INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Sentiment Analysis to Predict the Portuguese Economic Sentiment Based on Economic News

Cátia Daniela Lopes Tavares

Master Degree in Information Systems Management

Supervisor:

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,

ISCTE - Instituto Universitário de Lisboa

Co-supervisor:

PhD Fernando Manuel Marques Batista, Associate Professor,

ISCTE - Instituto Universitário de Lisboa

October, 2021

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Sentiment Analysis to Predict the Portuguese Economic Sentiment Based on Economic News

Cátia Daniela Lopes Tavares

Master Degree in Information Systems Management

Supervisor:

PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,

ISCTE - Instituto Universitário de Lisboa

Co-supervisor:

PhD Fernando Manuel Marques Batista, Associate Professor,

ISCTE - Instituto Universitário de Lisboa

October, 2021

Resumo

Medir o sentimento económico de um país é crucial para compreender e prever a sua condição económica de curto prazo. Este projeto propõe um indicador de sentimento automático, baseado em textos recolhidos de notícias económicas, que é capaz de medir com precisão o sentimento económico atual em Portugal e está altamente correlacionado com o Indicador de Sentimento Económico oficial, publicado pela Comissão Europeia algumas semanas depois e calculado com base em inquéritos.

Os dados utilizados nestas experiências consistem em cerca de 90 mil notícias económicas portuguesas, extraídas de dois jornais portugueses de renome, abrangendo o período de 2010 a 2020. Cada notícia foi automaticamente classificada com a polaridade de sentimento que tem associada, através de uma abordagem baseada em regras que provou ser adequada para detectar o sentimento das notícias económicas portuguesas. Para realizar a análise de sentimento das notícias económicas, também avaliámos a adaptação de módulos pré-treinados existentes e realizamos experiências com um conjunto de abordagens de Aprendizagem Automática. Resultados experimentais mostram que a nossa abordagem baseada em regras, que usa regras escritas manualmente específicas para o contexto económico, alcança os melhores resultados para detectar automaticamente a polaridade das notícias económicas, superando amplamente as outras abordagens.

O nosso estudo mostra que o sentimento expresso através das notícias económicas constitui uma forma promissora de prever o sentimento económico, permitindo entender a situação económica em Portugal quase em tempo real. O indicador desenvolvido, com base nas notícias, tem poder preditivo das flutuações económicas e do sentimento dos agentes económicos acerca do presente e o futuro da economia.

Palavras chave

Economia, Análise de sentimento, Sentimento económico, Indicador económico, Notícias económicas

Abstract

Measuring the economic sentiment of a country is crucial to understand and predict its short-term economic condition. This work proposes an automatic sentiment indicator, derived from collected economic news texts, that is able to accurately measure the current economic sentiment in Portugal and is highly correlated with the official Economic Sentiment Indicator, published a few weeks later by the European Commission, based on surveys.

The data used in these experiments consists of almost 90 thousand Portuguese economic news, extracted from two well-known Portuguese newspapers, covering the period from 2010 to 2020. Each document was automatically classified with the corresponding sentiment polarity, using a rule-based approach that proved suitable for detecting the sentiment in Portuguese economic news. In order to perform sentiment analysis of economic news, we have also evaluated the adaptation of existing pre-trained modules and performed experiments with a set of Machine Learning approaches. Experimental results show that our rule-based approach, that uses manually written rules specific to the economic context, achieves the best results for automatically detecting the polarity of economic news, largely surpassing the other approaches.

Our experimental results shows that the sentiment expressed through economic news constitute a promising way of predicting the economic sentiment, thus allowing to understand the economic situation in Portugal in almost real time. The developed indicator, based on the news, give us a predictive power of the economic fluctuations and the sentiment concerning the economic agents about the present and the future of the economy.

Keywords

Economy, Sentiment analysis, Economic sentiment, Sentiment indicator, Economic news

Agradecimentos
Acknowledgements

Aos meus orientadores, Prof. Doutor Ricardo Ribeiro e Prof. Doutor Fernando Batista, a quem agradeço pela transmissão de conhecimento, por toda a disponibilidade, ajuda e sugestões na supervisão deste trabalho.

Aos meus pais, pelos valores que me incutiram, pelo apoio constante e pela motivação na concretização de mais uma etapa.

À Tânia e ao Pedro por me fazerem sentir em casa, por toda ajuda e por tornarem esta etapa possível.

Ao Henrique, por acreditar em mim, por toda a ajuda e paciência ao longo deste percurso.

À minha restante família e amigos por acompanharem o meu percurso, por sempre me apoiarem e incentivarem.

Lisboa, outubro de 2021

Cátia Tavares

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Goals and Research Question 2

1.3 Methodology 3

1.4 Document Structure 5

2 Background **7**

2.1 Economic Indicators 7

2.1.1 Economic Sentiment Indicator 7

2.1.2 Gross Domestic Product 9

2.2 Text Mining 10

2.3 Natural Language Processing 10

3 Related Work **13**

3.1 Sentiment Analysis 13

3.1.1 Concept and applications 13

3.1.2 Techniques and Tools 14

3.2 News, Economic Sentiment and the Economy 17

4 Data **21**

4.1 Portuguese Economic News 21

4.1.1 Data Extraction and Collection 21

4.1.2 Data Understanding and Preparation 23

4.2 Manually Annotated Data 23

4.3 Economic Sentiment Indicator 26

5 Sentiment Analysis over Portuguese Economic News **29**

5.1 Pipeline . . . 29

5.2 Experiments and Evaluation . . . 30

5.2.1 Baseline/Translation-based Approach . . . 31

5.2.2 Rule-based Approach . . . 31

5.2.3 Machine Learning Approach . . . 32

5.3 Summary . . . 35

6 News-based Economic Sentiment Indicator **37**

6.1 News-based Economic Sentiment Indicator . . . 37

6.2 Correlation Between NESI and ESI . . . 39

6.3 Monthly Analysis . . . 40

6.3.1 Moving Averages . . . 40

6.3.2 Monthly Sentiment Calculation . . . 42

6.4 Weekly Analysis . . . 43

6.4.1 Weeks Back Indentation . . . 43

6.4.2 Weekly Sentiment Calculation . . . 44

6.5 Other indicators . . . 46

6.6 Discussion . . . 47

7 Conclusions and Future Work **49**

7.1 Main Conclusions . . . 49

7.2 Contributions to the Scientific and Business Community . . . 50

7.3 Limitations . . . 51

7.4 Future Work . . . 51

Bibliography **52**

A Web Scraping Ethical Issues **61**

List of Figures

1.1 General overview of the project steps 5

3.1 Sentiment analysis techniques [1] 14

4.1 Extraction and preparation of the news data 22
4.2 Word cloud for words in the headlines (left) and in the descriptions (right) . . . 24
4.3 Distribution of the number of news per month 24
4.4 Reference data polarity distribution 25
4.5 Economic Sentiment Indicator and Confidence Indicators 26

5.1 Automatic classification of economic news pipeline 30

6.1 Percentage of news associated to each polarity 38
6.2 Sentiment distribution between January 2010 to December 2020 38
6.3 NESI between January 2010 to December 2020 39
6.4 Correlation of NESI calculated through moving averages and ESI 40
6.5 ESI, NESI and NESI calculated through a 5 months moving average 41
6.6 Weeks back indentation indicator correlation 43
6.7 NESI and ESI since January 2010 to December 2020 44
6.8 Impact of the sentiment of the weeks on the correlation with ESI 45
6.9 Impact of the sentiment of the weeks on the correlation with confidence indicators 46

List of Tables

2.1 Types of surveys and related questions (source: European Comission Website) 8

2.2 Text mining techniques categories 10

3.1 Sentiment analysis in economic context 18

4.1 Number of words in the headlines and text descriptions 23

4.2 Economic sentences manually classified 25

4.3 Statistics related to the European Comission indicators 27

5.1 Expressions related to “Unemployment” 32

5.2 Model evaluation in our reference data 34

6.1 Correlation between the monthly news sentiment and ESI 42

1 Introduction

This chapter provides some insights about this project. First, in Section 1.1, we present the motivation for choosing this topic. Section 1.2 presents our objectives and our research questions. Section 1.3 explains the methodology adopted to reach our goals. And finally, in Section 1.4, we will present the structure of this work.

1.1 Motivation

Economic data and economic indicators are an important resource to reveal the true picture of the economic condition of a country. They allow us to understand the economic state and to determine our investment and consumption decisions. Also, forecasters and policy makers need information about how economy stands to take appropriate responses to their decisions. To respond to that and given the fact that economic indicators usually are published with a lag, having mostly a monthly and quarterly frequency, we should take advantage of the increasingly digital world that we have and transform the exponential amount of information available in an opportunity. Taking into consideration that news are the main form of transmission of information about the present, they can generate changes in the expectations of their readers. If the news are positive, the expectations of economic agents will also be positive and, consequently, the sentiment about the future of the economy as well. Otherwise, mistrust will be generated about the situation of the economy, which may have repercussions on economic agents investment and consumption actions. This can be a lever and generate aggregate economic effects and fluctuations in the Gross Domestic Product (GDP), the main indicator to the size of an economy [2, 3].

The Economic Sentiment Indicator (ESI) is an indicator published monthly by the European Commission, based on surveys about the sentiment and expectations of economic agents, both on the demand and the supply side [4]. It is an important indicator to monitor the current state of the economy and provide information about economic development [5].

Currently, a large amount of information is shared in news sites, blogs, and social networks. If processed timely and adequately, it can help to obtain key insights about the economic situation in almost real time. Extracting the sentiment expressed in economic texts can help us achieving such purpose. Sentiment analysis, also known as opinion mining, is

a well-known text mining task that consists of finding the overall sentiment expressed in a text [6]. It consists in identifying the polarity of a text span (e.g., a document or a sentence), in general, positive, negative, or neutral, that contains explicit opinions, beliefs, and views about specific entities (a subjective text span).

Economic news are now being produced everyday and everywhere in the world, and such unstructured data contains hidden information with the potential of producing real time powerful knowledge when combined. If correctly extracted, such knowledge can be extremely important for better understanding the current economic situation, being able to influence the actions taken either by political decision-makers, investors, or other economic agents. This on-the-fly usage of the available textual data contrasts with the existing ESI and GDP indicators, which do not allow to obtain real time information about the state of the economy [7], since they are published on a monthly and quarterly basis, respectively.

In recent years, sentiment analysis has contributed increasingly in various business areas and it also has been heavily applied in economic research, even in publications made by central banks of countries such as Spain and Germany. Given the low frequency and the delay in the publication of the official economic indicators, there were several studies in which indicators are created based on economic news to try to monitor the state of the economy in real time [8].

The economy is constantly changing and the creation of new economic indicators is under constant study. For example, with the crisis associated with COVID-19, the economy collapsed and GDP dropped significantly in a short period of time. As GDP is a quarterly indicator and is published with a significant lag, this drop was not immediately noticeable through the indicator, only in its previsions. According to an explanation of the Bank of England, central banks are constantly trying to create forecasts and new “fast economic indicators” to understand how economy is performing through big data analysis [9].

With this project, we intend to contribute to the field of study of sentiment analysis, a field of Natural Language Processing (NLP), and to study its application in the Portuguese language in the economic domain. Our goal is to develop an economic sentiment indicator based on the news in order to try to understand and analyze the state of the Portuguese economy through it.

1.2 Goals and Research Question

Considering the scenario presented before, it is important to reflect on the extraction and automatic sentiment analysis of economic news data, allowing us to obtain key insights about the economic situation in Portugal in almost real time.

Thus, the main research question of this project is the following:

> How, through sentiment analysis, can we construct a news-based economic sentiment indicator that could predict official economic indicators and the economic situation of the Portuguese economy?

Allied to this research question, the goals of this project, in addition to the construction of the news-based sentiment indicator, include:

1. Reflect on sentiment analysis and its application in the economic domain;
2. Reflect on approaches to perform sentiment analysis over the Portuguese language;
3. Correlate the developed indicator with ESI;
4. Correlate the developed indicator with the Industrial, Services, Consumer, Retail, and Construction Confidence Indicators;
5. Reflect on the use of Sentiment Analysis to understand the Portuguese economic situation.

The main objetive of this project is to understand how an indicator based on the sentiment present in the Portuguese economic news relates to the official Economic Sentiment Indicator based on surveys and if it could predict it and also predict the economic state and fluctuations.

In this work, we manually created a set of rules based on the economic context, and used a rule-based approach in order to classify our corpus of economic news. After evaluating other approaches to perform sentiment analysis, this approach has shown to be the most effective. After having the polarity of all the news, we develop an indicator of economic sentiment based on them. After that, we established a correlation between the developed indicator and the survey-based Economic Sentiment Indicator, showing how an indicator based on the news polarity allow us to understand the real sentiment of the economic agents.

In conclusion, the main benefit in this work is the possibility of obtaining an indicator based on the economic news that could portray the economy in almost real time, in contrast to official survey-based indicators, that are published with lag and with a low frequency. With our work we could be able to portray the sentiment of the Portuguese economic agents and the Portuguese economic fluctuations in almost real time.

1.3 Methodology

To achieve the proposed goals, we will adopt the CRISP-DM methodology (Cross Industry Standard Process for Data Mining). This methodology divides the projects into six phases. Although being essentially an iterative process, the sequence is not rigid [10].

According to this methodology, the project starts with a **business understanding** phase, where the problem, needs, and objectives are defined, as we have already done. Next is the **data understanding** phase, it includes data collection, its descriptive and exploratory analysis, and the validation of its quality to allow us to reach our goals. In this phase, we will collect and perform an initial exploratory analysis over our three datasets:

- Our corpus of Portuguese economic news with the title, description, and date of each news article, that will be used to develop our indicator;
- A reference data with the 400 more recent news that will be mannually classified and used to evaluate our sentiment analysis approaches;
- The portuguese Economic Sentiment Indicator data with the monthly indicator that will be used in our correlation and evaluation experiments.

After that, we have a phase of **data preparation**, in which data selection, cleaning, formatting and standardization must be carried out. It is necessary to be very careful in carrying out this step because if the data is not properly prepared, it can lead to wrong answers to the problem raised. Thus, this phase structures the textual data so that sentiment analysis algorithms can be applied. At this phase, we will uniformize some fields, do some cleaning and prepare our data, in order to make it ready for the next step.

Next is the **modeling** phase, where algorithms will be applied so that we have a model that classifies our news automatically. In this phase, we build and chose a model to perform the automatic classification of the polarity of each one of our news and, for that, we will evaluate several approaches. First, a baseline approach, where we translate our data and use existing pre-trained modules to perform sentiment analysis in the English language. Second, an approach based on rules created for the economic domain. And one last approach, where we do experiments with machine learning models in order to see if we could improve our results even further. The polarity of each news story will be calculated through the approach that obtained the best results, the rule-based approach, and, finnaly, we will construct our news-based sentiment indicator. To develop our indicator, we will take into account the polarities obtained with the rule-based approach and establish the difference between positive and negative news considering the total number of news.

After we have our sentiment classifier algorithm and our news-based economic sentiment indicator, the next phase is the **evaluation**. Once the official Economic Sentiment Indicators is published by the European Commission and is available for consultation and extraction, to evaluate the results obtained, correlations will be established between it and the developed indicator. This way, we will be able to understand the relationship between our indicator and the official one and understand how news allow us to know the current state of the economy and the sentiment of economic agents about the present and future of the economy. So, the result of this project will be a reflection about three approaches to perform sentimet analysis over the Portuguese language in the economic domain, a model

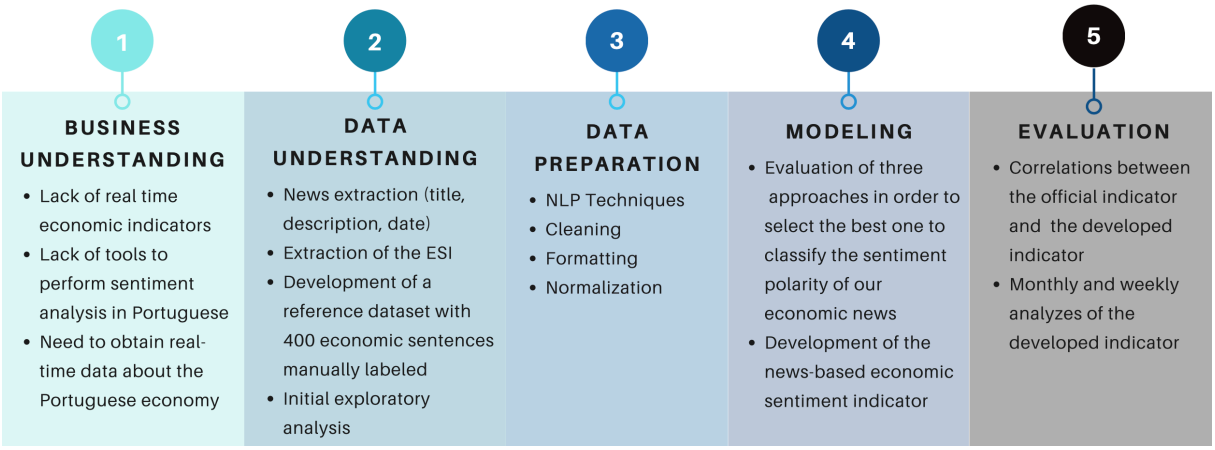

Figure 1.1: General overview of the project steps

that will automatically classify the sentiment of Portuguese economic news, and an indicator based on the calculated polarity of sentiment. In addition, we will be able to verify whether, through the sentiment present in the news, we can predict official indicators and the economic state in Portugal.

The last phase in this methodology is the **deployment**. The deployment phase of this research will be the full work made in this project and the analysis made. This research also intends to be an added value to the economic agents, and help them understand the benefits to analyze the news sentiment and their potential to predict and understand the Portuguese economy.

A summary of the steps performed in this work are presented in Figure 1.1.

1.4 Document Structure

This work is organized in seven chapters presenting all steps made in this study. Chapter 2 introduces some fundamental concepts related to the research topic; Section 2.1 focus on the understanding of the economic indicators related to our research; Section 2.2 and Section 2.3 present two topics related to sentiment analysis, the concepts of text mining and natural language processing, respectively. Next, in the Chapter 3, we present an overview on the related literature. Section 3.1 overviews the different strategies commonly used to perform sentiment analysis, while Section 3.2 focus on sentiment analysis in the economic context. Chapter 4 presents a description of our news data in Section 4.1 while Section 4.2 presents a reference data of 400 economic news manually annotated, and Section 4.3 describes the dataset with the survey-based Economic Sentiment Indicator. Chapter 5 outlines and evaluates three different approaches to perform sentiment analysis of Portuguese economic news. Section 5.1 presents the proposed pipeline of our sentiment analysis task; Section 5.2 describes the experiments performed with each one of our adopted approaches,

namely, the translation-based approach (Section 5.2.1), the rule-based approach (Section 5.2.2), and the Machine Learning approach (Section 5.2.3), with Section 5.3 presenting a summary of the results attained. Chapter 6 presents the construction of our news-based sentiment indicator (in Section 6.1) and its correlation with ESI (in Section 6.2), other experiments on a monthly basis (in Section 6.3) and on a weekly basis (in Section 6.4) and a final experiment where we consider other economic indicators, such as confidence indicators by sector (in Section 6.5) and, finally, Section 6.6 presents a discussion about our results. At last, Chapter 7 presents the major conclusions obtained from the experiences made throughtout this work and some recommendations for future work.

Background

2

This chapter introduces some fundamental concepts related with the research topic of this project. We start with a presentation of the economic indicators related to our work, the Economic Sentiment Indicator and Gross Domestic Product. After that, we give some context about text mining and natural language processing.

2.1 Economic Indicators

An economic indicator is a statistic about an economic activity. Economic indicators allow the analysis of the economic condition, performance, and fluctuations and give us insights into the health of an economy and help us to make the appropriate decisions in our current and future investment and consumption actions.

With the creation of the European Central Bank, the need for a range of monthly and quarterly statistics to measure economic and monetary developments was reinforced. Thereby, common economic indicators were created, allowing the assessment of the monetary policy and the economic situation in the euro area. In 2002, the European Commission produced a list of the 19 main European economic indicators, which include GDP (quarterly indicator) and ESI (monthly indicator) [11].

2.1.1 Economic Sentiment Indicator

The Directorate General for Economic and Financial Affairs of the European Commission calculates, based on responses to surveys conducted at each country, European Union (EU) and euro area level, the ESI and other monthly confidence indicators for industry, construction, retail, services, and consumers. A confidence indicator is calculated for each one of the sectors mentioned before, such as the arithmetic mean of the balance of survey responses (seasonally adjusted). The monthly and quarterly business and consumer surveys consists of a selection of variables such as the ones presented in Table 2.1. The confidence indicators for this sectors, and also for financial sector, which we will not focus on because it is not related with ESI, are calculated in order to reflect overall perceptions and expectations at the individual sectors.

Survey type	Monthly questions on	Quarterly questions on
Consumer	Financial situation General economic situation Price trends Unemployment Major purchases Savings	Intention to buy a car Purchase or build a home Home improvements
Industry	Production Employment expectations Order book levels Stocks of finished products Selling prices	Production capacity Order books New orders Export expectations Capacity utilization Factors limiting the production
Services	Business climate Evolution of demand Evolution of employment Selling prices	Factor limiting their business
Retail trade	Business situation Stocks of goods Orders placed with suppliers Firm's employment	-
Construction	Trend of activity Order books Employment expectations Price expectations Factors limiting building activity	Operating time ensured by current backlog

Table 2.1: Types of surveys and related questions (source: European Comission Website)

Depending to the type of survey, the statistical units are a firm (or enterprise) or a consumer. Each month, about 135,000 firms and 32,000 consumers are surveyed across the EU. The responses of the business and consumer surveys provide essential information for economic surveillance, forecasting, and economic research and it is usefull to detect economic fluctuations and turnings in the economic cycle.

ESI is a composite indicator, it has records since 1985 and is a weighted average calculated by assigning weights to each one of the sectors according to their market share, 40% for industry, 30% for services, 20% for consumers, 5% for construction, and 5% for retail [12]. This indicator aims to measure public perception and opinion about the current and short-term economy and can be considered as an early indicator of the economic future [11].

In summary, in the European Union, ESI is a weighted average of the balances of re-

sponses to surveys and its balance is constructed with the difference in the percentages of positive and negative responses, seasonally adjusted and with the objective of monitoring GDP growth at the member states, EU and euro area levels. It is a composite measure (with an mean of 100) that calculates the confidence level among the manufacturers, service providers, consumers, retailers, and constructors. If the indicator value is above 100, it indicates above-average economic sentiment and vice versa.

A limitation of this indicator is that the assigned weights are not continuously reviewed in order to represent changes in the economic system and there are several studies focusing on improving the predictive accuracy of the ESI [13]. Another limitation is the fact that the monthly surveys carried out to calculate the ESI are generally performed in the first two to three weeks of each month and the indicator only is published by the European Commission at the end of the month. Thus, we are facing a gap between data collection and publication [14].

2.1.2 Gross Domestic Product

GDP is the main measure of the size of an economy and represents the added value that is created in an economy, which can be calculated at the level of a country, a region, or a set of countries, as is the case of the European Union. The added value corresponds to the goods and services produced, after deducting the goods and services necessary to produce them [15]. Given that GDP is the main measure of a country's economic situation and its change over time, it is the most important indicator to show economic growths, and it is extremely important for making investment and consumption decisions. It is a monetary value of all finished goods and services made within a contry during a specific period and works as a scorecard of a given country economic health and provides a snapshot of the size of an economy and growth rate. GDP is the key tool to guide decision-making actions of policy-makers, investors, and businessess.

Considering the fact that GDP is a quarterly indicator and there is a lag in its publication, it does not give a real time overview of the economic situation, so, there are several studies in order to analyse and construct surrogate indicators. For example, Bortoli et al. [16] used data from social media as an alternative method to traditional indicators, concluding that this data is a promising tool to analyze the economy.

As conclusion, GDP is a measure of the economic production and is a general indicator of the development of an economy, being one of the most well-known and used statistics for combining production in an easily understood measure and for having a consistent structure and an international methodology, which makes it an easily comparable measure [11].

Category	Description
Information extraction	Consists of using a pattern-matching method to find specifics pieces of information such as key phrases and relationships in the text
Information retrieval	Task of searching interesting information from a collection of resources through the investigation of the appropriate mechanisms
Information visualization	Using visual representations to amplify human recognition and visualize information
Document classification	Finding patterns and features that allow grouping and assigning documents to known categories
Document clustering	Finding patterns and grouping similar documents based on their content
Document summarization	Preserving the meaning of the information but reduce the length and details of the source text

Table 2.2: Text mining techniques categories

2.2 Text Mining

Text mining is an important step in knowledge discovery, it consists in discovering new information through unstructured textual data [17].

Text mining tools range from simple statistics to more complex Natural Language Processing approaches [18]. These tools consist in the process of extracting quality information from texts and there is a wide range of topics and algorithms related to it [19].

Text mining objective is to extract patterns from large amounts of unstructured data and a lot of studies have discussed the applications of its techniques. The most popular areas of its application are: information extraction, identification of related topics, data/-document summary, categorization of themes/subjects, clustering of similar documents and obtaining answers to questions through a knowledge-based approach [20]. According to Feng et al. [21], the basic techniques identified in this task are divided into six main categories that are presented in Table 2.2.

2.3 Natural Language Processing

Natural Language Processing (NLP) is a set of computacional techniques that allow the analysis and representation of texts, with the goal of achieving a humam-like language processing [22].

According to Pereira [23], the sentiment analysis process depends on the use of NLP. NLP has tools that are used in the text pre-processing phase, in order to facilitate the text

analysis, such as tokenization, stemming, and lemmatization.

Tokenization is the task of finding the limits of a segment in a text, given a sequence of words, it divides the text into its words or sentences, so they can be analyzed individually, being one of the most important tasks of NLP. Each sentence or word is called token. An example of word tokenization is: “Unemployment is falling”, [“Unemployment”,”is”,”falling”].

Lemmatization is a pre-processing step used in text mining and NLP and consists of converting each word to its basic form, analyzing its vocabulary and morphology in order to remove the inflected endings, leaving only the lemma [24].

Stemming is a computational process in which a word’s suffixes and prefixes are removed and only the root stays [25].

Lemmatization and stemming are important steps to improve the results of document analysis, reducing the variation of word types as each word is represented by its lemma or stem. For example, a lemmatization algorithm will reduce the words “economic” and “economics” in the word “economic”, while a stemming algorithm will reduce them in the word “econom”.

NLP has a problem regarding the high dimensionality and processing time it has associated. Words like “and” and “the” do not have much meaning, do not influence the polarity of the sentence, and represent noise. These words are called stop words and they aggravate the problem mentioned above. They reduce performance and do not add relevant information to the sentences, so they are often discarded and there are pre-built stop word libraries that facilitate its removal [26]. Also, to reduce the complexity and the dimensionality of text, the removal of punctuation is a step performed in text analysis, in order to facilitate its treatment.

Some of the main NLP tools that work for the Portuguese language and have features for the techniques mentioned above are SPACY and NLTK [23].

3 Related Work

In this chapter, we present relevant literature and related work, which focus on economic and sentiment analysis domains. We address research studies done for sentiment analysis in general, and then we focus on its application in the economic context.

3.1 Sentiment Analysis

This section presents an overview about sentiment analysis. First, it will focus on its concept and applications and, after that, in the techniques and tools used to perform this task.

3.1.1 Concept and applications

Nowaday, with the World Wide Web and Big Data, a lot of information is available through news websites, blogs, and social networds. The large amount of data creates a difficulty: it is not possible to analyze it manually. Thus, it is necessary to have a set of systems that allow analyzing these textual data in massive amounts and extract relevant information from them. It is necessary to create solutions that try to extract public opinions and sentiments automatically and, sentiment analysis try to solve this problem. It is a tool that allow us to search opinions on a large scale, quickly, automatically, and saving resources [27].

Sentiment analysis is a text classification process and it refers to the use of NLP and computacional linguistics to identify and extract information from data . It is a research field within text mining and aims to identiy and extract information from data, with the goal of identify texts' opinions and polarity [28].

Sentiment analysis focuses on the analysis of users' expressions, classifying them according to the polarity. Data from different types of sources such as blogs, news, and social media, the use of different languages, non-standard words and the use of emojis and other symbols led to approaches with distinct complexity levels [29].

This process has gained an important role in the analysis and understanding of consumer communication in the media, allowing to provide key information about the public opinion on several subjects [30].

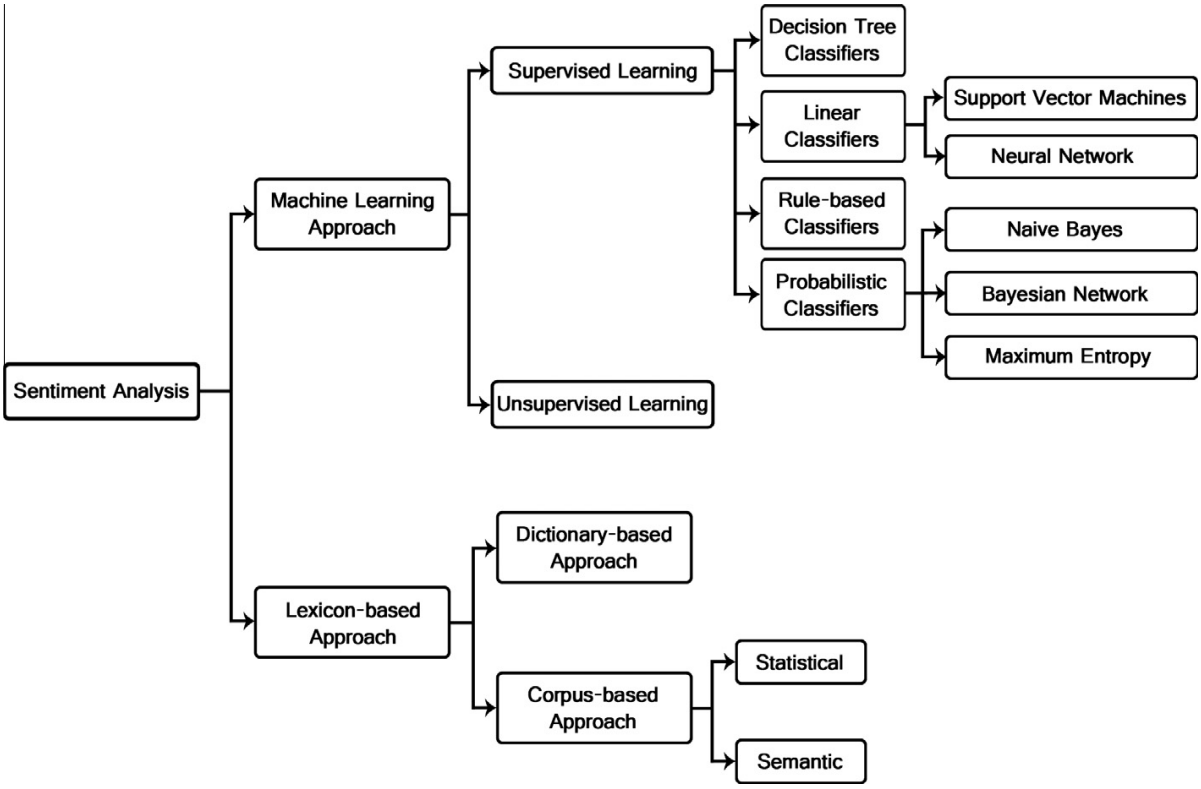

Figura 3.1: Sentiment analysis techniques [1]

At costumer behavior level, many efforts have been made in order to understand comments and reviews about products and services, to obtain information about public opinion, which can be useful for decision-making by brands, companies and consumers [31, 32, 33]. At organizational level, sentiment analysis offers the ability of monitoring various social media sources in real time and take rapid actions accordingly it. Also, at this level, its application in stock picking allows to obtain superior returns [6].

This task has been explored with data from various sorces and, in recent years, in addition to data from different news services, research has been carried out in several domains, focusing on the analysis of data from social networks such as Twitter [34]. The difference between these two types of textual data is that in the latter the opinion is generally clear, objective and is well defined in the text, while the first may cover several domains and may consist of more subjective texts and descriptions of complex and context-based events [35].

3.1.2 Techniques and Tools

There are several techniques for sentiment analysis as shown in Figure 3.1. There are approaches based on lexicons, which consist of predefined collections/dictionaries of terms and the associated sentiment/emotion; approaches based on Machine Learning (ML); and,

even hybrid approaches, in which both the previous approaches are combined. ML techniques can also be divided into two groups, supervised techniques and unsupervised techniques. In the first case, the data must be labeled, which is not the case with unsupervised techniques [1, 36].

NLP is highly used in approaches based on lexicons, as a pre-processing technique, to help find semantic relations and syntactic structure [37].

Sentiment dictionaries are dictionaries where each word is associated with and opinion/polarity (positive, negative, or neutral) and are very useful resources to classify the sentiment polarity. There are many sentiment dictionaries based on the English lexicon, however, for the Portuguese language these resources are scarce.

The literature about sentiment analysis focusing on the English language is vast, however, the linguistic resources available for sentiment analysis in Portuguese and other languages are still limited. Several studies adopt an approach based on the translation of the original data to English and after that an English sentiment analysis tool is applied. However, translation errors and language specific information can have a significant impact on the final result [23].

In the specific case of the Portuguese language, well-know sentiment analysis tools, like VADER [38], TextBlob, or Stanza [39], do not work. VADER combines a lexicon and a rule-based approach for sentiment analysis. VADER original experiments were performed only on English data. TextBlob is a Python library that provides several natural language processing modules, including one for sentiment analysis. TextBlob includes two sentiment analysis approaches, a rule-based model and a supervised ML model, based on a Naïve Bayes classifier. As provided, it only deals with the English language. Stanza toolkit also uses a ML model for sentiment analysis, in this case based on a Convolutional Neural Network classifier. Stanza is also a Python library and has models for English, Chinese, and German.

ML and Deep Learning (DL) techniques have shown good results in sentiment analysis and there are several studies to analyze the different SA techniques [40, 41, 42]. ML explores the study and application of algorithms to learn how to make predictions about data. On the other hand, DL is a particular area of Artificial Neural Networks, in which algorithms use several processing layers, acting as the input and output signal emitted betwen neurons. DL allows to predict, classify and group, reading the signals or data structure automatically. DL algorithms read the data, make assumptions, measure the error and automatically correct it, increasing accuracy and getting better results [43].

The ML approaches rely on ML algorithms to determine the sentiment as a text classification problem. Given a phrase/instance of unknown class, the model predicts the label/class to which it belongs. Supervised methods require the existance of labels in model training data, and example of supervised classifiers are Support Vector Machine (SVM) and Decision Trees [1, 44]. The unsupervised methods do not require the existence of labels

and have been subject of a lot of research [45, 46].

Depending on the approach used to perform sentiment analysis, before classifying the sentiment it is necessary to extract and select the features of the text. Feature extraction consists of performing a transformation to the original features and generate more significant ones, aiming to reduce complexity and provide simpler representations of the data [47, 1]. During the feature extraction process, useful features are identified and extracted, and analysis can also be performed to understand which features increase accuracy the most. To help assign weights to features, measures such as TD-IDF (term frequency-inverce document frequency) are used [42]. After extracting the features, the sentiment classification is performed.

Ahmed and Ahmed [48] used an approach based on lexicons to classify data from news. Firstly, they used text pre-processing techniques, such as pontuaction removal, stop-words removal, and stemming. After reducing the derivate words, they computed the polarity using TF-IDF to identify the most frequent words and to be able to assign them sentiment scores through dictionaries such as SentiWordNet. Finally, the news polarity was determined as positive, negative, or neutral, by the sentiment average of the total news words.

Mohamed [49] evaluated several algorithms to perform sentiment analysis based on ML approaches, concluding that SVM outperforms other methods such as Naive Bayes and Decision Trees. Also, Ahmed et al. [50], in order to analyze the sentiment expressed in Amazon food reviews, attained more than 80% accuracy when used Linear Support Vector Classifier (SVC), Logistic Regression, and Naive Bayes. They conclude that, in this case, Linear SVC performed better than Logistic Regression and Naive Bayes. However, each sentiment analysis technique will have differente performance and results depending on the data in which it is applied [51].

On the other hand, approaches that use Artificial Neural Networks can use word embeddings instead of feature extraction step. The word embedding consists in the projection of words in a continuous space in order to preserve semantic and syntatic similarities between them. This technique has been a great resource for NLP, in tasks such as entity recognition [52]. It represents a word form as a vector, and Word2Vec is an unsupervised method that uses neural networks to generate word embeddings, mathematically grouping similar words with equal context in the vector space. This method, based on the semantic relationship between words, improves the accuracy of the sentiment classification, showing the importance of word embeddings when dealing with NLP and sentiment analysis problems [46].

Recent research on sentiment analysis is advancing even further focusing on DL frameworks. The main conclusion is that DL presents better or comparable results to tradicional methods such as SVM and Logistic Regression [53].

In ML approaches, features are extracted and defined manually or with feature selection methods. In DL, features are automatically defined and extracted, improving perfor-

mance and accuracy. DL uses techniques such as Deep Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, among others [54].

3.2 News, Economic Sentiment and the Economy

The relationship between economic news, the economy and public perception, and opinion has been a subject of research for a while. The news have influence on the evaluations and opinions of economic agents about the economy. When they are negative, public opinion about future economic conditions is unfavorable and pessimism about the economy is generated [55, 56]. The importance of public opinion is due to the fact that changes in expectations about the economic future can be a source of economic fluctuations [57].

In order to understand the current state of the economy, high-frequency information is needed quickly and in real time [58]. This way, economic agents can use a multitude of high-frequency information in order to guide their actions, including news from the media [59].

There are several studies to understand the behavior of financial markets and stock values based on economic news [64] and that use news to perform sentiment analysis. Some of them are shown in Table 3.1. Mining the news plays an important role in designing strategies to predict market behavior and, based on events and news items, it is possible to predict market prices [65]. When there is pessimism in the media, patterns of falling stock prices and short-term returns are expected, concluding that the news information is useful for making predictions of market return and risk [66, 67].

In relation to the foreign exchange market, text mining is also a promising way to predict exchange rate movements based on the economic events present in the news, bringing benefits to investors and risk managers [68].

According to Huang et al. [60], traditional economic indicators based on surveys have been replaced by techniques for extracting sentiment from news texts and central bank statements, through the application of machine learning and other computational techniques. News-based sentiment indicators make it possible to predict periods of financial crisis, serving as early indicators of them. Nyman et al. [61] showed that periods of financial crisis can be detected in the news, being preceded by sentiments of anxiety. They were able to obtain information about episodes of risk and market volatility from the Bank of England news and publications data.

Aguilar et al. [63], when trying to monitor economic activity in Spain by building a sentiment indicator based on the news, found that the developed indicator has advantages over the indicator based on surveys in GDP forecasting and in forecasting the crisis related to COVID-19. Also according to Fraiberger et al. [62], the sentiment indicator based on the sentiment present in the news gives a direct and real time view of the aggregate sentiment of the current and future state of the economy, correctly portraying fluctuations in GDP, which allows policy makers to react more efficiently to economic conditions.

Ref.	Goals and Results	Data	Techniques	Metrics
[60]	Use news data to predict periods of financial crisis ——————— News sentiment index contains useful information to predict financial crises and market risk	Financial Times News	- Word vector representation - Semantic clustering - Sentiment of each cluster	Precision Recall F-score
[61]	Use text analysis to extract statistics about the economy, predicting important events and systemic risk ——————— Strong correlation with financial market events, such as structural breaks, and with other market measures such as sentiment, confidence, market volatility and systemic risk	- Comments about the bank of england market - Financial market research reports - Economic news	- Word count - Loughran and McDonald sentiment dictionary	Granger causality p-value
[16]	Create indicator to predict the state and evolution of the economy in France (GDP) ——————— Media are a promising tool for economic analysis and have made it possible to forecast French GDP	Le Monde News	- Construction of a sentiment dictionary - Logistic regression	RMSFE
[62]	Create sentiment index to predict economic fluctuations ——————— Index gives a direct and real time view of the aggregate sentiment of the current and future state of the economy, correctly portraying GDP fluctuations	Economic News	Loughran and McDonald sentiment dictionary (economy) and Young and Soroka dictionary (economy and politics)	Granger Causality p-value RMSE
[63]	Create sentiment indicator to monitor economic activity in Spain in real time ——————— Correlation of the indicator developed with the Economic Sentiment Indicator (ESI) of 0.8. Better performance than ESI in forecasting GDP and the economic crisis related to COVID-19. Better GDP forecast when we use the indicator developed compared to the ESI.	Economic News	Count words related to improvements and economic downturns	RMSE

Table 3.1: Sentiment analysis in economic context

According to Song and Shik Shin [69], when developing a sentiment indicator based on economic news, the developed indicator has a relationship with the indicator based on surveys. Also Shapiro et al. [70], while developing an approach to capture sentiments present in economic news to try to combat the limitations of survey-based sentiment indicators, found that daily news sentiment predicts official indicator movement.

4 Data

This work uses a large amount of data and the objective of this chapter is to present the three datasets used. First, a dataset that was created with Portuguese economic news that has been used to develop our indicator. Secondly, a dataset of 400 manually labeled news that was created to serve as a reference in our experiments to evaluate the sentiment analysis approaches presented in this work. Finally, a dataset of the survey-based Economic Sentiment Indicator and other confidence indicators that serves as a reference for the evaluation experiments and to establish correlations with our indicator.

4.1 Portuguese Economic News

In this section, we present the processes of extracting, understanding, and preparing our corpus of Portuguese economic news used to construct our indicator .

4.1.1 Data Extraction and Collection

In order to develop our indicator for a significant time period, we extracted economic news produced between January 1st, 2010 to December 31st, 2020, covering an 11 years timespan. The corpus was extracted from online news published by *Expresso* and *Público*, two well-known Portuguese newspapers.

Given the high volume of data that is available on the Web, the process of data collection can not be made manually. Technologies have been used to automate this process. This data collection process is called Web Scraping and it consists of automatically retrieving data from the Web. Web scraping needs to comply with a number of ethical and legal requirements and there is legislation that adresses topics related to it. According to Krotov and Silva [71], the topics to take into consideration when performing this task are the purpose of web scraping, copyrighted materials, damage on the website, terms of use, and ethics of web scraping. In Appendice A, we present our reflection about these topics regarding our context.

Before carrying out the data collection task, we analyzed which are the main Portuguese newspapers, of which *Expresso* and *Público* were part. After that, we analyzed

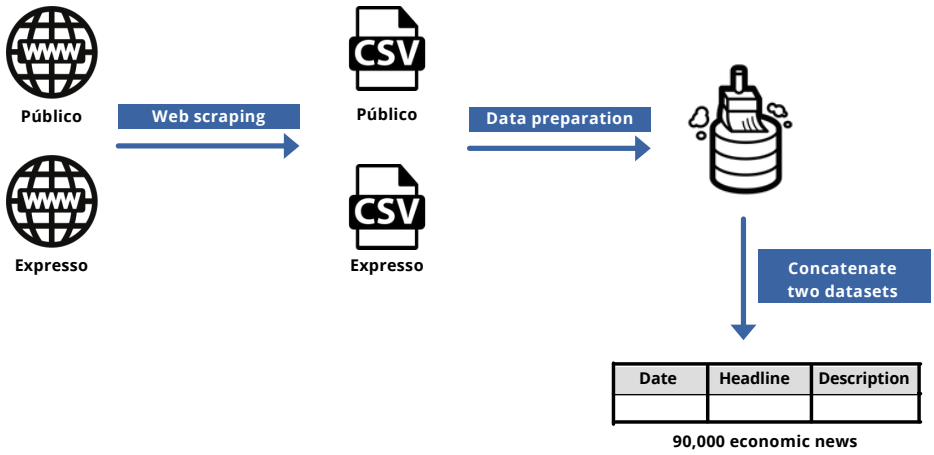

Figure 4.1: Extraction and preparation of the news data

the way how the various newspapers uploaded and presented information on their website. To facilitate the data collection process, we searched for the websites that load the news through an HTTP request, which happened in the two chosen newspapers. Our goal was to make a request in Python and get the data in a JSON format. We used the Chrome developer tools in order to intercept website requests to news endpoints. And, after getting the news endpoints, we manually tested their operation through Postman[1] (HTTP request tool), varying the parameters passed in the request. We noticed that there were certain parameters used to index the news (the date and page number). So, in order to get all the news published on the websites in the desired time frame we created a Python script that cycles through the required time interval, with a constant step (because each page has the same number of news). At each iteration, the script takes the start date and adds the offset generated by the cycle and sends it as a parameter to the journal's HTTP request. The endpoint's response is then parsed as JSON and the relevant fields are saved in a csv file.

For each article in the economic tab of the aforementioned newspapers, we extracted the date of the article, the headline, and the corresponding description, when available. So, we end up with a csv file per newspaper and respective data, that together had almost 90,000 economic news headlines, 62,326 of them also complemented with a textual description, which can contain one or more sentences.

In Figure 4.1, we can see the representation of the data collection process and the data preparation process. The last one will be presented in the following Section.

[1] Postman (https://www.postman.com/)

	Headline	Description
Average length	9.08	15.31
Median length	9	15
Maximum length	31	235

Table 4.1: Number of words in the headlines and text descriptions

4.1.2 Data Understanding and Preparation

To obtain our final corpus of Portuguese economic news, after the collection step, we needed to clean and prepare the collected data. So, a data preparation step was performed (as shown in Figure 4.1). Given the fact that we collected text data from two different newspapers, the data is not uniform and the data pre-processing and data preparation are very importante steps. This step is crucial to ensure that data has quality, is consistent, and can be used. For example, in our case, we needed to normalize fields like the dates that were in different formats in the two source newspapers. After normalizing the dates, we also cleaned the data in the text of the descriptions and the headlines. We removed some punctuation and HTML text formatting tags.

After that, we end up with a clean dataset of almost 90,000 economic news and we performed an initial exploratory data analysis.

From the extracted news, just over 62,000 have a description. Headlines are always made up of a single sentence, however, descriptions often contain more than one sentence. In that sense, we have over 90,000 news sentences coming from headlines and about 85,000 sentences from descriptions.

To understand our data, we calculated the average, median and maximum length of our headlines and descriptions for each article, as presented in Table 4.1. The headlines have an average of 9 words and the descriptions have an average of 15 words.

When considering all the headlines, we have 620,406 words and when considering the descriptions we have 991,069 words. In order to obtain the words with bigger frequency in a visual way, word clouds were created. Figure 4.2 shows the most common words in the headlines and in the descriptions.

The average number of news collected per month is 626. Figure 4.3 presents their distribution.

4.2 Manually Annotated Data

In order to evaluate the approaches under study to perform Sentiment Analysis, we created a dataset composed by economic sentences and their corresponding polarity. We have collected a sample of 400 sentences from recent economic news, and manually classified

Figure 4.2: Word cloud for words in the headlines (left) and in the descriptions (right)

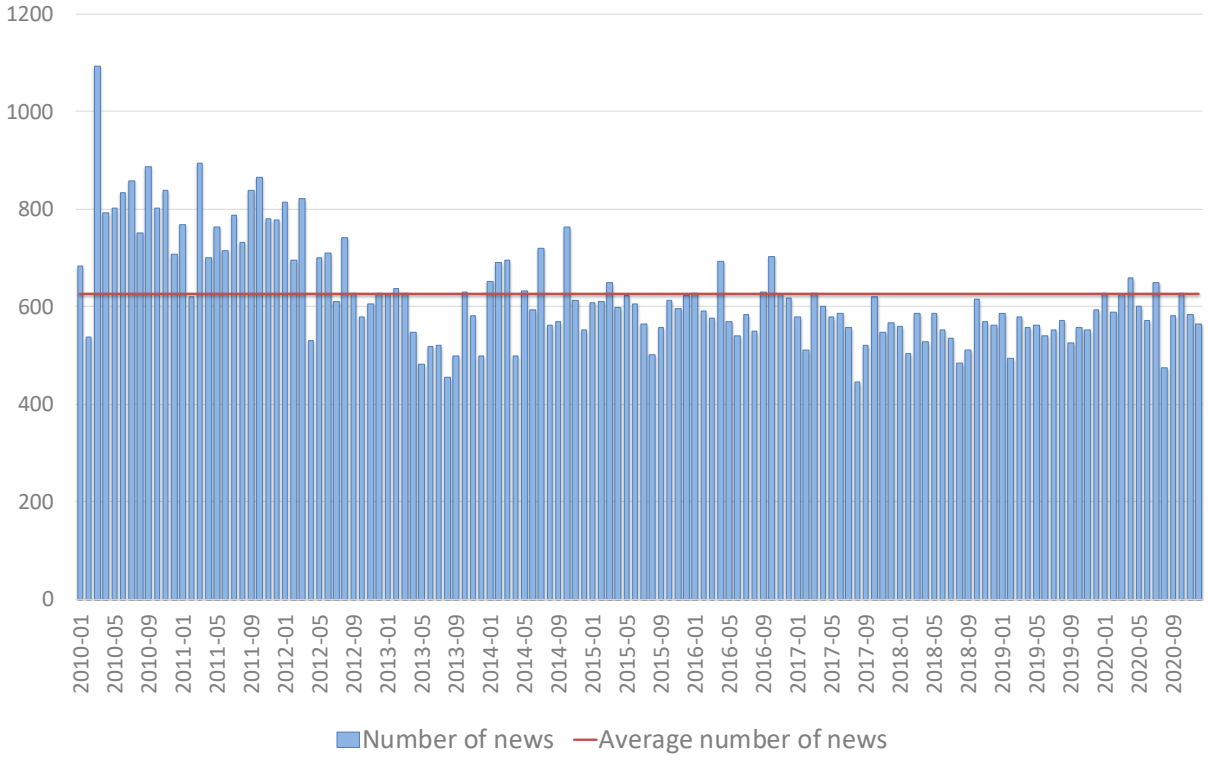

Figure 4.3: Distribution of the number of news per month

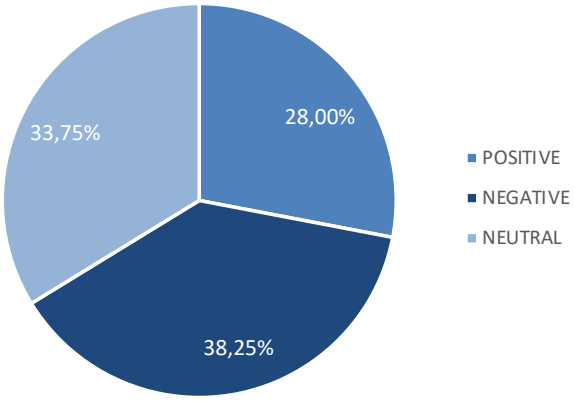

Figure 4.4: Reference data polarity distribution

Sentence	**Polarity**
Economia mundial recuou da ”beira do precipício”	1
Eurostat confirma que zona euro saiu da recessão no terceiro trimestre	1
Bolsa de Lisboa sobe dois por cento pelo segundo dia consecutivo	1
Grupo Michelin vai cortar 2300 postos de trabalho em França	-1
Portugal: risco e juros da dívida continuam a subir	-1
Japão acentua recessão no último trimestre de 2008	-1
Armando Vara substituído por Miguel Maya no BCP	0
Afinal, CTT mantêm entrega universal de cartas até ao fim de 2021	0
Graça Franco deixa direção de informação da Rádio Renascença	0

Table 4.2: Economic sentences manually classified

them with one of three possible labels, according to its corresponding polarity: Negative, Neutral, and Positive. Therefore, this dataset has two columns, one with the sentence and other with the sentiment associated to it. For the attribution of the sentiment, the economic meaning of the sentence was taken into account and, if it represents something good in the economy, it is positive (1), if it represents some bad thing that occured in the economy or that have a negative impact on it, it is negative (-1) and, if it has no impact on the economy, it is neutral (0). Table 4.2 shows some examples of sentences associated to each polarity class. Figure 4.4 presents the percentage of sentences belonging to each class.

Given the fact that we manually constructed this dataset, it has no missing or inconsistent data thus a cleaning and pre-processing step was not required.

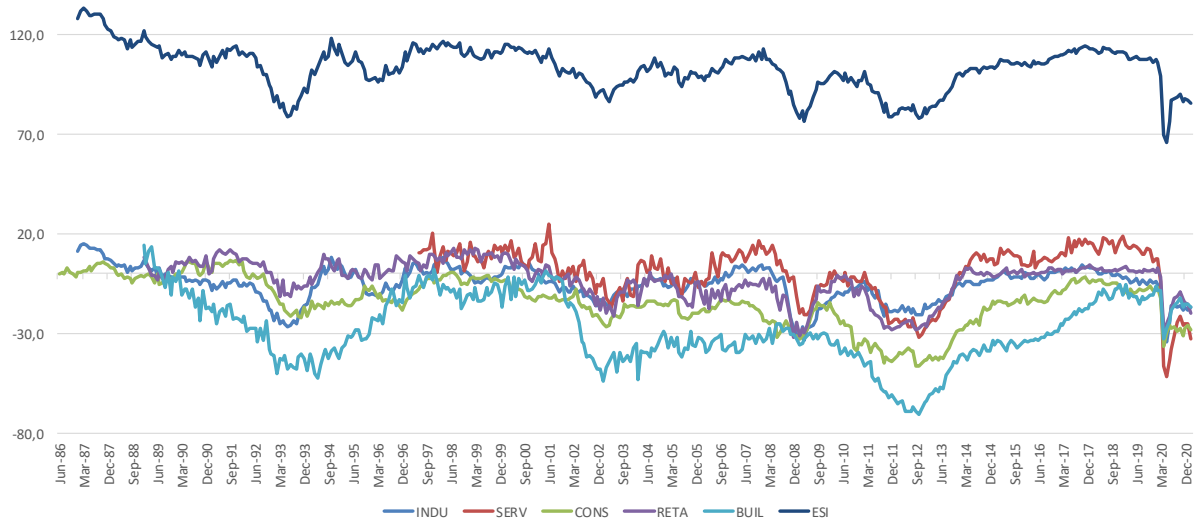

Figure 4.5: Economic Sentiment Indicator and Confidence Indicators

4.3 Economic Sentiment Indicator

As presented in Section 2.1.1, ESI data is published monthly by the European Comission and is available on its website[2]. The ESI data used in the development of this work was extracted from the European Comission website and the extracted file contains the following indicators based on monthly surveys:

- Industrial confidence indicator (INDU), which has a weight of 40%;
- Services confidence indicator (SERV), which has a weight of 30%;
- Consumer confidence indicator (CONS), which has a weight of 20%;
- Retail trade confidence indicator (RETA), which has a weight of 5%;
- Construction confidence indicator (BUIL), which has a weight of 5%;
- ESI which is a composite measure of the confidence indicators presented before and the percentages mentioned above and has an average of 100.

In the extracted file, there is data of the mentioned indicators for all euro area countries. For Portugal, we have the monthly indicators available from 1986 to 2020. We can see the distribution of the mentioned indicators for Portugal in Figure 4.5 and, in Table 4.3, we present some statistics about them. We can conclude that the ESI, our composite indicator, is very constant and its minimum value was obtained in May 2020 and it is not common to have values as low and as drastic collapse as happened in 2020. The remaining confidence indicators behave very similarly to the ESI, with the the construction confidence indicator

[2]European Comissions website (https://ec.europa.eu)

Indicator	25th Quartile	Median	75th Quartile	Min	Max
INDU	-9.40	-3.60	0.70	-34.00	14,90
SERV	-5.30	4.40	10.30	-51.20	24.60
CONS	-19.90	-13.50	-3.70	-46.30	6.50
RETA	-8.50	-0.80	2.80	-32.10	12.80
BUIL	-39.20	-30.20	-12.60	-70.20	14.10
ESI	97.20	105.40	110.90	66.20	132.90

Table 4.3: Statistics related to the European Comission indicators

being the least similar. When establishing correlations with ESI, we could see that with SERV we have an 93.47% correlation, 93.29% with INDU, 89.67% with RETA, 85.91% with CONS and 69.31% with BUIL.

The collected dataset was normalized and well structured and it did not contain missing or inconsistent data, thereby, a pre-processing step was not required.

We will use the data from the period between january 2010 to december 2020, to study possible correlations and take conclusions about our news economic sentiment indicator.

5 Sentiment Analysis over Portuguese Economic News

As mentioned before, the resources available to perform SA for the Portuguese language are scarce. It led us to try different approaches to perform this task. This chapter outlines the three different approaches that we used to perform SA of Portuguese economic news. Our baseline approach consists of translating Portuguese economic news data into English, and then apply well-known and widely used sentiment resources for English, such as VADER [38] and TextBlob[1]. In the second approach, we manually created a set of rules based on the economic context and used a rule-based approach. Finally, we trained different machine learning models, in order to try to improve our results even further. The performance of each one of the previous approaches was evaluated using a manually labeled dataset, containing 400 economic sentences, created in the scope of this work and presented in Section 4.2.

The proposed rule-based method for automatic polarity detection over economic news texts proved suitable for detecting the sentiment in Portuguese economic news and proved to perform well in our data.

5.1 Pipeline

In order to automatically classify the sentiment of Portuguese economic news, we have adopted the strategy represented in Figure 5.1, that consists of a data collection stage, an automatic classification stage, and an evaluation stage. We have started by collecting the data from online newspapers, as described in Section 4.1. Each one of the news stories was processed in order to extract the corresponding date, title (headline), and description. Additionally to collecting the data from 2010 to 2020, we also have selected 400 sentences, extracted from most recent news, that were manually annotated with the purpose of evaluating the approaches under study.

In terms of available NLP tools, the Portuguese language may be considered a low-resource language and during the course of this work, we could not find a sentiment analysis tool that could be directly applied to detect the sentiment of a sentence in Portuguese.

[1] TextBlob (https://textblob.readthedocs.io)

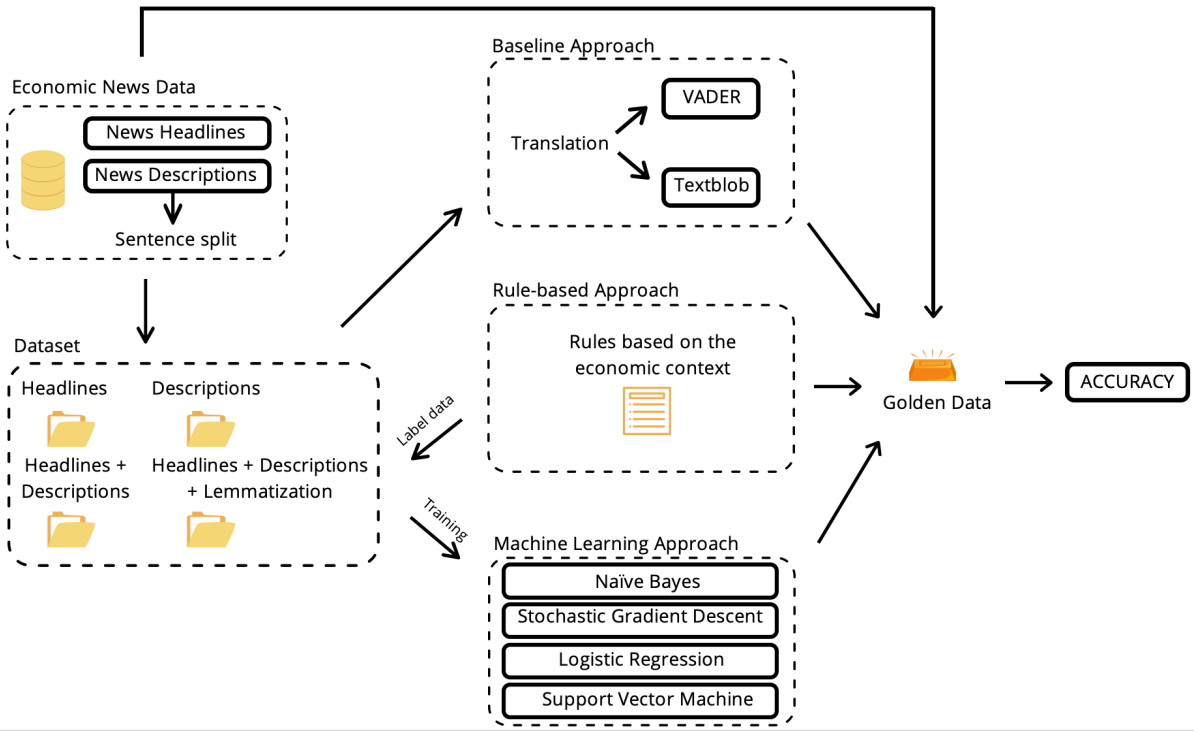

Figure 5.1: Automatic classification of economic news pipeline

For that reason, our initial strategy, represented in the top-middle box of Figure 5.1, consisted of translating the Portuguese sentences into English, and then using one of the existing English tools. However, sentiment analysis is known to be domain dependent, and soon we have realised that the commonly used tools could not be easily applied to the economic news domain. In order to overcome this problem, we have manually created a set of rules adapted to the economic domain, and we adopted the rule-based approach represented in the middle box of the figure. Finally, we have used our ruled-based approach to label our large dataset of economic news, and, in order to improve our results even further, we have trained several machine learning models, both in a supervised and semi-supervised way, to see if this models could generalise our rules and improve their performance, as represented in the bottom rectangle of Figure 5.1. All the described approaches are evaluated using the same manually labeled dataset, described in Section 4.2.

5.2 Experiments and Evaluation

In this section, we present the details about the three different approaches used to perform sentiment analysis. As previously mentioned, first we have tried to use existing tools to perform sentiment analysis, but we soon realised that the resources available for the Portuguese language are scarce. Thus, as a first approach, we translated our data into English and then used VADER [38] and TextBlob to perform the analysis. We concluded that this approach is limited when applied to the economic context. So, we tried a second approach

where we observed the most common patterns appearing in economy news stories and created a set of rules to classify each sentence, which proved to perform well in our data. In order to improve our results even further, we experimented a third approach where we trained different machine learning models. In the end of the section, we present a summary with the results obtained with the mentioned approaches.

5.2.1 Baseline/Translation-based Approach

When facing the lack of tools for a given language, one possible immediate solution is to translate the existing data to another language and then use the available tools for that language. In fact, during the course of this work, we did not find any available tools to directly perform sentiment analysis in Portuguese. As so, we have adopted TextBlob and NLTK VADER [38], two well-known tools for sentiment analysis, with the latter reported to perform well when applied to the finance domain [72]. So, after translating our reference data from Portuguese to English using Googletrans[2], a Python library that implemented Google Translate API, VADER achieved an accuracy of 46.5% and TextBlob achieved an accuracy of 32.0%. These results show that this approach is not suitable to the economic context, which was not an unexpected result since sentiment analysis is a domain-dependent task. The low accuracy could also have been aggravated by the loss of information in the translation.

We have then applied TextBlob and NLTK VADER to our unlabeled data in order to analyse the results in more detail. From the analysis we have observed that, for example, headlines with negative words like unemployment, crisis, deficit, etc., were classified incorrectly most of the times. In fact, the polarity associated with these words is negative, although we have seen that many news involving these words, such as “unemployment is decreasing” and “crisis is slowing down”, should be positive, and that VADER and TextBlob were not taking that into consideration.

5.2.2 Rule-based Approach

Our rule-based approach is similar to the approach proposed by Aguilar et al. [63] for classifying the polarity of economic news, where the sentiment attributed to each headline is also based on rules. For example, when combined with the word “economy”, the word “increase” becomes positive, and the word “decrease” becomes negative.

After identifying the errors and limitations in the classification performed by the previous approach, we have started looking at the more prominent words and combination of words in our unlabeled data. We have manually analysed through word frequency the set of words co-ocurring with words like “unemployment” to understand the most frequent patterns. As a result of the analysis of these associations of words, we constructed a list of expressions/rules related to the economic context, and labeled the sentiment associated to

[2]Googletrans (https://pypi.org/project/googletrans/)

Word 1	Word 2	Word 3	Sentiment
unemployment	reaches	minimum	1
unemployment	reaches	maximum	-1
unemployment	decreased		1
unemployment	increased		-1
...			
unemployment			-1

Table 5.1: Expressions related to “Unemployment”

them. We have observed meaningful combinations of two and three words, which derived in rules of one, two and three words, accordingly. For example, for the word “unemployment”, we can think of expressions involving words such as the ones presented in Table 5.1.

We did the same for other words related to the economic context such as “consumption”, “debt”, “economy”, “recovery”, and we ended up with approximately 600 rules with the corresponding associated sentiment (-1 if the expression is negative and 1 if it is positive).

Algorithm 5.1 details our rule-based classification process. First we try to match all the rules involving 3-words expressions. If more than one rule can be applied, we sum the sentiment associated with all the matching rules, and check if the resulting sum is positive, negative, or neutral. If none of the 3-words rules matches our sentence, we try to search all the rules involving 2-words, and again sum the sentiment of each one that we find. Finally, if none of the 2-words rules matches our sentence, we try to match 1-word rules. At the end of this process, if the sentence did not match any rule, then we assume it is neutral. When applied to our corpus of 90,000 news, 9.5% of the headlines where classified as negative and 7.5% as positive. Concerning the descriptions, 5.8% were classified as negative and 5.8% as positive. When applied to our reference data, our rule-based approach achieves an accuracy of 86.3%, a significant improvement over the baseline approaches.

5.2.3 Machine Learning Approach

In order to improve the results even further, we have performed additional experiments using our unlabeled economic news dataset for training our machine learning models. We wanted to see if these models could generalise our rules and increase their accuracy.

The dataset described in section 4.1 was used to create four different collections of texts that were used in our Machine Learning experiments: 1) sentences extracted from the headlines (about 90,000 sentences); 2) sentences extracted from the descriptions (about 85,000 sentences); 3) a combination of the previous two collections (about 175,000 sentences); and 4) a combination of the previous collection with its variant where lemmatization was applied to the words in the texts, in order to capture a broader set of economic terms (about 350,000 sentences). Lemmatization is a pre-processing step often used in text

Algoritmo 5.1 Classification of each sentence based in the rules created

```
input: sentence, rules

sentiment = 0
for rule in [rules with 3 words]:
  if rule.applies_to(sentence):
    sentiment += rule.sentiment
if sentiment = 0 then
  for rule in [rules with 2 words]:
    if rule.applies_to(sentence):
      sentiment += rule.sentiment
if sentiment = 0 then
  for rule in [rules with 1 word]:
    if rule.applies_to(sentence):
       sentiment += rule.sentiment

if sentiment < 0 then
  return −1
else if sentiment > 0 then
  return 1
else
  return 0
```

mining and natural language processing that consists of converting each word in its basic/-root form, analyzing its morphology in order to remove the inflected affixes, leaving only the lemma [24]. For this task we used SPACY [3] and, for example, this process will convert the words “increasing”, “increased”, and “increases” into the word “increase”.

Our rule-based classifier was used to classify all the sentences in each one of the collections. Then, we have converted all the labeled sentences into their corresponding document representation, using unigrams, bigrams, and trigrams.

We have applied the following classical supervised machine learning methods, used extensively for classification and regression tasks: Naïve Bayes (NB), Stochastic Gradient Descent (SGD), Logistic Regression (LR), and Support Vector Machines (SVM). Each one of the methods was applied to each one of the four previously described text collections, using their default parameters. Concerning the feature weights, we have used simple counts for Naïve Bayes and Term Frequency - Inverse Document Frequency (TF-IDF) weights for all the other methods. The corresponding evaluation results for our reference data are presented in Table 5.2.

[3] SPACY (https://spacy.io/)

Model	**Accuracy**	**Precision**	**Recall**	**F1**
VADER (baseline)	0.465	0.477	0.465	0.470
Textblob (baseline)	0.320	0.352	0.320	0.301
Rule-based approach	**0.863**	**0.863**	**0.863**	**0.863**
Naïve Bayes				
Titles	0.615	0.638	0.615	0.607
Descriptions	0.593	0.645	0.593	0.591
Titles + Descriptions	0.633	0.648	0.634	0.627
Titles + Descriptions + Lemmatization	0.663	0.675	0.663	0.661
Stochastic Gradient Descent				
Titles	0.740	0.739	0.740	0.739
Descriptions	0.730	0.743	0.730	0.733
Titles + Descriptions	0.763	0.763	0.763	0.762
Titles + Descriptions + Lemmatization	0.740	0.738	0.740	0.739
Logistic Regression				
Titles	0.738	0.737	0.738	0.737
Descriptions	0.693	0.718	0.693	0.698
Titles + Descriptions	0.758	0.764	0.758	0.759
Titles + Descriptions + Lemmatization	0.758	0.764	0.758	0.759
Support Vector Machine				
Titles	0.738	0.736	0.738	0.736
Descriptions	0.678	0.697	0.678	0.683
Titles + Descriptions	0.755	0.761	0.755	0.757
Titles + Descriptions + Lemmatization	0.770	0.773	0.770	0.771

Table 5.2: Model evaluation in our reference data

The results attained show that, in general, the text contained in the title is better for training than the text of the descriptions, this could be due the fact that we more more titles than descriptions. The best results are achieved when combining both fields. With NB and SVM we can conclude that the use of lemmatization contributed to a better result, we can not conclude the same when using SGD and LR.

Using only the title texts for training leads to a better accuracy with LR (74.0%), and using only the sentences from descriptions leads to a better performance with SGD (73.0%). The collection containing the titles and the descriptions led to better accuracy scores for SGD (76.3%), but the best accuracy (77.0%) was achieved by training the SVM with features produced with lemmatization.

We also have performed additional self-training classification experiments, considering all the texts labeled as positive or negative as the initial labels, and performing label propagation to all the remainder data. Nonetheless, the results achieved did not surpass our previous reported results.

5.3 Summary

We can conclude that the rule-based approach performs well in our context and is better when compared to others, as also showned in [73].

Table 5.2 summarizes the results achieved with each one of the approaches when evaluated using our reference data. The baseline approaches, using the NLTK VADER and TextBlob, performed poorly in the economic context, where accuracies of 46.5% and 32.0% were obtained, respectively. The best result was obtained with the rule-based approach with an accuracy of 86.3%. The machine learning approaches were not able to generalise and surpass our rule-based system: the best result of the machine learning approaches was achieved using SVM, with an accuracy of 77.0%.

6 News-based Economic Sentiment Indicator

In this chapter, we will present the details about the creation of our news-based economic sentiment indicator (NESI). And, after that, in order to take conclusions about it, we will calculate the correlation between NESI and ESI and perform other experiments on a monthly and weekly basis. In our last experiment, we will establish the correlation between NESI and other economic indicators, such as the confidence indicators of the industrial, services, consumer, retail and construction sectors.

Lets take in consideration that, when we refer "month t", it is the indicator reference month. In $W_{i,j}$, i presents the month and j represents the week number. For example, $W_{t-1,4}$ refers to the forth week of the month before the indicator reference month.

6.1 News-based Economic Sentiment Indicator

We decided to create a monthly indicator to facilitate the comparison with the survey-based ESI. With our rule-based approach, we classify the headlines and the description of each news article, and, to obtain the sentiment of each one of our news, we sum the sentiment in the title and the sentiment in the description and if it is greater than 0, we assume it is positive, if it is less than 0 it is negative, otherwise, it is neutral. Figure 6.1 shows the percentage of positive, negative and neutral news in our corpus of economic news. We can see that there is a high number of neutral news and the number of negative and positive news is close.

In Figure 6.2 we can see the distribution of news over time and the polarity that they have associated.

To obtain our news-based economic sentiment indicator for each month (period t), we subtract the number of negative news from the number of positive news and we divide it by the number of total news in period t.

$$NESI_t = \frac{Positive_t - Negative_t}{Total_t}$$

We can see the distribution of our indicator between January 2010 to December 2020 in Figure 6.3.

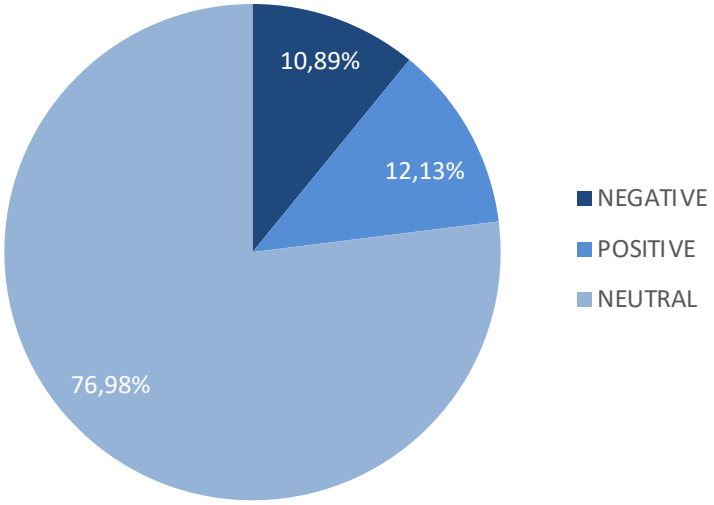

Figure 6.1: Percentage of news associated to each polarity

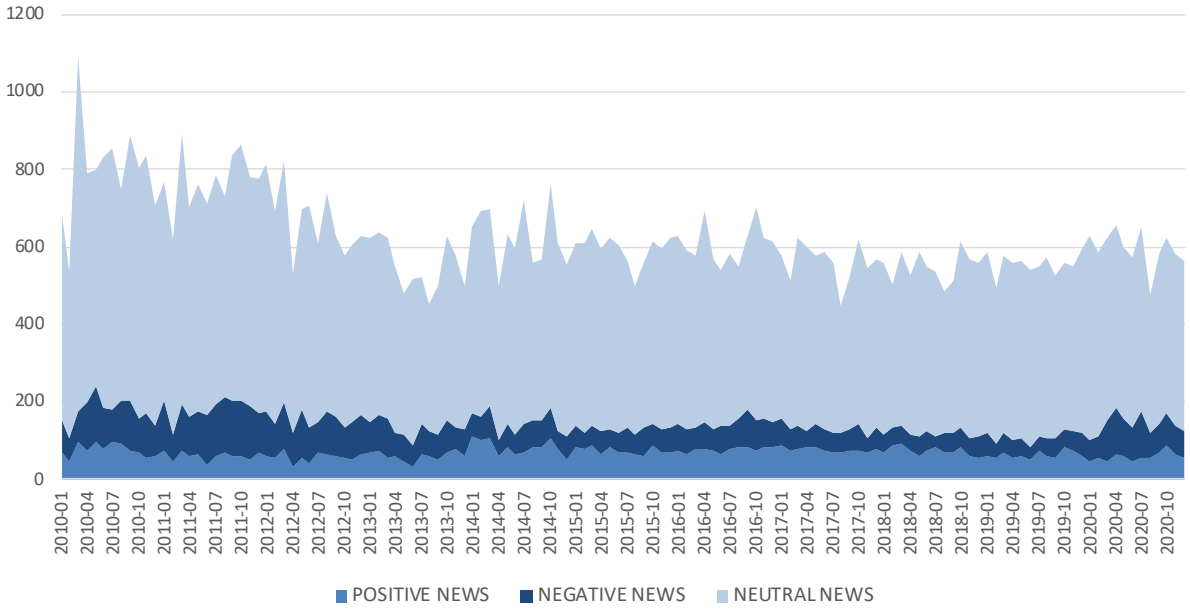

Figure 6.2: Sentiment distribution between January 2010 to December 2020

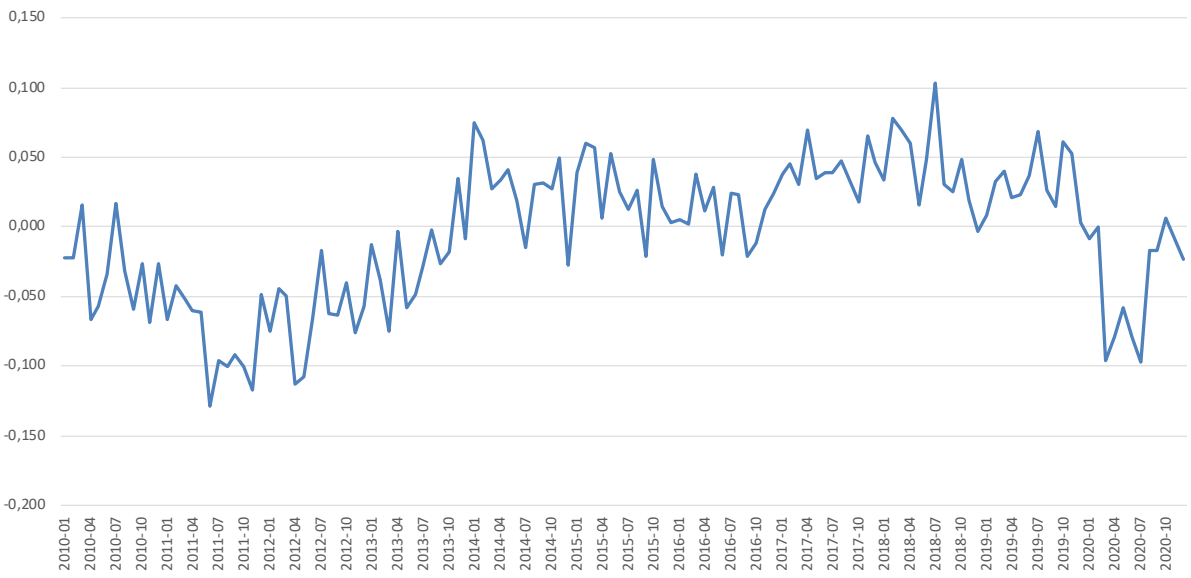

Figure 6.3: NESI between January 2010 to December 2020

6.2 Correlation Between NESI and ESI

The Pearson's correlation coefficient describes the direction and size of the relationship between two variables and can be a value between -1 (perfect negative relationship) and 1 (perfect positive relationship), 0 represents that there is no relationship between the variables [74]. In this work, we will use Pearson's correlation to measure the relationship between the ESI and our indicator, we will call this measure "correlation" from now on.

After calculating our indicator as presented before, we achieved a correlation of **76.1%** with ESI_t.

ESI is a monthly indicator for which the surveys are generally performed in the first two to three weeks of each month [14]. For example, the surveys for ESI of month t are done in the first two to three weeks of month t ($\text{W}_{t,2}$ to $\text{W}_{t,3}$) and the indicator only is published at the end of month. Thus, there is a a gap between the date of the surveys and the release date of the indicator.

To understand if we could see the impact of this lag in our indicator and to try to cancel it, we compared NESI of month t with ESI of month t and ESI of month $t+1$. Considering ESI_{t+1} the correlation increased to **79.9%**. We could conclude that with the sentiment present in the news of one month, we have better correlation with the official indicator of the following month, which could be justified by the lag between the surveys and the publish date. The economic events that happened in the one to two last weeks of the month ($\text{W}_{t,3}$ to $\text{W}_{t,4}$), could not influence ESI_t because the surveys for this month indicator are already done. It could better influence the sentiment of the following month's indicator, as reflected in our correlations. In conclusion, when considering the sentiment in the news of one month we have a better correlation with the next month ESI, so, we have a predictive

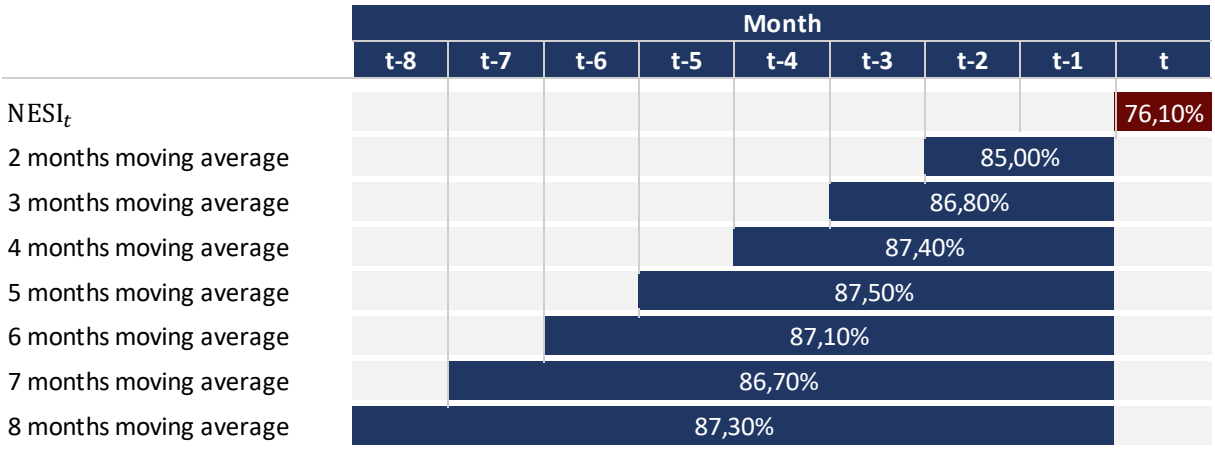

Figure 6.4: Correlation of NESI calculated through moving averages and ESI

power of the fluctuations of ESI through the news.

6.3 Monthly Analysis

As we saw through the previous correlations, the sentiment in the news of month t have better correlation with the ESI of the following month (ESI$_{t+1}$). So, we did some experiments based on a monthly analysis, in order to take more conclusions related to that. We started seeing the sentiment present in the news in the previous months, in order to see if their sentiment increases the correlation with ESI$_t$.

6.3.1 Moving Averages

In order to have smoother curves and to see if we could predict the ESI values, we calculated NESI with moving averages. For example, when considering a 2 months moving average, we calculate NESI for month t as the average of the values of NESI of month t-1 and NESI of month t-2. As we could see in Figure 6.4, the correlation increases when considering moving averages. Thus, we can conclude that it is better if we consider the previous months despite the current month to calculate the indicator. We also could conclude that the previous months affect the indicator and can be a way of predicting it.

We can also see that the best correlation with ESI is achieved when considering a 5 months moving average. This tells us that when we calculate NESI$_t$ through the average of the sentiment present in the news of month t-1, t-2, t-3, t-4 and t-5, we achieve our best correlation with ESI (87.5%).

As we can see in Figure 6.5, when we consider moving averages, we have smoother curves which justifies the greater correlation, however, we lose information. When considering NESI without moving averages, we reach the minimum sentiment value for 2020 in

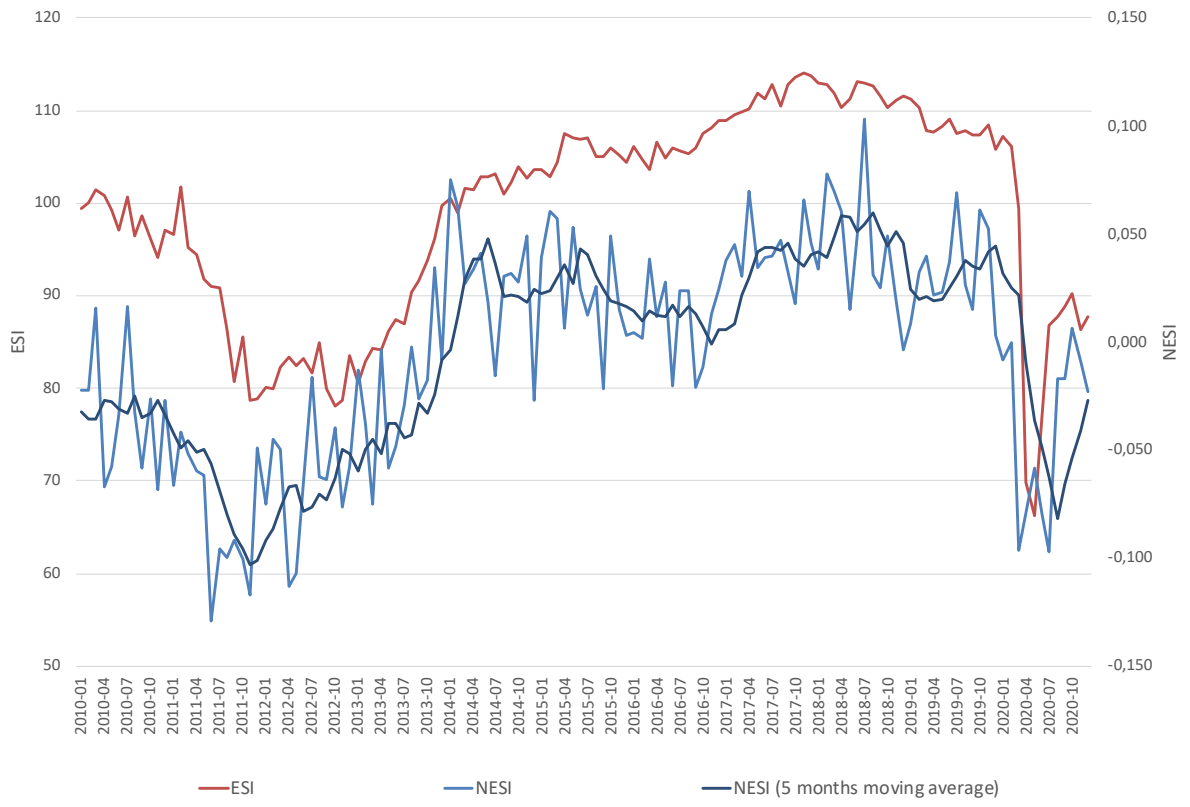

Figure 6.5: ESI, NESI and NESI calculated through a 5 months moving average

march. For ESI, the minimum value is in may, so, when not considering moving averages we have a 2 month advance in forecasting the economic recession in 2020. When considering our indicator with 5 months moving average, the correlation with ESI is higher, however, we lose the predictive power of our indicador and the minimum sentiment value in 2020 is delayed in relation to ESI.

With these experiments, we were able to conclude that the previous months impact the current month's sentiment and the correlation with the official indicator is greater when calculating NESI through moving averages. However, in order to not lose information, we can only consider a 2 month moving average because after that, we lose the predictive power of our indicator in relation to the economic recession in 2020 and the economic recovery in 2010, for example.

When considering a 2 months moving average, we calculate the NESI_t as the average of the sentiment in month t-1 and t-2. This value has an 85% correlation with ESI_t, so, we can say that we have a strong correlation and, as we can have the indicator value for month t at the end of month t-1 ($\text{W}_{t-1,4}$), we have a predictive power of ESI_t and an advance of 1 month, because ESI only will be released at the end of month t.

	Month correlation with ESI_t	Cummulative correlation with ESI_t
Month t	0.761	0.761
Month t-1	**0.798**	0.832
Month t-2	**0.796**	0.864
Month t-3	0.771	0.875
Month t-4	0.760	0.881
Month t-5	0.736	**0.883**
Month t-6	0.700	0.881
Month t-7	0.679	0.877
Month t-8	0.667	0.875

Table 6.1: Correlation between the monthly news sentiment and ESI

6.3.2 Monthly Sentiment Calculation

In this experiment, we calculate the sentiment for all the months in order to see the impact of each one in our indicator. We calculated the sentiment of each month (month t), the month before (month t-1) and we went till month t-8.

In the first column of Table 6.1, we can see the correlation of each month with ESI_t. In the second column, we can see the correlations, considering the cumulative sum of the sentiment in the news of the respective months mentioned before.

With the first column, we could conclude that month t-1 and month t-2 are the ones with the higher correlation with the ESI_t, and for the months before that, the correlation starts decreasing. If we established the correlation with ESI and the sentiment of all the news in month t (indicator month), we attain a correlation of 76.1% (first line in Table 6.1). If we consider the correlation of ESI_t and the sentiment in the news of month t-1 (79.8%), month t-2 (79.6%) and month t-3 (77.1%), all of them are higher than when considering the news of month t.

In the second column, we consider the cummulative sum of the sentiment in the months. For example, the second line consider the sentiment in the news for month t and t-1, and has a correlation of 83.2% with ESI_t.

If we start summing the sentiment of the previous months, as we can see in the second column of Table 6.1, the correlation increases till month t-5. And the best correlation is attained with the cumulative sum of the news sentiment from month t to month t-5 (88.3%).

As we saw in the other experiments, we can also conclude that the past influence the sentiment about the economic present and if we consider the sentiment of the previous months, the correlation with ESI_t is higher. This happens till month t-5, after that, we see a fall in the correlation. Also, we could see that month t-1 and month t-2 have the highest correlations with ESI_t, and are the ones that make the correlation grow faster.

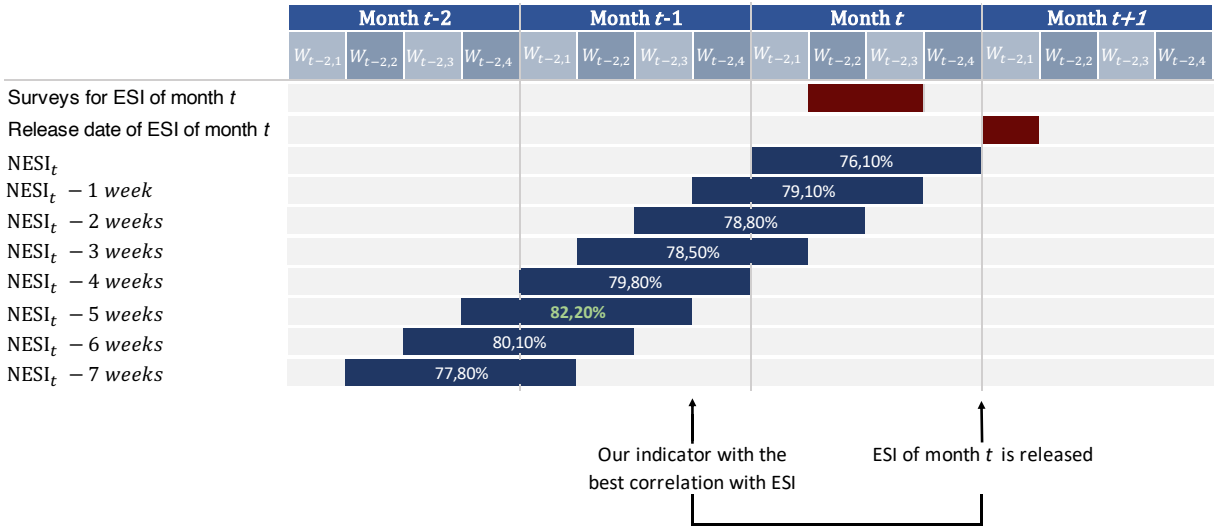

Figure 6.6: Weeks back indentation indicator correlation

6.4 Weekly Analysis

To go even further and to take conclusions with more details (weekly detail), we performed some experiments in order to understand the influence of each week and their impact in our indicator, calculating the sentiment present in their news.

6.4.1 Weeks Back Indentation

We compared ESI with NESI of month t and, from there, we started to calculate NESI going back a week, two weeks and so forth. In Figure 6.6 we can see the relevant dates related to ESI$_t$, NESI$_t$ and NESI$_t$ shifting weeks back. As already mentioned, the surveys for ESI are done in the first two to three weeks of each month and the indicator is published in the end of the month. When we calculate NESI$_t$, we summed the sentiment of all the news we have for month t. And, when going back one week in the NESI calculation (NESI$_t$ - 1 week), we stop considering the last week of month t and we start considering the last week of month t-1 (W $_{t-1,4}$). When we indent our indicator two weeks back, we stop considering the last two weeks of month t and start considering the last two weeks of month t-1. We did this and we end up going back 8 weeks.

We can see the correlations of our indicator shifting weeks back and ESI in Figure 6.6.

In general, our news-based indicator (with and without indentation of weeks) and the one based on surveys are highly correlated. We can see the effect of the lag in the indicator and the power of the past and see that the maximum correlation with ESI (82.2%) is reached when considering the sentiment of the indicator going back 5 weeks, so, considering the sentiment in the news from the W $_{t-2,4}$ to W $_{t-1,3}$ (purple curve in the bottom graphic in Figure 6.7). We can see that this curve and ESI have both a coincident minimum in May

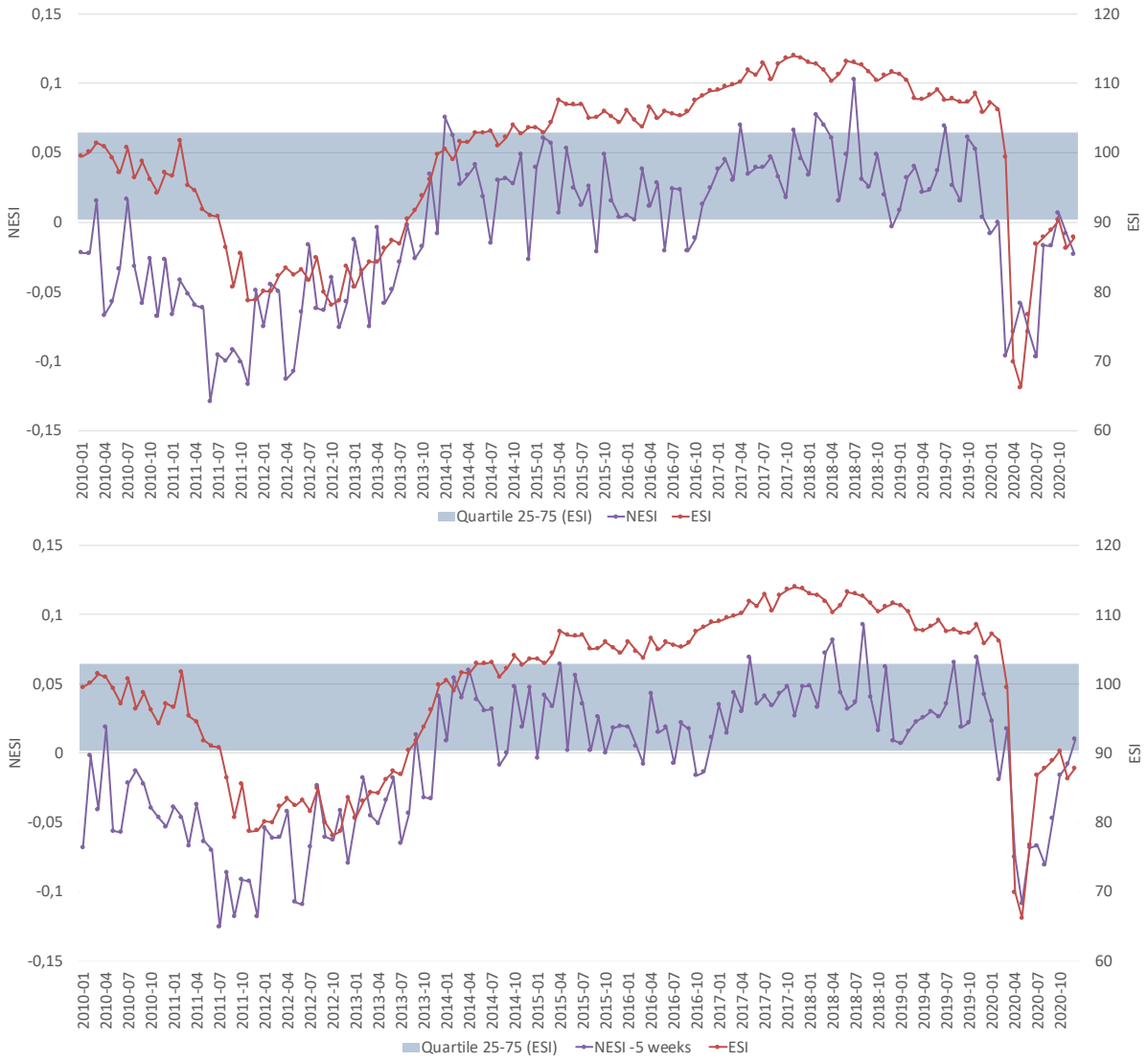

Figure 6.7: NESI and ESI since January 2010 to December 2020

2020. And that in this scenario, NESI presents well the recessions and the lower values above the quartile 25 of ESI between 2010-2014. We can also see that, the indicator going back 5 weeks is also coincident with ESI when referring to the economic recovery near 2011. When considering NESI going back 5 weeks, our news-based indicator turns out to lead ESI by 5 weeks (as shown in Figure 6.6). We have our indicator for month t in the end of week 3 of month t-1 ($W_{t-1,3}$), and the ESI only is released in the end of month t.

6.4.2 Weekly Sentiment Calculation

In our last experiment, we calculated the sentiment for all the weeks in order to see the impact of each one in our indicator.

As we already know, the surveys for ESI in month t are done in the second to third week

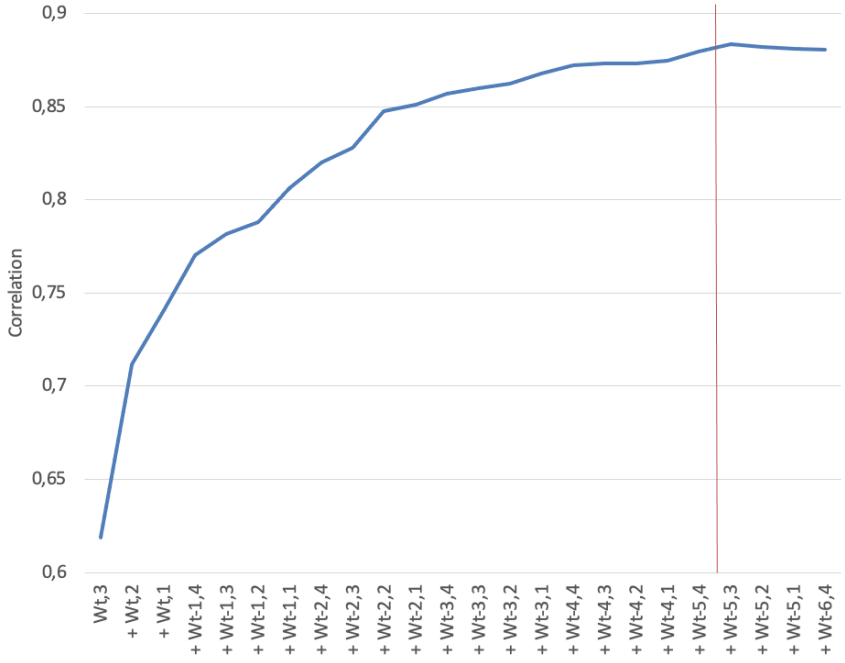

Figure 6.8: Impact of the sentiment of the weeks on the correlation with ESI

of each month (W $_{t,3}$ to W $_{t,4}$), so, the last week (W $_{t,4}$) will never influence the ESI of month t. Given that fact, in this experiment, we will calculate NESI for month t starting in W $_{t,3}$. If we establish the correlation with ESI, considering NESI as the sentiment of the news in the third week of month t (W $_{t,3}$), we attain a correlation of 61.9%. If we start summing the sentiment of the previous weeks, we can see that, when we consider more weeks, the correlation is higher. If we sum the sentiment in W $_{t,2}$ and W $_{t,1}$, the correlation is 74.1%. If we sum all weeks of month t-1 and month t-2, it raise to 85.1%. If we sum the sentiment of all weeks in the interval [W $_{t,3}$: W $_{t-5,3}$] we attain the maximum correlation of 88.4%. After this, the correlation starts decreasing slowly, as we can see in Figure 6.8.

As we saw in the other experiments, we can also conclude that the past influences the sentiment about the economy and if we consider the sentiment of older news, the correlation with ESI is higher.

To conclude, we saw that the last week of each indicator month (W $_{t,4}$) has a low correlation with ESI$_t$ (39.5%) and if we start considering the sentiment of this week and add it to sentiment of the news in the interval [W $_{t,3}$: W $_{t-5,3}$], the correlation drops from 88.4% to 87.9%. This reinforces the idea that the last weeks of each month do not have a good impact on ESI. In that sense, this experiment also reinforces the idea that the sentiment in the news of the past influences the sentiment about the present and future of the economy.

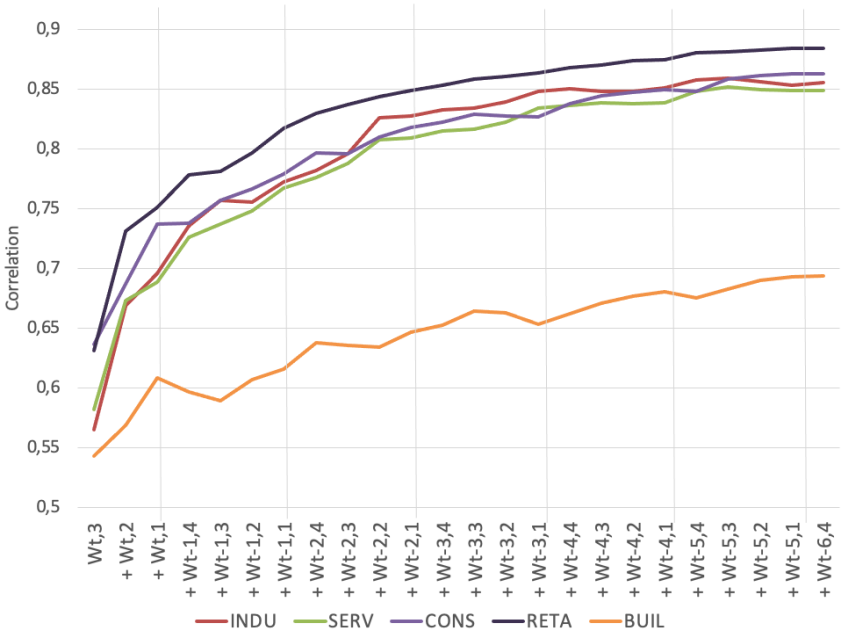

Figure 6.9: Impact of the sentiment of the weeks on the correlation with confidence indicators

6.5 Other indicators

After presenting and trying to understand the relation between NESI and ESI, in this section, we will reflect on NESI and its correlation with other economic indicators such as the ones presented in Section 4.3, the confidence indicators of the industrial (INDU), services (SERV), consumer (CONS), retail (RETA) and construction (BUIL) sectors.

We will focus on our weekly sentiment analysis and try to understand the NESI correlation with the aforementioned indicators and see if it behaves similarly to what we observed in its relationship with the ESI.

As we can see in Figure 6.9, all the confidence indicators are highly correlated with NESI. The lowest correlation is attained with the construction confidence indicator and the highest with the retail confidence indicator. The industrial, services and consumer confidence indicators have very similar correlations with NESI.

For all the mentioned indicators, the correlation with NESI is higher when we consider news from previous weeks. This correlation, similarly to our previous experiment, is higher if we consider until the second or third week of month t-5 and if we do not consider the last week of the month of the indicator. This is because the surveys have already been carried out and do not have any relation with the calculation of the indicator.

In conclusion, when considering the confidence indicators, NESI has a better correlation with the retail confidence indicator, followed by consumers, services, industry and construction. However, the highest correlation is still attained with the composite measure,

the ESI, and none of the indicators mentioned above have a major impact on it.

6.6 Discussion

In all our experiments we can see that our news-based economic indicator and ESI based on surveys are highly correlated. All our experiments shows that the past has an effect on the expectations about the economic present and that the economic news are a promising way to give us information about the sentiment and expectations about the economy. When we have positive news, NESI will be positive and the positive sentiment about the economy generated by the news in their readers will be reflected as a growth in ESI.

From our monthly analysis, we could conclude that when calculating ESI of month t, the news in the last five months are the ones that have more influence and add value to the calculation of the indicator. From our week analysis, we could conclude that we would have a better correlation with ESI if we shift five weeks back our indicator, that way, we have a five weeks advance in relation to ESI. Also, we saw that the best correlation with ESI is reached when not considering the news of the last week of the indicator month and when considering the sentiment of the previous months news, until the third week of month t-5. The last week of the indicator month not having impact on the correlation proves the effect of the lag between the surveys date (first two to three weeks of the indicator's month) and the publication date (end of month).

Both ESI and NESI can reflect the major economic issues such as the financial crisis related to COVID-19 in March, 2020 and the financial crisis in Portugal between 2010-2014, derived from the 2007–2008 global financial crisis. Also, our indicator turns out to lead ESI in predicting the lowest value of the economic sentiment related to the recession in 2020. In this recession times, the sentiment associated to the economic news was negative and also the sentiment generated on their readers. As the expectations of the economic agents were negative, it was reflected in the surveys done to calculate ESI and, consequently, in the ESI value.

Another advantage that our indicator could have is that, if we collect the news daily, we could have daily information about the economic sentiment and the economic agents expectations in almost real time, which could not happen if we have to wait to the release of the indicator based on surveys.

When analyzing the industrial (40%), services (30%), consumer (20%), retail (5%) and construction (5%) confidence indicators, we conclude that they are highly correlated with ESI and they behave very similarly. Although ESI is a composite measure of these indicators in the presented percentages, the correlations of confidence indicators with the ESI in descending order are the following: services, industry, retail, consumers and construction. When establishing correlations with the confidence indicators and our news-based economic sentiment indicator, the highest correlation between the confidence indicators

and NESI is attained with ESI, followed by retail, consumer, services, industrial, and construction confidence indicators.

7 Conclusions and Future Work

This chapter aims to present the conclusions regarding the overall research, namely, the main conclusions achieved, the limitations of the project, the contributions to the scientific and business comunity and, lastly, the proposals and future developments.

7.1 Main Conclusions

Nowadays, with the exponential amount of information available, it is difficult to analyse it. Sentiment analysis appears as a solution to this problem and it focus in the development of systems that can determine the polarity of a text. The lack of information about the state of the economy in real time and the high periodicity of the official economic indicators, led us to develop an indicator based on the sentiment present in the news to understand the economic sentiment and the economic situation of the Portuguese economy in almost real time. However, the sentiment analysis resources for the Portuguese language are scarse. To give response to that and given the complexity of the economic context, in this work, we analised different approaches in order to solve this practical problem. First, we have tried a baseline approach where we translated our texts into English and used well-known sentiment analysis tools, such as NLTK VADER and TextBlob. Given the poor results achieved in the economic context, we tried a rule-based approach for which we have created manual rules, based on the economic domain, and used those rules to classify the polarity of each economic text. Finally, we have created a set of machine learning models, based on the large amount of economic texts that we had available, aiming at improving our results even further.

In order to compare and evaluate the performance of the proposed approaches, we have also created a reference dataset, containing 400 economic sentences, manually classified. The performed experiments have shown that the baseline approach achieves poor results, when applied to the economic domain. The rule-based approach achieved an impressive performance of 86.3% accuracy, a significant increase of performance over the baseline approaches. The machine learning models that we have explored were not able to generalise and surpass the rule-based approach.

After classifying our news using the rule-based approach, we developed a news-based

economic sentiment indicator. We concluded that the news can give us knowledge about the economic sentiment and the developed indicator based on news is highly correlated with ESI and other confidence indicators based on surveys. This correlation increases when we consider the news of the previous months, which represent the effect of the events in the past on the expectations about the present of the economy. When calculating the indicator for month *t*, this effect is observed until *t*-5. After that, considering older news makes the correlation with ESI decrease. We also could see that the news in the last week of the indicator month have a bad impact in the correlation with ESI, and, is better to not consider them to have better values. It could be justified due the fact that the surveys for ESI are made in the first two to three weeks of the indicator month, so the events in the last week of the month will never influence the indicator.

Our news-based indicator turns out to lead ESI in predicting the economic recession in 2020 and other economic events and give us a predictive power of ESI.

In conclusion, the developed indicator can reflect the economic sentiment and the fluctuations in the Portuguese economy and could be a promising way to understand the state of the economy in almost real time if we collect the news and analyse it daily. That way, it could give us more real information about the Portuguese economy than official economic indicators that are published monthly or quarterly and with a lag.

7.2 Contributions to the Scientific and Business Community

The contribution of this work relies on the use of unstructured data from news articles. That way, this work fills the void in the sentiment analysis in the Portuguese language in the economic domain and gives a contribution to the literature in several ways. First, this work reflects on sentiment analysis and its application in the economic domain. Secondly, it reflects on approaches to perform sentiment analysis over the Portuguese language and presents an approach based on rules that proved suitable, which was the focus of an article published and presented in the SLATE symposium (Symposium on Languages, Applications and Technologies). Additionally, due to the lack of data, this work contributes with a construction of a Portuguese news dataset with 90 thousand news and a labeled set of 400 economic news.

Finally, this work is a great contribution to economic agents, as this work reflects on the use of sentiment analysis to understand the Portuguese economy and creates an economic indicator based on economic news that proved to represents the Portuguese economy and has a predictive power over the official Economic Sentiment Indicator and the economic fluctuations. Also, if we collect the news and analyse it in real time, we could have a real time economic indicator which brings an advantage over official indicators thar are published with a delay and based on a monthly and quarterly frequency.

7.3 Limitations

In the development of this work some difficulties were found. Most of them are related to the lack of data. On the one hand, it was difficult to obtain the news data. On the other hand, it was difficult to found a labeled dataset to allow us to use supervised approaches to perform sentiment analysis. And, building the reference dataset presented in this work was time-consuming and it was difficult to find more annotators.

Another limitation was the difficulty to found pre-trained modules to perform sentiment analysis in Portuguese and, the alternative approach that we propose based on rules, is not complete. Treating negation and adversative conjunctions is a difficult task and our rule-based approach still needs proper treatment to solve this cases.

7.4 Future Work

For future work, and given the fact that ESI is an indicator with the main objective of being an early indicator of the economic future and that there are several studies predicting GDP through the ESI, we could do the same with our news-based economic sentiment indicator and see which one (ESI or NESI) is better in predicting the GDP and the future of the economy.

Another recommendation of future work would be gather data through the Facebook and Twitter APIs in order to get a bigger amount of data and to see if we also could predict the official indicators through the social networks.

In the future, another recommendation is that the reference dataset presented in this work be labeled and classified manually by other annotators, so that there is no bias and, by improving the classification, the results could improve even more. In addition, correlation experiments can be done with sectorial confidence indicators independently, after segmenting the news by sector. And these experiments can also be performed with other economic indicators.

Finally, although our rules proved suitable for detecting the sentiment in Portuguese economic news, they are not complete. To increase the performance of our model, we should improve them, covering the greater number of economic expressions as possible. In addition, our rule-based approach still lacks proper treatment of the negation and adversative conjunctions. In the near future, in addition to the rules created, we plan to improve the classifier by treating differently words after a word such as "not" or "don’t", and consider ways of dealing with the classification of sentences with adversative conjunctions. Concerning the negation, we should classify each sentence with the opposite sentiment of the rule it matches, for example, “*unemployment did not increase*” should have a positive polarity, whereas “*unemployment increased*” has a negative one. Adversative conjunctions introduce additional challenges. They express opposition or contrast between two statements

and it is difficult even for a human, to tell the sentiment that it expresses. For example, in "*unemployment decreased but GDP increased*", the statement before the conjunction "*but*" has a positive sentiment, but the statement after it has a negative sentiment. Our rule-based approach would assign a neutral sentiment to this example, since it matches both negative and positive rules.

Bibliography

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[2] N. Ostapenko, "Macroeconomic expectations: news sentiment analysis," Bank of Estonia, Bank of Estonia Working Papers wp2020-5, 2020.

[3] C. Mendicino and M. Teresa, "Confiança E Atividade Económica : O Caso De Portugal," *Boletim Económico - Banco de Portugal*, pp. 43–53, 2013.

[4] S. Gelper and C. Croux, "On the construction of the European economic sentiment indicator," *Oxford Bulletin of Economics and Statistics*, vol. 72, no. 1, pp. 47–62, 2010.

[5] O. Claveria, E. Monte, and S. Torra, "Economic forecasting with evolved confidence indicators," *Economic Modelling*, vol. 93, no. April, pp. 576–585, 2020.

[6] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, p. 82–89, Apr. 2013. [Online]. Available: https://doi.org/10.1145/2436256.2436274

[7] L. A. Thorsrud, "Words are the New Numbers: A Newsy Coincident Index of the Business Cycle," *Journal of Business and Economic Statistics*, vol. 38, no. 2, pp. 393–409, 2018.

[8] K. Nguyen and G. L. Cava, "News Sentiment and the Economy," pp. 18–25, 2020. [Online]. Available: https://rba.gov.au/publications/bulletin/2020/jun/pdf/news-sentiment-and-the-economy.pdf

[9] A. Haldane and S. Chowla, "Fast economic indicators," *Nature Reviews Physics*, vol. 3, no. 2, pp. 68–69, 2021. [Online]. Available: http://dx.doi.org/10.1038/s42254-020-0236-y

[10] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, no. 24959, pp. 29–39, 2000.

[11] E. European Commission, *Principal European Economic Indicators*, 2009.

[12] Eurostat, "Short-term business statistics and the economic sentiment indicator," pp. 1–7, 2018. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Short-term_business_statistics_and_the_economic_sentiment_indicator

[13] P. Sorić, I. Lolić, and M. Čižmešija, "European economic sentiment indicator: an empirical reappraisal," *Quality and Quantity*, vol. 50, no. 5, pp. 2025–2054, 2016.

[14] D. G. f. E. Affairs and Financial, "The Joint Harmonised EU Programme of Business and Consumer Surveys," 2020. [Online]. Available: https://ec.europa.eu/info/files/user-guide-joint-harmonised-eu-programme-business-and-consumer-surveys_en

[15] Eurostat, "Beginners : GDP - What is gross domestic product (GDP)?" pp. 1–5, 2019. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:GDP_-_What_is_gross_domestic_product_(GDP)%3F&oldid=426966

[16] C. Bortoli, S. Combes, T. Renault, C. Bortoli, and S. Combes, "Nowcasting GDP Growth by Reading Newspapers," *Economie et Statistique / Economics and Statistics*, no. Part 1, pp. 17–33, 2019.

[17] S. VijayGaikwad, A. Chaugule, and P. Patil, "Text Mining Methods and Techniques," *International Journal of Computer Applications*, vol. 85, no. 17, pp. 42–45, 2014.

[18] L. K. Hansson, R. B. Hansen, S. Pletscher-Frankild, R. Berzins, D. H. Hansen, D. Madseni, S. B. Christensen, M. R. Christiansen, M. R. Christiansen, U. Boulund, U. Boulund, X. A. Wolf, S. K. Kjærulff, S. K. Kjærulff, M. V. D. Bunt, S. Tulin, T. S. Jensen, R. Wernersson, R. Wernersson, J. N. Jensen, and J. N. Jensen, "Semantic text mining in early drug discovery for type 2 diabetes," *PLoS ONE*, vol. 15, no. 6, pp. 1–18, 2020.

[19] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *arXiv*, 2017.

[20] T. Loya and G. Carden, *Business intelligence and analytics*. Pearson, 2015.

[21] L. Feng, Y. K. Chiam, and S. K. Lo, "Text-Mining Techniques and Tools for Systematic Literature Reviews: A Systematic Literature Review," *Proceedings - Asia-Pacific Software Engineering Conference, APSEC*, vol. 2017-December, no. May 2019, pp. 41–50, 2018.

[22] Y. Wilks, "Natural Language Processing," *Communications of the ACM*, vol. 39, no. 1, pp. 60–62, 2001.

[23] D. A. Pereira, "A survey of sentiment analysis in the Portuguese language," *Artificial Intelligence Review*, no. 0123456789, 2020.

[24] I. Boban, A. Doko, and S. Gotovac, "Sentence retrieval using Stemming and Lemmatization with different length of the queries," *Advances in Science, Technology and Engineering Systems*, vol. 5, no. 3, pp. 349–354, 2020.

[25] H. A. Almuzaini and A. M. Azmi, “Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization,” *IEEE Access*, vol. 8, pp. 127 913–127 928, 2020.

[26] D. Yuret and F. Türe, “Learning Morphological Disambiguation Rules for Turkish",” in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, Jun. 2006, pp. 328–334. [Online]. Available: https://aclanthology.org/N06-1042

[27] R. Martins, A. Pereira, and F. Benevenuto, “An approach to sentiment analysis of web applications in Portuguese,” *WebMedia 2015 - Proceedings of the 21st Brazilian Symposium on Multimedia and the Web*, pp. 105–112, 2015.

[28] S. Rani, “Sentiment Analysis: A Survey,” *International Journal for Research in Applied Science and Engineering Technology*, vol. V, no. VIII, pp. 1957–1963, 2017.

[29] A. Rajput, *Natural language processing, sentiment analysis, and clinical analytics*. Elsevier Inc., 2019.

[30] S. Yi and X. Liu, “Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers’ review,” *Complex & Intelligent Systems*, vol. 6, no. 3, pp. 621–634, 2020.

[31] M. Karamibekr and A. A. Ghorbani, “Sentiment analysis of social issues,” *Proceedings of the 2012 ASE International Conference on Social Informatics, SocialInformatics 2012*, no. SocialInformatics, pp. 215–221, 2012.

[32] R. D. Desai, “Sentiment Analysis of Twitter Data,” *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, no. Iciccs, pp. 114–117, 2019.

[33] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, “Opinion mining and sentiment analysis,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 452–455.

[34] E. Georgiadou, S. Angelopoulos, and H. Drake, “Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes,” *International Journal of Information Management*, vol. 51, no. November, p. 102048, 2020.

[35] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, “Sentiment analysis in the news,” *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, no. January, pp. 2216–2220, 2010.

[36] C. Catal and M. Nangir, “A sentiment classification model based on multiple classifiers,” *Applied Soft Computing Journal*, vol. 50, pp. 135–141, 2017.

[37] R. Sproat, C. Samuelsson, J. Chu-Carroll, and B. Carpenter, "Computational Linguistics," *The Handbook of Linguistics*, pp. 608–636, 2008.

[38] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*, E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, Eds. The AAAI Press, 2014. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

[39] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[40] D. Kapur, *Lecture Notes in Artificial Intelligence: Preface*, 2008, vol. 5081 LNAI.

[41] O. Kolesnikova and A. Gelbukh, "A study of lexical function detection with word2vec and supervised machine learning," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 2, pp. 1993–2001, 2020.

[42] K. Kinabalu, *Lecture Notes in Electrical Engineering 603 Computational Science and Technology*, 2019, no. August.

[43] P. Ongsulee, "Artificial intelligence, machine learning and deep learning," *International Conference on ICT and Knowledge Engineering*, pp. 1–6, 2018.

[44] Y. Villuendas-Rey, C. F. Rey-Benguría, Á. Ferreira-Santiago, O. Camacho-Nieto, and C. Yáñez-Márquez, "The Naïve Associative Classifier (NAC): A novel, simple, transparent, and accurate classification model evaluated on financial data," *Neurocomputing*, vol. 265, pp. 105–115, 2017.

[45] Y. Ko and J. Seo, "Automatic text categorization by unsupervised learning," in *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, ser. COLING '00. USA: Association for Computational Linguistics, 2000, p. 453–459. [Online]. Available: https://doi.org/10.3115/990820.990886

[46] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the Seventh International Conference on Information and Knowledge Management*, ser. CIKM '98. New York, NY, USA: Association for Computing Machinery, 1998, p. 148–155. [Online]. Available: https://doi.org/10.1145/288627.288651

[47] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," *Proceedings of 2014 Science and Information Conference, SAI 2014*, pp. 372–378, 2014.

[48] J. Ahmed and M. Ahmed, "A framework for sentiment analysis of online news articles," *International Journal on Emerging Technologies*, vol. 11, no. 3, pp. 267–274, 2020.

[49] A. Mohamed, "An Evaluation of Sentiment Analysis and Classification Algorithms for Arabic Textual Data," *International Journal of Computer Applications*, vol. 158, no. 3, pp. 29–36, 2017.

[50] H. M. Ahmed, M. J. Awan, N. S. Khan, A. Yasin, and H. M. F. Shehzad, "Sentiment Analysis of Online Food Reviews using Big Data Analytics," *Ilkogretim Online*, vol. 20, no. 2, pp. 827–836, 2021.

[51] L. G. Singh and S. R. Singh, "Empirical study of sentiment analysis tools and techniques on societal topics," *Journal of Intelligent Information Systems*, 2020.

[52] S. Ghannay, B. Favre, Y. Estève, and N. Camelin, "Word embeddings evaluation and combination," *Language Resources and Evaluation*, vol. 1, pp. 1–14, 05 2016.

[53] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep convolution neural networks for twitter sentiment analysis," *IEEE Access*, vol. 6, pp. 23 253–23 260, 2018.

[54] C. Dang, M. Moreno García, and F. De La Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, p. 483, 03 2020.

[55] A. Damstra and M. Boukes, "The Economy, the News, and the Public: A Longitudinal Study of the Impact of Economic News on Economic Evaluations and Expectations," *Communication Research*, p. 009365021775097, 2018.

[56] J. B. Hester and R. Gibson, "The Economy and Second-Level Agenda Setting: A Time-Series Analysis of Economic News and Public Opinion about the Economy," *Journalism & Mass Communication Quarterly*, vol. 80, no. 1, 2003.

[57] I. Fujiwara, Y. Hirose, and M. Shintani, "Can News Be a Major Source of Aggregate Fluctuations? A Bayesian DSGE Approach," *Journal of Money, Credit and Banking*, vol. 43, no. 1, pp. 1–29, 2011.

[58] M. Stanger, "A Monthly Indicator of Economic Growth for Low Income Countries," *IMF Working Papers*, vol. 20, no. 13, 2020.

[59] L. A. Thorsrud, "Nowcasting using news topics. Big Data versus big bank," Centre for Applied Macro- and Petroleum economics (CAMP), BI Norwegian Business School, Working Papers No 6/2016, Nov. 2016. [Online]. Available: https://ideas.repec.org/p/bny/wpaper/0046.html

[60] C. Huang, S. Simpson, D. Ulybina, and A. Roitman, "News-based Sentiment Indicators," *IMF Working Papers*, vol. 19, no. 273, 2019.

[61] R. Nyman, S. Kapadia, D. Tuckett, D. Gregory, P. Ormerod, and R. Smith, "News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment," *SSRN Electronic Journal*, no. 704, 2018.

[62] S. P. Fraiberger, "News Sentiment and Cross-Country Fluctuations," *SSRN Electronic Journal*, pp. 1–18, 2016.

[63] P. Aguilar, C. Ghirelli, M. Pacce, and A. Urtasun, "Can News Help Measure Economic Sentiment? An Application in COVID-19 Times," *SSRN Electronic Journal*, 2020.

[64] R. Yadav, A. V. Kumar, and A. Kumar, "News-based supervised sentiment analysis for prediction of futures buying behaviour," *IIMB Management Review*, vol. 31, no. 2, pp. 157–166, 2019.

[65] A. Mahajan, L. Dey, and S. M. Haque, "Mining financial news for major events and their impacts on the market," *Proceedings - 2008 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2008*, no. December 2008, pp. 423–426, 2008.

[66] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *Journal of Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.

[67] C. W. Calomiris and H. Mamaysky, "How news and its context drive risk and returns around the world," *Journal of Financial Economics*, vol. 133, no. 2, pp. 299–336, 2019.

[68] H. Naderi Semiromi, S. Lessmann, and W. Peters, "News will tell: Forecasting foreign exchange rates based on news story events in the economy calendar," *North American Journal of Economics and Finance*, vol. 52, no. December 2018, p. 101181, 2020.

[69] M. Song and K. Shik Shin, "Forecasting economic indicators using a consumer sentiment index: Survey-based versus text-based data," *Journal of Forecasting*, vol. 38, no. 6, pp. 504–518, 2019.

[70] A. H. Shapiro, M. Sudhof, and D. Wilson, "Measuring News Sentiment," *Federal Reserve Bank of San Francisco, Working Paper Series*, pp. 01–22, 2020.

[71] V. Krotov and L. Silva, "Legality and ethics of web scraping," *Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*, pp. 1–5, 2018.

[72] S. Sohangir, N. Petty, and D. Wang, "Financial Sentiment Lexicon Analysis," in *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, vol. 2018-January. Institute of Electrical and Electronics Engineers Inc., 4 2018, pp. 286–289.

[73] C. Tavares, R. Ribeiro, and F. Batista, "Sentiment Analysis of Portuguese Economic News," in *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*, ser. Open Access Series in Informatics (OASIcs), R. Queirós, M. Pinto, A. Simões, F. Portela, and M. J. a. Pereira, Eds., vol. 94. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, pp. 17:1–17:13. [Online]. Available: https://drops.dagstuhl.de/opus/volltexte/2021/14434

[74] L. D. Goodwin and N. L. Leech, "Understanding correlation: Factors that affect the size of r," *Journal of Experimental Education*, vol. 74, no. 3, pp. 249–266, 2006.

Web Scraping Ethical Issues

Purpose of Web Scraping

In the context of this project, this process was carried out due to the lack of data to develop our research and, we will not use the extracted data for any illegal or fraudulent purpose, it will only be used for academic and research purposes.

Copyrighted materials

With our work, we will not republish any data or information that is owned and explicity copyrighted by the website.

Damage on the website

In order to not cause problems and degradation of the websites, the scraping process performed by us was done with time intervals of 2 seconds between each request and we did simple requests as the ones performed when we manually consult the website through the web browser.

Terms of use

According to the terms of use of the websites of the two newspapers in question, they do not explicitly prohibit programmatic access to the website. In the terms of use of *Público*, "denial of service" attacks are expressly prohibited, but, as we have already mentioned, we made simple requests with time intervals, so we do not attack and create problems on the website.

Ethics of Web Scraping

In addition to legal questions, ethical questions such as those presented below can be raised in relation to the Web Scraping process. With our scraping we do not:

- Compromise privacy of any individual;
- Compromise the confidentiality and privacy of the organizations;
- Create any content or product that compete with the organizations or decrease their value. In fact, our goal is to reinforce the importance and usefulness of the news produced by them.