LUISS

Master's Degree in Data Science & Management

Course of Data Visualization

Corporate Language as a Market Indicator: A Computational Study of Textual Data from S&P 500 Companies

Prof. Blerina Sinaimeri

SUPERVISOR

Prof. Paolo Spagnoletti

CO-SUPERVISOR

Francesco Tramotano
ID 758191

CANDIDATE

Academic Year 2023/2024

ACKNOWLEDGEMENTS

I do not know where to begin, but I have been fortunate to be surrounded by people who have enriched me during these long years of study. First and foremost, I would like to thank the dear Professor Blerina Sinaimeri for her kind guidance and availability, as well as my co-supervisor Paolo Spagnoletti and Professor Jonathan Berkovitch for their valuable insights into this thesis.

Behind me, I am lucky to have family and friends who have always supported me and encouraged me to do more. I dedicate this journey to my beloved grandparents, who cannot attend my thesis defence due to logistical and health reasons, and whom I love dearly...

ABSTRACT

Analysing the annual management disclosure reports of S&P 500 companies provides valuable insights into corporate communication strategies. Using the EDGAR database, we examined these reports to study sentiment and boilerplate language over the past two decades. The Central Index Key (CIK) numbers were extracted by web scraping, and the reports were downloaded using the EDGAR package in R. The data was then processed and analysed using the Loughran-McDonald sentiment dictionary.

The analysis included examining trends and correlations between word count and sentiment scores and detecting boilerplate language. Using the predictive models, we forecasted future trends in boilerplate language usage. The findings highlight the relationship between disclosure practices and market sentiment, offering practical insights and a general view that investors, analysts, and regulators can use. This study underscores the importance of transparency and the strategic use of language in corporate disclosures, demonstrating the potential of textual analysis in market prediction and corporate governance.

INDEX

1. Introduction

1.1. Background

Financial markets host a complicated environment and communication is an essential and crucial part of this environment. Depending on the case, communications can influence the activities of investors, analysts, or other stakeholders positively or negatively.

The language through which companies communicate is very varied, and numerous financial information channels are available from which you can access the communication texts freely provided. Corporate communications often 'hide' critical aspects of a company's financial status and action plans.

Among the several documents that are periodically released, central to this thesis project are the so-called corporate documents of the Management's Discussion and Analysis (MD&A) section, which simply mirror the company's status in the year preceding the date of release of the selected document.

The EDGAR system, which stands for *'Electronic Data Gathering Analysis and Retrieval'*, plays a vital role in collecting, analysing and archiving these documents through listed companies on behalf of the Securities and Exchange Commission (SEC) of the United States. “This system is an incredible source of information, including both annual (10-K) and quarterly (10-Q) reports, among other documents” *(Loughran, T., & McDonald, B., 2017)*. The extensive archive of financial documents offered by the EDGAR database is a valuable resource for those dealing with financial research or market analysis.

This system represents an incredible source of information, including both annual reports (10-K) and quarterly reports (10-Q), among other documents. The extensive archive of financial documents offered by the EDGAR database constitutes a valuable resource for those involved in financial research or market analysis.

We must also remember that the world is constantly developing and updating, as is communication and the language companies use. At the same time, analysis, particularly textual analysis, offers powerful means of analysing specific patterns or insights that characterise different aspects of this financial world.

Today, thanks to natural language processing (NLP) and machine learning advances, we can process virtually unlimited amounts of text, spotting patterns and trends in seconds. Leveraging these technological developments, this project offers an in-depth analysis of the MD&A sections of *S&P 500* companies, spanning a period of the last 20 years, from 2004 to 2023. By analysing the language used in these communications, the thesis aims to offer a perspective on the changes in corporate communication strategies and their possible impact on market perception. Sentiment analysis, standardised sentence detection, readability indices, and predictive modelling will be the subjects of this study to gain a broad view of corporate disclosure practices.

1.2. Objectives of the Study

A crucial part of this project is conducting sentiment analysis. Sentiment analysis involves implementing techniques that help quantify the positivity and negativity of language used within the MD&A sections of corporate disclosures *(Chan, S.W.K., & Chong, M.W.C., 2016)*. By doing sentiment analysis, we can observe changes in sentiment over time and assess the reasons for shifts in this sentiment. Interestingly, for the analysis of these texts, it is undoubtedly crucial to establish a quantitative measure of the emotional tone present in the text.

Another crucial component of this research centres on the examination of boilerplate text. Boilerplate refers to standardised and often repetitive text that appears in multiple filings. We attempt to quantify the prevalence of boilerplate in corporate disclosures and, more importantly, to consider the consequences of boilerplate for information quality. The prevalence of boilerplate language could lead to the conclusion that companies might want to mitigate the risks using

standardised language or are trying to ‘hide’ something. Thus, investigating the boilerplate's extent and impact on reports is another crucial component of this study.

Another important component of this analysis are the readability indices which are focal to understand how well the stakeholders are able to perceive and understand the information that is communicated. The indicators of readability, such as the Flesch Reading Ease, the Flesch-Kincaid Grade Level, and the Gunning-Fog Index, assess the comprehensibility and accessibility of such documents over time. This helps determine whether firms are increasing the accessibility of their reports to a broader audience or not.

Lastly, predictive models have instead the role of estimating the future trends of corporate disclosure practices. Using methods such as random forest, LSTM, and Prophet, we attempt to foresee how sentiment, boilerplate language, and readability will change in the following years. These models could give investors, analysts, and regulatory bodies valuable intuition to foreknow the transformed situation and make decisions based on these trends.

As we mentioned before, the information is regulated by the U.S. Securities and Exchange Commission (SEC) and continuously registered by listed companies through the EDGAR database, which is an electronic tool for data distribution, analysis and retrieval.

Over the past decade, the financial analysis sector has seen significant transformations thanks to the advent of new computing technologies. What was once an expensive process based on subjective interpretations has now become more accessible, and these technological advancements have made it possible to make decisions based on data analysed by different algorithms and specific methodologies.

Therefore, summarising the practical objectives of the project, the thesis explores in depth the MD&A sections of the S&P 500 companies, analysing how the language used in reports has changed over time and how this can influence market perception. The study dedicates itself, after providing a precise exploratory overview of the data and their trends over the years, in detail to aspects such as

sentiment analysis to examine the recognition of standardised sentences, the evaluation of readability and predictive modelling. The main aim is to offer a complete and detailed vision of corporate communication practices from a general perspective and then to delve into the company's details, showing how the texts released influence the company's image.

1.3. Research Questions

To achieve the objectives outlined above, this study seeks to answer the following research questions:

1. *How has the sentiment in the MD&A sections of S&P 500 companies evolved over the years?*

The analyses conducted on the tone of the language used in the reports can offer indications about the variations over the years and the trends of different companies. Therefore, the words chosen in the documents released help us better understand the economic and corporate context.

2. *What is the extent of boilerplate language in these disclosures, and how has it changed over time?*

Evaluating how many companies use boilerplate text and how much they use it can provide several interesting insights.

3. *How readable are the MD&A sections, and have there been significant changes in readability over the years?*

Verify if today's texts are more readable and less complex than in the past or if the trend has reversed to notice the transparency provided by companies.

4. *Can we predict future trends in boilerplate language?*

This point questions the possibility of obtaining future estimates based on values analysed in the past through machine learning methods.

1.4. Structure of the Thesis

This thesis is structured into several chapters, each covering different parts of the research. *Chapter 2* reviews the literature on management disclosure, textual analysis in accounting and finance, and sentiment research in finance.

We will discuss web scraping in Chapter 3, describing how the research data was collected and prepared for the study, including the data sources, data extraction methods, and data cleaning techniques.

Chapter 4 deals with the methodologies employed for sentiment analysis, boilerplate text analysis and detection, readability indices and predictive modelling. We will also discuss the implementation in Python from a slightly more technical perspective.

Visualizations and results are central in *Chapter 5*; by applying the methodologies described in the previous chapter, we will assess some insights from a general viewpoint and a more detailed one, selecting specific companies of interest.

Chapter 6 discusses different interpretations of the results, analysing the trends and the information deduced from the visualisations described in the previous chapter. Additionally, this chapter attempts to give a practical impression of the possible behaviour of investors and analysts based on the results obtained. Following this is a brief section on the limitations of our study.

We will then draw our conclusions, first providing a summary of the research conducted and then giving a final general consideration of the entire project.

Finally, the thesis concludes with an Index of Figures, tables, and References, which includes all the sources consulted for its writing.

2. Literature Review

In the realm of financial reporting, the following literature will provide, from a general perspective, an overview of the numerous studies already conducted on topics related to those addressed in this thesis, namely Management Disclosure, Textual Analysis, and Sentiment Analysis. We will analyse the results and insights that have emerged from these studies, highlighting their strengths, and we will continue these studies in depth with our subsequent analyses.

2.1. Management Disclosure in Financial Reporting

The crucial element of financial reporting is, therefore, Management Disclosure, which enables companies to communicate financial and non-financial information to their stakeholders or other companies and markets *(Li et al., 2010).* Management Disclosure is characterised by multiple nuances and sometimes provides details of different kinds that can be very important. These details may relate to the company's future outlook, managerial perspectives on financial performance, potential upcoming events, or specific strategies.

Research indicates that Management Disclosure promotes transparency, builds stakeholder trust, and can be critical in positively or negatively influencing stock prices in the market *(Huang et al., 2016).* Given this particular aspect, it is clear that this is a highly sensitive form of communication, capable of shifting financial balances, and whose language must be used with absolute precision.

Studies and research mentioned in the final references section also suggest that effective communication is associated with reduced information asymmetry and lower capital costs, thereby implying greater overall market efficiency *(Tetlock, 2007).* Given the significant power of these communication channels, there is a need for regulatory oversight to ensure that these channels are used efficiently. These regulations are overseen by the Securities and Exchange Commission (SEC),

which establishes general guidelines to promote consistency, reliability, and accountability in financial communications.

2.2. Overview of Textual Analysis in Financial Studies

Over the years, textual analysis has become increasingly prevalent in financial studies due to its ability to quickly analyse a substantial amount of data, extracting relevant and valuable information. The methods by which researchers have studied financial communication have also evolved over time with the advent of Natural Language Processing (NLP) techniques and the digitalisation that has impacted every sector, with data being employed in various contexts and for numerous uses *(Kearney & Liu, 2014)*. Through NLP, researchers have been able to extract information from corporate communications, including earnings announcements, press releases, and regulatory filings, with high scalability and depth.

These studies highlight how stakeholders' perceptions can be altered and significantly influenced by the details found in texts, which can be observed through analyses of sentiment, text complexity, frequently used keywords, and other research parameters *(Feldman, 2013)*. Correlations have indeed been found between the sentiment expressed in corporate texts and subsequent shifts in stock prices, emphasising how language can play a critical role in understanding and predicting stock behaviour.

In conjunction with textual analyses, new research is emerging that includes, as in this thesis study, machine learning algorithms to understand what patterns certain stock trends might follow based on accumulated past and present data, thereby paving the way for more accurate and reliable predictions *(Yu, X., 2014)*.

2.3. Sentiment Analysis and Its Importance in Finance

The emotional tone that characterises texts is captured through sentiment analysis, a specialised form of textual analysis that has become increasingly popular over the years because it suggests to stakeholders the direction in which certain companies and markets are heading, allowing them to identify potential opportunities *(Loughran & McDonald, 2016).*

The research conducted by Loughran, M., and McDonald, B. enabled them to construct a specific sentiment dictionary to evaluate the positivity or negativity of financial texts. The dictionary, which also bears their name, contains an extensive range of finance-specific terms with corresponding sentiment scores, which are then used to calculate overall sentiment. These calculations serve as leading indicators of stock performance.

The dictionary, also adopted in this study, represents a significant advancement in financial textual analysis and, when combined with other analytical techniques, can lead to more detailed analyses involving time series to examine changes in emotional tone over the years within the market and specific companies, likely due to influential events or circumstances.

3. Data Collection and Preparation

Delving into the practical details of the thesis project, we will closely examine the entire process covering data extraction and collection and its manipulation for our research purposes. Firstly, it is necessary to obtain a precise index of the S&P 500 Companies that includes all the information we are looking for, namely the company name, central index key, and also the company ticker, which is the corporate abbreviation. The next step involves extracting the documentation for the desired companies through EDGAR modules and archives, which provide us with detailed textual information at regular intervals for each CIK included in its functions.

3.1. Data Sources

3.1.1 S&P 500 Companies

Although it is a very popular and globally recognised index, obtaining information about the companies that are part of it is relatively simple; however, obtaining detailed documentation of the list of companies and their respective CIKs is not a trivial challenge. Moreover, it must be taken into account that it is an index in constant flux. The S&P 500 index is based on strict inclusion criteria that must be met, and there have been numerous removals and additions of companies over the years because it is a dynamic index that reflects market developments and the companies that comprise it. There are various sources to obtain this updated list, but no source offers a precise downloadable table (at least for free). Therefore, thanks to the kind help of my supervisor, whom I sincerely thank, we will perform a web scraping of a website that offers a decent visualisation of the data we desire.

3.1.2 EDGAR System

Analysing the EDGAR system from a more technical perspective, unlike the general definitions already given previously, it is globally used as a database to allow stakeholders to access documents for financial analysis and market research, providing interested parties with information such as annual reports (10-K), quarterly reports (10-Q), and communications on significant events (8-K). Moreover, the EDGAR system provides a package usable on Rstudio with specific functions for retrieving and extracting data based on the CIK input. Among the various functions, you can retrieve documents with the 'getFilingsHTML()' function, extract a specific section with the 'extract()' function, search for keywords with the 'searchFilings()' function, or perform Sentiment analysis (with pre-set parameters of the package) with the 'getSentiment()' function.

3.2. Data Extraction Methods

3.2.1 Web Scraping S&P 500 CIKs using Beautiful Soup library

As already defined, the first objective is to extract a valid table for exporting the companies in the S&P500 index. This table will serve as a guide to obtain the CIKs to be used later in the document extraction functions in RStudio.
To achieve this extraction, we will perform web scraping, using Python code, on the website *'https://primeaim.wordpress.com/2014/09/24/central-index-keys-cik-for-dow-30-and-sp-500/'.*

In the code, we mainly use three libraries: the *pandas* library, globally used mainly for building and manipulating DataFrame, and the *requests* and *BeautifulSoup* libraries. The *requests* library is well-known because it allows you to make HTTP requests simply and intuitively, thus sending GET requests to our website of interest, managing sessions, and handling any cookies. For HTML parsing, *BeautifulSoup* does a job that allows you to navigate within an HTML document by accessing it as a tree of objects, simplifying access to information.
In short, the code starts with the URL definition, to which we send an HTTP GET

request. We then parse the HTML content, proceed with data extraction from the table, and finally create the data frame and save it as ‘sp500_scraped.csv’.

Table 1. Example output of ‘sp500_scraped.csv’

Index	Ticker	CIK	Company name	Exchange
1	A	1090872	Agilent Technologies Inc.	Reports
2	AA	4281	Alcoa Inc.	Reports
3	AAPL	320193	Apple Inc.	Reports
4	ABBV	1551152	AbbVie Inc.	Reports
5	ABC	1140859	AmerisourceBergen Corporation	Reports
6	ABT	1800	Abbott Laboratories	Reports
7	ACE	896159	ACE Limited	Reports
8	ACN	1467373	Accenture plc	Reports
9	ACT	1578845	Actavis plc	Reports
10	ADBE	796343	Adobe Systems Inc.	Reports

3.2.2 File Extraction Using RStudio

Now that we have obtained the information regarding the CIKs, the second step involves extracting the files through the Rstudio code and the EDGAR package already mentioned earlier. With a brief and intuitive R code, we can download the 'Management’s Discussion and Analysis' sections for a sample of 100 companies (out of a total of 500) covering a 20-year interval from 2004 to 2023, thus considering the last two decades of reporting. This extraction provides us with a multitude of ‘.txt’ files.

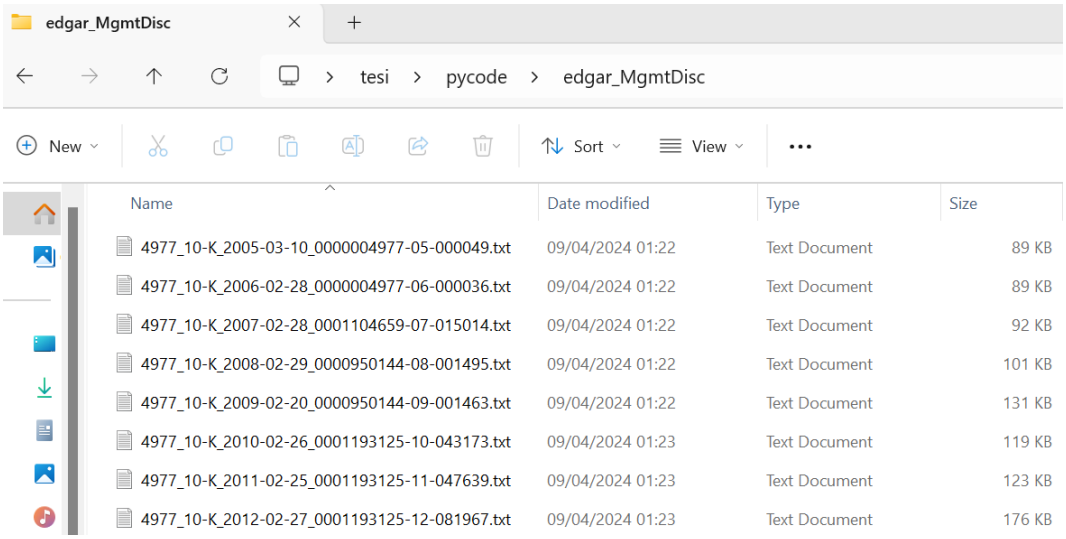

Figure 1. Example output of the RStudio extraction process.

3.3. Data Cleaning and Preparation

In this section, we analyse the entire data cleaning process and the subsequent manipulation to conduct our analysis. We load the folder containing all the previously downloaded documents and begin defining all the necessary functions for the textual analysis.

3.3.1 Text Cleaning and Tokenization

Text cleaning can be done in various ways; we first clean the text with the 'clean_text()' function, removing non-alphabetic characters from the texts to ensure that the text used for analysis is free of noise or unwanted formats. Subsequently, we tokenise the text with the *ntlk* library, dividing the text into words or 'tokens'. This step is crucial for performing the lemmatisation process, where we map the parts of speech of words into base forms. For example, if we have the words 'analysed' and 'analysing', they will both be reduced to the base form 'analyse'. The following sentiment analysis will require this data dimensionality reduction to improve its accuracy.

3.3.2 Preparing the Loughran-McDonald Sentiment Dictionary

The *Loughran-McDonald Sentiment Dictionary* is an essential tool for sentiment analysis in the financial context and is used on a global scale. It was possible to download it via its official link in '.csv' format, and it contains words in the English language that are classified as positive and negative, with a value assigned to them based on the intensity of the word itself. Therefore, we divide the words into *'positive_words'* and *'negative_words,'* with which we analyse the text of the documents and calculate a sentiment score based on the presence of words within the dictionary. To build the final dataset, we thus need to determine the overall tone of the text to extract valuable information about corporate communications.

3.4. Description of the Dataset

To start our analysis, we, therefore, need a clean and complete dataset that contains all the variables and values we require. Our dataset will be used as a *Pandas DataFrame* within the Python code and consists of 1676 occurrences, which correspond to 1676 distinct documents that refer to a specific company in a specific year. The total number of companies is 100, and, as mentioned earlier, they cover the two decades from 2004 to 2023. Through data manipulation, we obtained additional variables within our dataset, including the analysis of keywords or readability scores, to ascertain whether there has been any change in lexical complexity over time. The dataset is thus structured with the following columns:

- *'company'*: name of the company
- *'cik'*: central index key, a specific identifier for the company
- *'date'*: day, month, and year when the document is released
- *'word count'*: total number of words contained in the document
- *'sentiment score'*: numerical score relating to the overall positivity or negativity of the sentiment in the text

- *'sentiment category'*: sentiment category, which can have five different values based on the overall score of the document
- *'positive words'*: dictionary listing the most relevant positive words within the document
- *'negative words'*: dictionary listing the most relevant negative words within the document
- *'key sentiment words'*: words that most influenced the sentiment score of the document
- *'boilerplate percentage'*: percentage of boilerplate text about the total text within a single document
- *'sentence count'*: total number of sentences within a document
- *'max sentence length'*: number of words in the longest sentence of a document
- *'Flesch Reading Ease'*: an index that evaluates the readability of the text; as the value increases, the ease of reading and understanding the text increases
- *'Flesch-Kincaid Grade Level'*: another readability indicator that estimates the school grade level necessary to understand the text
- *'Gunning-Fog Index'*: this readability indicator instead considers the length of words and sentences to categorise the difficulty of the text
- *'Year'*: the year of publication of the document
- *'Company Presence by Year'*: number of companies out of the total of 100 analysed companies for which we collected reports in a specific year
- *'Total Reports by Company'*: total number of documents released by a single company over the last two decades.

Table 2. Example output of the Dataframe of the study

Company	CIK	Date	Word Count	Sentiment Score	Sentiment Category	Positive Words	Negative Words	Key Sentiment Words	Boilerplate Percentage
NETAPP INC	1002047	29/06/2004	10591	-0.147013783	Negative	{'great': 1, 'enabl	{'defer': 13, 'unpaid	[('restructuring', 49), ('lc	0
NETAPP INC	1002047	07/08/2005	12147	-0.117395945	Negative	{'enable': 1, 'favo	{'disqualify': 1, 'defe	[('restructuring', 42), ('lc	0
NETAPP INC	1002047	07/12/2006	14114	-0.135009311	Negative	{'enable': 1, 'ben	{'correction': 2, 'dis	[('restructuring', 34), ('lc	0
NETAPP INC	1002047	26/06/2007	13610	-0.140503876	Negative	{'great': 3, 'enabl	{'defer': 15, 'forfeitu	[('restructuring', 38), ('lc	0
NETAPP INC	1002047	24/06/2008	15746	-0.153732758	Negative	{'great': 6, 'enabl	{'defer': 17, 'investi	[('restructuring', 33), ('lc	0
NETAPP INC	1002047	17/06/2009	15212	-0.247708424	Negative	{'great': 4, 'enabl	{'curtail': 1, 'defer':	[('loss', 38), ('restructuri	0
NETAPP INC	1002047	18/06/2010	11668	-0.224529891	Negative	{'great': 3, 'enabl	{'curtail': 1, 'defer':	[('loss', 28), ('impairmer	0
NETAPP INC	1002047	23/06/2011	14002	-0.172252168	Negative	{'great': 2, 'enabl	{'curtail': 1, 'defer':	[('loss', 27), ('benefit', 2	0
NETAPP INC	1002047	19/06/2012	12196	-0.110269696	Negative	{'great': 1, 'enabl	{'curtail': 1, 'defer':	[('benefit', 22), ('impairn	0.262915736
NETAPP INC	1002047	17/06/2013	10617	-0.098802395	Slightly Negative	{'enable': 1, 'favo	{'curtail': 1, 'defer':	[('defer', 17), ('benefit', 1	0
NETAPP INC	1002047	17/06/2014	10796	-0.087937857	Slightly Negative	{'favorably': 7, 'er	{'curtail': 1, 'defer':	[('effective', 18), ('defer'	0
NETAPP INC	1002047	21/06/2021	10181	-0.074015748	Slightly Negative	{'great': 1, 'enabl	{'curtail': 2, 'defer':	[('benefit', 13), ('gain', 1	0.319004705
YAHOO INC	1011006	27/02/2004	28940	-0.190408253	Negative	{'great': 3, 'enabl	{'prejudice': 1, 'defe	[('loss', 138), ('impairme	0.769908752
YAHOO INC	1011006	03/11/2005	28862	-0.122503776	Negative	{'great': 7, 'enabl	{'weaken': 2, 'preju	[('loss', 81), ('defer', 43),	0
YAHOO INC	1011006	03/03/2006	10203	-0.008159269	Neutral	{'great': 3, 'enabl	{'correction': 1, 'del	[('impairment', 18), ('los	0
YAHOO INC	1011006	23/02/2007	10656	-0.026194145	Neutral	{'great': 3, 'enabl	{'defer': 7, 'forfeitur	[('impairment', 17), ('ber	0.325708787
YAHOO INC	1011006	27/02/2008	12786	0.011506009	Neutral	{'great': 3, 'enabl	{'defer': 7, 'forfeitur	[('benefit', 34), ('loss', 19	0.659099231
YAHOO INC	1011006	27/02/2009	12409	-0.109775222	Negative	{'great': 2, 'enabl	{'nonfunctional': 1,	[('impairment', 38), ('res	0.83744766
YAHOO INC	1011006	26/02/2010	13407	-0.13796442	Negative	{'great': 2, 'enabl	{'nonfunctional': 2,	[('restructuring', 50), ('in	0.898928201
YAHOO INC	1011006	28/02/2011	14138	-0.142394073	Negative	{'great': 2, 'enabl	{'nonfunctional': 2,	[('restructuring', 54), ('in	0.876122879
YAHOO INC	1011006	29/02/2012	13510	-0.141629343	Negative	{'great': 1, 'innov	{'nonfunctional': 1,	[('restructuring', 56), ('d	0.951772967

Sentence Count	Max Sentence Length	Flesch Reading Ease	Flesch-Kincaid Grade Level	Gunning-Fog Index	Year	Company Presence by Year	Total Reports by Company
445	257	37.1	12.4	8.92	2004	67/100	12/20
473	278	36.79	12.5	8.89	2005	74/100	12/20
559	275	36.79	12.5	8.86	2006	84/100	12/20
563	220	46.78	10.7	8.21	2007	83/100	12/20
670	260	46.47	10.8	8.21	2008	83/100	12/20
549	273	43.63	11.9	9.41	2009	87/100	12/20
393	279	43.53	12	9.59	2010	89/100	12/20
484	280	43.43	12	9.5	2011	92/100	12/20
427	260	43.73	11.9	9.45	2012	94/100	12/20
357	189	43.93	11.8	9.48	2013	94/100	12/20
358	207	42.82	12.2	9.87	2014	97/100	12/20
324	247	32.57	16.2	14.04	2021	78/100	12/20
1284	361	32.94	14	9.5	2004	67/100	14/20
1218	383	39.47	13.5	10.31	2005	74/100	14/20
453	147	29.99	15.1	11.49	2006	84/100	14/20
460	234	30.4	14.9	11.34	2007	83/100	14/20
552	206	39.06	13.7	11.03	2008	83/100	14/20
525	240	38.66	13.8	11.22	2009	87/100	14/20
553	271	38.15	14	11.33	2010	89/100	14/20
567	286	36.22	14.8	12.02	2011	92/100	14/20
547	273	36.22	14.8	12.06	2012	94/100	14/20

4. Methodology

Central to this research is sentiment analysis, which evaluates the emotional tone expressed in the documents of a sample of companies that are part of the S&P500 index, specifically within the *'Management's Discussion and Analysis'* (MD&A) sections. This study, therefore, offers an objective measure of the tone of corporate communication based on the quantification and evaluation of positive or negative expressions found in the text.

4.1. Sentiment Analysis

4.1.1 Calculation of Sentiment Score

The sentiment score is calculated through a complex process involving identifying and counting words considered positive or negative according to the values in the *Loughran-McDonald* dictionary. The final sentiment score is obtained through the use of a normalised formula that balances positive and negative words, taking into account the difference between the frequencies of positive and negative words relative to the total number of sentiment words.

4.1.2 Categorizing Sentiment

After obtaining the sentiment score, we chose to categorise each document into five different sentiment categories based on the sentiment score, as follows:

- **Positive**: score greater than 0.1
- **Slightly Positive**: score between 0.05 and 0.1
- **Neutral**: score between -0.05 and 0.05
- **Slightly Negative**: score between -0.1 and -0.05
- **Negative**: score lower than -0.1

These thresholds are designed primarily to simplify the analysis of results and provide an intuitive interpretation of the predominant tone in the documents.

4.1.3 Frequently Used Words

This section is the main focus for the subsequent analysis concerning word clouds for each company. It aims to identify the most frequently used words and their frequencies. This can give us further insights into whether, for example, a company overuses certain words over the years or tends to change vocabulary more dynamically.

4.2. Boilerplate Text Analysis

4.2.1 Definition and Identification of Boilerplate Text

By *boilerplate text*, we refer to portions of text that are standardised and repeated within one or more documents. It is text with little significance since it is repeated and used to fill apparent gaps or to omit and mask relevant information. In this corporate context, boilerplate text is usually related to legal information, procedural descriptions, or disclaimers that have relatively little impact in terms of market utility and influence. Therefore, identifying and even trying to quantify this boilerplate text can be an advantageous tool for potential investors or stakeholders.

Identifying it is not trivial, as there is a risk of easily misidentifying important text as boilerplate. Our approach is based on the frequency of phrases and n-grams repeated in the reports.

4.2.2 Categorizing Sentiment

As mentioned earlier, the *nltk* library, or *Natural Language Toolkit*, does an excellent job of extracting *n-grams* (groups of *n* consecutive words) from documents to identify and analyse the frequency of standardised words. Subsequently, the frequency of these repetitions is analysed, and a minimum threshold is established—set at 1% of the text—to identify such text as boilerplate.

4.2.3 Implementation in Python

Without delving into overly technical details, text cleaning was performed earlier, followed by our defined function *'find_repeated_phrases'*, which uses *n-grams* to find word sequences and how many times the same sequence appears within the document.

Then, through the subsequent function *'calculate_boilerplate_percentage'*, we can define a percentage value of boilerplate language using the minimum threshold of 1% described in the previous chapter.

These functions are executed in a *for* loop that processes each document individually and stores the data within the DataFrame that contains all the textual analysis metrics of interest.

4.3. Readability Analysis

4.3.1 Readability Indices

The textual analysis we are conducting also raises the question of the complexity of the text used in reports to understand its accessibility from the perspective of reader comprehension and clarity. Therefore, we have explored corporate documents' readability using these three well-known readability indices, which perform similar work but differ from each other in rather interesting nuances.

The indices *(Lahmar, O., & Piras, L., 2023)* we analyse are:

- **Flesch-Kincaid Grade Level**: Provides a U.S. school grade level; this tells us an estimate of the grade level someone needs to be at to understand the text.
- **Flesch Reading Ease**: Rates text on a 100-point scale; the higher the score, the easier it is to understand the text.

- **Gunning-Fog Index**: Estimates the years of formal education a person needs to understand the text on the first reading.

4.3.2. Calculation Methods

These indices are calculated following predefined formulas influenced by the total number of words, sentences, and syllables within a document. The formulas are as follows:

- Flesch Reading Ease is calculated as:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

- Flesch-Kincaid Grade Level is calculated as:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

- Gunning-Fog Index is calculated using:

$$0.4\left[\left(\frac{\text{total words}}{\text{total sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{total words}}\right)\right]$$

4.3.3. Implementation in Python

The indices can be calculated using the *textstat* library, which provides the tools to perform various readability tests. Using these functions, we can automatically obtain the number of syllables, complex words, and other metrics necessary for the indices.

In the following sections, we will visualise the different readability thresholds of the indices and define which S&P500 companies write in a more or less complex manner and how easy their text structure and comprehension are.

4.4. Statistical and Predictive Analysis

4.4.1 Correlation Analysis

Among the various analyses we conducted, we also delved into the relationship between variables, specifically *Word Count* and *Sentiment Score*. To verify this correlation, we used Pearson's correlation coefficient, a robust statistical measure that has existed since the late 19th century and is still widely used by analysts for its intuitive interpretation. This correlation results in a value ranging from -1 to +1. A value close to -1 indicates a negative correlation, meaning that as one variable increases, the other decreases proportionally. The opposite occurs when the values are close to +1, while if the values settle near 0, no significant linear correlation can be observed, and thus, the two variables do not show any direct linear relationship.

4.4.2 Time Series Analysis

Time series are central to the project as they offer the most explicit possible representation of a value's trend over time. Noticing the temporal evolution of the data collected from the MD&A reports helps us identify significant periodic fluctuations that characterise different variables in certain periods of the year or specific years. Time series is crucial in observing potential shifts, particularly from 2020 due to the COVID-19 pandemic. Additionally, we will add graphical details to the plots that will provide more information about each year's corporate reporting relationship and other important details.

4.4.3 Predictive Models (Random Forest, LSTM, Prophet)

In the final part of the project and code, we attempt to generate predictions about the *boilerplate language* value in the coming years. This estimate will use various sophisticated predictive methods based on past data. We will use the following:

- **Random Forest**: This is an ensemble predictive model that operates through a multitude of decision trees during training. This series of trees

tends to reduce the variance of the model while keeping the bias as low as possible. It is usually used when a robust method is needed to avoid overfitting, especially when the data is complex to process, however, it also requires good computing power and considerable execution time to make the prediction.

- **LSTM (Long Short-Term Memory)**: LSTM refers to a model based on recurrent neural networks that perform training through data sequences. It is highly recommended in the field of temporal data prediction because it effectively calculates the prediction while avoiding the frequent problem of gradient vanishing, often present in many other cases where simpler recurrent architectures are used.
- **Prophet**: This model is one of the most modern, as it was introduced and designed by Facebook. It focuses on decomposed seasonal components and trends. It is particularly recommended for variables that exhibit strong seasonality and with numerous accumulated past data to improve the prediction.

From a general perspective, predictive modelling typically involves three macro steps: data preparation, model training, and model validation.

The performance of the models is verified using cross-validation methods and standard error measurement, examples of which are MSE and MAE, respectively, Mean Squared Error and Mean Absolute Error. In machine learning, models are not generally superior to others, but we will see which of the three models best fits our data and gives us the best results and estimates.

5. Visualizations & Results

5.1. Exploratory Data Analysis (EDA)

In this chapter, we finally delve into the first results of this analysis and the first concrete findings of everything we have carefully prepared and explained. The interpretation of the data is critical to obtaining information about the many metrics we will analyse, using the most straightforward and intuitive method—graphs and visualisations that give us an interesting perspective to evaluate and judge the results. Through EDA, we aim to define and understand the relationships between variables while identifying the dynamics of variables like *word count* and *sentiment score*, which characterise corporate communication over time. As already mentioned, the dataset contains a total of 100 companies. However, there will also be more in-depth analyses of smaller groups of companies that perhaps have a similar or competitive market.

In the following sections, we will analyse all the aspects we have tried to investigate, with results supported by numerous consistent graphs and tables built in Python, culminating in interpreting the resulting data or possible implications or hypotheses.

5.1.1 Correlation Analysis Between Word Count and Sentiment Score

As usual in data analysis, we initially try to understand whether there is a correlation between the main variables, in this case, *word_count* and *sentiment_score*, and what type of correlation it is.

The goal of this correlation analysis is to understand whether there is a relationship between the amount of text used by companies and the emotional tone employed—essentially, whether more words in a text correspond to a text with a higher degree of positivity or the opposite.

The following graph in *Figure 2* clearly shows the negative average correlation, reporting a value of -0.31. This value is not very explanatory in itself, but it can provide us with insight, as it indicates the presence of a negative correlation of moderate intensity.

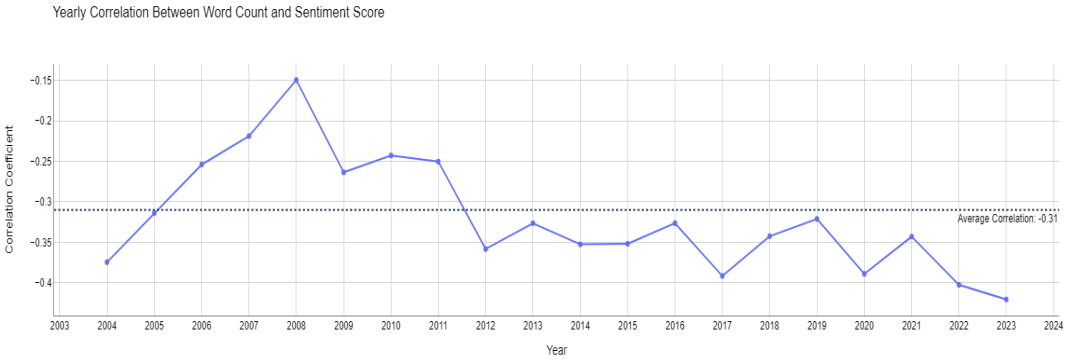

Figure 2. Yearly correlation between Word Count & Sentiment Score

In essence, on average, the more words a report contains, the more negative the emotional tone tends to be. If we divide the two decades, we can also see that there is more neutrality between the two variables in the first decade analysed, but this changes in the second decade, and then the value remains relatively stable.

5.1.2 Word Count Analysis

5.1.2.1 Word Count Distribution and Trend Over Time

Analysing the distribution of *word_count* is one of the primary and essential tasks for subsequent analyses. The following two graphs show the distribution of values and the trend of total words over time.

These observations can reveal shifts or trends that may reflect changes in the way companies tend to communicate, which could be influenced by external factors such as regulatory, cultural, or technological ones.

In *Figure 3*, we notice how the distribution is very similar to a normal distribution with a long right tail. The graph shows how most documents have a total word count slightly above 10,000, as we can see from the peak within the barplot.

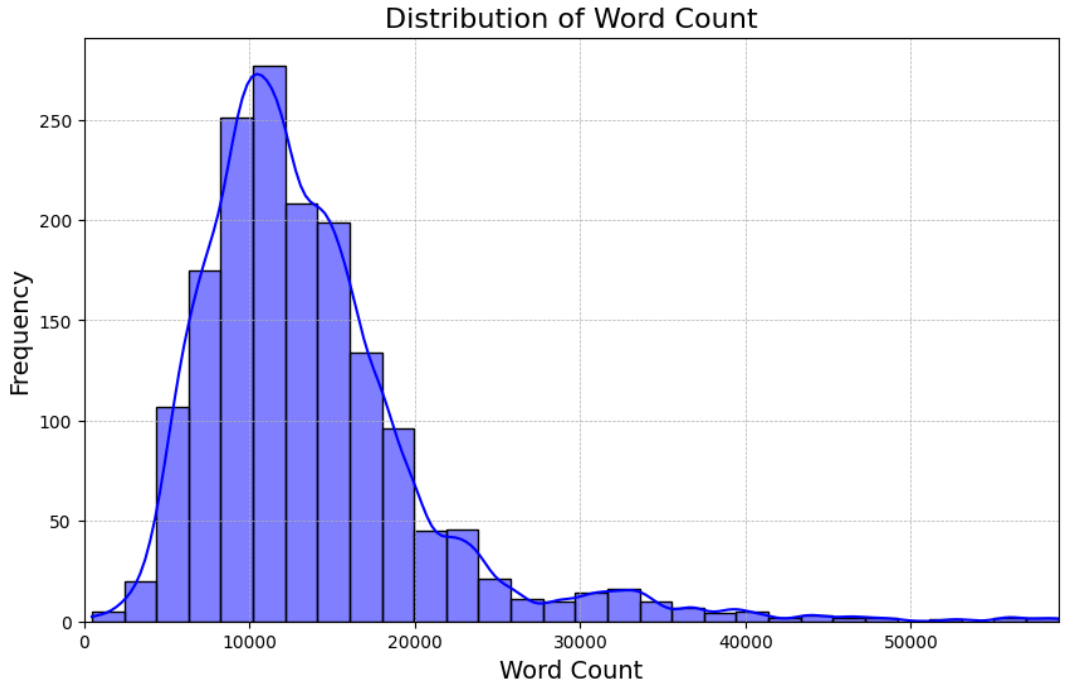

Figure 3. Word Count distribution

In the following graph in *Figure 4*, illustrating the value over time, we first have evidence that the average value is around 13,000 total words per single document, but we also get a look at the evolution of the value over time. As we can see in the line plot, the annual trend of the average word count in the reports is shown, with a 95% confidence interval and the relative margin of error expressed with the red error bar, which symbolises the *SEM* (Standard Error of the Mean) used to represent the uncertainty associated with the sample mean estimate.

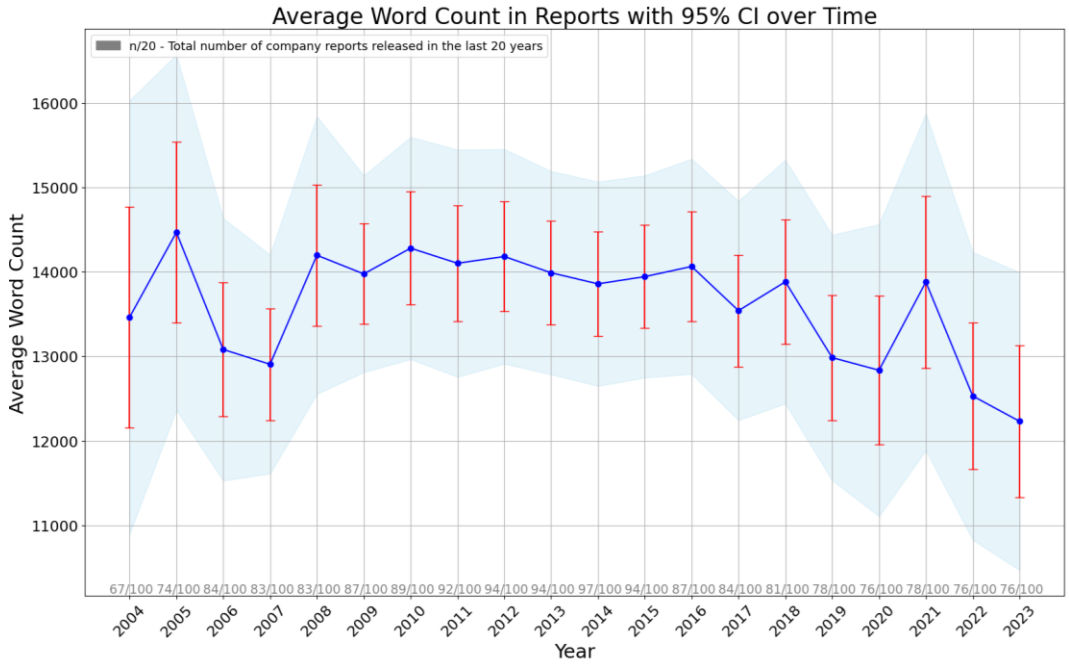

Figure 4. Yearly Distribution of Average Word Count

The graph shows that in the decade covering the years from 2008 to 2018, the value seems relatively stable, then it reaches a small peak in 2021 and then drops sharply in the last two years. This may be due to numerous factors that we will discuss further in subsequent analyses involving other variables.

5.1.2.2 Top Companies by Average Word Count

This section aims to address the questions regarding which of the 100 companies in our sample tend to write more and which companies tend to write less. To answer these questions, two different types of graphs were developed: the first type includes the top companies by the total number of words written, while the second type of graph highlights which companies write more or less in a specific year. Digging into the results, we have the following visualisations to determine the companies that are more or less verbose in their reports:

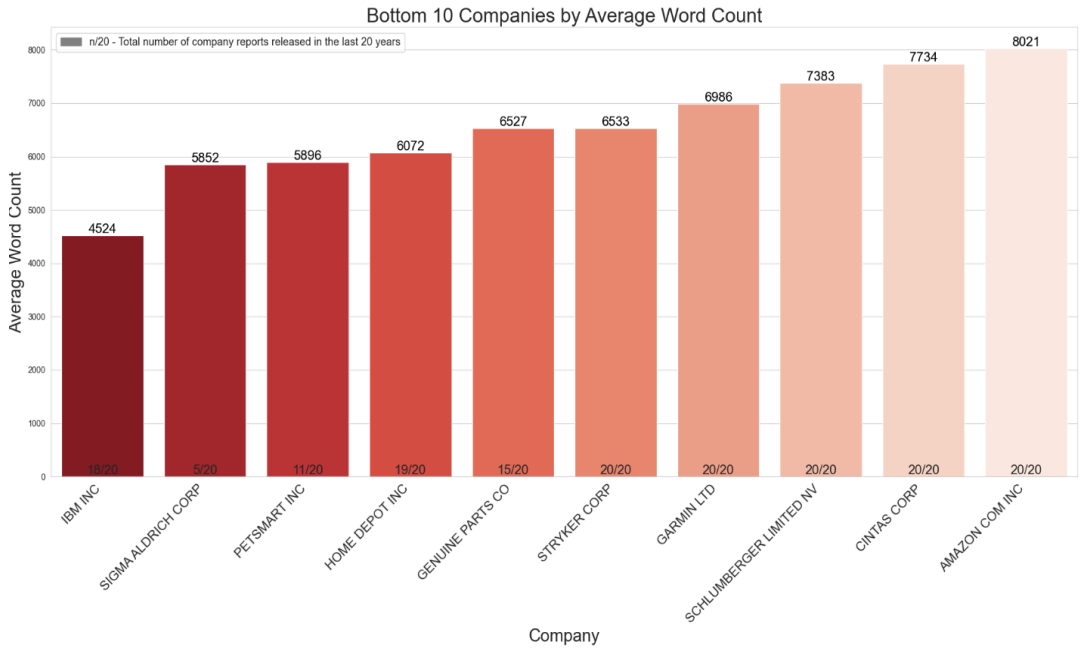

Figure 5. Companies with the lowest values of Avg. Word Count

We start by studying the companies that can be defined as more 'concise,' such as the well-known IBM, which stands out in *Figure 5*. We indeed notice a marked difference in word count between these companies and the total average word count, which is slightly above 10,000.

These companies opt for a more concise approach to communicating their financial information. Among these companies, it is also interesting to note giants like Amazon that are part of this interesting ranking, and a note should be made for SGMA-Aldrich Corp, which records a relatively low number of documents. However, as we can see at the base of its bar, it has only released five documents in the last 20 years.

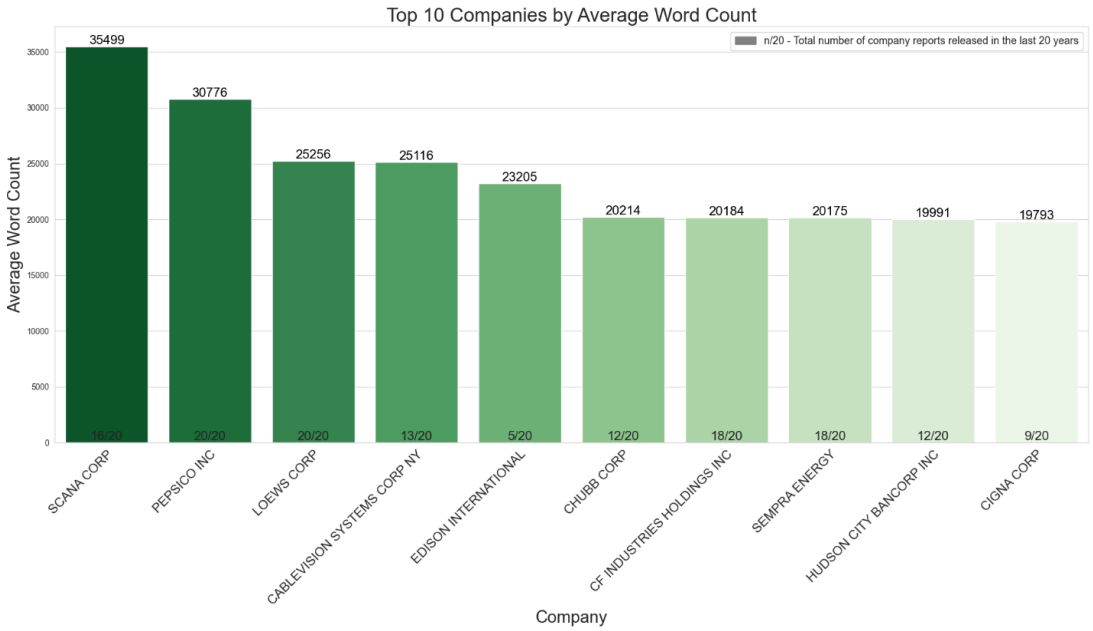

Figure 6. Top 10 Companies by Word Count

On the other hand, SCANA Corp is recorded as the most verbose in its documents, followed by major companies such as PepsiCo Inc. and Loews Corp, with values of around 30,000 words per report. According to the previously analysed correlation coefficient, which is -0.31, these companies should generally use a more negative tone in their reports. However, in the following sections, we will see if this correlation holds true to the reported values.
As mentioned earlier, an analysis was also conducted for each year with an interactive graph that allows selecting a specific year to understand which companies wrote more or less in that particular year. Analysing the year 2021, represented in *Figure 7*, we notice some truly interesting information.

The year 2021 follows 2020, the year that unfortunately marked history across all sectors due to the COVID-19 Pandemic and massively impacted financial history. From this barplot, we can see, in addition to the top 5 and bottom five companies by total word count, the percentage change in the total words compared to the previous year.

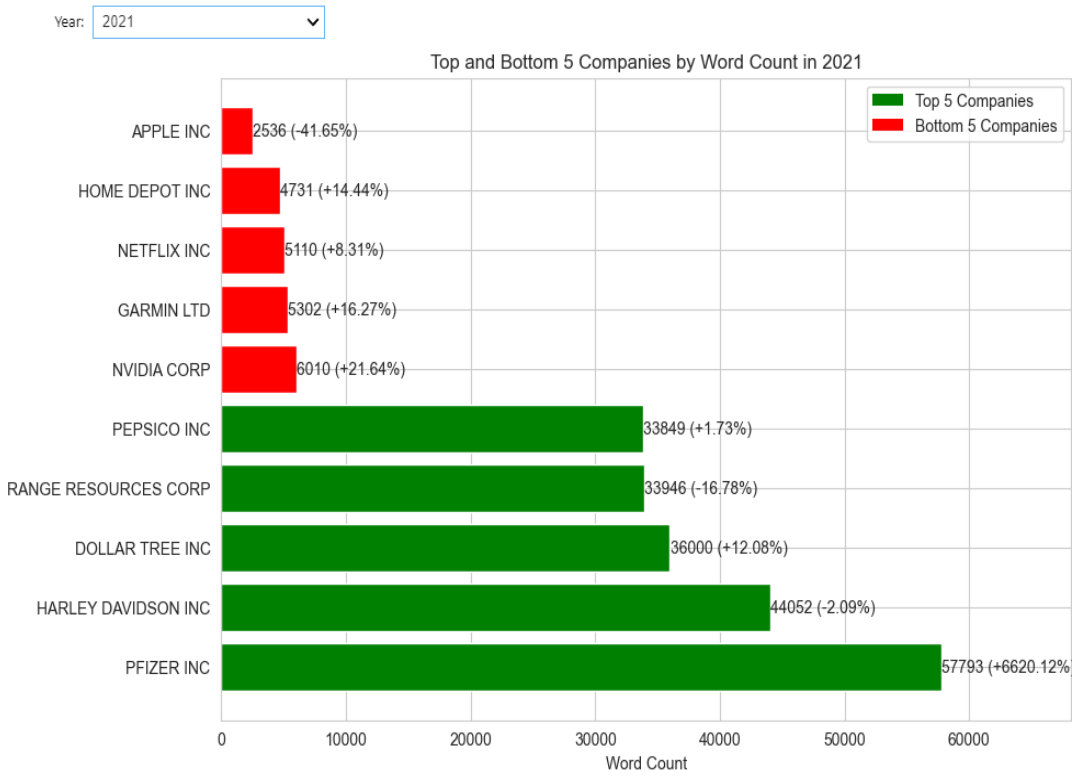

Figure 7. Top and Bottom 5 Companies by Word Count in 2021

Unfortunately, we know well why Pfizer Inc. is the company that wrote the most, with a staggering percentage change exceeding 6000% compared to the previous year; at the same time, we see tech companies like Apple, Netflix, or Nvidia at the bottom of the list, which were obviously of secondary priority in the early stages of the global pandemic.

5.1.2.3 Word Count Over Time by Tech Companies

After analysing the entire sample of 100 companies, we delved deeper by examining a smaller number of companies related to the tech sector.

In this sub-sample, we find a total of 10 companies listed in alphabetical order as follows:

- Adobe Inc.
- Amazon Com Inc.
- Apple Inc.

- Ebay Inc.
- Electronic Arts Inc.
- Intel Corp.
- Meta Platforms Inc.
- Netflix Inc.
- Nvidia Corp.
- Oracle Corp.

The following graph, Figure 8, shows the time trends in the word count of these tech companies' annual reports.

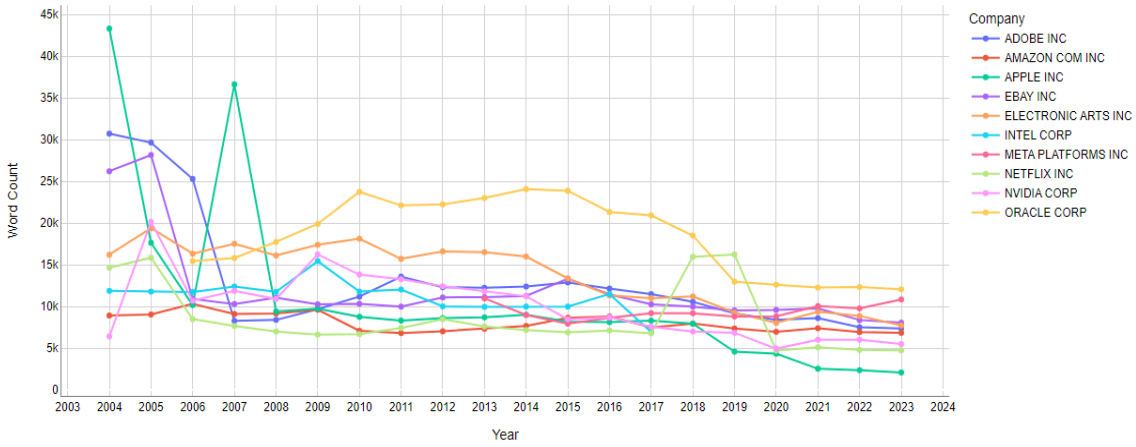

Figure 8. Word Count Time Series for Tech Companies

It is interesting to note that since 2008, excluding Oracle, all tech companies have basically followed the same trend. Oracle tends to write consistently a few thousand more words over time, while until 2008, the year of the famous 2008 financial crisis induced primarily by the American subprime mortgage crisis, Apple, Adobe, and eBay wrote significantly more than their competitors. These trends can give us insights into which companies tend to be less susceptible to sector volatility, following a more consistent and moderate approach over time, aligning with a business model that we can define as more stable. Additionally, it is important to consider how these years have seen a rapid technological evolution that has given rise to new fields that have marked the markets, such as cloud computing, artificial intelligence, and cybersecurity, with the associated regulations that may have contributed to more detailed disclosures in the annual reports.

However, the number of words and, thus, the length of the reports is not definitively explanatory, and to understand more about these shifts, it is necessary to analyse more deeply to generate concrete hypotheses and explanations.

5.1.3 Sentiment Score Analysis

5.1.3.1 Sentiment Score Distribution and Trend Over Time

To delve deeper into the reasons or factors that influenced certain changes over the years, it is also necessary to explore the *sentiment_score* variable in detail, similarly to what was done with *word_count*. Since the visualisations are similar to those used for *word_count*, it is interesting to conduct a comparative analysis with the previous results.

From *Figure 9*, it is clear that the histogram distribution, in this case as well, follows a normal distribution, but in this case, the long tail does not extend to the right but to the left. This is an important finding as it suggests that there is an intrinsic caution in the language used in corporate communications, which could be driven by the desire to mitigate negative market reactions or avoid overly optimistic comments that could imply a lack of confidence if expected results are not met.

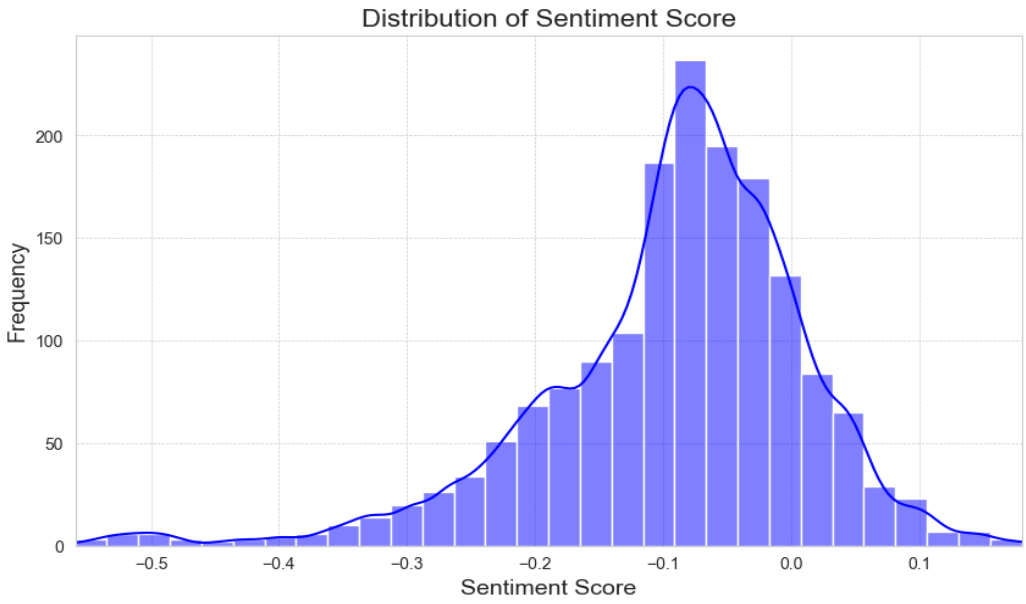

Figure 9. Sentiment Score Distribution

However, it should be remembered that the *sentiment_score* is defined by the values contained within the *Loughran-McDonald* dictionary and, therefore, depends on its structure and evaluation. The dictionary we used is commonly known for being "less optimistic" and for highlighting negative tones more than positive ones within texts.

In Figure 10, we can verify that the correlation is indeed negative. The trends over the years, especially before 2010 and from 2021 onwards, are clearly inversely proportional.

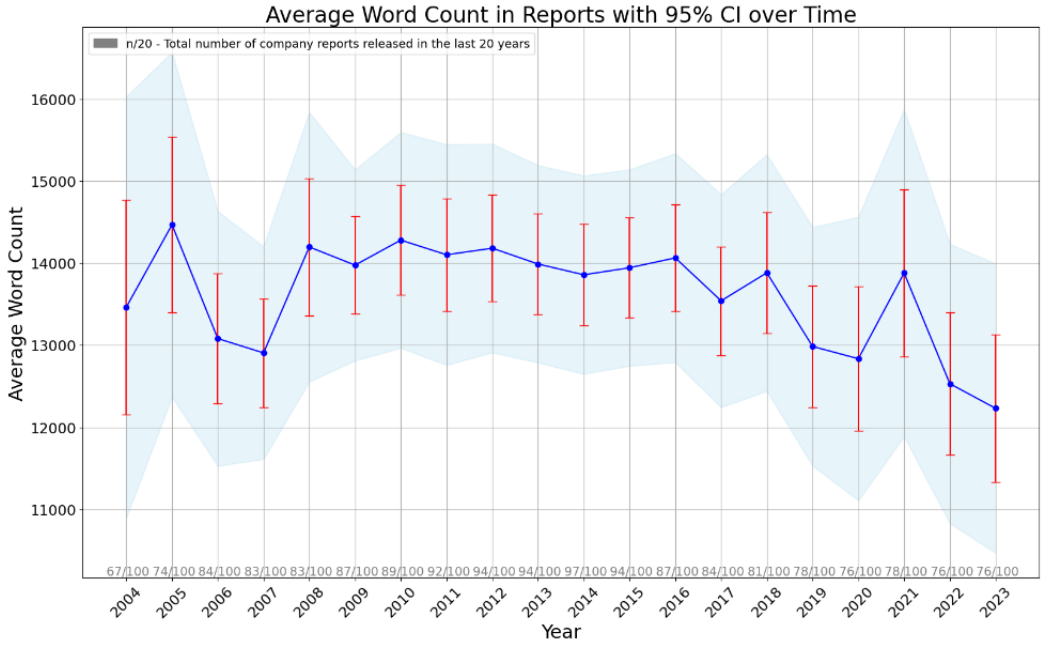

Figure 10. Yearly Distribution of Average Sentiment Score

We can clearly observe that in the trend of the *average sentiment_score* in 2005, the value is low, in contrast to the peak in *average word_count*. A similar situation also occurs in 2021, where a higher average word count corresponds to a more negative sentiment, as noted by the dip in *Figure 10*.

5.1.3.2 Top Companies by Average Sentiment Score

In this section, we aim to provide a quick overview of companies classified according to the tone of their communicative language. Visualising the graphs also

highlights how public perception and communicative tone can vary significantly between different companies.

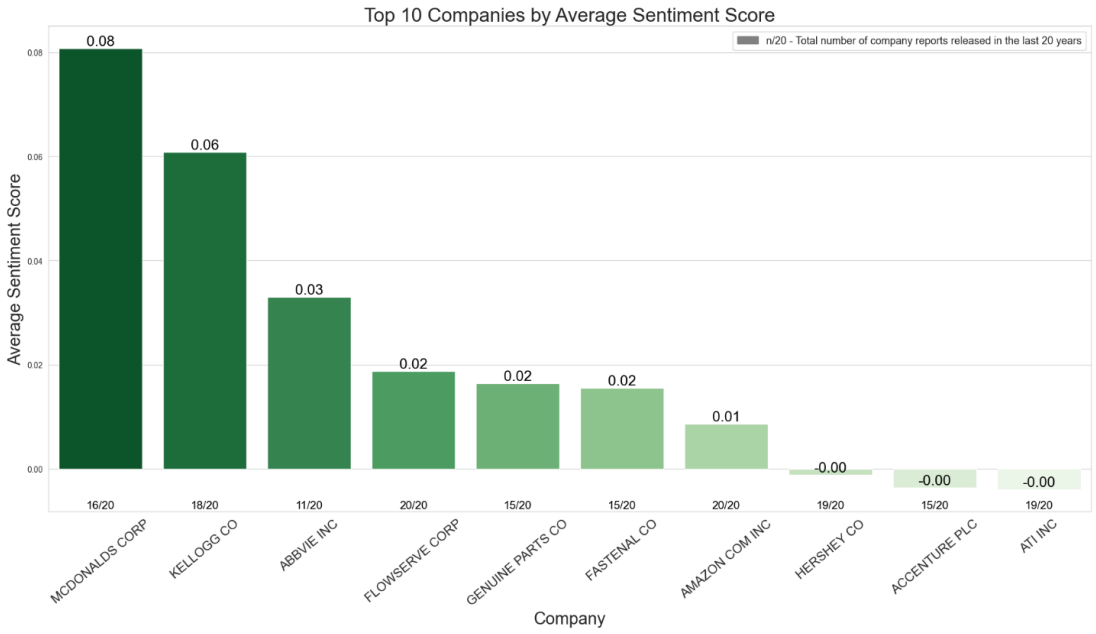

Figure 11. Top 10 Companies by Sentiment Score

As we can see from *Figure 11*, despite the values hovering around 0, McDonald's stands out for its slightly positive language, which is not a trivial result, given that the dictionary tends to highlight negative terms over positive ones. Consequently, it is worth considering that all values are slightly positive, close to 0. On the other hand, companies that do not excel in communicative tone include Chubb Corp. (an insurance company) and Berkshire Hathaway Inc. (a holding company), with rather low values, -0.51 and -0.41, respectively.

Let's take a closer look at the companies that tend to write with a less positive tone. To do so, unlike the previous case with word_count, where we analysed the year of the pandemic, we will analyse the year 2019. This date precedes the unfortunate event and should give us a more general view of the market in those years of financial stability following Trump's tax reforms in the USA in 2017, which allowed companies to benefit from significant tax breaks.

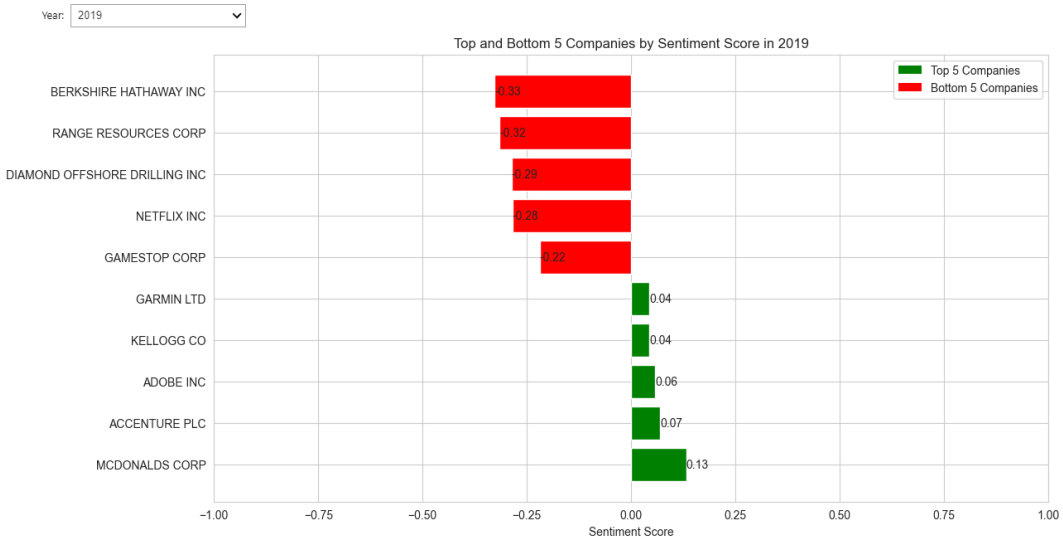

Figure 12. Top and Bottom 5 Companies by Sentiment Score in 2019

In the previous *Figure 12*, it is immediately evident that companies like Netflix and Gamestop are not in a flourishing moment in their history, at least according to what they are communicating. The video game industry has shifted towards digital products rather than physical ones in the last decade, while Gamestop has always had large volumes of physical sales in the gaming sector. This data can provide us with similar reflections on certain companies' growth or decline over the years.

5.1.3.3 Tech Companies – Trend in Average Sentiment Score

Let’s briefly return to our small sample of 10 tech companies, examining how they behave specifically to notice some differences within the same sector.

The figure provides us with very interesting information: first, we notice how eBay used a very negative communicative tone in 2004, despite being founded in 1995 and listed in 1998. However, from 2006 onwards, it used a predominantly neutral language.

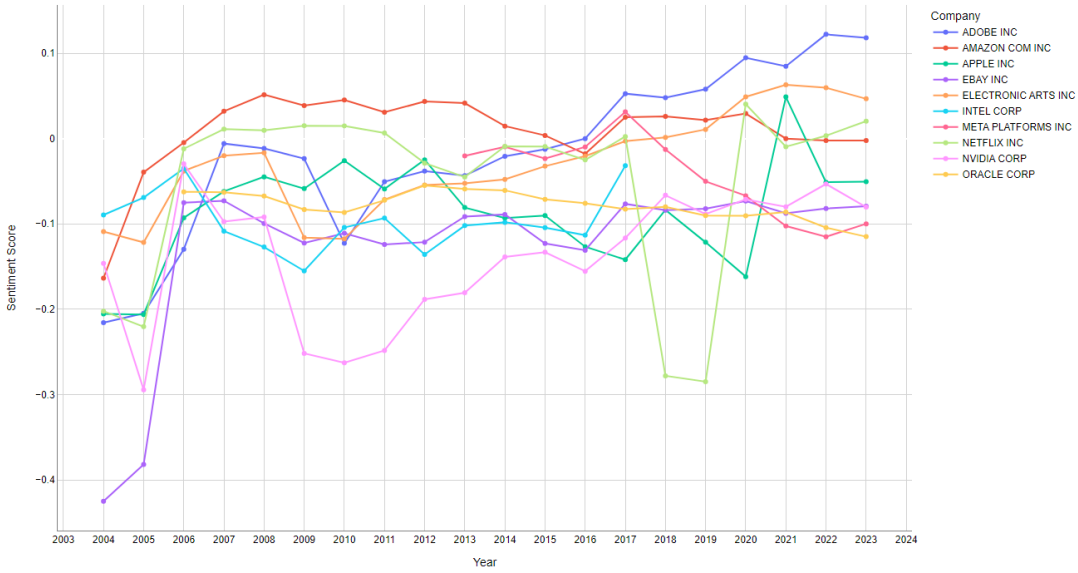

Figure 13. Sentiment Score distribution for Tech Companies

Nvidia shows low values starting from 2009, which may reflect a challenging period that lasted until the end of the second decade, where there was a recovery due to a decline in Intel from the years of the COVID-19 pandemic onwards. This recovery is also seen in Adobe, which started in 2016, regarding the emotional tone of language. Finally, in this graph, we also observe the 2018-2019 biennium, where Netflix's language was tinged with terms inclined towards financial negativity, but this was followed by a decent recovery and stabilisation from 2020 onwards.

5.1.4 Boilerplate Text analysis

5.1.4.1 Distribution of Boilerplate Percentages

As previously explained, we collected these *boilerplate* values by considering the repetition of phrases composed of at least four words. To be considered *boilerplate* language, these repetitions must cover at least the 1% threshold within the document.

However, our analyses did not detect widespread use of this type of standardised language; rather, we found that only specific companies use it. Most companies recorded a null percentage of *boilerplate* text, resulting in a relatively low average value, as shown in *Figure 14.*

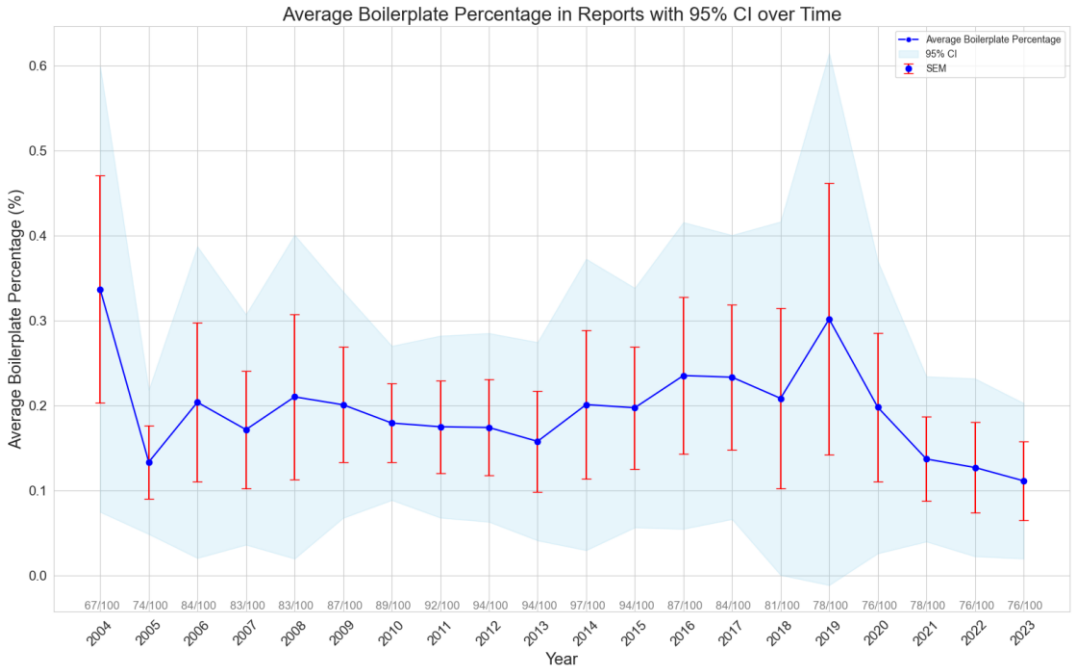

Figure 14. Yearly Distribution of Average Boilerplate Percentage

The average value of our sample is slightly above zero, although it should be noted that, generally speaking, a *boilerplate* language usage of over 3% is considered significant. From *Figure 14*, at first glance, we notice how the use of *boilerplate* language increased in 2019, which may be caused by numerous factors or the presence of some outliers.

5.1.4.2 Average Boilerplate Language Over the Years

Boilerplate language characterises each company; even when a company does not use it, it still communicates something to us. Before discovering which companies stand out for their consistent use of *boilerplate* language, let's try to understand which sectors might be more affected by this phenomenon. *Boilerplate* language is almost devoid of financial significance, and its use should somehow protect large companies with a substantial number of investors from possible SEC regulations or other applicable laws.
The most regulated sectors involve finance, energy, and utilities markets, which force companies to use a significant amount of standardised language to comply with strict regulations.

Another critical driver of *boilerplate* language is certainly risk management, which is reduced by repeating specific phrases that allow companies to avoid some information being misinterpreted and causing significant damage. From the analyses in *Figure 15*, it is evident that the company that uses it the most is IBM Inc., which reports an average value of 5.15%, which is a very high result in itself.

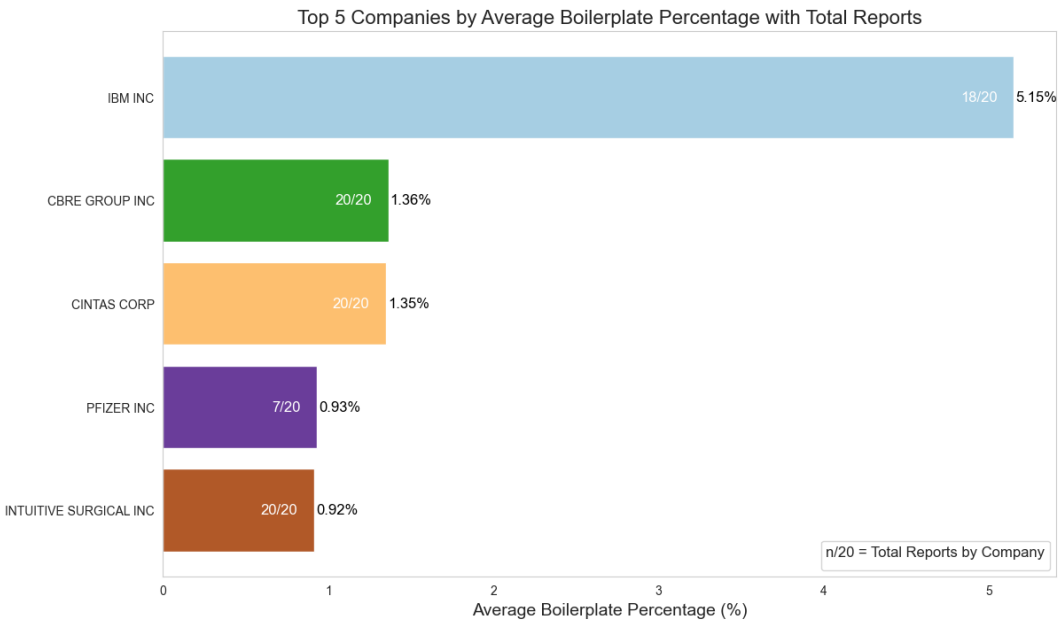

Figure 15. Top 5 Companies by Avg. Boilerplate Percentage

The global pharmaceutical company Pfizer also appears in this particular ranking, likely using *boilerplate* language as a precautionary measure during the pandemic. In the following scatterplot in *Figure 16*, we notice how Pfizer was the company with the highest *boilerplate* text value in 2020 due to the global pandemic and its central role in vaccinations.

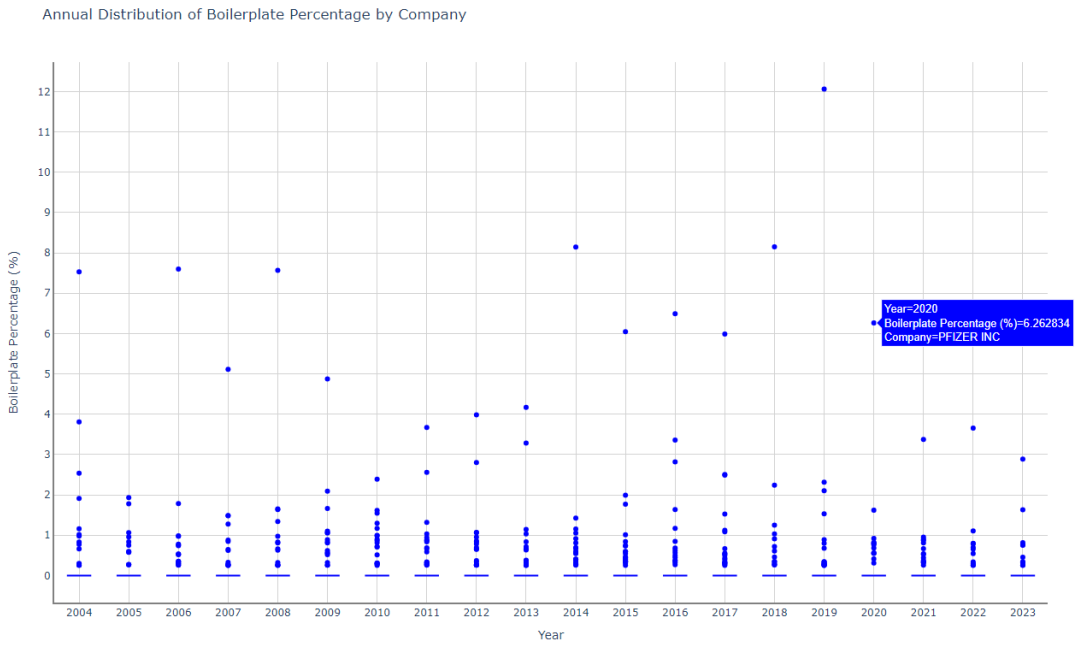

Figure 16. Yearly distribution of Boilerplate Percentage

However, Pfizer has released only 7 of the 20 annual reports, and for this reason, this should not be considered a decisive result.

5.1.4.3 Annual Distribution of Boilerplate Percentage by Company

Let's take a look at our hi-tech companies by analysing their behaviour, particularly in *Figure 17*.

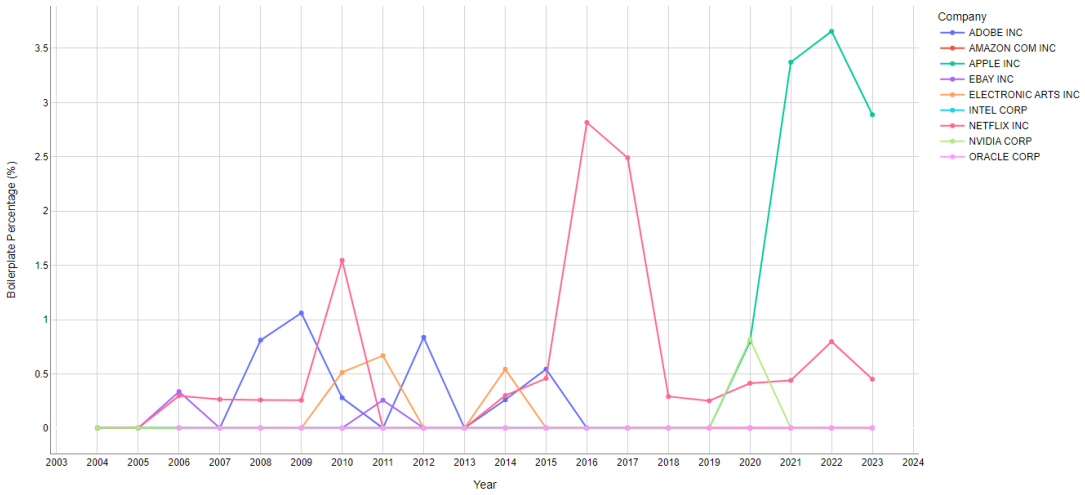

Figure 17. Boilerplate Percentage for Tech Companies

This time series is interesting as it reveals several curiosities; we immediately notice how Netflix had a 3-year interval during which it wrote more standardised text compared to other years. In recent years, the protagonist has been Apple, with a sharp increase in *boilerplate* content starting from 2020 onwards. As previously defined, 2020 is a delicate year due to the great economic and market uncertainties caused by the pandemic. It is noticeable that Apple tends to use a more conservative approach in terms of communicative language toward its investors.

On the other hand, it is also important to note that companies of the calibre of Amazon, Intel, and Oracle have not registered any cases of boilerplate text in their documents. Finally, Nvidia, except for the year 2020, joins this group of companies with this peculiarity.

5.1.5 Most Frequent Words - Tech Companies

This section will delve into an analysis that takes a more practical approach. Instead of dealing with pure numerical data, we will explore which words are most frequently used among tech companies.

We will use word clouds to visualise the most frequently used words by each company. This will help us understand, through this textual investigation, the most dominant themes within corporate dynamics. With word clouds, we can immediately visualise the most frequently used terms, allowing us to understand a company's priorities and areas of interest over two decades. The following visualisations refer to the most commonly used terms by Apple, Nvidia, and Intel.

Focusing first on Apple, in *Figure 18,* we notice how Apple uses terms related to risk and performance, as revealed by the graph with the negative words ‘decline,’ ‘impairment,’ and ‘loss.’ This, as also defined earlier, is a clear case where the company wants to maintain caution in its language to avoid challenges or potential threats to corporate performance.

.

Figure 18. Sentiment Word Cloud for Apple Inc.

However, the presence of positive terms such as ‘strong,’ ‘gain,’ and ‘benefit’ is also notable, suggesting excellent opportunities and financial successes at the same time. There are some cases where competitor companies may use diametrically opposite communicative language in terms of emotional tone. To analyse such a phenomenon, let’s see if this occurs with competitor companies Nvidia and Intel.

From the subsequent *word clouds* in *Figure 19*, we observe that Intel’s word cloud has a smaller number of words than Nvidia's and a higher *positivity ratio*. Referring again to Intel, the company uses positive terms like ‘gain,’ ‘benefit,’ and ‘effective,’ suggesting an effective strategy or benefits regarding corporate operations. The negative terms, on the other hand, suggest operational challenges or value reductions. Words like ‘restructuring’ and ‘divestiture’ suggest a possible response to external or internal pressures through strategic corporate changes. In contrast, analysing Nvidia’s *word cloud*, we immediately notice that with terms like ‘improvement,’ ‘enable,’ or ‘progress,’ the company aims toward innovative horizons and is more future-oriented. Looking at the negative words, ‘loss’ and ‘impairment’ also emerge here. However, we also see words like ‘litigation’ and

'investigation' likely related to legal challenges or investigations that may have influenced Nvidia's communicative tone.

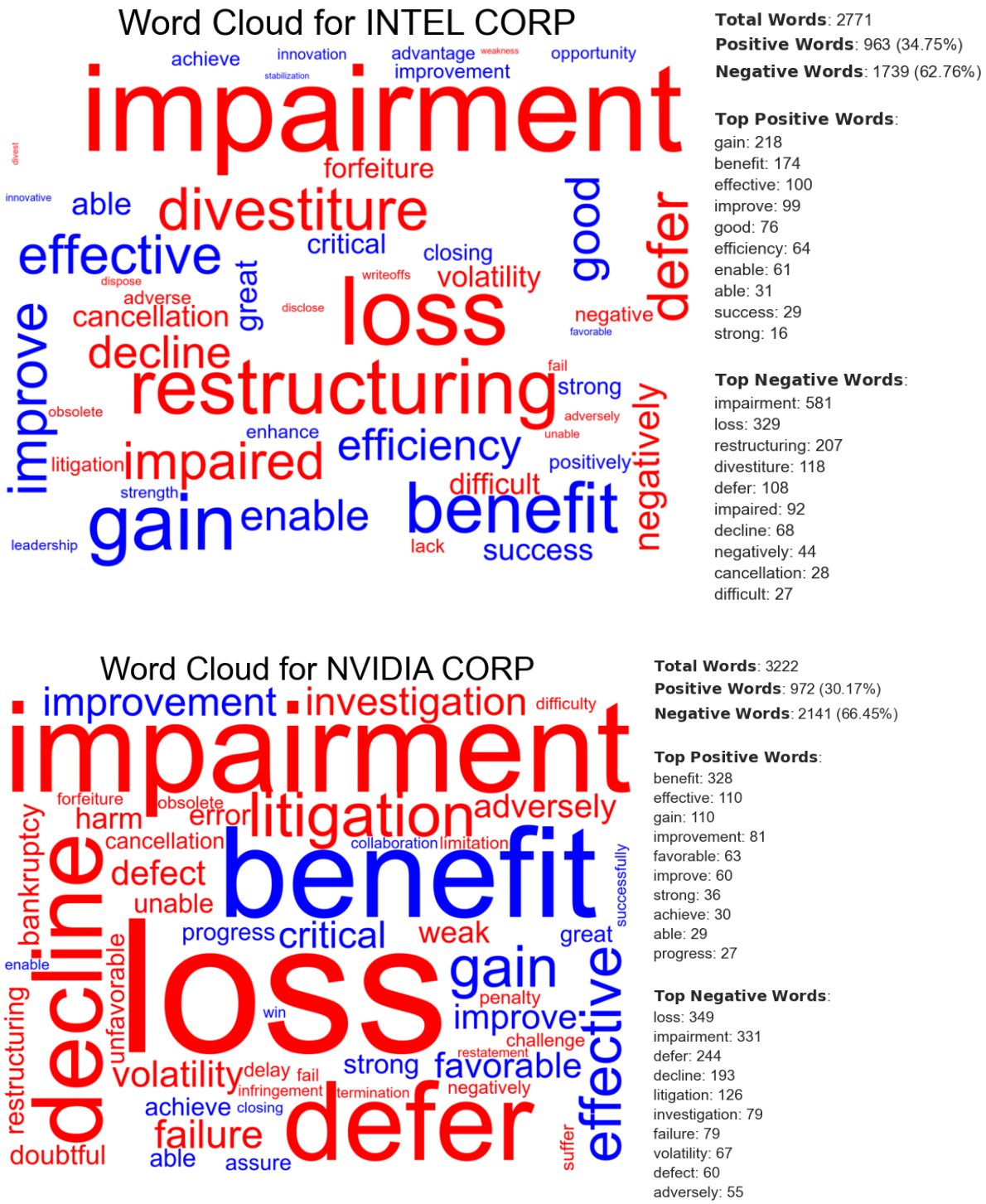

Figure 19. Sentiment Word Cloud comparison between Intel & Nvidia

To summarise the comparison, from the *key words*, it emerges that Nvidia focuses more on innovation and growth opportunities, as confirmed by the terms related to the technological field. In contrast, Intel focuses more on managing operational and financial challenges, simultaneously seeking a balance between future growth and risk management. This analysis not only reveals the emotional tone and most frequently used words by the two companies but also offers a glimpse into their strategic priorities, the challenges these companies may have faced, and the opportunities they are seeking to seize.

5.1.6 Readability Analysis

It is now time to visualise the results of the readability metrics we constructed during the preprocessing phase. We will analyse the different indices and the complexity of the text, identifying which companies tend to write in a more complex manner compared to others.

5.1.6.1 Sentence Analysis

The syntactic structure of texts characterises the communication language, as it influences the average sentence length and the grammatical constructions' complexity. Generally, a text composed of shorter sentences and minimal use of subordinate clauses corresponds to more straightforward and more easily interpretable language, thus reducing the likelihood of misunderstandings and enhancing the transparency and accessibility of textual communication. *Figure 20* presents a specific ranking based on the average, grouped by company, of the longest sentences in each report (expressed in words). Among all 100 companies in the sample, *Cablevision Systems Corp. NY* and *Harley Davidson Inc.* dominate all others with an average of 387.62 and 370 words, respectively, within their longest sentences in each report released. This is a very high value, suggesting at first glance that these two companies release the most complex and articulated documents, possibly enriched with lengthy texts on current regulations, such as those concerning the American Securities and Exchange Commission.

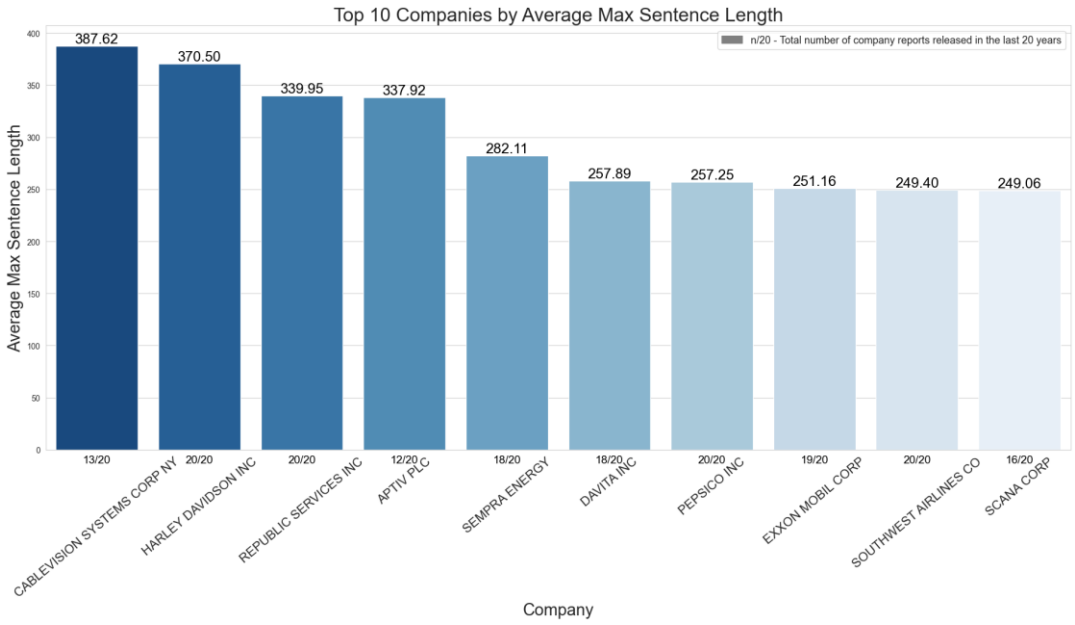

Figure 20. Top Companies by Avg. Max Sentence Length

Following the American motorcycle manufacturer are other companies operating in the energy and environmental services sectors, such as *Sempra Energy* and *Republic Services Inc.*

An important point that generates much curiosity is that the *word_count* and *max_sentence_length* metrics, while seemingly related, demonstrate that companies that write longer documents do not necessarily produce more complex documents.

5.1.6.2 Readability Indices – All companies

We now analyse the three metrics that allow us to get an understanding of the readability levels of the MD&A sections of companies within the S&P500 index. Previously, the values related to the number of words contained in the longest sentences were relatively high, and indeed, the idea of language complexity that had emerged earlier is confirmed here with the *Flesch-Kincaid Grade Level* index in *Figure 21*. The *Flesch-Kincaid Grade Level* is an index that estimates the school grade level required to comprehend the text on the first reading attempt. It is based

specifically on the number of syllables per word and the average number of words per sentence.

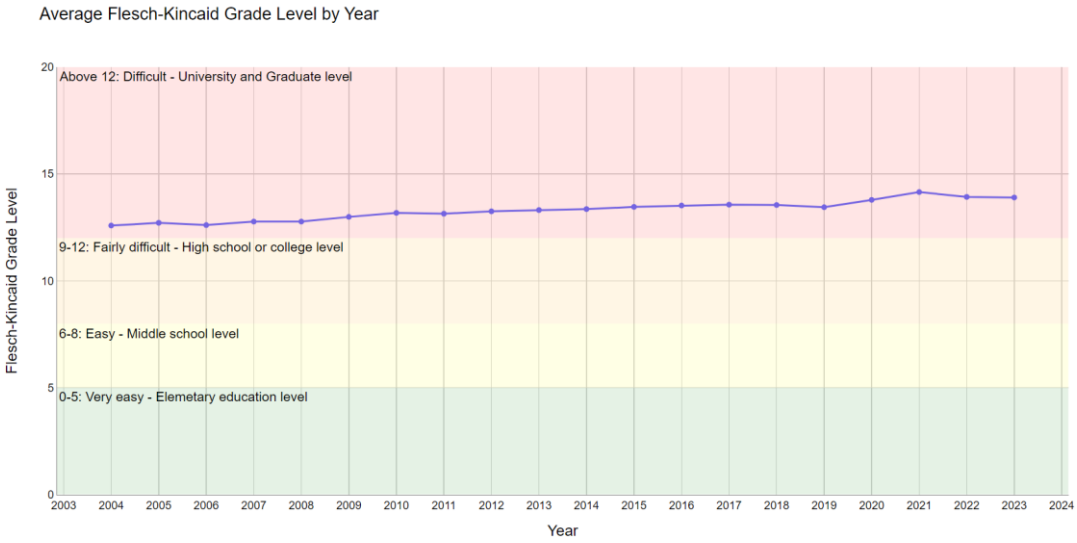

Figure 21. Trend of Avg. Flesch-Kincaid Grade Level

We can immediately observe how the readability level thresholds are delineated by the coloured areas, with four overall bands covering the intervals in which the text may require an elementary, middle school, high school, or university education level. The line plot clearly indicates that the language companies use is, on average, complex. This factor should not entirely surprise us, as the topics addressed by S&P500 companies are exceptionally delicate, and any misunderstanding could cause significant losses for many investors. Therefore, companies may strive to specify many details, and sentences may generally be richer in subordinate clauses, making them more complex.

In the following line plot, we analyse the *Flesch Reading Ease* index, which measures how easy it is to read a text. In this case, unlike the previous index, a higher value corresponds to a syntactically more straightforward text to read. All values are within the range of 0 to 100, and general guidelines suggest identifying seven readability level bands, which we can distinguish in *Figure 22*.

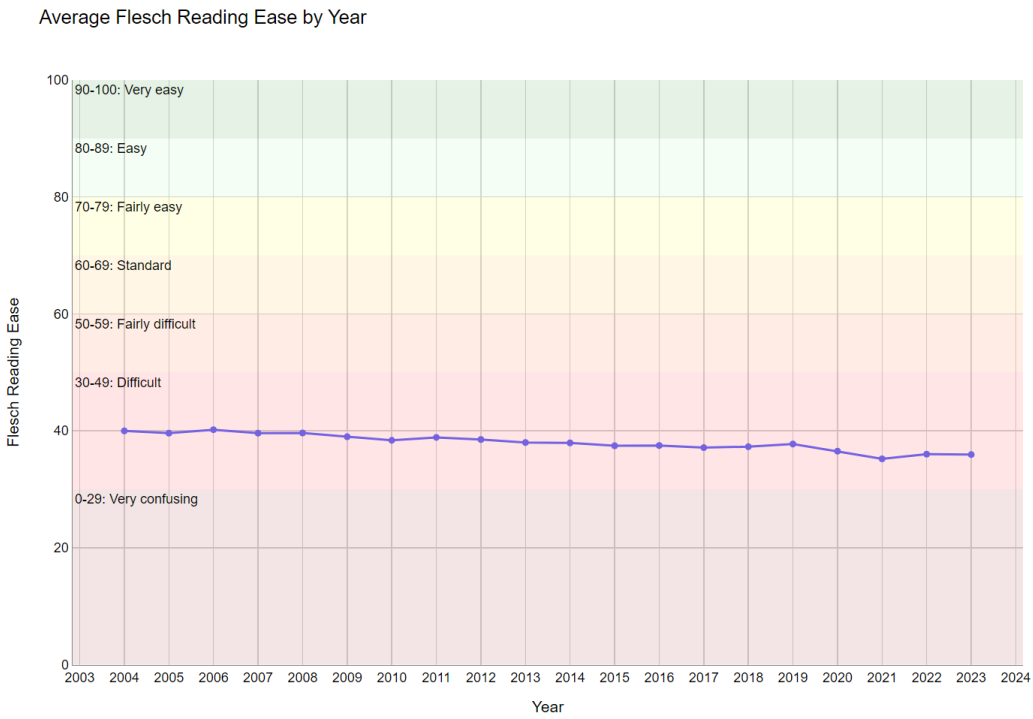

Figure *22. Trend of Avg. Flesch Reading Ease*

Unlike the Flesch-Kincaid Grade Level, this index provides a general value of text complexity, aiming to assess how easily anyone can read the text. On the other hand, the previous index is more commonly used in educational contexts to verify that certain materials are appropriate for specific education levels, ultimately providing a school grade level as a result. However, even in the *Flesch Reading Ease* index, the trend remains the same, and we notice that texts are becoming slightly more complex over the years.

The last readability index we will analyse is the *Gunning-Fog Index*, which differs slightly from the other two because it is based on the percentage of words with three or more syllables and the average sentence length. Even here, however, the results are expressed in terms of formal education level, starting from 0 and following the years of American school education—for example, a value of 10 indicates that the reading is suitable for individuals with schooling up to the 10th grade (which corresponds to high school in Italy).

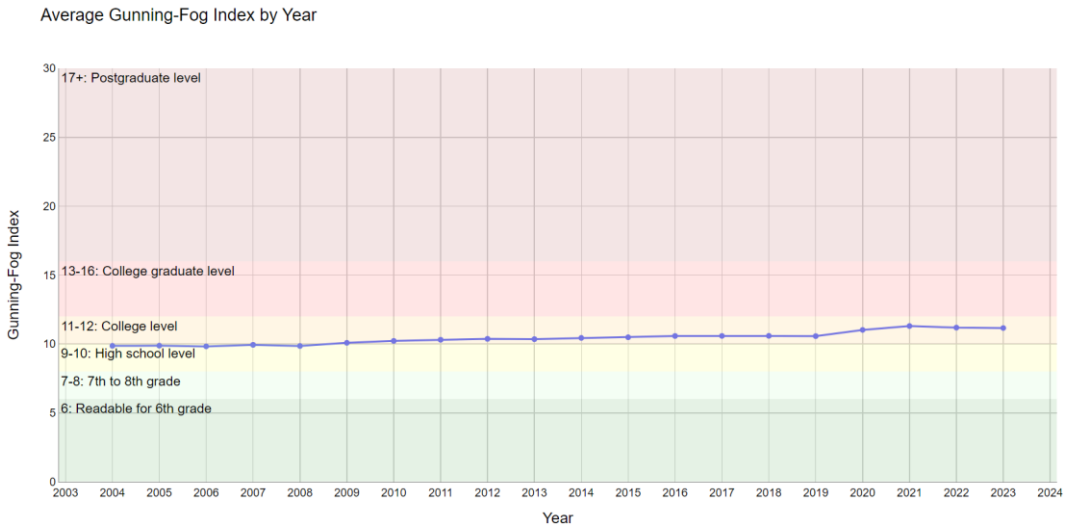

Figure 23. Trend of Avg. Gunning-Fog Index

The preceding *Figure 23* provides results consistent with those of the previously analysed indices and confirms how the trend of complexity has been gradually increasing over the years, although this index highlights vocabulary complexity more than the other two.

5.1.6.3 Tech Companies - Readability Indices

Taking once again our subsample of tech companies as a reference, let's take a look at how they perform in terms of language complexity by observing the comparisons in the following graphs in *Figure 24*, *Figure 25*, and *Figure 26*, which depict how these companies are ranked over the years according to the three indices: *Flesch-Kincaid Grade Level*, *Gunning-Fog Index*, and *Flesch Reading Ease*.

In all three representations, it immediately becomes apparent that Oracle stands out for the complexity of its texts, which require even a graduate level of education to be understood on the first reading, as we can observe from the following *Figure 24*.

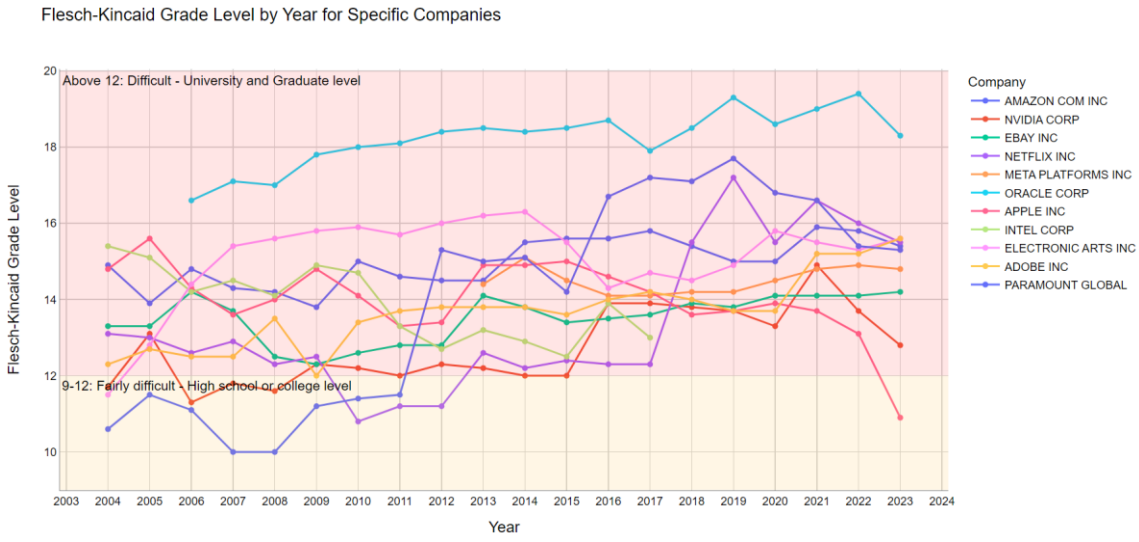

Figure 24. Trend of Flesch-Kincaid Grade – Tech Companies

Another noteworthy detail emerges from the second graph concerning the *Gunning-Fog Index*: Paramount Global was able to write reports with relatively simple propositions up until 2011, but from 2012 onwards, in terms of communicative style, it aligned with other tech companies by using more complex sentences.

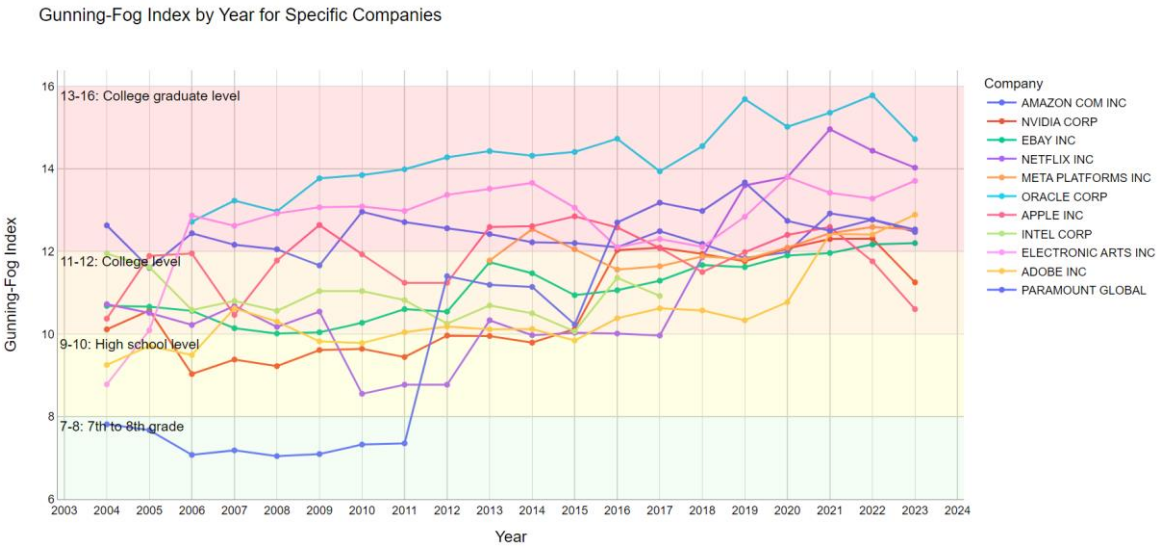

Figure 25. Trend of Gunning-Fog Index – Tech Companies

In the next graph for the *Flesch Reading Ease* index, we mainly notice that, overall, companies operating in the tech sector are, over time, raising the bar of

communicative language complexity, reaching levels of education and difficulty that are not accessible to everyone.

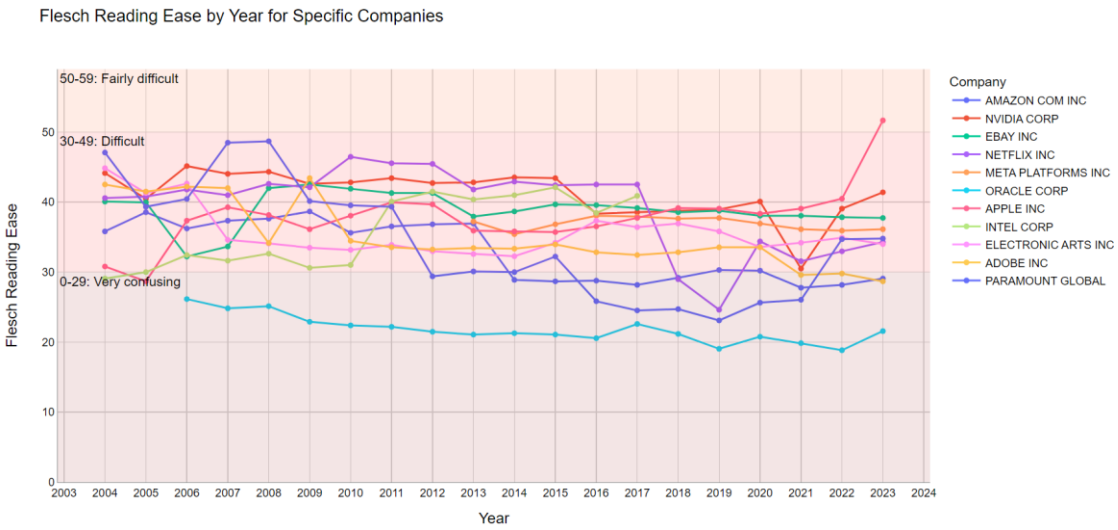

Figure 26. Trend of Flesch Reading Ease – Tech Companies

In *Figure 26*, we can also see the only company that tries to act as a standalone case, namely Apple. The company from Cupertino, although slightly different, seems to be the only one among the ten companies to work towards greater clarity in its reports, focusing on transparency, as we can observe the upward shifts in recent years and a generally positive trend.

5.2. Predictive Modeling Results

In the previous sections, we thoroughly examined the information provided by the data and their most relevant associations, identifying how the text of S&P500 companies is characterised. Based on the vast amount of data obtained and analysed, the question now arises: "How will this data behave in the future?" To answer this question, predictive models come to help, which can be of different

types depending on the specific use cases, as we explained in more detail in section 4.4.3.

5.2.1 Boilerplate Percentage Forecast

Estimating the behaviour of a variable in the future is a challenge that is far from trivial. In fact, we will use three different predictive models to then evaluate which of the three provides the most accurate results. It should also be specified that we are talking about estimates, not precise values; however, despite being estimates, having a look and getting a general idea of the future trend of the data can prove interesting and provide numerous points for reflection.
The variable under consideration, in this case, is the *boilerplate_percentage* to understand whether we are heading towards a future where corporate communication is characterised by useless and formal language, i.e., standardised. In *Figure 27*, the future trend of the *boilerplate_percentage* variable is represented according to the three different predictive models. The blue line represents the real trend, which refers to the past, and records the minimum value in 2023.

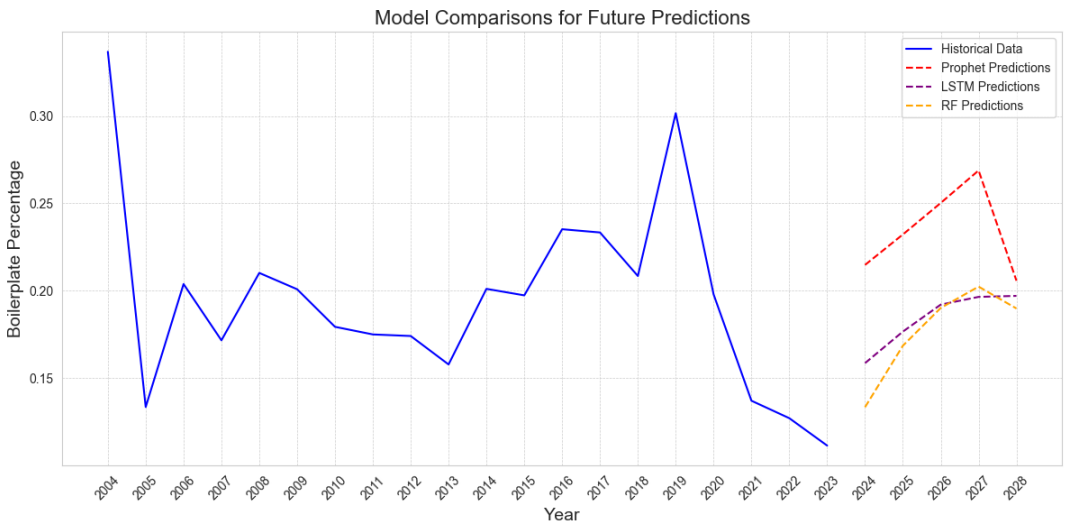

Figure 27. Comparison of Predictive Models for Future Boilerplate Percentage

However, the three models register an increase and thus a shift in the trend for the future years from 2024 to 2028. The most "optimistic" model is the *Prophet Model*, which registers an increase of over 100% compared to the previous year. The *LSTM*

and *Random Forest* models, on the other hand, register a lighter and more constant increase in the following years.

5.2.2 Model Performance Metrics

In order to choose the best model, we must consider the error metrics that we printed to the screen from the Python code: we refer to *MSE* (mean squared error), *MAE* (mean absolute error), and R^2 (coefficient of determination). The outputs of these metrics in Figure 28 suggest that the most performant model is the *Random Forest*.

```
Prophet Model - MSE: 0.0023, MAE: 0.0370, R²: 0.2070
LSTM Model - MSE: 0.0013, MAE: 0.0262, R²: 0.3132
Random Forest Model - MSE: 0.0004, MAE: 0.0159, R²: 0.7851
```

Figure 28. Performance Metrics Comparison Across Predictive Models

Let's analyse the models one by one:

- **Prophet Model**: Based primarily on models that follow trends and seasonality, a generally increasing trend is observed until 2027, with a subsequent decrease in 2028. The low values of MSE and MAE indicate a decent level of accuracy, but the R^2 value of 0.2 shows that only about 20% of the data variability is explained by the model, implying a not very precise prediction.

- **LSTM Model (Long Short-Term Memory)**: This model shows a better ability to adapt to temporal variations compared to the *Prophet model*, with slightly better values for MSE and MAE. The R^2 value of 0.32 is also better, making it overall a more performant method than the *Prophet*.

- **Random Forest Model**: This ensemble model reports the best values in all metrics, with lower errors and a higher coefficient of determination. The *Random Forest* simply outperforms both the *Prophet* and *LSTM* models. It

is the most reliable model that best represents the future values of the sample, explaining approximately 78.5% of the variance in the data.

Through these analyses, it becomes clear that the *Random Forest* is the best choice among the three, thanks to its complexity, as it manages non-linear relationships without incurring overfitting and presents an almost negligible error.

6. Discussion & Conclusions

6.1. Summary of Key Findings

6.1.1 Interpretation of Sentiment, Boilerplate, and Readability Trends

Throughout this study, the in-depth analysis of the MD&A sections of companies within the S&P 500 index has highlighted trends and patterns regarding corporate communications, specifically indicating how variables related to sentiment, boilerplate language, and readability indices of corporate communication language have primarily behaved.

Observing the sentiment analysis first, there is a general decline in corporate tones towards slightly more negative ones, particularly in 2021, which was influenced by the global pandemic. With the scores derived from the Loughran-McDonald dictionary, we observe a linguistic and communicative approach that leans more towards caution. This is clearly evident following the unfortunate events of 2008 with the financial crisis and in 2020-21 with the aforementioned pandemic, which highlights a much more cautious and less "optimistic" language regarding economic recovery.

On the other hand, during periods characterised by a more stable economy, we see that the sentiment reflects this stability, reaching values close to neutral, which could be considered partially negative, as it should be noted that the Loughran-McDonald dictionary tends to highlight negative terms over positive ones.

It is interesting to note that, among the Tech companies analysed, the emotional tone of Nvidia Corp. was particularly negative at the beginning of the second decade, unlike its competitor Intel Corp. From the middle of the decade, around 2015, Nvidia's sentiment gradually improved, eventually surpassing its competitor. These improvements are evident during the COVID-19 years, when there was a massive demand for Graphic Processing Units (GPUs), with Nvidia becoming the leading company in the sector, effectively outpacing the competition.

Regarding boilerplate language, the analysis determined a low frequency of purely boilerplate language, showing that many companies hardly use it. At the same time, a few, such as IBM, make extensive use of it. Excluding this company, the primary sectors characterised by a significant presence of boilerplate language are the pharmaceutical and energy industries, likely due to the need to comply with legal and regulatory requirements. Generally, when an economic sector is associated with sensitive issues or where detailed accountability is required, it tends to use more boilerplate language as a precaution and to list applicable rules and regulations.

Indeed, the pharmaceutical sector faces considerable challenges regarding the commercialisation of drugs and their safety, as approval procedures require a rather lengthy legal process, obliging companies to fully comply with current laws. Examples of boilerplate in this specific case can, therefore, include legal disclaimers and standard statements on compliance, which serve the additional purpose of protecting the company from potential legal issues and meeting regular regulatory requirements. In the energy sector, the situation appears similar, with the inclusion of environmental and governance regulations. These stringent regulations thus lead to a gradual increase in boilerplate language, as companies tend to use precise communication and may repeat phrases to ensure complete clarity, guaranteeing compliance with all regulatory requirements.

However, even here, by analysing the trend over the last twenty years, the shifts in the amount of boilerplate language around 2008 and 2020 stand out, once again confirming how events have impacted all financial aspects, including communication, in a comprehensive manner.

From the analysis of Tech companies, it also emerges that Netflix made intensive use of boilerplate language only in the biennium 2016-2017. This period coincides with a delicate time for the streaming company. Indeed, during this period, following the launch of the TV series "Narcos" and "Stranger Things" (some of the most famous and successful series in the world), Netflix faced the most significant challenge in its history, coinciding with its global expansion, making its service available in 130 countries worldwide. This step represented a bold move by the company, which had to navigate a challenging period due to high costs and incurred

debts, leading the company to likely opt for caution by extensively using boilerplate language, as seen in *Figure 17*.

The predictive models have indicated an increase in the use of boilerplate language in the coming years, driven by companies' desire to protect themselves, especially as the market becomes increasingly volatile over time. However, this also leaves less room for stakeholders, who may lose confidence in companies that adopt a less transparent position.

The trends in readability indices follow a similar pattern, moving in tandem with boilerplate language, as readability also tends to become increasingly complex over time. The reasons for this could similarly be linked to regulations, which are becoming increasingly stringent and numerous over time. As a result, the texts released require more excellent 'expertise' in the field of financial communication to be fully understood, inhibiting the interest and engagement of individuals who are less specialised in this sector. This outcome raises numerous concerns regarding the accessibility of these texts and the issues related to balancing the precision required in providing detailed information with the goal of allowing communication to remain widely accessible and comprehensible.

6.1.2 Implications for Market Perception and Corporate Communications

This text analysis has clearly highlighted how language represents a particularly delicate tool for companies, as it significantly influences market perception and investor decisions. Shaping market perception is key for a company aiming for long-term success, as is providing transparent communication.

At the end of the previous section, we discussed the balance between providing detailed information and using simple language to ensure reasonable accessibility to financial content and how communication must be adequate to achieve the optimal balance. Therefore, a communication strategy must be implemented to resolve these comprehension issues while adhering to specific criteria.

First and foremost, communication should focus on conveying the company's current reality and the market it operates in, covering recent challenges, successes, or failures. Special attention should also be given to the company’s future expectations, at least outlining the general directions it plans to take, highlighting plans for expansion into new markets or the desire to invest in new technological innovations that could improve or revolutionise the company itself. Another key factor is the frequency of document releases, specifically annual reports, which are not necessarily released every year by companies due to various dynamics, such as temporary delisting or other issues.

In this case, a significant concern is transparency, which is equally essential. Trust from investors can be gained through honest communication that avoids downplaying obvious problems, threats, or challenges the company faces and avoids excessively emphasising its achievements. Stakeholders appreciate reputations built on integrity. However, it is essential to consider that companies must still fulfil the crucial task of protecting sensitive information so as not to put the company at a competitive disadvantage.

Thus, communication must include accessible information, but comprehension is equally critical. Complex, convoluted, and unclear propositions should be avoided for simple, direct, clear, and unequivocal discourse. By following these steps, companies can achieve more from their corporate communications, earning the complete trust of investors and reaching a wider audience, positively influencing market perception.

6.2. Practical Implications and Limitations

6.2.1 Insights for Investors and Analysts

Having discovered through this analysis how financial communication represents a significant indicator capable of influencing the investment decisions of various stakeholders and impacting market trends, in this section, we will discuss the

practical implications that may arise from this analysis, providing several insights for practical advice to refine investment and analysis strategies.

Sentiment analysis is undoubtedly one of the most important factors to consider before choosing to invest in a particular stock. For example, in a scenario where a company increases the negative tone in its texts, this should generally be interpreted as a likely precursor to a decline in stock prices. Consequently, stakeholder behaviour in this type of scenario might be linked to an opportunity for timing sales or taking a defensive position.

An investor must also take into account the level of transparency a company maintains. Using boilerplate language can serve as a double-edged sword, signalling a lack of specificity and confidence. Just as companies tend to adopt a more neutral and cautious stance, stakeholders should emulate this behaviour when faced with an increase in standardised text in MD&A documents. Such an approach may conceal potential issues, and an increase in the boilerplate text is expected to correspond to a delicate and particularly challenging period for a company or market, leading to increased uncertainty.

Looking at the readability of reports, it is clear that the more complex a company's writing, the lower the level of transparency. Reports that are easier to read correspond to greater investor confidence, and through this correlation of factors, investors may prefer such companies as they feel more involved and engaged. Analysts can interpret how open and engaged the company is with its investors based on how clearly and simply a company communicates. This factor, therefore, influences the stocks selected in investors' portfolios, improving the relationship between the company and its stakeholders, making it more direct and reducing the risk of potential misunderstandings.

The approach to predictive models discussed in this thesis provides interesting insights into language trends in the coming years. Investors can organise various appropriate and reactive strategies in response to changes in language over time. A valid strategy could involve anticipating market movements with the associated risks of such an operation. In the case of a shift towards increasingly negative tones

in language, investors may consider anticipating a decline in stock prices. Another strategy could be portfolio diversification, which would protect investors from the consequences of uncertainty perceived in corporate texts, reducing the portfolio's overall risk.

On the other hand, an intelligent move could be to identify opportunities from potentially favourable stock prices: if a report shows a general increase in transparency and positive tones without market evidence to justify such improvement, purchasing shares in that company could be a winning choice. Conversely, predicting an imminent decline in a stock would warrant selling shares at the most opportune moment.

6.2.2 Consideration of Study Limitations

This thesis has provided a detailed insight into financial communication, highlighting how various factors characterising language can be used as market indicators. Typically, the data speak for themselves and are explanatory; however, it must be acknowledged that the market is subject to external forces that are beyond control, which can distort or complicate the interpretation of the data. The analysis should be conducted and viewed as a first evaluation, which subsequently requires deeper and more targeted analyses, as the financial market can be influenced by events such as regulatory interventions or catastrophes like COVID-19, which can completely overturn previously established trends.

One limitation of this study can be the assumption that past trends and behavioural patterns will remain consistent over time, which could lead to a failure to evolve corporate strategies and, consequently, yield inaccurate results in response to new challenges or opportunities. Predictive models, by their very nature, include a certain margin of error and may not adequately capture the complexities and dynamics of the market. The model processes numbers and data and finally provides an output, but if an external factor undergoes a significant change, the model is unable to detect it.

On the other hand, it must be recognised that the insights provided by this study remain valuable. They offer a general overview and serve as a reference point that can help analysts and investors understand and interpret a market trend, assess how a company communicates relative to the average, and examine its specific linguistic details. Overall, this thesis promotes a cautious yet informed approach that can provide numerous clues for developing strategies or interpreting trends, with its associated potential and limitations.

6.2.3 Contribution of the Study to Previous Works

Using a multi-disciplinary perspective, we study how communication in S&P500 reports evolved in terms of sentiment analysis, boilerplate text and readability indices. Although there has been much discussion of these metrics in the extant literature, they have never before been combined into an integrated longitudinal analysis over a 20-year timeframe to build an understanding of how corporate communications dynamics shift over time.

Foundational methodological and theoretical contributions come from seminal studies by authors such as Loughran and McDonald (2016) and Yu (2014) that largely focus on the creation of sentiment dictionaries and the efficacy with which sentiment can forecast movements in the market. But what makes this study unique is that it combines these with modern predictive analysis techniques such as Random Forests, LSTM models and Prophet.

This methodological innovation provides a more comprehensive analysis of boilerplate text trends and their connection to market perceptions and investor decisions, ultimately expanding the boundaries of how corporate communication can affect organisations.

It was also conducted a readability analysis of corporate reports, examining the extent to which financial communications are accessible by a variety of readability indices. In this case, the impact on the non-specialized reader of the complexity in language inherently present in previous similar studies that were considered as such was likely neglected or omitted during the review by Kearney and Liu (2014) topic. As such, this study directly aids in closing a fundamental void of financial

communications research by offering evidence on the crucial role that structural and visuospatial streamlining plays in making finance-related information more accurately interpretable to market constituents.

This thesis therefore extends the existing literature by arguing that qualitative text features such as sentiment and boilerplate language play a significant role not only in impacting market perceptions but also can be used as powerful indicators for investment decisions. These results are useful in the academic discussion around textual analysis, from a financial perspective and aid analysts, who can consider this new information to interpret market dynamics better, and investors gather more information.

6.3. Final Remarks

This study about how language in corporate communications works as a leading market indicator gives us a significant perspective on what drives investment decisions and shapes market conditions. Consequently, language presents itself as a dynamic and revealing instrument; however, analysts should consider these findings within a broader context, incorporating a variety of tools and analytical approaches to proceed with their studies. Thus, in order to perform a comprehensive study, the best approach is recommended to combine textual analysis with other quantitative methods as well as qualitative insights into market conditions and concrete investment ideas.

Anyway, the research illustrates that scrutinising financial communication offers an unlimited universe of perspectives and information which, if well interpreted, can be precious in understanding market trends. This type of analysis suggests a bright future for the use of language as a market analysis tool and shows how valuable information and details can be extracted from every characteristic, which can be crucial for uncovering specific data. Moreover, with the advancement of technologies, computational resources, and techniques for conducting analyses, we

can expect a future in which companies have more tools at their disposal to succeed in the global economic landscape.

This case study shows that the proper use of text analysis, exploited correctly, provides a competitive advantage for companies and investors. Companies can use these data to gain insights into the quality of their communication and adjust accordingly to improve their engagement with stakeholders, enhancing transparency, trust, and accessibility. By adopting this method, market perception is improved, and investment choices are channelled towards scenarios of long-term success and financial stability.

Index of Figures

Index of Tables

References

1. *Loughran, T., & McDonald, B. (2017). The use of EDGAR filings by investors. Journal of Behavioral Finance*.
2. *Chan, S. W. K., & Chong, M. W. C. (2016). Sentiment analysis in financial texts. Decision Support Systems, 91, 75-87.*
3. *Securities and Exchange Commission. (2023). EDGAR database. Retrieved from www.sec.gov/edgar/searchedgar/companysearch.html*
4. *Feldman, R. (2013). "Techniques and applications for sentiment analysis." Communications of the ACM.*
5. *Yu, X. (2014). Analysis of news sentiment and its application to finance. Doctoral thesis, School of Information Systems, Computing and Mathematics, Brunel University.*
6. *Kearney, C., & Liu, S. (2014). "Textual sentiment in finance: A survey of methods and models." International Review of Financial Analysis.*
7. *Li, F. (2010). "The Information Content of Forward-Looking Statements in Corporate Filings." The Accounting Review.*
8. *Loughran, M., & McDonald, B. (2016). "Textual Analysis in Accounting and Finance: A Survey." Journal of Accounting Literature. https://doi.org/10.1111/1475-679X.12123*
9. *Tetlock, P. C. (2007). "Giving content to investor sentiment: The role of media in the stock market." The Journal of Finance.*
10. *Lahmar, O., & Piras, L. (2023). Making sense and transparency in finance literature: Evidence from trends in readability. Journal of Finance and Economics, 12(1)*