Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

An Evaluation of The Impact of Negative Sentiment from Francophone Africa towards France on the French Stock Market

Joseph McInerney

Supervisor: Prof. Khurshid Ahmad

April 15, 2024

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Computer Science and Language

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

I consent / do not consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

I agree that this thesis will not be publicly available, but will be available to TCD staff and students in the University's open access institutional repository on the Trinity domain only, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. **Please consult with your supervisor on this last item before agreeing, and delete if you do not consent**

Signed: _____ Date: _____

Abstract

This project investigates the relationship between sentiment expressed in news media from Francophone Africa towards France and its impact on the French stock market.

A French language corpus of 1,678,267 words across 52 newspapers and 23 countries was collected and constructed. The negative sentiment of these articles was measured over the time span 2022-2023. 5 different sentiment lexicons were used to compare their negative sentiment measurement's impact on the French stock market.

Financial data from the CAC 40 French financial index was also collected, and the impact of the negative sentiment on the financial market was statistically evaluated using multivariate vector autoregressive econometric models.

The negative sentiment measurements of each lexicon were compared, and it was found that the lexicons designed for analysing financial media appear to be more appropriate in detecting a relationship between the negative sentiment news media proxies and financial market proxies.

Acknowledgements

I would like to thank my supervisor, Professor Khurshid Ahmad, for his continued support throughout this project.

Contents

List of Figures

List of Tables

1 Introduction

The current accessibility and volume of news media contrasts with that of the past. Before the introduction of the internet, news was consumed in print form, disseminated by local and national print media. Nowadays, with just an internet connection, news from around the world can be viewed. It is important to remember that news media does not just report information, but also expresses feelings and sentiment in the way in which the information is conveyed. This sudden availability of widespread subjective information sparked the interest in the field of sentiment analysis, where, as of 2018, 99% of papers had been written after 2004 [1]. Here, sentiment analysis aims to measure and quantify subjective information from text. With recent research showing negative sentiment as factor in price movement in financial markets [2] and a changing relationship between France and its former colonies in Francophone Africa, this project aims to explore a potential relationship between the sentiment of Francophone Africa towards France and the French stock market.

1.1 Motivation

In my time studying computer science and language at Trinity College Dublin, I have developed skills in the French language, computer science and linguistics. Therefore, his project was of particular interest to me as it allowed me to utilise all the multidisciplinary skills of this degree in the context of a real world application.

This involved the collection and processing of textual data but also an evaluation in relation to financial markets. This introduced me to the field of finance and econometrics, which was completely alien to me before undertaking the project. Therefore, in the process of this project, I had the opportunity to learn a lot about the properties of financial markets and methods of statistical evaluation. The methodology of this project in investigating hypotheses through the collection and curation of data and statistical testing showed me the importance of caution and precision when working with large corpus's of data. It equally showed me the significance of the scientific method in developing a sound methodology that is focused on forming a justifiable procedure and not centred on generating results.

1.1.1 Project Overview

Chapter one, the literature review section, relates previous research in sentiment analysis and econometrics to the practical implementation of the project. Chapter two, the methodology section, provides an overview of the approach taken to complete the sentiment analysis and evaluate its impact on financial markets. Chapter 3, the evaluation, presents the results of the statistical analysis and a critical assessment of possible interpretations. Chapter 4, the conclusion, aims to contextualise the results in relation to the thesis, acknowledging limitations and proposing future work.

The following sections provide a more detailed overview of the project structure, chapter by chapter.

Literature Review

Previous research in sentiment analysis has used different approaches in an attempt to quantify the subjective information conveyed in text. Recently, machine learning techniques have been applied to sentiment analysis in an attempt to recognise patterns and analyse words within their context. However, For the purposes of this project the classic sentiment lexicon approach was used. A sentiment lexicon is a dictionary where terms are defined in relation to categories that are of relevance to sentiment analysis. So, for this project, the terms and whether they were classified as negative in the lexicon was considered. Then a comparison of the available techniques used to analyse French text was undertaken, comparing their vocabulary.

As the sentiment measurement was to be externally evaluated with financial market data, modelling of macroeconomic trends was described. This looked at different approaches that aim to identify and quantify the factors that cause movement in the financial markets. In particular, the difference in model specification between structural and reduced form models. This was important to motivate the choices taken in model specification to evaluate the data. The economic relationship between France and Francophone Africa was then reviewed. This provides a context to the project, understanding the evolving nature of the bilateral relations between the two entities. It also detailed the large variation within the Francophone African countries in terms of their relationship with France and how it is currently changing. This caveat is important to take into consideration, as for all purposes in the analysis carried out in this project the countries are treated as a single entity.

Methodology

In this section, an overview of the project structure is given. The specification of proxies for the negative sentiment of Francophone Africa towards France and the French financial market was detailed. A proxy is a representation of something that is not measurable, it

serves as a measurable entity of the variable of interest. The data collection and processing of these proxies was described in order to create time series to model their relationships. Data collection involved filtering data collected to match the definition of the proxies. For the proxy representing news media from Francophone Africa towards France, it was important to ensure that the topic of the articles was indeed France. As for the French financial market proxy, it was important to detrend the data in order to meet the assumptions of the statistical evaluation.

A Vector Autoregressive Model (VAR) was used to model the impact of the negative sentiment proxy on the financial market proxy. Particular caution to assumptions made and requirement for accurate model estimations were adhered to. This involved ensuring that all the time series were stationary, removing the presence of autocorrelation from financial market data and accounting for heteroscedasticity to improve the robustness of the model estimations.

Evaluation

In presenting an analysis of the sentiment proxy, an acknowledgement of the limitations in aggregating all of the countries was made. Then, in methodically presenting the results of the statistical evaluation of the data, a further understanding of the properties of the data and their interactions was quantified. The stationarity assumption of the time series in the model is tested, and the absence of autocorrelation in returns is shown. Additionally, the impact of the negative sentiment proxies measured by 5 different sentiment lexicons is evaluated and compared.

Conclusion

Finally, the results of the project are discussed, thus quantifying the impact of negative sentiment from Francophone Africa towards France and its impact on the French financial market. Limitations and future work are also discussed in relation to the curation of the sentiment proxy.

2 Literature Review

2.1 Sentiment Analysis

Framing is a key element in understanding sentiment analysis. Framing denotes a subjective value that can be attributed by the manner in which information is presented. Sentiment analysis aims to quantify this subjective value that has been added. The understanding of how sentiment can be a factor that influences investor behaviour provides a theoretical basis in interpreting possible results acquired from the statistical evaluation.

Subjective information can irrationally impact the actions of investors. Thaler's [3] research showed that one's adversity to risk can be swayed by how something is framed. The subjective behaviour towards money contrasts with the notion of fungibility. This is the key concept underpinning our monetary system that denotes that a dollar will always be a dollar. Thaler demonstrated that this principle does not always hold in analysing patterns in the irrational way in which people use their money. He reasoned that the way in which money is framed leads people to hold mental accounts of money. This personal organisation of money reflects a consciousness of self-control. Suboptimal approaches are exercised in order to minimise risk, such as relinquishing discretionary contributions in pension plans. This behaviour, though anomalous to economic theory of personal financial management, reflects the condition of human irrationality with regard to handling money. Therefore, we can understand that just like economically irrational pension contributions, investors can exhibit irrational behaviour towards the stock market.

2.1.1 A Comparison of Approaches

The measurement of qualitative data, like sentiment, is a task that has been undertaken with different approaches, varying from machine learning to lexicons to hybrids of the two. In comparing the two approaches, before specifying the method by which sentiment is measured, an informed selection as to which approach is more appropriate for the analysis at hand was made. Machine learning approaches aim to spot patterns in data to classify sentiment, whereas sentiment lexicons are specified *a priori* in order to serve as references when measuring sentiment.

Machine Learning

Machine learning approaches when implemented correctly can serve as strong sentiment analysis tools. Qi et Al. [4], compared the effectiveness of sentiment classification of various machine learning classifier such as Random Forest, Naïve Bayes, and Support Vector Machine (SVC) to classify the sentiment of tweets. They found that the SVC model using a Bag of Words (BoW) approach to be the most accurate when compared with the other machine learning and lexicon approaches. However, they acknowledged the limitations of employing a lexicon approach for the data that they were analysing. They recognised that the lexicons that they used had not necessarily been designed to handle informal textual data such as that found on Twitter. They also note the relatively restricted domain of the data analysed of tweets by English users on the topic of Covid-19. These two limitations recognised in their research, underline the key consideration of the appropriateness of a lexicon to the text it analyses in sentiment analysis.

The advantage of using a machine learning approach is that words can be represented in their context. Using word embeddings like word2vec in accompaniment with sophisticated statistical learning models such as neural networks, the context of words and their sentiment classification can be learned.

Lexicon

In the BoW approach to sentiment analysis, lexicons are used to assign sentiment to words. BoW unlike machine learning does not consider context but merely occurrences of terms. However, Lexicon approaches, unlike machine learning approaches, do not require huge data sets to identify patterns of classification, as is the case with unsupervised machine learning models. Then, with regard to supervised models, the production of annotated data required to train a proposed model is a laborious endeavour. As for using readily annotated data, this choice limits the possible implementations of the model relative to the domain the data has been trained on.

On the other hand, lexicons can be implemented for the task at hand with relative technical ease, essentially embodying a dictionary lookup as functionality. Lexicons also have the advantage of being easily interpretable. If one doubts the classification of sentiment by a lexicon, the doubt in question can be easily verified in examining the assignment of sentiment of individual terms in the lexicon. For lexicons, the terms and their assigned polarity are often derived from linguistic theory and additions to such lexicons are reviewed manually.

If one were to use machine learning techniques to measure the sentiment, the task would become much more complicated in the design of complex machine learning architecture. The use of lexicons however is a well researched topic in measuring sentiment and its impact on

the stock market [2,5]. This permits the testing of novel hypotheses and the reproduction of results to be straightforward. However, where appropriate, machine learning models can be very powerful when dealing with textual data. That is why, for the task of lemmatization and tokenization, I did make use of an appropriate pre-trained language model.

2.1.2 An Overview of the Lexicons Used

The lexicons employed in this project vary by vocabulary, sentiment assignment and language. 5 different sentiment analysis lexicons were selected to compare their performance in capturing the relationship between sentiment and financial markets. As of the date of completion of this project, I could not find any sentiment lexicons that were specifically designed to measure sentiment of French language text. The Lexicons most adapted to readily analyse French language text were English language lexicons that had been professionally translated to French. The lack of development of new French language specific lexicons could indicate a trend in favour of machine learning research for sentiment analysis. For example, the French language, large language model CamemBERT [6] which was released in 2020 with sentiment analysis functionality.

Coming back to the lexicons used, two lexicons were professionally translated French lexicons, whereas the three others were English language lexicons that were translated automatically for the implementation of this project. The contrast of automatically and professionally translated lexicons serves as a way to investigate whether automatically translated English lexicons can be used to accurately analyse sentiment. This is evaluated externally by evaluating their capability of measuring negative sentiment in relation to its impact on financial markets. Next, some relevant properties of each lexicon are analysed.

Harvard General Inquirer (GI)

The GI [7] is widely used for sentiment analysis tasks and has been used in previous research capturing the relationship between sentiment and financial markets [2,5]. It provides disambiguation functionality to discriminate sentiment attribution of homographs in English. It also contains a wide array of other categories of interest for each of its terms, however, they were not considered in this project, as the lexicon was only used to measure negative sentiment.

Loughran and McDonald (LM)

The second lexicon selected was that proposed by Loughran and McDonald [8] which claims that the GI lexicon misattributed negative sentiment to almost 75% of the words examined in a financial context. Thus, they proposed their own lexicon to be more suitable for

accurately categorising sentiment in the financial domain. This lexicon was therefore selected to capture negative sentiment of news media relating to finance.

Oil and Economics (OIL)

The final English language lexicon used was the Oil and Economics lexicon specifically designed for financial forecasting and contains oil industry related terms. This lexicon was also selected for use to see how it compared to the other lexicons. This lexicon was provided by my supervisor, Professor Khurshid Ahmad of Trinity College Dublin.

Lexicoder (LEX)

One of the professionally translated lexicons was the French version of the Lexicoder English language lexicon. This lexicon was evaluated externally with sentiment from public opinion on elections in Quebec. As well as this, it was internally evaluated with an annotated data set by Duval and Pétry [9].

FEEL

Finally, the FEEL lexicon [10] was conceived in validating an automatic translation of the English Emotional Lexicon NRC-Canada [11] where the terms' polarity was derived through human crowdsourcing. The paper also provided a table of the currently available French lexicons at the time that the paper was published in 2016. This detailed 4 lexicons, of these, two used automation in their compilation of terms in the lexicon, one was not publicly available, and the final lexicon was used to analyse facial expressions. The automated compilation contrasts with the more careful derivation of terms by the other lexicons, and so these lexicons were not selected. The lexicon used for facial expressions was not applicable to textual data and the final lexicon was not publicly accessible, so none of these lexicons were used for the sentiment analysis in the project.

2.2 Econometric Analysis

Modelling macroeconomic trends is not an easy task as there are so many factors influencing the stock market. Bachelier [12] was the first to apply probability theory as a tool to analyse the financial markets, assuming that prices follow normal distribution in his thesis in the year 1900. His stochastic approach provided a framework to calculate the price of stock options. The prices of derivatives in the options market at the time were set via negotiations between buyer and the seller. Bachelier, however, provided a way to calculate the true price under the assumption that it was just as likely for the stock to increase as it was to decrease in a financial market. This idea that each price is independent of previous prices and that all previous information has been factored in predates the Efficient Market Hypothesis proposed

much later by Eugene Fama [13]. The properties of the fluctuations in stock market prices is of particular interest as it informs the curation of the data to be used in the statistical evaluation.

2.2.1 Structural and Reduced Form Models

To develop a macroeconomic model in order to test our hypothesis, we lean on the notion of vector autoregression (VAR) and multivariate time series analysis, first introduced by Christopher Sims in 1980 [14]. Previously, economists used *a priori* economic theory to undergo structural identification. Sims, argues that these models are over-identified and constructed based on false restrictions. However, Sim's reduced form model is estimated by solving for each independent variable. This approach allows us to explore relationships between economic variables without imposing strong theoretical assumptions. By using reduced form models, we can better understand the dynamic interactions within the system. Thus, any conclusions reached from analysing the data collected in the project can be justified empirically.

2.2.2 The Relationship between France and Francophone Africa

To justify the exploration of the hypothesis that there exists a relationship between the sentiment of the linguistic zone of Francophone Africa towards France and the French financial market itself, it is important to understand the economic ties that these two entities have. While many of these countries ascended to independence in the 1960s, strong economic ties remain. The analysis of the strength of the economic relationship between each of the constituent countries of the sentiment proxy demonstrates the potential variance in the weighting of sentiment towards France with regard to their possible impact on the French stock market.

Currency: The CFA Franc

One particular economic tie is observable in the form of currency. Taylor [15] discusses the maintained economic relations between France and Francophone Africa evoking the fact that 14 countries maintain membership of the Communauté Financière Africaine (CFA) economic zone and how France continues to exercise a certain level of influence in printing the currency, managing Treasury contributions and making changes without consulting the African member countries in only asking them to sign changes to proposed agreements. Of these countries, news media from 11 of the countries was collected. A synopsis of the current arrangement is available from the French Ministry for Europe and Foreign Affairs [16]. It details how the CFA Franc is a currency that has a fixed parity agreement with the Euro guaranteed by France and is printed at Chamalières by the Bank of France

since 1945. Within this monetary group, there are subgroups with varying levels of other economic obligations. Notably, the new agreement in 2019 between France and the UEMOA (West African Economic and Monetary Union) demonstrates the recent evolutions in the economic zone. Concerning the 8 west African countries that make up these unions, news media from all is collected for this project. An important economic change was ratified by the aforementioned agreement in 2019 ending the centralisation of reserves and contributions of these countries to the French Treasury. This demonstrates the continued but changing involvement of France with these countries.

Bilateral Trade

Outside of clear a clear economic relationship in relation to currency, there exists also strong trade relations between France and Francophone Africa. The French Treasury [17] provides an overview of the trade relations between France and the UEMOA in 2022, comparing it with other regions, including North African Francophone countries. France's trade with UEMOA increased by 9.8% in 2022, with exports up by 6.8% and imports by 18.6%. Côte d'Ivoire and Senegal are the main trade partners within the union. French exports to UEMOA are significantly higher than to CEMAC (Economic and Monetary Community of Central Africa) but much lower than to North Africa (€17.3 billion). Morocco is the leading North African trade partner (€6.5 billion), followed by Algeria (€4.5 billion) and Egypt (€2.2 billion) all of which are Francophone countries. The level of exports to UEMOA is comparable to that of Tunisia alone (€3.8 billion). France has a trade surplus with UEMOA, which remained stable at €2.3 billion in 2022.

These trade figures for the relevant year 2022 further describe that France's propensity to trade with Francophone countries was on the rise over the time frame examined in this project. This is captured in the elevated trade levels of North and West African Francophone countries when compared to those in Central Africa. These statistics are important to understand the potential impact that negative sentiment from these countries would have on the French stock market. Conceivably, countries with higher levels of trade and thus more substantial economic ties with France could have more of an impact on the French stock market. It also shows that while even though economic obligations of the UEMOA towards France have been weakened, the trade between the two increased afterwords.

Military

Equally, many of these countries have military ties and agreements with France. Including military bases in Senegal, Côte d'Ivoire and Djibouti, and 1,000 troops deployed in Chad according to the French Ministry of Defence [18]. However, in the President of France Emmanuel Macron's speech [19] before his visit to Central Africa, he acknowledged how France is open to changing its current military involvement in Africa. He laid out a

road-map whereby a more accompanying role is to be played by France, this involves reducing the number of French troops deployed and integrating their African counterparts into the existing military bases. While this recent speech does indicate France's willingness to change military policy in Africa, it also describes the continuing military involvement of France in Africa by the French president.

Change

As shown, the relationship between France and the Francophone linguistic zone is evolving. Its policy of a common currency with fixed parity to the euro, trade relations and military partnerships all remain, but have experienced changed in recent years, which is particularly relevant to the time frame examined in this project being 2022-2023.

2.3 The Impact of Sentiment on Financial Markets

The existence of a relationship between the financial markets and sentiment is often met with scepticism, stating that news media does not provide new information to investors and thus, the content can not have an effect in the movements of financial markets. However, Tetlock [2], showed that we can predict stock market movement to a certain degree using news sources as a proxy for investor sentiment. He demonstrated that words associated with pessimism (weak / negative) are associated with negative investor sentiment and thus prices fall accordingly, followed by a reversion to fundamentals. To do this, he used financial news media as a proxy for the investor sentiment and stock indexes as proxies for financial markets. Thus demonstrating that the content of media reports serves as a proxy for investor sentiment and not novel information regarding fundamental asset values. These irrational investors, motivated by factors unrelated to fundamental principles of investing, are why in financial markets, there exists a buyer for every seller. When the irrational investor sentiment is pessimistic towards a stock, they sell it, while arbitrageurs are there to buy it. As these arbitrageurs understand that this transient increase in negative investor sentiment towards a stock is not indicative of the fundamentals of said stock, they can, thus, expect it bounces back accordingly.

Ríordan [20] examined the impact of newswire messages on market quality, particularly focusing on how the sentiment of messages classified as either positive, negative, or neutral affect prices and trading behaviour. He found that negative messages convey more information than positive or neutral one's leading to higher adverse selection costs and a decrease in liquidity around negative news. This assertion that negative sentiment is more important when measuring the impact of investor sentiment is similar to that of Tetlock [2].

The decision therefore to measure only negative sentiment of the news media was taken to

measure its impact on the stock market using proxies.

3 Methodology

This section aims to both describe and justify the method used to investigate the hypothesis, while also providing enough technical detail for the method undertaken to be reproducible. It describes the process of selecting and curating data for statistical models, which requires a lot of attention to detail in order to satisfy the assumptions of the model and the specification of the variables.

To commence, data was collected in order to represent the sentiment and financial market proxies. The sentiment proxy was analysed using each of the 5 sentiment lexicons to measure negative sentiment from Francophone Africa towards France. The statistical significance of the impact of the negative sentiment proxy and the financial market proxy was then examined. To do this I developed a Python program to collect and process the data of the proxies and then used the GRETL econometric analysis tool to model the impact employing parsimonious vector autoregressive models. The GitHub repository containing the Python code can be found here:

https://github.com/jmcinern/CapstoneProject

3.1 Visual Overview

This section aims to provide a visual overview of the method used to carry out the project to provide context for the more detailed description of methodology that follows.
The project has 3 main parts, each with their own constituent processes that will be described in more detail in this chapter.

1. Constructing a Time Series of the Sentiment Proxy (in green).
2. Constructing a Time Series of the Financial Market Proxy (in yellow).
3. Evaluating the Impact of the Sentiment Proxy on the Financial Market Proxy (in orange).

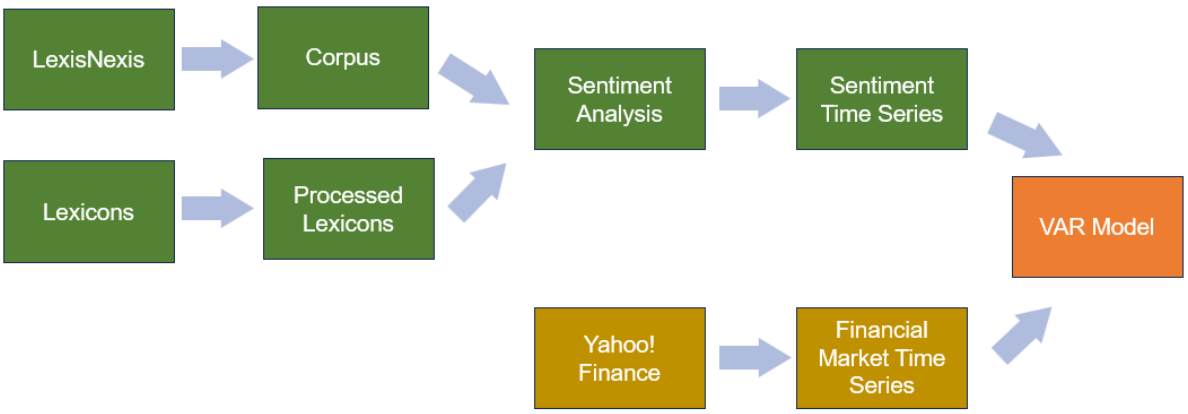

Figure 3.1: An Overview of the Project System

3.2 Sentiment Proxy

The aim of the data collection was to establish a corpus of French language newspaper articles from Francophone countries in Africa pertaining to France. This proxy represents a measurement of negative sentiment of Francophone Africa towards France. The proxy must therefore be:

1. Representative
The proxy ensures that the data collected covers the breadth of Francophone Africa proportionally and that the topic of the news articles is accurately classified as having the subject of France.

2. Measurable
The negative sentiment of the news articles is quantifiable for sentiment analysis using sentiment lexicons.

3.2.1 Data Collection

Data was collected from the LexisNexis News and Business Library, with access provided by Trinity College Dublin. LexisNexis provides the functionality of filtering sources selected before download. First, the source continent was selected as Africa, then for each country it was determined whether there were news sources available in the French language. The following list of 23 countries with at least one French news source is as follows: Algeria, Benin, Burkina Faso, Cameroon, Chad, Comoros Islands, Republic of Congo (ROC), Côte d'Ivoire, Djibouti, Egypt, Gabon, Guinea, Madagascar, Mali, Mauritania, Mauritius, Morocco, Niger, Rwanda, Togo, Senegal, Sudan, Tunisia. The attempt to identify the topic of France was specified by the presence of two keywords in an article: "France" AND "français", (*français* meaning French and France being the same in English). The logical

operator AND here is used in the LexisNexis keyword specifier to only show articles with both keywords present.

Considering that LexisNexis limits the number of downloads to 500 articles per day, the time span of 4 years recommended by Taylor [21] for a time series analysis was reduced to 2 years to provide ample time to complete the project. Data availability in these countries also appeared to reduce the further in the past that the time frame selected covered. A time span of starting from 01/01/2022 to 31/12/2023 inclusive was therefore used. This data collection took approximately 2 weeks. Data was available to download in PDF, RTF and DOCX format. DOCX was selected as PDF was slower to download and the RTF encoding prevented simple copy and paste to TXT file, which was the file format used in the Python program developed.

The Data Collected

The total number of articles of the timespan 2022-2023 was recorded as well as the number of articles with the keywords "France" and "français" for each newspaper, where the mean proportion of articles containing the keywords to the total articles was 2.8%. The total number of articles downloaded containing the keywords was 7,709 from 52 newspapers across the 23 countries. A table displaying the newspapers accessed and corresponding countries is available in the appendix A1.1 that displays the news media data collected. The newspaper names are displayed as they appear in LexisNexis where sometimes the source location, language or institution is included in the brackets after the newspaper name.

3.2.2 Data Preprocessing

Having downloaded the files of relevant news articles, I developed a program in Python to process these articles and create a corpus as an organised representation of the data collected.

Metadata

The program reads all the files, extracting the metadata in order to allow for manipulation and analysis of the data. Metadata was extracted using parsing techniques dependent on the formatting of the desired information in the file. This was done by identifying text patterns in the files using regular expressions (REGEX) and capturing the consistent patterns in the format that LexisNexis displayed the information. For example, The number of words in the article was provided consistently by LexisNexis in the same format and as such was extracted using the REGEX:

"Length: (\d+) words"

Where:
\d+ denotes the presence of 1 or more digits.

The title of the article was extracted based on its position as either 1 or 2 lines above the name of the newspaper. The date component was extracted in relation to its position following the newspaper name. Dates were parsed in both French and English with dynamic handling of different day and month ordering as neither the language nor the order was consistent across all files even though the language of the articles themselves was French. The text of the article was extracted as all text between the keyword "Body" and the upload date of the article.

Handling Data Noise and Silence

In any data collection process, it is important to be aware of the presence of noise and silence. Silence in this data collection would manifest in having a large proportion of days over the time frame where no data is available implying an absence of negative sentiment, when in fact there is just an absence of data. As such, a large set of articles were collected to prevent gaps in the time series. Whereas, noise would be observed by the presence of irrelevant newspaper articles downloaded where the topic is not France. To reduce noise in the data, the filter of articles of where the subject is France was tightened after downloading.

The program thus filtered also by the presence of keywords relating to France in the article titles. Additionally, as some of the data downloaded from LexisNexis contained duplicate news articles, these articles were identified by their title and removed accordingly. This careful processing and construction of the corpus is important to justify the use of this corpus as a proxy of the sentiment of Francophone countries in Africa towards France. The proxy must be representative of what it claims to represent. The following section further outlines the list of keywords proposed to classify articles as having France as topic.

France Topic Identification

As mentioned, further filtering was also applied to include only articles with specific keywords present in the title. It was noticed that the initial filter in downloading articles containing the keywords "France" and "français" was insufficient in determining that the subject of the article was in fact, France. The choice to utilise a list of keywords that ought to be present in the title was justified due to the elevated importance of the words contained in the title serving to represent the subject of the article.

The first method explored was to use the online French dictionary LeRobert as this dictionary provides not just a definition of a word examined, but also a list of synonyms.

However, there was no entry in the lexicon for France. Being also that as of the date of this project, no appropriate France related keyword list was found in the literature, the decision to propose my own list was taken. How to derive this list posed a question. A machine learning approach was considered, but was deemed to be outside the scope of the project. Instead, upon reading the titles and their corresponding articles, keywords that indicated that the subject of the article was France were identified. The following is the proposed list:

[“france”, “hexagone”, “paris”, “marseille”, “français”, “française”, “françaises”, “franco”]

The *hexagone* being a term used to describe French territory in mainland Europe. Then we have the two largest cities in France, being *Paris* and *Marseille*. Additionally, a complete list of the inflected forms of the demonym *français* meaning French was added to the list. Finally, the prefix *franco* which indicates a relationship with France was added. This further filtering reduced the amount of articles in the data collected that were not of the topic of France.

Handling Variance of Article Lengths

It was important to analyse the distribution of article lengths within the corpus for the calculation of an article sentiment score. Before any manipulation of the data, the standard deviation of the articles was 799 which was greater than the mean article length of 590. With a large kurtosis of 129, the data displayed a right-tailed non-Gaussian distribution. This is due to the variation in article style across newspapers, upon inspection of the large outliers with over 10,000 words, it was found that The Rwandan News Agency had published court transcriptions of a trial taking place in Paris. This variation across newspapers and articles if not unhandled would skew the sentiment analysis to overemphasise the sentiment of larger articles. Furthermore, as Excel has a maximum amount of text permitted per cell, in order to do a statistical analysis of the countries article counts, I removed outliers. An outlier threshold was determined as follows:

$$OT = 3\sigma + \mu$$

Where:
OT corresponds to the outlier threshold.

The threshold was 2460, so all articles greater than this were not included in the corpus. After implementing this removal of outliers and further topic classification filtering, the mean number of words per article was 499 decreasing from 590. Furthermore, the number of articles processed decreased from the original 7,709 downloaded to 3,363 articles with a total 1,678,267 words. The distribution of the token lengths after the preprocessing of the

data can be seen in figure 3.2 below.

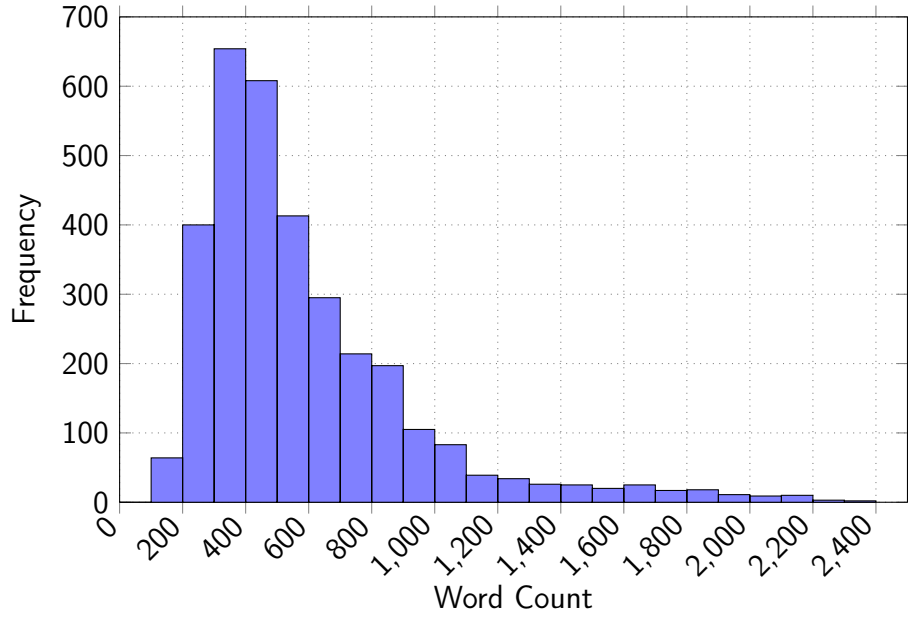

Figure 3.2: Histogram of Article Length Distribution

Text Preprocessing

The article text was processed to permit the computation of negative sentiment in the articles. Firstly, as the identification of terms in the lexicon sentiment was not case-sensitive, so the text of the article was normalized to lower case. The text was then tokenized, as the sentiment analysis employs a Bag of Words approach. The text was finally lemmatized to account for lexical variation, which will be described in more detail later. Tokenization and lemmatization were accomplished using the spaCy fr_core_news_sm (release 3.7.0) language model. This free, open-source model was pretrained on French newspaper data and was therefore appropriate to use with the data collected. According to the release details, the model has a tokenization accuracy of 99.8% and a lemmatization accuracy of 90.8%.

3.2.3 Analysing Differences in Lexicon Vocabulary

As multiple lexicons are being tested in measuring the sentiment proxy, it is important to analyse some differences between them. The lexicons all share the common function of measuring sentiment in assigning sentiment at term level. The lexicons, however, differ in terms of vocabulary used. This leads to differences in the count of words in the lexicon that assign sentiment and also their negative sentiment proportion. The lexicons were analysed after having been preprocessed and can be seen in table 3.1.

Table 3.1: The Proportion of Negative to Positive Terms in Each Lexicon

Lexicon	Positive Count	Negative Count	Negative Proportion
FEEL	8237	5554	40%
General Inquirer	1620	1931	54%
Lexicoder	1188	2572	68%
Loughran McDonald	126	820	87%
Oil and Economics	265	1605	86%

As my aim is to measure negative sentiment, the observation that some of the lexicons skew negative is not a concern. What is interesting is the actual counts of negative sentiment affect terms. This number of terms in each lexicon impacts the number of words in the text that are identified as having sentiment (relative to term frequency). The goal of the lexicon is to accurately measure sentiment, but this is difficult to achieve when the meaning of terms can vary in relation to context. The use of too many terms can lead to noise in the analysis, whereas too few can lead to silence. Taking the example of the FEEL lexicon, which boasts over 13,000 terms, potential noise can be identified upon a quick inspection. With words such as *jour* meaning *day* in French classed as positive, the question of precision is brought into question. A word such as *day* is a very common word which does not seemingly appear to be of positive affect. The domain relevance of the terms in relation to the hypothesis is also considered.

Analysing the Proportion of Vocabulary of the Lexicoder (LEX) Lexicon Appropriate for Financial Media Sentiment Analysis

Another metric to examine is the aptness of the lexicons to analyse sentiment of a financial nature. As in this sentiment analysis, while the topic is France and not necessarily financial news relating to France, a lexicon with terms designed to analyse sentiment of financial news media could potentially better capture sentiment relevant to investors.

I chose the LEX lexicon by process of elimination. The classification of Financial terms by the GI lexicon has already been examined by Loughran and Mcdonald, ruling these two out for novel vocabulary understanding. The FEEL lexicon has over 13,000 terms and appears to have noise in the data in relation to sentiment assignment, so an analysis of its finance related terms appears to be less pertinent. As for the OIL lexicon, this lexicon is specifically designed for financial analysis, so a further investigation of the proportion of financially related terms in the lexicon does not seem of great importance. However, the assumption that the OIL and LM lexicons contain vocabulary selected for financial sentiment analysis is

useful to determine the proportion of the LEX dictionaries terms that are relevant for sentiment analysis in a financial context. Thus, to classify the LEX terms, I used the French language translated LM and OIL lexicons as reference. This provides an external justification of the French terms in relation to their classification as financially related or not, saving a manual classification of the over 2,000 terms.

To do this, the lexicons are represented in Python as sets of unique terms. The set of LEX terms explicitly suitable to financial analysis is the terms present in the LEX lexicon that are also present in either the LM or the OIL lexicon.

$$LEXFIN = LEX \cap (OIL \cup LM)$$
$$proportion = \frac{|LEXFIN|}{|LEX|}$$

Where:
LEXFIN corresponds to the set of terms in the Lexicoder Lexicon also in either the OIL or LM lexicons.
OIL corresponds to the set of terms in the Oil and Economic Lexicon.
LM corresponds to the set of terms in the Loughran McDonald Lexicon.

Upon calculation, 55% of the terms in the LEX Lexicon are present in the two other lexicons previously identified as suitable for sentiment analysis of financial media. Limitations to this approach instead of manual by hand classification is that some of the words presented in their stemmed form in the LEX lexicon were not accurately lemmatized using the spaCy lemmatization function and thus were not matched in the other dictionaries. Also, as neither LM nor OIL are exhaustive in listing terms potentially applicable to financial sentiment analysis, this can be interpreted as at least 55% of the terms in the LEX lexicon are suitable for financial sentiment analysis.

3.2.4 Preprocessing Multiple Sentiment Lexicons

The goal of processing the lexicons is to render them interpretable by the same sentiment analysis function. This involves ensuring that the language of all the lexicons is French, that the detail of all the lexicons remains after manipulation, and that the format of all the lexicons is the same.

Common Formatting

All the lexicons were reformatted to obtain a common formatting, which enabled sentiment analysis of the articles using the same function regardless of the lexicon in use. The format was based on the General Inquirer format, where the first column represent the terms of the

lexicon. There also, then, exists a negative sentiment column, where terms are indicated to be of negative polarity. The Python pandas library proved particularly useful in manipulating the lexicons stored as Excel files in representing them as data frame objects.

Automatic Translation

Of the 5 different lexicons were used to measure the sentiment in the articles, 3 of the lexicons had the source language English and 2 French lexicons were selected. The 3 English lexicons were translated using the automatic GOOGLETRANSLATE function on Google Sheets. The column of English terms was translated to generate a new column of French terms at the same index, maintaining the assigned polarity information of the term.

Handling N-Grams

As the LEX, GI and OIL lexicons contain n-grams (multiple word terms) in their term list, not only unigrams were considered but n-grams of up to 5 tokens. This renders the processing of the lexicons like LM and FEEL to not be optimal, as token combinations in the text were considered that are impossible to be found in the lexicon were looked up. However, this did not have a significant effect on the time complexity of the program.

Limitations in Lexicon Preprocessing

The disambiguation functionality of the General Inquirer lexicon depreciated following automatic translation. For example, the English lexicon differs between 5 different interpretations of the word *content*, varying by syntactic part of speech tag and also word meaning. The French translation however is always the same, being *contenu* meaning the noun *content* in the sense of *constituent elements of information*.

As for the LEX lexicon, as of the completion of this project the domain Lexicoder.com is no longer in use to conduct the sentiment analysis, so the raw terms were then processed. Some of the words had an asterisk as the final character of the word, which I identified as indicating the stem form of the term. To handle this, and lemmatize the term, the asterisk was removed, and the term was lemmatized using the spaCy language model. Terms that are already present in their lemmatized form are left unchanged. A limitation of this approach is that the model sometimes does not recognise the stemmed form of the word, which prevents lemmatization. This lemmatisation is then applied to the other lexicons in order to provide consistency in the sentiment analysis.

Accounting for Lexical Variation in French

Just like the corpus, the lexicons were lemmatized to account for variation in terms at a lexical level. This prevented the sentiment lexicons from under classifying inflected forms of

words. In examining lexical variation of French adjectives, variation in relation to the gender and number of the noun that they modify is observed, which is not the case in English. Take for example, the English word *bad* which is automatically translated to the default masculine singular inflection *mauvais*. Thus, the plural *mauvaises* and the feminine *mauvaise* will not be classified even though they all point to the same lexeme.

This lemmatization process does have limitations as the syntactic and semantic context of the word is not considered by the language model as the words are presented in the lexicons as a list and not as occurrences in natural language. This can be observed in the processing of the word *dû* in French, meaning a *debt* marked as negative in the LEX lexicon. *Dû* is also the past tense of the modal verb *devoir* meaning *to have to/ must* in English. Therefore, the spaCy language model without the context of *dû* as a noun lemmatizes the term as *devoir*. This causes the analysis using the LEX lexicon to falsely attribute negative to all inflected forms of *devoir* in the news media articles, which is a frequent misclassification error.

3.2.5 Calculating Negative Sentiment

To analyse the sentiment at article level, sentiment lexicons were used in order to assign polarity to the words in the article. The lexicon was represented as a python dictionary object to facilitate constant time look up for determining the sentiment of words. The negative sentiment for an article was then calculated as essentially the average negative sentiment of each word in the article. For each word in the article, a classification of *negative* or *none* is attributed to each word in the article with reference to the sentiment lexicon. The count of negative words is then divided by the total number of words in the article to generate a sentiment score representing the proportion of negative sentiment in the article. The score is then normalized by article length to prevent the magnitude of negative sentiment to be related to the article length. Then, as there are often multiple articles per day due to the proportion of articles collected and days of sentiment analysed, sentiment at a given date was computed as the mean of the sentiment across all articles for that day. In order to see this computation of the negative sentiment per day more clearly, the following mathematical overview of the process is outlined.

To start, the relevant elements and their corresponding representation are described:

- N as the count of negative words in an article.
- T as the total number of words in the article.
- S as the sentiment score, representing the proportion of negative sentiment in the article.
- D as the set of all articles for a given day.

- $\text{Sentiment}_{\text{date}}$ as the sentiment at a given date.

Then, the equation for computing the sentiment score for each article is:

$$S = \frac{N}{T}$$

To compute the sentiment at a given date, we calculate the mean sentiment across all articles for that day:

$$\text{Sentiment}_{\text{date}} = \frac{1}{|D|} \sum_{\text{article} \in D} S_{\text{article}}$$

Where $|D|$ represents the total number of articles for that day.

Optimisation of the Program Developed

As the program required the processing and analysis of a large amount of data. The decision to serialize some of the python objects generated was taken. Serialization is the process of storing these objects as files. I used the Pickle module in python, this permitted me to encode the objects as a byte stream and subsequently decode them when running the program. The corpus and all the lexicons were serialized in order to reduce the time complexity. Consequently, the time taken to run the sentiment analysis and construction of both the sentiment and financial market time series proxies was reduced by 94% from 13 minutes and 41 seconds to just 49 seconds.

3.2.6 Representation of the Sentiment Proxy

So, having constructed the time series for proxy of negative sentiment. It remains pertinent to assess whether each of the countries is represented proportionally. To do this, counts of articles contributed to the time series by countries are calculated and divided by the total count of articles. The magnitude of the bilateral relationship between each country and France is not considered. This political and economic analysis was outside the scope of this project. Instead, the countries are aggregated to produce one single source aiming to represent Francophone Africa.

3.3 Economic Proxy

The aim of the construction of a time series for the financial market proxy, just like that of the sentiment proxy had to be:

1. Representative
The proxy should represent the movements of the French stock market.
2. Measurable
The proxy should be measurable and suitable for the statistical analysis.

To model the fluctuations of the French stock market, the CAC 40 index was selected as a proxy. The CAC 40, or *Cotation Assistée en Continu* is a benchmark French stock market index representing the top 40 stocks among the 100 highest market capitalizations in France. The French Ministry of Economics and Finance [22] provides an overview of its history and current function, where the dates and figures have been accessed for this section. The CAC 40 was established in 1987 and has been operational since June 15, 1988, it reflects the performance of the French economy through its listed companies. The index is weighted by market capitalization on the NYSE Euronext, with a limit of 15% per company to maintain diversification.

The CAC 40's composition is regularly updated to ensure sector representation, with changes decided by the independent scientific council of NYSE Euronext. It serves as a basis for various financial products and is updated every 15 seconds during trading hours. The index's value fluctuates continuously, indicative of the general trend of France's major companies' market performance.

Since December 1, 2003, the CAC 40 has used free-float market capitalization, aligning with major global indices such as the S&P500, favouring companies with more shares available for purchase. This switch to a free-float market capitalization metric renders the CAC 40 index more suitable to evaluate the impact of negative sentiment. This is due to the fact that free-float market capitalization considers the amount of shares a company has available in the market. Therefore, public opinion is weighted more heavily than private investors in valuing the companies.

The CAC 40 is crucial for investors, reflecting the broader market trend and serving as a foundation for financial products tracking its performance.

3.3.1 Data Collection

Data of the CAC 40 index was collected using Yahoo! Finance to access historical data. A time period of two years from 2022 to 2023 was selected to match that of the sentiment proxy. This process was more straightforward than the data collection for the sentiment proxy, owing to the fact that the data collected was qualitative and thus more easily measured.

3.3.2 Data Preprocessing

Once downloaded, the financial market data was manipulated to make it suitable for the statistical evaluation. One of the key steps was to compute a time series of returns from the time series of prices.

Financial Returns

Fluctuations in prices stock market prices can be represented by returns. The returns of a stock market index for a given day is essentially the profit or loss an investor incurs if they buy the index one day and sell it the next. Thus, returns are calculated in the data set as:

$$r_t = Log(p_t) - Log(p_{t-1})$$

More precisely for the data, the return for a day was calculated as the difference in the natural logarithmic price of the CAC 40 index between that day's close and the previous day's close, not considering derivatives. The price was always the closing price to keep the return calculation consistent. Then, weekends and bank holidays where no trading occurred were removed from the data set. Returns are examined instead of prices due to the stationarity and autocorrelation of prices. Taylor [21], noted that his price correlation obscures the changes in the market which we would like to measure. This detrending of data is important to ensure that time series have the property of stationarity.

Stationarity

If a time series is stationary, statistical properties like standard deviation and mean do not change over time. This permits the covariance being measured between variables to not be influenced by seasonality or trends in the time series, and thus accurately estimate the coefficients. Therefore, all the time series were tested for stationarity using the Augmented Dickey-Fuller test for up to 5 lags. The null hypothesis of this test is that the time series has a unit root, implying non-stationarity. The tests were run using the GRETL console and the following command:

$$? \quad adf \quad 5 \quad TimeSeriesName$$

Where:
adf specifies the Augmented Dickey Fuller test.
5 specifies the number of lags.
TimeSeriesName corresponds to the name of the time series being tested.

3.4 Statistical Evaluation

From the data, I run an initial Pearson correlation test for the time series of closing prices recorded and the time series of returns (the change of these prices).
The test for autocorrelation involves a Pearson correlation test between a time series of today's prices with yesterday's prices, and the same for returns. The Pearson coefficient ranges from -1 to 1, with 1 being highly positively correlated and 0 being no correlation. The covariation of the series is divided by the standard deviation of the series to get the coefficient representing the strength of the relationship.

The following equation calculates the Pearson coefficient in testing for autocorrelation of a time series S, the formula is the same for testing the presence of autocorrelation in the time series of returns and the time series of prices:

$$r = \frac{\sum_{t=2}^{T}(S_t - \hat{S})\sum_{t=2}^{T-1}(S_{t-1} - \hat{S})}{\sqrt{\sum_{t=2}^{T}(S_t - \hat{S})^2 \sum_{t=2}^{T-1}(S_{t-1} - \hat{S})^2}}$$

Where:

T represents the number of observations present in the time series.
$\hat{S}$ represents the mean of the time series.

The time subscript of the S vector shows that the formula is testing for autocorrelation, as the two vectors differ only in their index accessed, representing the one observation lag.

3.4.1 Linear Regression

Linear regression models the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting line that minimizes the difference between the observed values of the dependent variable and the values predicted by the model. The equation of the line in a linear regression model illustrates this.

$$y = a + mx + \epsilon$$

Here, the dependent variable y is estimated with relation to the data. a represents the y intercept (where all independent variables are 0).

The independent variable m represents the rate of change of the independent variable x. When interpreting m we can make inferences of the relationship between the independent variable(s) and the dependent variable. The sign of m, being either positive, negative or

zero, corresponds to the direction of the relationship. If m is positive, then the relationship between the independent variable and the dependent variable is positive, meaning that an increase in the independent variable is reflected with an increase in the dependent variable. The inverse is true for a negative slope, whereas no change in the dependent variable is reflected by the dependent variable if m is 0. The magnitude of m corresponds to the strength of the relationship, a significant magnitude of m corresponds to a strong relationship that is not the result of random variation in the data.

The error term ϵ indicates the level of noise in the model, if the error term is zero, the model does not systematically over or underestimate the dependent variable. If the error term is not zero, this indicates heteroscedasticity. The presence of heteroscedasticity in a model undermines the statistical significance of the coefficient estimation, as it is an assumption of the model. Properties of the data such as having large variation can cause heteroscedasticity, this is why returns are calculated as the difference in the natural logarithm of prices.

Another useful measure of how well the model fits the observed data is R^2, the coefficient of determination. It describes how much of the variation of the dependent variable is described by the independent variable(s). It ranges from 0 to 1 however negative values are possible if the model is worse than just using the mean value of the dependent variable.

3.4.2 Vector Autoregression (VAR)

VAR, like simple linear regression, serves to model relationships between variables. However, its assumption of stationarity is what makes it appropriate for time series analysis. VAR owing to its property of extrapolation is appropriate for financial market forecasting by estimating the dependent variable as a linear combination of past values of itself and other variables.

A VAR model of lag p often notated as a VAR(p) model has the following equation which demonstrates its use of lagged variables. All VAR models used Newey-West [23] standard errors by selecting *HAC* in the GRETL program with a bandwidth of 5 and Bartlett kernel. This takes into account both heteroscedasticity and autocorrelation when estimating standard errors for the regression coefficients.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \epsilon_t$$

VAR, like linear regression, allows the use of multiple variables. This permits the analysis of the impact of multiple time series on the dependent variable in question, including the dependent variable itself.

3.4.3 Model Specification

The following section justifies and describes the VAR models used to evaluate the thesis of this project.

Testing for Autocorrelation of Returns

It is important to ensure that returns are not autocorrelated in the VAR model with up to 5 lags so that relationships with other independent variables can be evaluated. Therefore, a VAR model evaluating the autocorrelation of returns is specified in order to justify the subsequent models.

Testing Correlation of Negative Sentiment Time Series

In order to inform the specification of the statistical models, the correlation between the negative sentiment time series was tested. Initially, it was considered to include all 5 time series of the negative sentiment proxy measured by the lexicons as independent variables. This, potentially, would have allowed the determination as to which sentiment lexicon had measured the most significant relationship of the negative sentiment proxy with the financial market proxy.

However, the fact the lexicons were all used to measure the negative sentiment from the same text brought into question their independence. If the independent variables show a strong linear dependence with each other, this would indicate the presence of collinearity in the model. Collinearity of independent variables in a model can lead to large standard errors and inaccurate coefficient measurements. As such, a Pearson correlation test between each of the time series of the negative sentiment proxy was conducted in order to justify the model specification, whereby each proxy negative sentiment time series was tested individually.

System of Models

To explore the relationships between the various time series, 6 parsimonious VAR models were run in the following order. Model (1) had to be evaluated first, whereas the order of models 2-6 is arbitrary. A lag value of 5 is chosen in accordance with the previous studies [2, 5] to represent at least one trading week.

Model (1) tests for autocorrelation of returns. This is to verify the VAR assumption that returns are not autocorrelated and that values are independent over time.

$$(1) \quad r_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i r_{t-i} + \epsilon_t$$

Model (2-6) evaluate the impact of the negative sentiment proxy on the financial market proxy relative to the lexicon used to measure the negative sentiment. Each equation below represents a different VAR model, whereby the negative sentiment proxy has been measured using a different sentiment lexicon. This allows the comparison of each of the sentiment lexicons in capturing a relationship with the negative sentiment proxy and the financial market proxy. Each model estimates different coefficients for the negative sentiment proxy. This can be seen below, represented by the coefficient preceding the negative sentiment variable, which indicates which lexicon has been used.

$$(2) \quad r_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i r_{t-i} + \sum_{i=1}^{5} \gamma_i LM_{t-i} + \epsilon_t$$

$$(3) \quad r_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i r_{t-i} + \sum_{i=1}^{5} \omega_i OIL_{t-i} + \epsilon_t$$

$$(4) \quad r_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i r_{t-i} + \sum_{i=1}^{5} \delta_i FEEL_{t-i} + \epsilon_t$$

$$(5) \quad r_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i r_{t-i} + \sum_{i=1}^{5} \zeta_i GI_{t-i} + \epsilon_t$$

$$(6) \quad r_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i r_{t-i} + \sum_{i=1}^{5} \lambda_i LEX_{t-i} + \epsilon_t$$

Where:
r corresponds to the time series of returns.
FEEL corresponds to the FEEL lexicon.
GI corresponds to the General Inquirer lexicon.
LEX corresponds to the Lexicoder lexicon.
LM corresponds to the Loughran and McDonald lexicon.
OIL corresponds to the Oil and Economics lexicon.

4 Evaluation

The representation of the sentiment proxy from Francophone Africa towards France was first evaluated. Then the VAR models were used to evaluate the significance of the impact of one or multiple time series on another. Attention was paid to verifying the assumptions of the VAR model in order to ensure the reliability of statistical results.

4.1 Proportionality of the Corpus to the Sentiment Proxy

Figure 4.1 and figure 4.2 shown on the following page display the Francophone countries with their number of articles, showing the proportion of countries' article count in relation to the total count of articles in the corpus. Geographically, it shows the tendency of Francophone African countries to be located to the north and the west of the continent, reflecting the linguistic zone. The 3 largest contributing countries to the corpus, Mali, Morocco and Tunisia make up 46% of the total corpus.

With regard to its aim to represent the sentiment of Francophone Africa in a measurable capacity, the proportionality of the sentiment proxy and thus precision of representation of the corpus is brought into question. This is seen in the variation of each countries' contribution to the corpus, this limitation was accepted in the evaluation of the hypothesis.

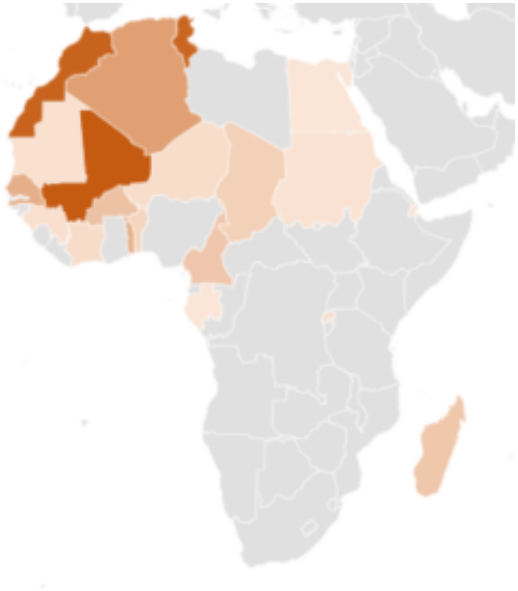

Figure 4.1: Map View Proportion of Article Count by Country in Corpus

Country	Count	Proportion (%)
Mali	549	16%
Morocco	515	15%
Tunisia	502	15%
Algeria	277	8%
Senegal	217	6%
Republic of Congo	188	6%
Togo	213	6%
Cameroon	127	4%
Madagascar	126	4%
Burkina Faso	147	4%
Chad	87	3%
Benin	72	2%
Mauritius	71	2%
Comoros	49	1%
Côte d'Ivoire	42	1%
Guinea	29	1%
Mauritania	28	1%
Niger	38	1%
Rwanda	38	1%
Sudan	19	1%
Djibouti	12	0%
Egypt	9	0%
Gabon	8	0%
Total	3363	100%

Figure 4.2: Proportion of Article Count by Country in Corpus

4.2 Stationarity of the Time Series

A visualisation of the seasonality and trend in prices over time is displayed in figure 4.3 on the following page. Here it can be observed that for prices, the mean value between 2022 and 2023 is not the same, indicating an upward trend in the prices over time and therefore a change in the mean over time. Where the price time series visually emulates a sinusoidal wave, the returns times series resembles that of white noise. Stationarity can be examined visually in time series plots like this, but it is important to statistically test for it in order to have a higher certitude of the assumption.

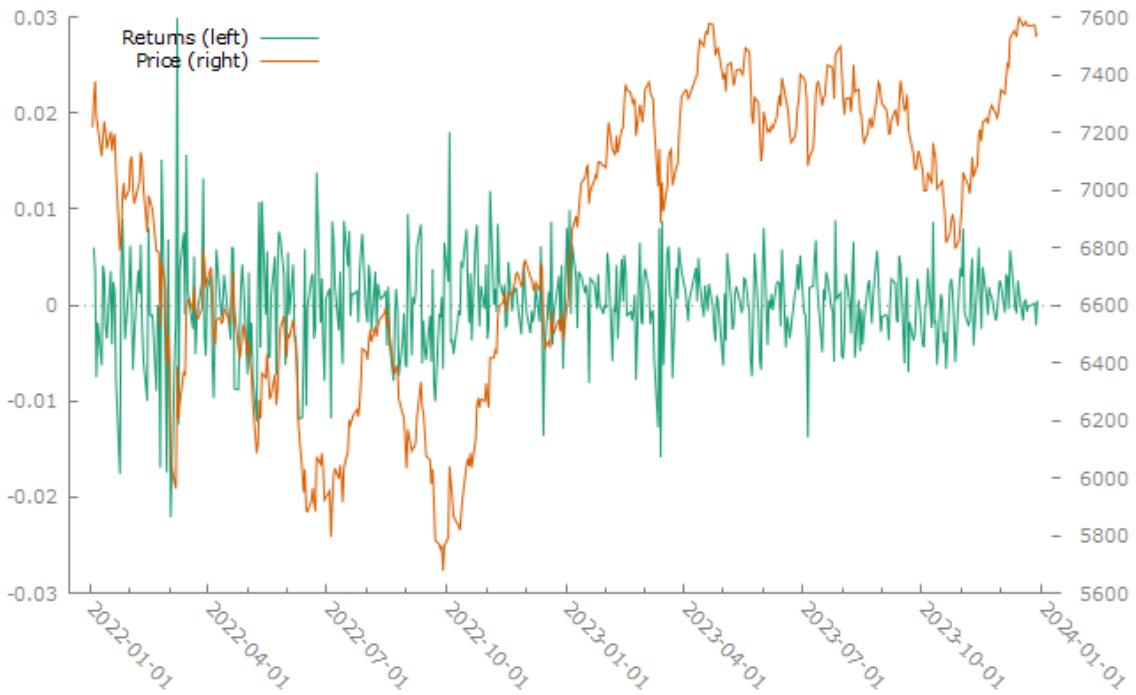

Figure 4.3: 2022-2023 CAC 40 Index Time Series of Prices and Returns (Daily)

The following table 4.1 displays the results from the Augmented Dickey-Fuller (ADF) test. The null hypothesis is that the time series has a unit root and is, thus, not stationary. The p value shown in the table is asymptotic, meaning that it represents the probability that the null hypothesis is true using an approximation of the true distribution. The prices time series is also displayed to provide an example of when the null hypothesis can not be rejected, indicating stationarity. Therefore, in examining the p values of the other time series which will be included in the VAR models, the ADF test indicated with strong confidence that these time series are stationary. This justifies their use in the VAR model, which assumes stationarity of variables.

Table 4.1: ADF Stationarity Test Results

Time Series	**P value**
Prices	0.5702
Returns	6.28×10^{-16}
LEX	7.16×10^{-15}
LM	3.18×10^{-15}
OIL	6.05×10^{-15}
GI	3.14×10^{-13}
FEEL	4.19×10^{-15}

4.3 Correlation of Negative Sentiment Time Series by Lexicon

The correlation matrix displayed in table 4.2 shows that the proxy negative sentiment time series are highly correlated. This justifies the choice of specifying VAR models, whereby the impact of each time series measured by each lexicon is evaluated separately to avoid collinearity of independent variable. This correlation matrix also serves to evaluate the similarity in sentiment analysis across the lexicons. The LM lexicon provides the most independent analysis, only showing particularly strong correlation with the OIL lexicon which could be due to both lexicons' vocabulary being specified for use in financially related sentiment analysis. The rest of the lexicons appear to have quite strong positive correlations with each other, indicating not much variance in their sentiment analysis of the text.

Table 4.2: Correlation Matrix of Negative Sentiment Measurements by Lexicon

	FEEL	GI	LEX	LM	OIL
FEEL	1.0000	0.7822	0.8368	0.6037	0.7442
GI		1.0000	0.8004	0.6759	0.7953
LEX			1.0000	0.6361	0.7635
LM				1.0000	0.8044
OIL					1.0000

4.4 Autocorrelation of Returns

As for the proxy of the French financial markets, the use of the CAC 40 index proved appropriate in representing the French financial markets and measurable in the availability of quantitative data.

As shown in the table 4.3 prices are positively correlated with yesterday's price whereas there is no significant correlation of returns. Here, Taylor's [21] justification of using returns instead of price is initially examined as the financial metric of choice in time series analysis, owing to the property of not being autocorrelated from one day to the next.

	Prices	Returns
Pearson Coefficient (r)	0.99	-0.03

Table 4.3: Correlation Coefficients for Prices and Returns

VAR model (1), only including the financial returns of the CAC 40 index, was evaluated further assessing the autocorrelation of the returns with a lag order of 5.

This is a key step in vector autoregression, as the time series are assumed to not be autocorrelated in order to accurately characterize the coefficients of the independent variables. The reason therefore that this vector autoregression was evaluated first was to provide validity in the coefficient estimation of the subsequent models. The following equation represents the modelling of the returns time series at a given time in relation to the 5 previous values of returns in the series.

(1) $r_t = \alpha_0 + \sum_{i=1}^{5} \alpha_i r_{t-i} + \epsilon_t$

Where:
α_0 corresponds to the Y intercept, i.e., when X is zero.
α_i corresponds to the weight of the coefficient.

The following table 4.4 represents the coefficients of the lagged values of returns in predicting returns at time subscript t. The hypothesis of this statistical model is that there exists an autocorrelation of returns. Significance is conveyed by the mark of an asterisk for the weight of a coefficient. This asterisk corresponds to the degree of confidence in the hypothesis, where the confidence levels are expressed as 90%(*), 95%(**) and 99%(***). This notation is used for all tables of the subsequent vector autoregression results, models (2-6) that included the negative sentiment proxy.

As observed here, the null hypothesis that there is no autocorrelation of returns is maintained as none of the lagged coefficients are significant, indicating correlation and potential forecasting capabilities of lagged values of returns. This provides a sound basis as we further explore the relation of other time series with returns.

The R^2 coefficient indicates the amount of variance of the dependant variable explained by the independent variable is not significant for this model.

Coefficient	Weight
α_0	$8E-05$
α_1	-0.045
α_2	-0.011
α_3	-0.030
α_4	$0.1E02$
α_5	$0.3E02$
R^2	0.04%

Table 4.4: Results of VAR Model Evaluating Autocorrelation of Returns

4.5 Lexicon Comparison

To compare the performance in capturing the relationship hypothesised. The significance of using the lagged measures of negative sentiment provided by each lexicon was compared. It is important to note that each lexicon measured sentiment from the same corpus, so the source itself does not change, but merely the reference that was used in order to classify words and their sentiment. The following table 4.5 displays the negative sentiment scores as a proportion of negative sentiment per article of the total words in an article. The FEEL Lexicon has the highest mean negative sentiment score, while the LM lexicon has the lowest mean sentiment score. Equally, variations of the maximum sentiment score are evident across lexicons.

Table 4.5: Sentiment Lexicon Descriptive Statistics of Negative Sentiment Time Series

	Mean	Median	S.D.	Min	Max
FEEL	5%	5%	2%	0%	14%
GI	2%	2%	1%	0%	5%
LEX	2%	2%	1%	0%	8%
LM	1%	1%	1%	0%	3%
OIL	3%	3%	1%	0%	7%

The time series are then visualised in figure 4.4 where the elevated negative sentiment of the FEEL dictionary is observed. This variation in negative sentiment proportion per article can help the understanding of the performance of each of the sentiment lexicons in capturing the relationship between the negative sentiment proxy and the proxy of financial returns.

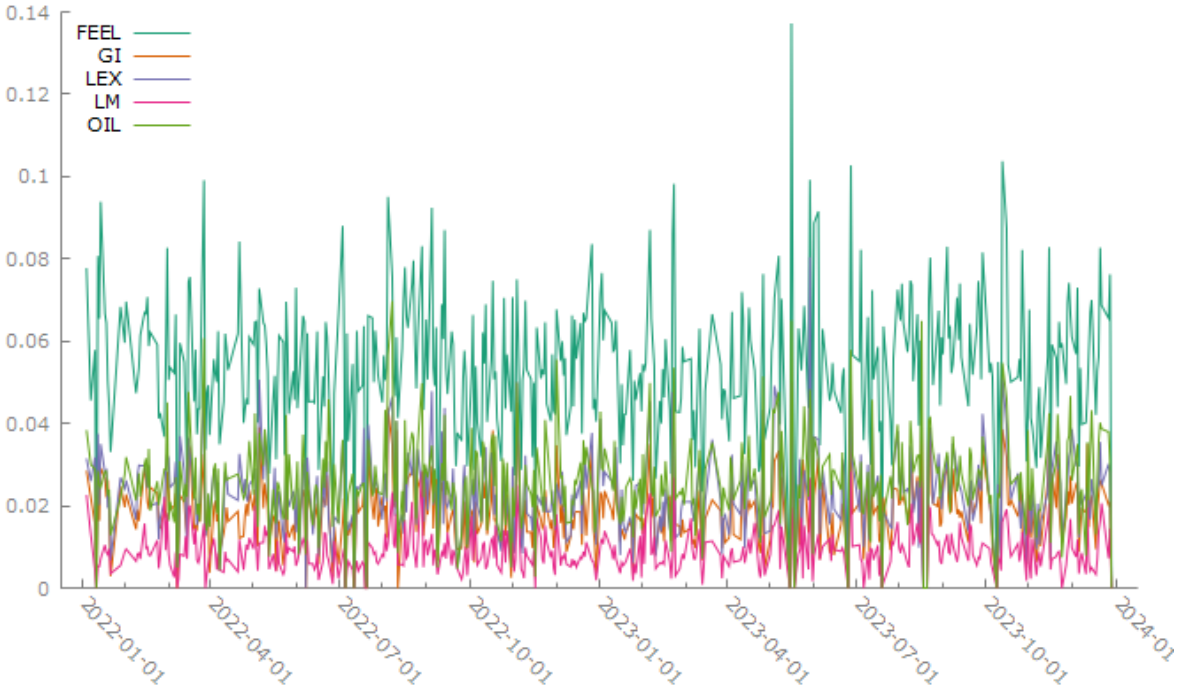

Figure 4.4: Time Series of Negative Sentiment Measured Using Different Lexicons

4.5.1 VAR Results

The results of the VAR(5) models (2-6) were used to compare the relationship between the financial market proxy, and that of each of the sentiment time series generated using the different lexicons. These results are shown in table 4.6 below. α_0 corresponds to the constant of the model and the α coefficients from lag 1 to 5 correspond to the lagged values of financial returns

Table 4.6: A Comparison of Impact of the Negative Sentiment Proxy on the Financial Market Proxy as Measured by Each Lexicon

	LM(2)		OIL(3)		FEEL(4)		GI(5)		LEX(6)
α_0	0.003***	α_0	0.002	α_0	0.002	α_0	0.001	α_0	0.001
α_1	−0.061	α_1	−0.048	α_1	−0.049	α_1	−0.046	α_1	−0.047
α_2	−0.020	α_2	−0.013	α_2	−0.016	α_2	−0.013	α_2	−0.012
α_3	−0.039	α_3	−0.031	α_3	−0.034	α_3	−0.033	α_3	−0.031
α_4	0.014	α_4	0.015	α_4	0.011	α_4	0.012	α_4	0.016
α_5	0.036	α_5	0.029	α_5	0.028	α_5	0.027	α_5	0.028
γ_1	−0.054	ω_1	−0.009	δ_1	−0.012	ζ_1	−0.014	λ_1	−0.034
γ_2	−0.042	ω_2	−0.023	δ_2	−0.011	ζ_2	−0.019	λ_2	−0.013
γ_3	−0.069*	ω_3	−0.014	δ_3	−0.003	ζ_3	−0.010	λ_3	0.008
γ_4	−0.125**	ω_4	−0.037*	δ_4	−0.013	ζ_4	−0.034	λ_4	−0.007
γ_5	0.005	ω_5	0.014	δ_5	0.004	ζ_5	0.017	λ_5	0.023
R^2	3.0%	R^2	1.4%	R^2	1.0%	R^2	0.9%	R^2	1.1%

As shown, the LM lexicon model displays some statistical significance in its relationship with the financial market proxy. The constant (α_0) in the model has a confidence level (1 - p value) of 99.7%. Understanding that returns are not autocorrelated, we can interpret the positive coefficient 0.003, as representing the case where there is no negative sentiment detected in the sentiment proxy. This describes a positive relationship with returns when there is an absence of negative sentiment.

The mean proportion of 1% negative sentiment in the LM negative sentiment time series indicates that it detects fewer words of sentiment in the corpus when compared to the other lexicons, as shown previously in table 4.5. This could be due to the method of constructing a corpus where the news media topic is not restricted to the financial domain but the topic of France. The LM lexicon may therefore be more selective at assigning sentiment relative to the domain of finance, thus having more days when the negative sentiment is zero. In turn, this provides meaning to days when no negative sentiment is detected in relation to the domain of finance.

Then, in further examining the statistically significant relationship of the LM lexicon at lags 3 and 4, we can see the impact of the negative sentiment proxy on financial returns. With confidence levels of 91% and 97% respectively, these lags indicate that the presence of negative sentiment signifies a decrease in returns just under a week later. A higher proportion of negative sentiment indicates a larger decrease in returns.

Then, in examining the fourth lag of the OIL lexicon, we see that it too has a statistically significant relationship with financial returns with a confidence level of 92% at lag 4. The similarity in these two models at detecting a relationship at the fourth lag could potentially be due to their strong correlation in measuring negative sentiment, related to their specification as models intended for sentiment analysis in the financial domain. This in turn can explain why it was possible for the other negative sentiment time series measured by the other lexicons to not follow suit as they were less strongly correlated with the LM lexicon as shown in the correlation matrix in table 4.2. It is noted that both of these lexicons were also automatically translated from English, which could justify the use of automatic translation in sentiment analysis of language for which, there are less resources available

In relation to describing the variance of financial returns as indicated by R^2, the negative sentiment proxy measurement by the LM lexicon explains the variation of the financial returns more than any of the other sentiment. This is interpreted as the LM negative sentiment measurement explaining approximately 3% of the variation of the fluctuations in the financial returns. The LM's R^2 value of 3% is approximately twice as strong as that of the OIL lexicon and three times that of the other three lexicons. When compared with model (1) including financial returns but excluding sentiment, an R^2 value of 0.4% is observed. This indicates that the inclusion of the negative sentiment proxy does provide

some explanatory power in the variation of the financial market proxy.

5 Conclusion

The low R^2 (3%) coefficient of the most statistically significant model using the LM lexicon indicated that little of the variation of returns is captured by negative sentiment. This could be interpreted as, while there is some relationship of covariance between the negative sentiment towards France from Francophone Africa, there exists other explanatory factors of the variance of returns in the French financial market that were not measured in this project.

The comparison of the lexicons in the statistical significance of their negative sentiment measurements indicates some advantage of employing lexicons designed for the use in financial domains when analysing the impact of sentiment on financial markets.

5.1 Future Work

As shown in the overview of economic and military relations between France and Francophone Africa, the magnitude of the relationship of between each country and France varies significantly. Equally, in constructing the corpus, the amount of data collected from each country varied as well. I suggest, therefore, for future work, to measure individually the impact of each negative sentiment proxy from each country towards France on the French stock market.
Then, motivated by the statistical significance observed when lexicons designed for financial media analysis were used to measure the negative sentiment, the collection of data from more financially focused media sources could be more rewarding in the capturing of a relationship between sentiment and financial markets.

Bibliography

[1] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[2] P. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, vol. 62, pp. 1139–1168, 05 2007.

[3] R. H. Thaler, "Anomalies: Saving, fungibility, and mental accounts," *Journal of Economic Perspectives*, vol. 4, pp. 193–205, 02 1990.

[4] Y. Qi and Z. Shabrina, "Sentiment analysis using twitter data: a comparative application of lexicon- and machine-learning-based approach," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 31, 2023.

[5] S. Kelly and K. Ahmad, "The impact of news media and affect in financial markets," in *Intelligent Data Engineering and Automated Learning – IDEAL 2015 16th International Conference*. Wroclaw, Poland: Springer, 2015, pp. 535–540.

[6] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7203–7219. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.645

[7] B. F. Green, "Reviews: Stone, philip j., dunphy, dexter c., smith, marshall s., and ogilvie daniel m. the general inquirer: A computer approach to content analysis. cambridge, mass.: M.i.t. press, 1966." *American Educational Research Journal*, vol. 4, pp. 397–398, 11 1967.

[8] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, pp. 35–65, 01 2011.

[9] D. Duval and F. Pétry, "L'analyse automatisée du ton médiatique : construction et utilisation de la version française du lexicoder sentiment dictionary," *Canadian Journal of Political Science*, vol. 49, pp. 197–220, 06 2016.

[10] A. Abdaoui, J. Azé, S. Bringay, and P. Poncelet, "Feel: a french expanded emotion lexicon," *Language Resources and Evaluation*, vol. 51, pp. 833–855, 06 2016.

[11] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, pp. 436–465, 2013.

[12] L. Bachelier, "Théorie de la spéculation," *Annales scientifiques de l'École normale supérieure*, vol. 17, pp. 21–86, 1900.

[13] E. F. Fama, "Random walks in stock market prices," *Financial Analysts Journal*, vol. 21, pp. 55–59, 09 1965.

[14] C. A. Sims, "Macroeconomics and reality," *Econometrica*, vol. 48, p. 1, 01 1980.

[15] I. Taylor, "France à fric: the cfa zone in africa and neocolonialism," *Third World Quarterly*, vol. 40, pp. 1064–1088, 04 2019.

[16] M. de l'Europe et des Affaires étrangères, "La coopération monétaire entre l'afrique et la france : le franc cfa," France Diplomatie - Ministère de l'Europe et des Affaires étrangères, 12 2021. [Online]. Available: https://www.diplomatie.gouv.fr/fr/dossiers-pays/afrique/la-cooperation-monetaire-entre-l-afrique-et-la-france-le-franc-cfa/#sommaire_6

[17] D. générale du Trésor, "Commerce bilatéral entre la france et les pays de l'uemoa en 2022," www.tresor.economie.gouv.fr, 07 2023. [Online]. Available: https://www.tresor.economie.gouv.fr/Pays/CI/commerce-bilateral-entre-la-france-et-les-pays-de-l-uemoa-en-2021

[18] M. des Armées, "Opex - opérations extérieures," Gouv.fr, 2022. [Online]. Available: https://www.defense.gouv.fr/terre/missions-larmee-terre/opex-operations-exterieures

[19] E. Macron, "Discours du président de la république." 02 2023. [Online]. Available: https://www.elysee.fr/emmanuel-macron/2023/02/27/discours-du-president-de-la-republique-dans-la-perspective-de-son-prochain-deplacement-en-afrique-ce

[20] R. Riordan, A. Storkenmaier, M. Wagener, and S. Sarah Zhang, "Public information arrival: Price discovery and liquidity in electronic limit order markets," *Journal of Banking and Finance*, vol. 37, no. 4, pp. 1148–1159, 2013.

[21] S. J. Taylor, *Asset price dynamics, volatility, and prediction*. Princeton University Press, , Cop, 2007.

[22] M. de l'Économie des Finances et de la Souveraineté Industrielle et Numérique, "Le cac 40," 2023. [Online]. Available: https://www.economie.gouv.fr/facileco/cac-40

[23] W. Newey and K. West, "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, vol. 28, no. 3, pp. 777–87, 1987.

A1 Appendix

A1.1 French Language Newspaper Data Collected from LexisNexis

Table A1.1: French Language Newspaper Data Collected from Lexis Nexis with Proportion of Articles Containing Key Words France and Français in Relation to Total Articles of the Time Frame Displayed

Country	Newspaper	Total	France AND français	Proportion
Algeria	Ech-Chorouk El Yaoumi (French)	4,648	458	10%
Benin	L'Evénément Précis	3,632	101	3%
	Fraternite	1,263	23	2%
	La Nation (Benin)	3,772	61	2%
Burkina Faso	Burkina 24	10,357	207	2%
	L'Observateur Paalga (Ouagadougou)	813	59	7%
Cameroon	Cameroun24	2,783	184	7%
Chad	Al Wihda	14,154	157	1%
Comoros Islands	Al Watwan (French)	4,582	171	4%
Republic of Congo	Agence Congolaise de Presse	33,679	279	1%
	La Prospérité	6,814	121	2%
Côte d'Ivoire	Fraternite Matin	5,100	67	1%
	L' Intelligent d' Abidjan	1,271	14	1%
	Le Patriote	1,185	21	2%
Djibouti	La Nation	1,308	24	2%
Egypt	Wutani (French)	98	20	20%
Gabon	Infos Plus Gabon	301	10	3%
	Infos Plus Gabon (Libreville)	147	1	1%
Guinee	Aminata (French)	2,492	23	1%
	Focus Guinee	792	16	2%
	Guinee7.com	1,476	10	1%
	loupeguinee.com	1,013	14	1%
	Mediaguinee.com	2,428	36	1%
Madagascar	L'Express de Madagascar (Antananarivo)	8,139	223	3%
	Midi Madagasikara (Antananarivo)	13,901	251	2%
Mali	Agence Malienne de Presse et de Publicité (AMAP)	1,536	36	2%
	Mali Actu	19,989	1,117	6%
Mauritania	Agence Mauritanienne d'Information (AMI - French)	9,912	35	0%
	Points Chauds (Mauritania)	497	12	2%
Mauritius	L'Express (Port Louis)	10,972	189	2%
	Le Mauricien	8,418	172	2%
Morocco	Al Bayane	20,140	242	1%
	La Nouvelle Tribune	5,607	85	2%
	La Releve.ma	11,537	263	2%
	L'Economiste	5,100	40	1%
	Le Desk (French)	8,831	378	4%
	Le Matin	7,427	127	2%
Niger	Le Sahel	1,934	61	3%
	Tamtaminfo	839	41	5%
Rwanda	Rwanda News Agency (Agence Rwandaise d'Information)	2,176	88	4%
Togo	IciLome.com	10,769	471	4%
	Agence Togolaise de Presse (ATOP)	2,580	15	1%
	Togo-Presse	1,349	21	2%
Senegal	Agence de Presse Africaine (APAnews) (French)	7,690	301	4%
	Le Soleil (Dakar)	5,469	128	2%
	Sud Quotidien (Dakar)	8,123	107	1%
	Walfadjri	1,710	33	2%
Sudan	Sudan News Agency (SUNA) (French)	2,458	19	1%
Tunisia	La Presse (Tunis)	10,628	272	3%
	African Manager(French Language)	25,925	467	2%
	Agency Tunis Afrique Press (French)	9,168	203	2%
	Journal la Presse & Assahafa AL YAOM	7,032	80	1%
	Leaders	2,373	155	7%
23	52	336,337	7,709	2.8%