Sastry, G. (2020). *Language Models are Few-Shot Learners* (Vol. 33). Curran Associates, Inc.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance, 52(1)*, 57-82.

Chan, W. C. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics, 70(2)*, 223–260.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.*

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance, 47(2)*, 427-465.

Fama, E. F., & French, K. R. (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies, 29(1)*, 69-103.

French, K. R. (2023). Retrieved from Darmouth Archive: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Habimana, O., Li, Y., Li, R., Gu, X., & Yu, G. (2020). Sentiment analysis using deep learning approaches: an overview. *Science China Information Sciences, 63*.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance, 48(1)*, 65-91.

Jivani, A. (2011). A Comparative Study of Stemming Algorithms. *Int. J. Comp. Tech. Appl, 2*, 1930-1938.

Johri, P., Khatri, S. K., Al-Taani, A., Sabharwal, M., Suvanov, S., & Chauhan, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. 365-375.

Khyani, D., & Siddharta, B. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology, 22*, 350-357.

Levene, H. (1960). Robust tests for equality of variances. (H. H. Ingram Olkin, Ed.) *Stanford University Press*, pp. 278-292.

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The journal of finance, 20(4)*, 587-615.

MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of economic literature, 35(1)*, 13-39.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The annals of mathematical statistics*, 50-60.

OpenAI. (2023). GPT-4 Technical Report. *ArXiv*.

Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global Vector for Word Representation* (Vol. 14). EMNLP.

Pham, H., Manzani, T., Liang, P., & Poczos, B. (2018). *Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis .*

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*, 130-137.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.*

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591-611.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance, 19(3)*, 425-442.

Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems, 4.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention Is All You Need.

Welch, B. L. (1947). The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika, 34(1-2),* , 28-35.

Willett, P. (2006). The Porter stemming algorithm: Then and now. *Program electronic library and information systems, 40.*

Yan, H., Deng, B., Li, X., & Qiu, X. (2019). *TENER: Adapting Transformer Encoder for Name Entity Recognition.*

Yergin, D. (1991). *The Prize : the Epic Quest for Oil, Money, and Power.* New York: Simon & Schuster.