NOVA IMS Information Management School

MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Using Twitter News Sentiment Analysis to Forecast US GDP

José Miguel Pereira de Matos

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Using Twitter News Sentiment Analysis to Forecast US GDP

by

José Miguel Pereira de Matos

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Bruno Damásio

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Vila Nova de Famalicão, 19 of June 2024

ABSTRACT

Sentiment analysis is an emerging subject which has been growing in popularity and applicability with the growth in availability of accurate pre-trained machine learning models fine-tuned for a given domain and task (sentiment classification in this case). GDP is one of the variables which are attempted to forecast more often. Given its relevance and aggregate scope, it makes it both one of the most important to accurately forecast and also one of the hardest to do so.

This research, therefore, attempts to use sentiment analysis of news articles to forecast GDP and assess if it produces forecasting accuracy gains. Previous studies have used either news articles datasets from single sources or heuristics-based sentiment classification models. Here, a dataset composed of thousands of news articles extracted from several sources present in Twitter/X will be used and classified with pre-trained context-aware machine learning models.

An ARIMAX enriched with exogenous variables produced by such sentiment analysis will produce rolling forecasts alongside a base ARIMA and both forecasts will be compared to assess the existence of forecasting accuracy gains.

The findings show that this approach does generate statistically significant forecasting accuracy gains for the one and four steps-ahead forecasting horizons. Such findings point to the possibility of including exogenous variables coming from news articles sentiment classification in the models currently in use for GDP forecasting by policy-making agents. Additionally, using sources like the one explored in this research allows an easy process of fine-tuning the specific type of news articles used as a basis for such a forecast, making this approach very flexible to be applied across adjacent scopes.

KEYWORDS

Sentiment analysis; ARIMAX, Time Series Forecasting; Twitter

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

GLOSSARY OF ABBREVIATIONS

GDP	Gross Domestic Product
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average Exogenous
AIC	Akaike's Information Criterion
BIC	Bayesian Information Criterion
HQIC	Hannan-Quinn Information Criterion
DM test	Diebold-Mariano test
RMSE	Root Mean Squared Error
MSE	Mean Squared Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
VAR	Vector Autoregressive
MIDAS	Mixed-Data Sampling

1 Introduction

This thesis seeks to ascertain the extent to which the sentiment analysis of economic news article tweets can contribute to forecast GDP. The United States was selected as a case study, due to its status as the largest economy in the world and, perhaps even more relevant, due to its prominence in economic discourse, ensuring a substantial volume of relevant news articles.

The selection of Twitter/X as the source of these news articles may appear to be a risky decision, given the reputation of social media as an unreliable and sensationalist source of information. However, such drawback may be offset by its potential to provide a more accurate indication regarding the general tone underlying economic news. Additionally, it provides relatively easy access to a large number of news articles from very diverse sources across a sufficiently large timeline in one single platform. Furthermore, its use as a source is also often cited in suggestions for further work, in literature similar to the research at hand, such as in Jiang & Zeng (2023).

Drawing from the literature gathered, two aspects about similar work became obvious, most of it had been concerned with analysing the titles of the news articles and most of them had performed sentiment analysis using dictionary-based techniques on a word or a sentence level. This research seeks to perform context-aware sentiment analysis, by using pre-trained machine learning (ML) models, on the textual corpus (rather small, coming from tweets, nevertheless richer in content and less prone to ambiguities in its tone than the title alone).

The assessment of the relevance of the technique previously described is to be tested regarding its ability to generate short-term forecasting power gains. Given a short forecasting horizon and the economics field's familiarity with it, the base ARIMA model will be built for US GDP and used as a benchmark. Due to its simple way of incorporating exogenous regressors for time series forecasting and its similar nature, ARIMAX will be used as the model to incorporate sentiment analysis for GDP forecasting.

A comparison between the forecasts produced by the base and the sentiment analysis-enriched approaches is to be performed. In the event of its results indicating a statistically significant difference in forecasting power, this thesis aims to both expand the body of research suggesting the inclusion of news sentiment analysis for the purposes of forecasting economic variables and serve as a framework on how to apply it for professionals who could benefit from it.

The remainder of the document will be organized as follows: Section 2 will go through the literature most similar to this research and draw recommendations from it to assist the decision making and elaboration of the methodology, Section 3 will present the methodology used to achieve the empirical results, Section 4 will present said empirical results and Section 5 will draw conclusions from those results and make suggestions for further research.

2 Literature Review

2.1 General Context Regarding Economic Forecasting

Forecasting economic cycles and accurately detecting turning points is one of the most aggregate forecasts attempted, given the variety and number of factors that influence it and the variety of components of GDP, especially in modern economies. Its importance is also relevant for professional use in the private sector, as it can be crucial for a banker to make timely investment decisions or to adjust its lending practice (Lai & Ng, 2020). Policy makers also give it vital importance, as they are probably the most interested in accurate forecasts of the business cycle to help guide decisions regarding public spending, taxation and monetary policy, as described in Urasawa (2014) and Giannone et al. (2008).

According to both Hawkins (2005) and Bob Namvar (2010), macroeconomic forecasting has had a long history. With the first mentions coming from religious texts (namely the Bible) and encompassing the prediction of future food production in Ancient Egyptian, Greek and Roman times. Although ambiguous and more heavily based on superstition than on data, these highlight the relevance macroeconomic forecasting has been given since ancient times.

Since its initial works from William Petty, Gregory King and Charles Davenant in the 16th century (Geweke et al., 2006), however, forecasting of macroeconomic events and tendencies based on data has risen to dominance over the field.

As per Hawkins (2005), the 1910-1930 did see the rise of a forecasting industry in the US, but also its demise after failing to foresee the Great Depression of 1929. It was not until the post Second World War period that macroeconomic forecast as we know it today came into existence, following the Keynesian Revolution.

2.2 General Context Regarding Sentiment Analysis

To understand sentiment analysis and sentiment analysis models, it is important to, first, understand the algorithm that is the base for all of them, Artificial Neural Networks (ANN's).

According to Wang & Raj (2017), Warren McCulloch and Walter Pitts modelled a primitive neural network based on their findings after speculating the inner workings of neurons (McCulloch & Pitts, 1943). Their model came to be known as the McCullogh-Pitts model (MCP) neural model and it resembled the modern perceptron, although the former's weights between neurons were fixed as opposed to the adjustable weights of the latter. Donald O. Hebb introduced the Hebbian Learning Rule in Hebb (1949), laying the groundwork for the future development of the modern perceptron.

Still as per Wang & Raj (2017), Frank Rosenblatt further developed the Hebbian Learning Rule with the introduction of perceptrons (Rosenblatt, 1958). Unlike Hebb and previous theorists who focused on the biological system, Rosenblatt built an electronic device which demonstrated the ability to learn in accordance with associativism.

The 1970's were a period of frustration regarding Artificial Neural Networks, as Minsky & Papert (2017) described the limitations of single layer perceptrons to solve non-linear problems, which kickstarted what some authors call an "AI Winter". The book suggested the need to use Multi-Layer Perceptrons to address those kinds of problems, although there was still no way to train them. The development of backpropagation for Neural Networks (Werbos & John, 1974) and its posterior real-world implementation (LeCun et al., 1989) allowed the use of multi-layer Perpetrons, which, in turn, put an end to this "winter".

As described by Otter et al. (2019), the field of Natural Language Processing (NLP) concerns a variety of topics involving computational processing and the understanding of human languages. The work in this field can be divided into two main sub-areas: core areas and applications. In this research, the emphasis will be put on the sub-area of applications, more specifically regarding document classification as in classification of its general sentiment.

Since the 1980's the field has increasingly relied on data-driven computation involving statistics, probability and ML approaches such as Naive Bayes, K-Nearest Neighbours, decision trees and support vector machines.

Nevertheless, recent increases in computational power and parallelization, consubstantiated by Graphical Processing Units (GPU's) now allow the usage of ANN's with billions or trillions of parameters. These advances have allowed for a transformation in which previously used techniques have either been entirely replaced or enhanced by neural models.

2.3 State-Of-The-Art Regarding Sentiment Analysis for Economic Forecasting

In this research, we propose to combine the power of previously discussed recent developments in NLP for sentiment classification of documents, more specifically news article tweets to address the issue of forecasting GDP.

A rather simple approach to our topic is documented by Ferrari Minesso et al. (2023). It gathers articles from major US-based newspapers exclusively from Factiva Analytics and filters it for the ones written in English and covering domestic news (from the American perspective). The sentiment score for the approach in question is based on an index that tracks the ratio between the number of articles discussing a recession (this is achieved by filtering news articles for the keywords "recession" and "slowdown") relative to the total number of articles published about the economy. The forecasting framework discussed by Ferrari Minesso et al. (2023) is based on probit regressions where the probability of being in a recession at a future horizon t + k is explained by the slope of a curve (originally the yield curve), in this case, the curve of the newspaper-based index.

Nowcasting, as suggested by Bańbura et al. (2013) concerns the prediction of the present, very near future and very recent past and has recently started to be used in economics. Ashwin et al. (2021) attempts to address such issue using daily-frequency sentiment scores (extracted from the Factiva database) and monthly indicators such as the Purchasing Manager's Index (PMI), separately, for Euro-Area countries. The news dataset for this approach is derived from several European newspapers and processed to be attributed a sentiment metric. Such research obtains larger forecasting gains with the textual sentiment metrics at the beginning of

each quarter, as compared to the official European Central Bank (ECB) and the PMI-based projections, when other indicators are not yet available. The article also hints at a solution for a particular obstacle that is raised when attempting to perform news sentiment analysis for cross-national and, specifically, non-English speaking geographies. Most of the NLP literature and pre-built and pre-trained models for sentiment classification available are built on the English language. After trying three different solutions, the one employed by Ashwin et al. (2021) is to standardize its article dataset to the English language by making use of the Google Translate API for Python. Regarding the classification of the overall sentiment of each news article, this article follows through with seven different approaches, some using word level and one using sentence level sentiment classification (Hutto & Gilbert, 2014).

Huang et al. (2018) attempts to assess whether the inclusion of a news sentiment metric is useful to improve the performance of current economic downturn forecasts. Two components of news are considered. The first and most obvious is the general sentiment (positive or negative) of each news article being published. But it also addresses a second and less obvious component, the overall concentration of topics in the news. This second consideration comes from speculating that during recessions, a majority of news articles will be dedicated to the downturn itself, as opposed to when the economy is in an upswing, news stories related to financial markets or general economic concerns are sparser.

Its approach to build a time series is similar to Ashwin et al. (2021), in which it seeks to make a daily aggregation of the general sentiment of the news article of that day. For each of the news articles themselves, it suggests calculating the overall sentiment (positive or negative), first scoring each individual word that composes it as -1, 0, 1(for negative, neutral and positive, respectively) and averaging that score over the number of words that composes it.

Both Huang et al. (2018) and Ashwin et al. (2021) used relatively simple dictionary-based methods to measure the sentiment of the news corpuses and build their sentiment score time series. A reason for this can be found in Chong et al. (2022), in which the importance of the consideration of nuance and context beyond heuristic rules for accurate sentiment classification is discussed, but its work still follows through with a dictionary-based approach. It mentions several examples of successful applications of ML models trained in large corpus of text mapped between textual utterances and human-labelled sentiment rating. Nevertheless, these techniques are dismissed as labelling such a large dataset would be very time-consuming and expensive.

A promising alternative can, however, be found in Lukauskas et al. (2022), which leverages pre-trained ML models to build its sentiment score. Its approach is based on making used of ML models from the Hugging Face repository (and taking advantage of its Transformers python library) which have already been trained, either on a general corpus or, sometimes already re-trained in domain-specific corpus. Given that the models in question can be directly deployed before further training, they solve the obstacle regarding the costs of labelling large enough text datasets and using them for training.

Jiang & Zeng (2023) attempt to leverage this approach for prediction of Stock Market Movement. Despite a relatively different domain, their research provides useful hints. It starts by using the pre-trained sentiment analysis model FinBERT (Araci, 2019) (domain specific) to classify news from several publicly available financial news datasets and posteriorly build a sequence model with a Long-Short Term Memory network (LSTM) with the previous results merged with financial price information. As a baseline for the usefulness of the domain specific

FinBERT, this research used a Base BERT (Devlin et al., 2019) pre-trained model and a LSTM with the same approach. This baseline helped conclude that the power of FinBERT was not taken advantage of, as its approach barely outperformed the one using Base BERT, leaving them with the suggestion to try the use of full-text news articles and alternative information sources like social media (Twitter/X, etc.). For context, domain adaptation is proved effective by Peng et al., (2021).

Nisar & Yeung (2018) adopt a rather different approach to sentiment analysis, by attempting to prove the "Twitter Effect" in the stock market. This Effect corresponds to the possible influence social media (in this case Twitter/X) may have over real events, almost as a self-fulfilling prophecy. Its approach is based on performing sentiment analysis on common tweets, rather than news texts or headlines. The tweets on which the analysis was performed were filtered by keywords relevant to the indicators which this study attempted to predict.

Apart from specific practical examples, a comprehensive overview of the attempt to now- or forecast economic and financial variables can be found in Algaba et al. (2020). Three relevant topics for the research to be developed are discussed in the referenced literature. The gain in performance between lexicon-based and machine learning-based sentiment classification approaches is further recognized as the benchmark provided in Ribeiro et al. (2016) justifies. Different types of aggregation of sentiment data are discussed. Within-Document or within-text aggregation is a natural step for lexicon-based approaches, but only relevant for ML-based ones if large individual news texts (possible paragraph-level sentiment analysis to further aggregate) are used for the analysis. Across documents aggregation at a given frequency will be particularly relevant as it is necessary to build a sentiment measure time-series. Across-time aggregation aimed at smoothing the obtained time-series after across-document aggregation is discussed and justified as a way to remove outliers, especially for high-frequency series (daily or weekly). Smoothing the sentiment time-series through a moving average as in Thorsrud (2020) or using a Kalman filter is suggested.

The most relevant insight from the paper above is, nevertheless, about how to model the sentiment measure or measures to the variable to be predicted. The simpler way described is to use an Ordinary Least Squares (OLS) regression when possible. When not possible, due to either multicollinearity or a high ratio of dimensionality relative to sample size, Ridge (Hoerl & Kennard, 2000) and LASSO (Tibshirani, 1996) approaches either individually or in combination are suggested.

A promising alternative to incorporate exogenous regressors in GDP forecasting is found in Andrianady et al. (2023), which uses the ARIMAX model to incorporate a *CRISIS* dummy variable (referring to existence or lack thereof of a political crisis in a given period) as the exogenous regressor into a base ARIMA model of Madagascar's GDP. In a paper by the same author (Andrianady, 2023), an overview of econometric models for the same forecasting goal can be found, comparing base ARIMA, VAR and MIDAS. Such overview reaches the conclusion that the ARIMA model was superior for forecasting Madagascar's GDP.

3 Materials and Methods

3.1 Datasets

The main dataset, used in the context this research, is composed of news article tweets from the Twitter/X social network obtained through web scraping, made possible through the use of a Python library, Twscrape, proposed by vladkens (2023).

The web scraping process was directed towards tweets posted by several reputable American newspapers and media outlets focused on economic and financial topics. The following is a list of those newspapers and media outlets: Bloomberg, Fox Business, Seeking Alpha, CNN, Financial Times, the Wall Street Journal, NBC News, CBS news, CNBC, Barron's, Reuters, Sky News and The Street.

To further restrict the topics of the tweets scrapped, filtering by Keyword was used, in a similar way to Araci (2019), in which the same type of filtering was performed for his financial communication training dataset. The Keywords used in this filter were the following: "recession", "economy", "unemployment", "stock market", "GDP", "inflation". The rationale behind using "recession", "economy" and "GDP" is quite obvious, research attempting to predict the evolution of GDP based on economic news will inevitably look for news articles which mention it more or less directly. Regarding the other three keywords used, those were picked as they look for news articles mentioning variables which can in a larger or smaller way impact economic performance. Mohseni & Jouzaryan (2016) conduct a test on the effect of both unemployment and inflation on the Iranian economy (particularly its GDP), concluding both variables did prove relevant both in the short and in the long term. Sa'idu & Muhammad (2015) examine the effect of the same variables on the Nigerian economy and find significance only for the effect of inflation.

A date filter was also used, restricting the tweets scrapped from the 1st January of 2007 to the 31st December of 2021, allowing the dataset to encompass the periods of both the 2008 Financial Crisis and the 2020 Covid-19 related Crisis, while preventing the capture of periods of Twitter's very early adoption phase, which would have led to very low volumes of news articles extracted for those same periods.

This news dataset suffered several cleaning and formatting operations. These consisted in the following: eliminating entries with duplicate news texts, cleaning the news texts from non-textual noise (links, hashtags, and so on), eliminating news articles extracted from sources or by keywords not included in the ones previously mentioned (this only happened for a few dozen news articles, so a bug on the part of the extraction library's code is assumed) and restricting the news texts to the relevant geographic scope of this research (the US). The latter was implemented by looking for geographical "cue keywords", which could indicate the geographical scope they refer to, specifically: "US", "U.S.", "America", "American", "United States", "USA", "U.S.A". The entries whose news texts contained the keywords mentioned were kept as pertaining to the geographical scope appropriate to this research. Figure 1 provides a comprehensive flowchart, regarding the filtering and preprocessing performed on the original news article dataset and described above.

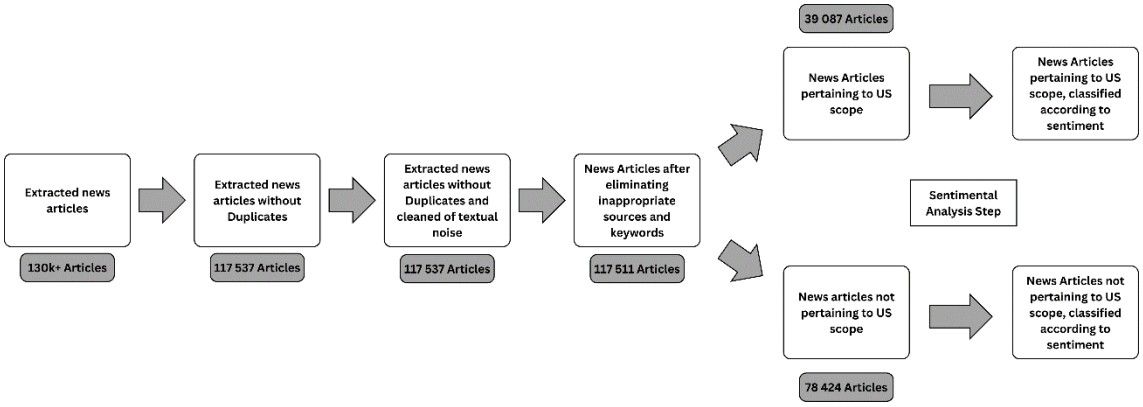


Figure 1: Flowchart for News Dataset Preprocessing

The volume of news, discriminated by source and keyword is provided in Table 1. While the volume of news per year and keyword or per year and source can be found in the Appendix section.

Table 1: Volume of News Sources vs Keywords

Sources	GDP	economy	inflation	recession	stock market	unemployment	Total
BBC News (World)	5	395	52	85	65	70	672
Barron's	36	379	132	90	268	74	979
Bloomberg	281	3052	1240	669	629	523	6394
CBS News	11	409	31	56	50	186	743
CNBC	199	1773	341	213	614	390	3530
CNN	21	727	39	135	100	249	1271
FOX Business	156	1075	164	54	177	190	1816
Financial Times	144	885	308	109	118	107	1671
MarketWatch	363	1743	730	317	1332	660	5145
NBC News	8	399	17	52	28	239	743
Reuters	291	2511	796	455	309	404	4766
Seeking Alpha	561	1121	744	771	1038	164	4399
Sky News	37	620	108	112	117	82	1076
The Economist	187	971	177	162	68	126	1691
The Wall Street Journal	153	1420	285	189	255	354	2656
TheStreet	115	631	223	84	388	94	1535
Total	2568	18111	5387	3553	5556	3912	39087

A second important dataset obtained corresponds to a dataset which contains the seasonality-adjusted real (inflation discounted) quarterly GDP figures for the US and serves as a target dataset for this forecasting research. This dataset was obtained from the U.S. Bureau of Economic Analysis (1946). Since this indicator is available quarterly since 1947, the complete series from 1947 until 2023 has been extracted.

3.2 Pre-trained Models and the Inference Process

As previously discussed, building on the suggestion from Peng et al. (2021) that domain adaptation does generate gains in terms of sentiment classification quality, this research aims to use such approach, particularly selecting one of the models examined in said paper (FinBERT by Araci (2019)). FinBERT is a Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) model further pre-trained in a massive corpus of financial communication and fine-tunned for sentiment classification, with the last activation function being SoftMax, for multiclassification purposes.

Another similar model was employed in this research, FinBERT-Tone (Huang et al., 2023). Also a BERT model, further pre-trained on three different types of texts: corporate annual and quarterly fillings, financial analyst reports and earnings conference call transcripts and fine-tunned for sentiment classification on 10 000 sentences from analyst reports. This model was chosen as an alternative to the first one as it is further pre-trained on the same base (BERT) but on a more diverse dataset.

The inference process was performed with the help of the Transformers library (Wolf et al., 2020) for Python. It provides easy to use methods such as *pipeline*, used for inference in the research at hand, which allows the user to specify the desired tokenizer, inference model and task at hand. It takes a string (our news article text) as input and retrieves a key-value pair, corresponding to the sentiment classification (positive, neutral and negative) and respective probability. For context, this tokenizer takes a chunk of text and divides it into its smallest possible parts and transforms them into their numerical representation, so that they can be inputted into the classification model at hand.

A general idea about the proportions of such classification for the whole period for which economic news articles are available can be observed in Table 2, discriminated by the model which performed the classification. A significant dissimilarity can be inferred about both sentiment analysis models in what regards to sentiment classification, given the significant differences in the volume of news classified as neutral and as negative by both.

Table 2: Sentiment Classification for the Whole Period (FinBERT vs FinBERT-tone)

label_FinBERT	count	label_FinBERT-Tone	count
negative	17021	**neutral**	18123
neutral	12754	**negative**	12629
positive	9312	**positive**	8335

Overall, 13603 out of the 39087 news articles have been classified differently by the two models used regarding the sentiment tone expressed in them. The level of agreement between the two models for each type of classification can be observed on Table 3. A higher level of agreement is observed for neutral classifications. However, when disagreement happens in case of a neutral classification, it tends to divide itself into the positive and negative categories equally, unlike the disagreement in case of a positive or a negative classification, which tends much more heavily to flow to a neutral classification. The latter kind of disagreement between

models can be understood as indicating a lack of certainty regarding the classification, rather than a tendency to produce perfectly opposite classifications for the same news article. Given this level of disagreement between the models in question, two classifications for each news article have been kept, one performed by each sentiment classification model.

Table 3: FinBERT vs FinBERT-Tone Sentiment Classification

label_FinBERT	label_FinBERT-tone	count	FinBERT-Tone/FinBERT
negative	**negative**	10157	59,67%
	neutral	5356	31,47%
	positive	1508	8,86%
neutral	**neutral**	9892	77,56%
	negative	1470	11,53%
	positive	1392	10,91%
positive	**positive**	5435	58,37%
	neutral	2875	30,87%
	negative	1002	10,76%

An additional feature was created based on the previous classification, a sentiment score. This feature is obtained through multiplying the probability of a tweet being the label it was given (the value in the key-value pair mentioned above) by 0 if the label is neutral, by -1 if the label is negative or by 1 if the label is positive.

3.3 Grouping and Feature Engineering

As mentioned on the previous section, the output of the sentiment analysis is a key-value pair for each input row of the dataset, in this case for each tweet. However, to be able to build a predictive approach for a time series, the processed dataset needs to be standardized for a given periodicity, therefore creating a time series to forecast the target one. Datasets standardized for a daily, monthly, quarterly and yearly periodicity were created. Such has been achieved by calculating the percentage of news tweets classified as negative, neutral or positive for each time unit (day, month, quarter or year), for the case of the base features. For the case of the sentiment score, all the sentiment scores of all the news articles contained within a given time unit have been averaged, being said average the aggregated version of the sentiment score feature.

Some descriptive statistics regarding both sentiment scores can be observed in Table 4. The same descriptive statistics, now calculated after excluding neutrally classified news articles (classified as such by either one of the two sentiment classification models), are provided in Table 5.

The two models' disagreement in sentiment classification, inevitably spilled over to the distributions of the scores that are built on top of it. Such a discrepancy between the two can be further analysed in Tables 4 and 5 through some descriptive statistics of those same

sentiment scores, consubstantiated by a less dispersed distribution of FinBERT's version of the sentiment score as compared to FinBERT-Tone's (as can be witnessed by their Quartiles and their standard deviations), suggesting a higher degree of conservatism in its predictions. Given such differences, two instances of the sentiment score were kept, one for the results of each sentiment classification model.

Table 4: Descriptive Statistics for Sentiment Scores before Neutral Exclusion

Descriptive Statistics	score_FinBERT-Tone	score_FinBERT
count	39087	39087
mean	-0,102	-0,180
std	0,689	0,667
min	-1,000	-0,977
Quartile 25%	-0,940	-0,889
Quartile 50%	0,000	0,000
Quartile 75%	0,000	0,000
max	1,000	0,959

Table 5: Descriptive Statistics for Sentiment Scores after Neutral Exclusion

Descriptive Statistics	score_FinBERT-Tone	score_FinBERT
count	18102	18102
mean	-0,218	-0,269
std	0,931	0,820
min	-1,000	-0,977
Quartile 25%	-1,000	-0,950
Quartile 50%	-0,927	-0,804
Quartile 75%	0,997	0,777
max	1,000	0,959

Neutrally classified news article tweets compose the majority of the dataset used in this research, when looking at FinBERT-Tone's classification and the second largest cohort when looking at FinBERT's. Since they can undershoot the relevance of positive or negative sentiment in percentage-based features and dramatically reduce the value of the sentiment score at a given time step, the decision was made to test the effect and relevance of their inclusion in the calculation of the features already mentioned.

Henceforth, removing the neutrally classified news articles, before performing the calculations which result in the mentioned features, can tackle such issue. Two datasets which exclude neutrally classified news articles have been built (removing the news articles classified as neutral by one or the other model), allowing for feature calculation without the mentioned adverse effect. Keeping and following along with the transformations of such alternative datasets will allow the comparison between forecasting performance resulting from this approach and the approach which keeps neutrally classified news articles, therefore providing insights to whether this issue significantly impacts forecasting power.

Given that until this stage, only eight features had been created (and that, in the approaches where instances classified as neutral were removed, the negative and positive percentages would become highly correlated and so, not very advisable to use together), additional features had to be created. A difference variable (which is called diff, for simplicity) was created for all the eight previous variables, that represents the difference in percentage points in the percentage of news tweets of a given sentiment classification from a time unit t and a time

unit $t-1$, in the case of percentage based features, or a simple difference between the sentiment score observed in time unit t and in time unit $t-1$, in the case of sentiment score features. Given that this research concerns time series forecasting and that lag variables are widely used for the task, lag variables for each of the difference variables were created up to $t-5$. This means that for every row of the dataset (which represents a time unit t), diff variables pertaining to time units $t-1, t-2, \ldots$ until $t-5$ are present.

Given the presence of lag variables, the series was truncated to the third quarter of 2008 at the beginning to eliminate the presence of null values.

3.4 ARIMA Benchmarking

Naturally, to be able to assess the usefulness of a given new model or of a given set of new variables to forecast a given time series variable, one or several benchmark models need to exist to compare the goodness of fit and the out-of-sample forecasting performance.

Given this and following the work developed by Andrianady (2023), in this study, a base ARIMA model will be used as the benchmark.

$$\Delta^d y_t = c + \phi_1 \Delta^d y_{(t-1)} + \cdots + \phi_p \Delta^d y_{(t-p)} - \theta_1 \varepsilon_{(t-1)} - \cdots - \theta_q \varepsilon_{(t-q)} + \varepsilon_t \qquad (1)$$

Equation 1 defines the general form of such model, in which y_t is the original time series, c is a constant term, $\phi_1, \ldots, \phi_p$ are the coefficients of the autoregressive (AR) terms, $\theta_1, \ldots, \theta_q$ are the moving average (MA) terms and ε_t is the error term at time t.

Regarding the (p, d, q) terms to be defined by the user, p is the order of the AR terms, q is the order of MA terms and d represents the order of differencing. Δ^d represents the differencing operator applied d times as in $\Delta^d y_t = y_t - y_{t-1}$ for $d = 1$.

The GDP time-series used for this purpose corresponds to the previously mentioned target time-series of seasonally adjusted real GDP with quarterly frequency (U.S. Bureau of Economic Analysis, 1946). This series has been extracted without null values, so inputting these will not be necessary. From this point on, the GDP series used will be the subset between the third quarter of 2008 and the last quarter of 2019. The motivation behind the cut in the beginning of the series is to match the temporal horizon of the sentiment analysis variables' series available. The GDP series was also truncated at the end as to avoid modelling during the Covid-19 crisis, since it introduces too much volatility which led to a situation in which the random walk process would be the most significant model (lowest AIC and BIC measures).

The Box-Jenkins methodology (Box & Jenkins, 1976) is used to specify the benchmark ARIMA model. According to said methodology as described by Makridakis & Hibon (1997), five steps should be followed before selecting an ARIMA model for forecasting. Firstly, trend and seasonal stationarity must be verified in the working time-series. Alternatively, the series should be differenced as to obtain said stationarity conditions.

After stationarity is verified, model specification should ensue with the differencing (d) parameter having presumably been estimated already as to guarantee stationarity. Next, the moving average (q) and auto-regressive (p) parameters should be defined. These can be inferred from the Autocorrelation and the Partial Autocorrelation Function plots, respectively, by selecting the lags which fall outside the boundaries of statistical insignificance.

Having found a set of possible ARIMA specifications, all of these can be fitted, and diagnostic checks should be performed on them. All the models which pass the diagnostic tests should then be compared by goodness of fit measures and variables' significance (through checks of the P-Values for the variables which compose them) and the best one should be selected as the most accurate ARIMA. This most accurate ARIMA is then used as the base specification for the following sentiment analysis-enriched ARIMAX models. The optimal ARIMA will also serve as a benchmark against which all sentiment analysis-enriched models explored in this research will be compared, in terms of out-of-sample forecasting performance.

3.5 Input Filtering by Source and Keyword

Given that the original news dataset used to create the sentiment analysis series was extracted from specific sources on the Twitter/X platform and by specific keywords, it is a sensible approach to test if all this sources and keywords really worked to extract news article tweets relevant to the scope of this research. Such is justified by the fact that including a large volume of news articles which reside outside of our scope in the calculation of the time-series of sentiment analysis variables can heavily affect the performance of the forecasts produced and, therefore, lead us to the wrong conclusions.

The method used in this research to deal with this issue will consist in creating several different groups of datasets (each containing three, corresponding to the different approach explored to deal with neutrally classified news articles) consisting in different source and keyword filtering combinations performed before the grouping stage and after the news sentiment classification stage.

The following sections will describe an approach to determine the optimal neutral handling approach and the optimal combination of exogenous variables for each group of datasets, producing an optimal model for each input filtering combination.

The input filtering approaches which will be attempted are the following:

1. No US scope filtering;
2. Only US scope filtering (every approach after this one will contain US scope filtering);
3. Removing news article tweets extracted from Seeking Alpha;
4. Removing news article tweets extracted by the "stock market" keyword;
5. Removing news article tweets extracted by the "stock market" or "inflation" keywords;
6. Removing news article tweets extracted by the "stock market" keyword and/or from a group of more generalist sources ("CNN", "CBS News", "NBC News" and "Sky News");

7. Removing news article tweets extracted by the “stock market” or “inflation” keywords and/or from a group of more generalist sources (“CNN”, “CBS News”, “NBC News” and “Sky News”);
8. Removing news article tweets extracted by the “stock market” or “inflation” or “unemployment” keywords;
9. Removing news article tweets extracted by the “stock market” or “inflation” or “unemployment” keywords and/or from a group of more generalist sources (“CNN”, “CBS News”, “NBC News” and “Sky News”)

Despite extensive filtering performed in the case of some of the approaches described above, the volumes of news articles kept after said filtering are still sizeable and sufficient for modelling, as can be observed in Figure 2.

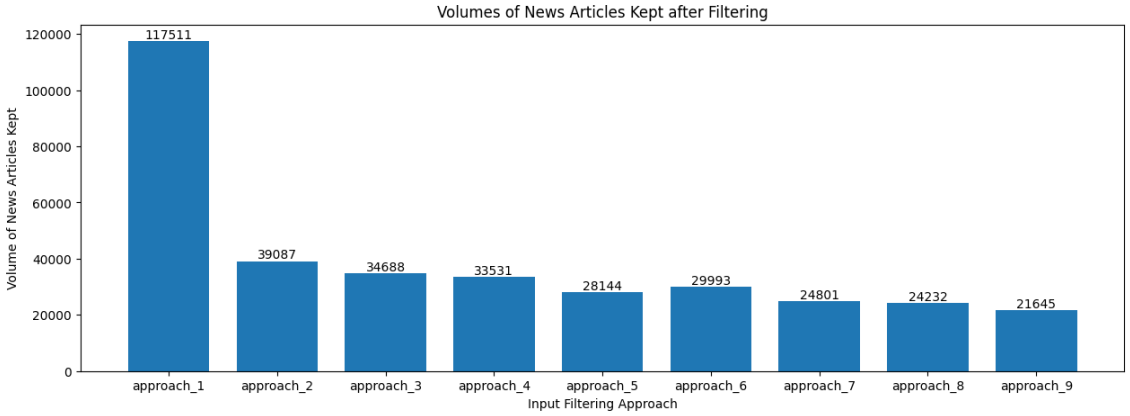


Figure 2: Volume of News Articles Kept after Filtering by Approach

3.6 Model Comparison and the Diebold-Mariano Test

Having defined a benchmark (ARIMA) for comparing the significance of the sentiment analysis exogenous variables in improving the forecast of GDP and aspiring to discriminate between several competing models, concerns regarding this comparison itself start to arise. Classically, out-of-sample forecasting accuracy measures such as RMSE, MSE, MAE and MAPE could be compared between models, but the possibility that one model had produced one such metric marginally better than the other purely by chance cannot be discarded, as mentioned by Mohammed & Mousa (2020). Therefore, a statistical test is warranted to validate statistically different forecasting power between two models, which is where the Diebold-Mariano test (Diebold & Mariano, 1995) comes into the picture.

The DM test provides a null hypothesis of equal out-of-sample accuracy of the forecasts of two models and an alternative hypothesis of inequality, in its two-sided version. In its one-sided

version, the alternative hypothesis is of the forecasts of one model having greater accuracy than the forecasts of the other.

The DM test allows for the specification of the power of the test, so the power to which errors will be raised. In this research, the conventional power of 2 will be used, therefore comparing squared errors.

Equation 2 formulates the DM test:

- Actual values: $\{y_t : t = 1, \dots, T\}$
- Two forecasts: $\{\hat{y}_{1,t} : t = 1, \dots, T\}$ and $\{\hat{y}_{2,t} : t = 1, \dots, T\}$
- Forecast error: $e_{i,t} = \hat{y}_{i,t} - y_t$, $i = 1, 2$
- Loss function: $g(e_{i,t}) = e_{i,t}^p$, $p =$ power of the test (2)

- Difference between forecasts: $d_t = g(e_{1,t}) - g(e_{2,t})$
- Diebold-Mariano two-sided test: $H_0 : E(d_t) = 0$ and $H_1 : E(d_t) \neq 0$
- Diebold-Mariano one-sided test: $H_0 : E(d_t) = 0$ and $H_1 : E(d_t) < 0$ or $H_1 : E(d_t) > 0$

This research used the Python implementation for the two-sided DM test proposed by Tsang (2017) and the R implementation proposed for the one-sided DM test proposed by Hyndman & Khandakar (2008).

3.7 ARIMAX

An ARIMAX model, built as an extension of the previous ARIMA enriched with exogenous sentiment analysis variables will be used as the model which incorporates the sentiment analysis exogenous regressors, in a similar approach to Andrianady et al. (2023).

$$\Delta^d y_t = c + \phi_1 \Delta^d y_{(t-1)} + \cdots + \phi_p \Delta^d y_{(t-p)} - \theta_1 \varepsilon_{(t-1)} - \cdots - \theta_q \varepsilon_{(t-q)} + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t \quad (3)$$

Equation 3 defines the general form of the ARIMAX model and expands Equation 1 with exogenous variables $x_{1,t}, \dots, x_{k,t}$ at time t, $\beta_1, \dots, \beta_k$ representing the coefficients of such exogenous variables, with k being the number of exogenous variables.

For out-of-sample forecasting, both the log-GDP series and the series of the exogenous regressors have been divided in a kind of a train-test split, with a cut-off on the second quarter of 2017. Out-of-sample forecasting will be performed for up to the four steps-ahead forecasting horizon. For each one of these forecasting horizons a rolling window approach will be employed with the initial "training set" in which every model will be fitted corresponding to the first part of the previous split. The window will then shift forward by one quarter and the

models will again be fitted in this new window and perform a new forecast until there are not enough records available in the second split of the datasets to perform this process again.

As discussed previously, several variables and its lags have been created using the original sentiment analysis variables as a basis for them and these must be filtered for the statistically significant ones, furthermore, feature selection should also be pursued as to create a more parsimonious model. Additionally, three different datasets were created for each time unit of aggregation, regarding the handling of neutrally classified news article tweets, so one of these approaches should also be selected. In this section, only the quarterly aggregation is relevant given that the models from the ARIMA family do not support exogenous variables of frequencies different from the one of the target variable.

In this research, this variable filtering and selection is performed in two steps, one in-sample step and one out-of-sample step.

Sentiment analysis exogenous regressors will be included in the ARIMAX model both alone and in pairs.

It is important to note that only lagged exogenous variables can be used. ARIMAX models demand observations of the exogenous regressors for the same timestep for which forecasts are to be performed. This would require knowing the future observations of the sentiment analysis variables, which would not be possible in a typical forecasting scenario, as it would imply being able to scrape and classify news tweets which have not yet been posted at the time of the forecast. To address this issue, one can restrict the exogenous regressors used in a forecasting exercise to the ones whose observations in the future are merely lagged feature transformations of base features observed in the present. In the case of a one step-ahead forecast, the ARIMAX model would demand the observation of exogenous regressor x at timestep $t+1$, which is only problematic if x is really an observation at timestep $t+1$. If x is the lagged version of an exogenous variable by 1 timestep, its observation at timestep $t+1$ will already exist at timestep t. As proposed by Chu & Qureshi, (2023), forecasts are produced for each period $t+h$ for up to four quarters ahead ($h = 1, 2, 3, 4$) and the information used to forecast for the period $t + h$ is done exclusively using information available at t.

For the in-sample step, the ARIMA model selected as described in the previous sections will be enriched with all the possible pair combinations of exogenous variables and with all the same variables individually.

This process will be repeated once for every neutral handling approach (so three times). All the possible models described above will be fitted and only the ones whose AIC and BIC metrics are inferior to those of the base ARIMA model and whose P-Values for the significance of the variables which compose them are all below 0.05 will be selected for each neutral handling approach. If one of the neutral handling approaches does not contain any selected model, it will be automatically discarded. Equations 4 and 5 represent the formulas for the AIC and BIC measures, respectively.

$$AIC = 2k - 2\ ln(\hat{L}) \tag{4}$$

$$BIC = k\ ln(n) - 2\ ln(\hat{L}) \tag{5}$$

Where k represents the number of parameters estimated in the model, n represents the sample size and $\hat{L}$ represents the maximized value of the likelihood function.

For the out-of-sample step, it will be divided into two parts: selecting an optimal ARIMAX model for one, two, three and four steps-ahead forecasting, with the eventual possibility that the optimal models for some of those horizons will be the same. To note that, the forecasts are relative to the log-GDP series as this is the one inputted into the models. Nevertheless, to compute meaningful forecasting accuracy metrics, which can be interpreted, it is useful to reverse the natural logarithm transformation and transform the forecasts into the original scale of the forecasted variable (GDP), by exponentiating (e^x) the forecasted values. All the models which survive the previous step will go through the previously described rolling window out-of-sample forecasting process and the Root Mean Squared Error (RMSE) of their forecasts will be calculated against the real values of the GDP series.

The model with the lowest RMSE will be selected as optimal and its forecasts will be compared to the forecasts of the base ARIMA model and to the actual GDP series through a Diebold-Mariano test. Therefore, the optimal neutral-handling approach, the optimal exogenous variable or set of variables will be known and its forecast improvement significance will be known.

The last step is relevant especially given the fact that this whole process will be repeated for every input filtering approach and so the DM test P-Values of the optimal ARIMAX models can be compared between the results obtained through the different input filtering approaches. Therefore, if any P-Value's below 0,1 are produced by the one-sided DM test (rejecting its null hypothesis for the less conservative confidence level of 10%, therefore indicating statistically significant better forecasts produced by the ARIMAX as compared to the base ARIMA) based on any input filtering approach, it can be said that such approach can produce statistically significant forecasting power gains. Otherwise, the comparison between DM test P-Values can indicate where to continue input filtering further.

The formula for the RMSE can be observed in Equation 6.

$$RMSE = \sqrt[2]{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \quad (6)$$

Where y_t and $\hat{y}_t$ represent, respectively, the actual value of the GDP and its forecast at period t of the test dataset and T represents the number of forecasts performed.

4 Results and Implementation

4.1 ARIMA Benchmarking

Given that the extracted GDP series extracted is seasonally adjusted, seasonal stationarity is assumed to exist.

The GDP series is not trend stationary as it contains an upward trend, which calls for differencing to make it as such. Log-differencing was used to achieve stationarity, as it is a common methodology when working with economic and financial data, consisting of transforming the original series into its natural logarithm and performing a simple differencing of order 1. The logarithmic transformation is used to address stationarity in the series' variance and the differencing is used to address stationarity in the mean as suggested by Makridakis & Hibon (1997). After said transformation, the series passes an Augmented Dickey-Fuller test (Dickey & Fuller, 1979) (the Null hypothesis is rejected with a P-Value of 0,003, thus the alternative hypothesis of stationarity is accepted) and visual inspection for stationarity.

The d parameter for the ARIMA models tested can therefore be assumed to be 1 and the base series to be used will be the logarithm transformed one.

As can be seen in figures 3 and 4, the plots for both functions mentioned before show that the possible values for the q parameter can be 0 or 1 and for the p parameter can also be 0 or 1.

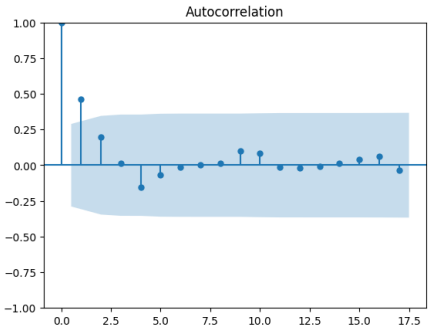


Figure 4: Autocorrelation Function Plot

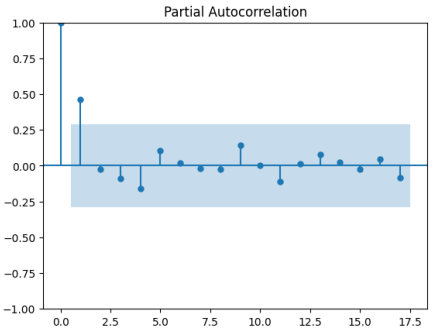


Figure 3: Partial Autocorrelation Function Plot

Thus, leaving as possible ARIMA models to test: ARIMA (0, 1, 0), ARIMA (1, 1, 0), ARIMA (0, 1, 1) and ARIMA (1, 1, 1).

All possible ARIMA specifications were fitted on the now truncated GDP log series and tested for residuals autocorrelation, through both the Breusch-Godfrey (Godfrey, 1996) and the Ljung-Box (Ljung & Box, 1978) tests, for whose P-Values were always above 0.99, thus not rejecting the null hypothesis in every case, proving the absence of autocorrelation of residuals for every ARIMA model proposed, thus validating them.

The valid ARIMA specifications were ordered by AIC and BIC, and the P-Values for all their respective components were checked, leaving the ARIMA (1, 1, 0) and the ARIMA (1, 1, 1) specifications in a tie for the first place, since they displayed slightly conflicting goodness-of-fit measures. The former displaying a lower BIC (327.8 VS -329.6) and HQIC (-331.2 VS -331.8) and the latter displaying a lower AIC (-333.22 VS -333.19). Both specifications displayed P-Values for regressor significance below 0.05 for all the regressors they incorporate.

The out-of-sample forecasting performance can be measured in several metrics. In the context of this model being built as a benchmark, RMSE will be used as it is the one in which the candidate exogenous-enriched models will be measured as well. When looking at out-of-sample forecasting accuracy, it becomes clear that the ARIMA (1, 1, 1) specification outperforms the ARIMA (1, 1, 0) specification for all forecasting horizons relevant to this research as can be observed in Table 6. Therefore, the former is chosen as the specification to proceed with both as a benchmark and as a basis to the exogenous-enriched models discussed in the following section.

Table 6: Out-of-Sample Forecasting Accuracy Measure for ARIMA Specifications

ARIMA (1, 1, 1)	RMSE	ARIMA (1, 1, 0)	RMSE
1 steps-ahead	63,94	**1 steps-ahead**	81,44
2 steps-ahead	123,77	**2 steps-ahead**	188,10
3 steps-ahead	164,98	**3 steps-ahead**	294,99
4 steps-ahead	207,55	**4 steps-ahead**	380,75

4.2 ARIMAX Results and Implementation

As discussed in the Methodology section, ARIMAX models with one and two exogenous regressors have been tested and meaningful models have been produced for all approaches of input filtering, in both cases.

4.2.1 Single Exogenous Variable Approach

Using only one Exogenous Regressor on the ARIMAX model built on top of the previously discussed base ARIMA, all input filtering approaches have produced models which survive the in-sample first phase of model selection and can perform forecasts for up to four steps-ahead, with the exception of input filtering approach 1, which was only able to produce models for up to two steps-ahead forecasting.

Regarding the handling of neutrally classified news tweets, most of the optimal ARIMAX models for every input handling approach for any steps-ahead forecasting horizons were built with features which used some approach of neutral removal.

In Figure 5, a comparison of out-of-sample RMSE for one step-ahead forecasting can be observed between the optimal models of every input filtering approach. The models built with input filtering approaches 3 and 4 register the lowest RMSE's for one-step ahead forecasting with values of 56.3 and 56.8, respectively. These were also the only models whose forecasting residuals registered a P-Value inferior to 0.1 in a one-sided DM test against the base ARIMA model for the forecasting horizon in question, with values of 0.065 and 0.084, respectively, therefore rejecting its null hypothesis for the less conservative 10% significance level. These models were built with features which used neutral removal as classified by the FinBERT and FinBERT-Tone sentiment classification models, respectively.

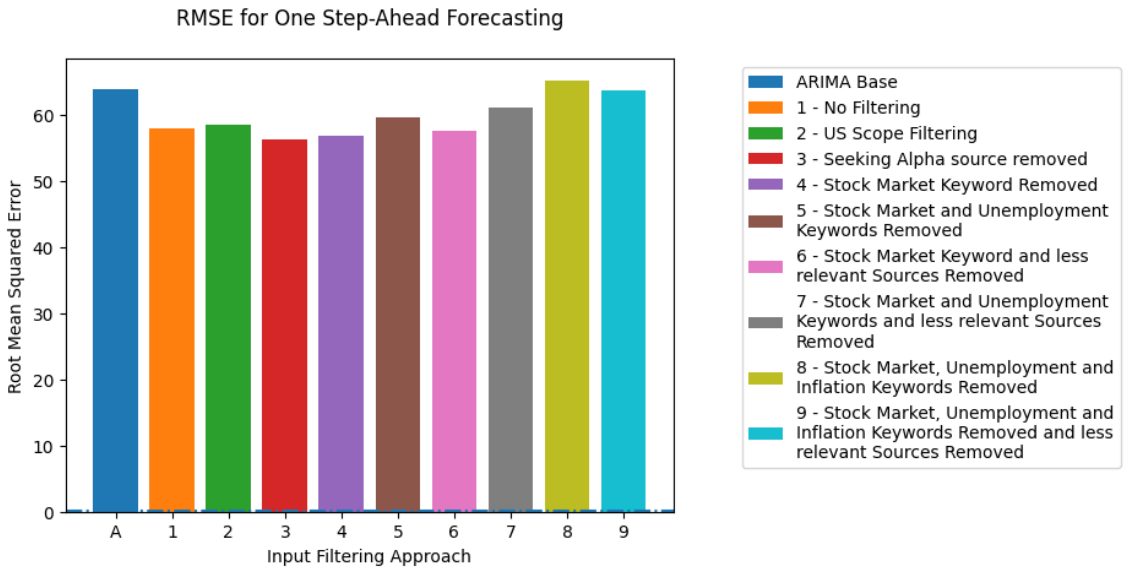


Figure 5: One-Step Ahead RMSE with Single Exogenous Regressor

The exogenous regressors used in such models can be observed in Table 7.

Table 7: Exogenous Regressors Used in Relevant ARIMAX model for One Step-Ahead Forecasting

Input Filtering Approach	Variable Used	Neutral Handling Approach
3 - Seeking Alpha source removed	ln_negative_percentage_tone_n-4	Neutral_Removal_by_FinBERT
4 - Stock Market Keyword Removed	ln_negative_percentage_tone_n-4	Neutral_Removal_by_FinBERT-Tone

For the 2 and 3 steps-ahead forecasting horizons, no model's forecasting residuals have been able to reject the null hypothesis of the one-sided DM test against the base ARIMA model even for the less conservative 10% significance level. For the two steps-ahead forecasting horizon, the model built with input filtering approach 1 registered the lowest RMSE of 104.1. For the three steps-ahead forecasting horizon, the model built with input filtering approach 3 registered the lowest RMSE of 138.8.

A comparison of out-of-sample RMSE for four steps-ahead forecasting can be observed in Figure 6. In this instance, the results are not as straightforward as in the previous. The lowest

RMSE's can be found for the models built with input filtering approaches 7, 9 and 5, with values of 136.2, 138.2 and 142.9, respectively. Nevertheless, the only model whose forecasting residuals registered a P-Value inferior to the less conservative 0.1 threshold was the one built input filtering approach 6, with a value of 0.079. The latter was built with features whose calculation used neutral removal as classified by the FinBERT sentiment classification model.

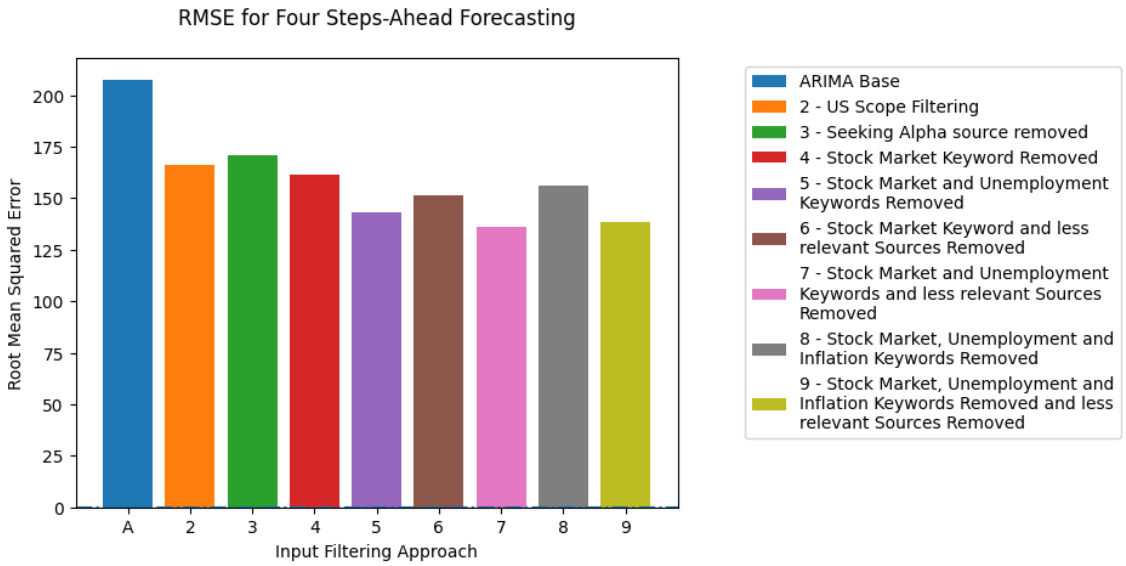


Figure 6: Four Steps-Ahead RMSE with Single Exogenous Regressor

The exogenous regressors used in such models can be observed in Table 8.

Table 8: Exogenous Regressors Used in Relevant ARIMAX model for Four Step-Ahead Forecasting

Input Filtering Approach	variable_used	neutral_handling_technique
5 - Stock Market and Unemployment Keywords Removed	tone_score_n-4	Neutral_Removal_by_FinBERT
6 - Stock Market Keyword and less relevant Sources Removed	tone_score_n-4	Neutral_Removal_by_FinBERT
7 - Stock Market and Unemployment Keywords and less relevant Sources Removed	tone_score_n-4	Neutral_Removal_by_FinBERT
9 - Stock Market, Unemployment and Inflation Keywords Removed and less relevant Sources Removed	tone_score_n-4	Neutral_Removal_by_FinBERT

4.2.2 Pairs of Exogenous Regressors Approach

Using pairs of Exogenous Regressors on the ARIMAX model built on top of the previously discussed base ARIMA, all input filtering approaches have produced models which survive the in-sample first phase of model selection and can perform forecasts for up to four steps-ahead.

Regarding the handling of neutrally classified news article tweets, unlike the previously described approach of using only one exogenous regressor, in this case the situation is more mixed, with about half of the optimal models being built with features which used some approach of neutral removal and half being built with features which did not.

A comparison of out-of-sample RMSE for one step-ahead forecasting can be observed in Figure 7. This is the only forecasting horizon for which more than one sentiment analysis enriched ARIMAX model can significantly outperform the base ARIMA in terms of RMSE and for which any came close to rejecting the null hypothesis for the one-sided DM test. The model built with input filtering approach 6 can easily be selected as the one with the lowest value for the metric, at 50.2. This was not the model which came the closest to rejecting the null hypothesis for the one-sided DM test, with a P-Value of 0.156, but the one built with input filtering approach 7, with a P-Value of 0.147. Nevertheless, no model's forecasting residuals registered a P-Value on the one-sided DM test capable of rejecting the null hypothesis, even for the less conservative 10% significance level. Both models were built with features which used neutral removal as classified by the FinBERT sentiment classification model.

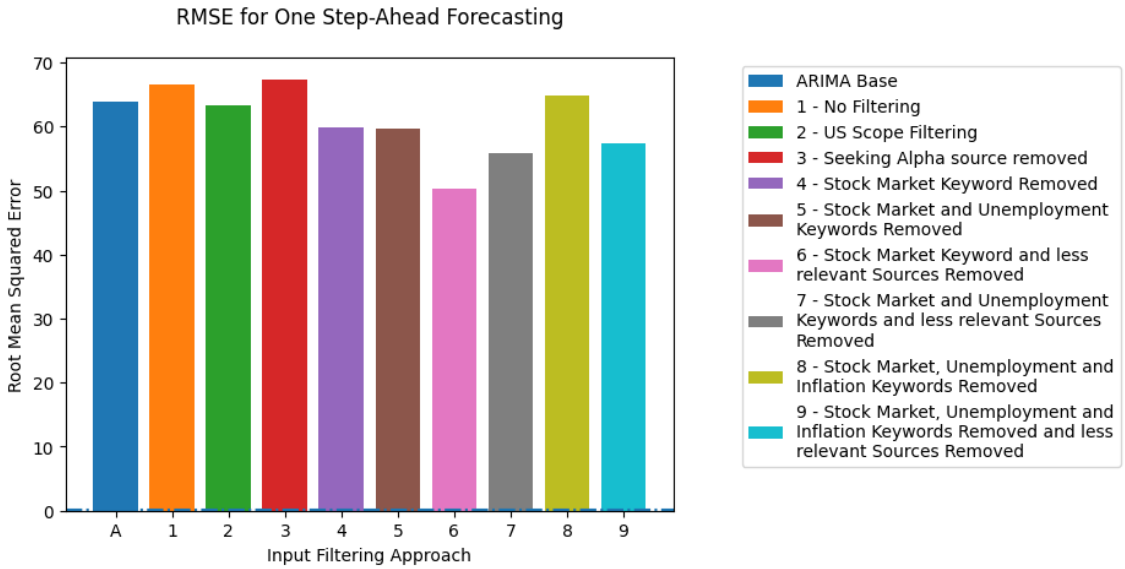


Figure 7: One Step-Ahead RMSE with Pairs of Exogenous Regressors

The exogenous regressors used in such models can be observed in Table 9.

Table 9: Pairs of Exogenous Regressors Used in Relevant ARIMAX model for One Step-Ahead Forecasting

Input Filtering Approach	variables_used	neutral_handling_technique
6 - Stock Market Keyword and less relevant Sources Removed	tone_negative_n-5, var_tone_positive_n-1	Neutral_Removal_by_FinBERT
7 - Stock Market and Unemployment Keywords and less relevant Sources Removed	tone_negative_n-5, var_tone_positive_n-5	Neutral_Removal_by_FinBERT

As for the results regarding the other forecasting horizons, the model built with the input filtering approach 6 registered the lowest RMSE for the two steps-ahead forecasting horizon at 114.34. The model built with the input filtering approach 7 registered the lowest RMSE for the three steps-ahead forecasting horizon at 164.5, the only to slightly beat the base ARIMA for this forecasting horizon, which registered an RMSE value of 164.98. Finally, for the four steps-

ahead forecasting horizon, the model built with the input filtering approach 3 registered the lowest RMSE at 183.4.

5 Conclusions

After exploring the results of the proposed approach, some conclusions about the following topics can be reached: overall forecasting capacity of news articles sentiment analysis, inclusion of neutrally classified news and choice of Keywords and Sources for news article extraction. The following section will also feature an overview of some of the limitations faced throughout the research at hands.

5.1 Forecasting Power of News Article Tweets Sentiment Analysis

As explored earlier, using the single exogenous regressor approach, it can be concluded that news sentiment analysis does add forecasting power to a traditional forecasting method such as an ARIMA. This is true, at least for the one and four steps-ahead forecasting horizons. Such a conclusion is based on the results of the one-sided DM test which, for the forecasting horizons mentioned, did reject its null hypothesis at the 10% confidence level.

As for the approach using pairs of exogenous regressors, such a conclusion did not materialize, as no sentiment analysis enriched model produced forecasts which could reject the null hypothesis of the one-sided DM test.

As can be observed in the previous section, all the models which produced meaningful results, so statistically significant forecasting power gains, were built with exogenous regressors whose values were calculated on top a dataset which used some approach of removal of neutrally classified news article tweets.

Additionally, an interesting effect can be observed in which all the exogenous regressors incorporated in the models which proved meaningful were calculated with sentiment classification performed by the FinBERT-Tone model, while having had neutrally classified news article tweets removed from the dataset they used based on FinBERT's classification. This effect is a statement to the complementarity of different sentiment classification models for the same original dataset and task.

The effect excluding some of the sources from which news articles are extracted or keywords according to which they are extracted was also explored. The first step in such exploration was an attempt to circumscribe the geographical scope of the news articles extracted. Such a step can be concluded as successful, since the approach which used the dataset prior to it did not serve as the base for any meaningful models. Removing the news article tweets extracted through the "stock market" and "inflation" keywords has also proven effective given such steps were applied to the datasets used in four and three out of the five meaningful models produced, respectively. Therefore, it can be concluded that geographical circumscription of the scope of the news article tweets extracted in the terms performed does indeed produce more meaningful models. Additionally, it can also be concluded that sentiment classification of news article tweets extracted through the "stock market" and "inflation" keywords does not improve GDP forecasting.

5.2 Limitations and Suggestions for Further Work

5.2.1 Time Scope Limitation

One important limitation of this study is the limited time scope of the exogenous regressors' dataset, which has hindered the capacity of the models created to be fitted in a big enough dataset which could have allowed a better comprehension of patterns between them and the variable being forecasted (GDP), therefore allowing it to achieve better out-of-sample forecasting accuracy and, subsequently, more significant differences between them and the base ARIMA's. This limitation stems from an availability issue, nevertheless, making it hard to overcome, as this study already tried to extract news from Twitter/X from as far back as possible. Nonetheless, assuming a continued publication of news article tweets from reputable sources in said social media platform, the availability of longer and longer datasets for the same regressors built on top of said news article tweets will only grow.

5.2.2 Model Selection Methodology and Computing Power

Although using only one sentiment analysis variable at a time produced the best results at hands, testing the incorporation of more than two exogenous regressors could turn out to be a good way to get even more significant results, especially for other forecasting horizons. Nevertheless, such an attempt would require either a different model selection methodology than the one used or more computing power than the one which was available for this research. This is the case, since the model selection methodology used exhaustively tests all the possible exogenous regressors combinations for goodness-of-fit and increasing the size of the combinations to three or more would cause an exponential increase in the number of combinations to test and, therefore, an exponential increase in the machine time it would take to perform such a selection.

5.2.3 Econometrics Based Approach and Linearity Limitations

In the context of this research, an ARIMAX model was used for forecasting due to its similarity to the benchmark model, therefore ease of comparability of the exogenous regressors added and due its simplicity and widespread knowledge among professionals of the economics and statistics field.

Nevertheless, due to its construct, it only allows for linear relationships between the regressors and the regressand, thus losing the potential gains in predictive power that allowing for a non-linear relationship (perhaps closer to their true relationship) between the two could have. Such non-linearities could be modelled in future projects using hybridization techniques with ML models for forecasting purposes, such as the case described by Yucesan et al. (2018), which models an ANN on the residuals of the ARIMAX model previously used, as those residuals are assumed to contain the non-linearity information which was not captured by the ARIMAX.

5.2.4 Additional Scopes for Future Research

An interesting project to develop following the approach laid out in this research, relates to the application of the described methods to forecasts other macroeconomic variables, such as unemployment.

Business-related matters, interesting for corporations, such as the real-time monitoring of the popularity of one's competitors can be improved through sentiment classification of tweets selected by keywords referring to such competitors' brands and products. Such monitoring could produce popularity metrics which could be further used to forecast the competitors' gain of market share. Such an approach could also be experimented to track the users' own firm, allowing for a better understanding of consumer's opinions about the products and services they provide, in a larger way than surveys could.

BIBLIOGRAPHY

Algaba, A., Ardia, D., Bluteau, K., Borms, S., & Boudt, K. (2020). Econometrics Meets Sentiment: An Overview of Methodology and Applications. *Journal of Economic Surveys*, *34*(3), 512–547. https://doi.org/10.1111/joes.12370

Andrianady, J. R. (2023). Crunching the Numbers: A Comparison of Econometric Models for GDP Forecasting in Madagascar. *MPRA Paper*, Artigo 116916. https://ideas.repec.org//p/pra/mprapa/116916.html

Andrianady, R. J., Ranaivoson, M. H. P., Randriamifidy, F., & Miora, T. (2023). *Econometric Analysis and Forecasting of Madagascar's Economy: An ARIMAX Approach* (SSRN Scholarly Paper 4593283). https://doi.org/10.2139/ssrn.4593283

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models* (arXiv:1908.10063). arXiv. https://doi.org/10.48550/arXiv.1908.10063

Ashwin, J., Kalamara, E., & Saiz, L. (2021). Nowcasting Euro Area GDP with News Sentiment: A Tale of Two Crises. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3971974

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Chapter 4—Now-Casting and the Real-Time Data Flow. Em G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 2, pp. 195–237). Elsevier. https://doi.org/10.1016/B978-0-444-53683-9.00004-9

Bob Namvar, P. (2010). Economic Forecasting. *2000 Volume 3 Issue 1*, *1*. https://gbr.pepperdine.edu/2010/08/economic-forecasting/

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis, forecasting and control* (Rev. ed). Holden-Day.

Chong, E., Ho, C. C., Ong, Z. F., & Ong, H. H. (2022). Using news sentiment for economic forecasting: A Malaysian case study. *IFC Bulletins Chapters*, *57*. https://ideas.repec.org//h/bis/bisifc/57-17.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Em J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dickey, D., & Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *JASA. Journal of the American Statistical Association*, *74*. https://doi.org/10.2307/2286348

Diebold, F., & Mariano, R. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Ferrari Minesso, M., Lebastard, L., & Le Mezo, H. (2023). Text-Based Recession Probabilities. *IMF Economic Review*, *71*(2), 415–438. https://doi.org/10.1057/s41308-022-00177-5

Geweke, J., Horowitz, J. L., & Pesaran, M. H. (2006). *Econometrics: A Bird's Eye View* (SSRN Scholarly Paper 947529). https://doi.org/10.2139/ssrn.947529

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676. https://doi.org/10.1016/j.jmoneco.2008.05.010

Godfrey, L. G. (1996). Misspecification tests and their uses in econometrics. *Journal of Statistical Planning and Inference*, *49*(2), 241–260. https://doi.org/10.1016/0378-3758(95)00039-9

Hawkins, J. (2005). Economic forecasting: History and procedures. *Economic Round-Up: Journal of the Department of the Treasury*, 1–10.

Hebb, D. O. (1949). *The organization of behavior; a neuropsychological theory* (pp. xix, 335). Wiley.

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.2307/1271436

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text*. *Contemporary Accounting Research*, *40*(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Huang, M. Y., Rojas, R. R., & Convery, P. D. (2018). *News Sentiment as Leading Indicators for Recessions* (arXiv:1805.04160). arXiv. https://doi.org/10.48550/arXiv.1805.04160

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Artigo 1. https://doi.org/10.1609/icwsm.v8i1.14550

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*, 1–22. https://doi.org/10.18637/jss.v027.i03

Jiang, T., & Zeng, A. (2023). *Financial sentiment analysis using FinBERT with application in predicting stock movement*.

Lai, H., & Ng, E. C. Y. (2020). On business cycle forecasting. *Frontiers of Business Research in China*, *14*(1), 17. https://doi.org/10.1186/s11782-020-00085-3

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). *Handwritten Digit Recognition with a Back-Propagation Network*. Neural Information Processing Systems. https://www.semanticscholar.org/paper/Handwritten-Digit-Recognition-with-a-Network-LeCun-Boser/86ab4cae682fbd49c5a5bedb630e5a40fa7529f6

Ljung, G. M., & Box, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, *65*(2), 297–303. https://doi.org/10.2307/2335207

Lukauskas, M., Pilinkienė, V., Bruneckienė, J., Stundžienė, A., Grybauskas, A., & Ruzgas, T. (2022). Economic Activity Forecasting Based on the Sentiment Analysis of News. *Mathematics*, *10*(19), Artigo 19. https://doi.org/10.3390/math10193461

Makridakis, S., & Hibon, M. (1997). ARMA Models and the Box–Jenkins Methodology. *Journal of Forecasting*, *16*(3), 147–163. https://doi.org/10.1002/(SICI)1099-131X(199705)16:3<147::AID-FOR652>3.0.CO;2-X

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Minsky, M., & Papert, S. A. (2017). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press. https://doi.org/10.7551/mitpress/11301.001.0001

Mohammed, F. A., & Mousa, M. A. (2020). Applying Diebold–Mariano Test for Performance Evaluation Between Individual and Hybrid Time-Series Models for Modeling Bivariate Time-Series Data and Forecasting the Unemployment Rate in the USA. Em O. Valenzuela, F. Rojas, L. J. Herrera, H. Pomares, & I. Rojas (Eds.), *Theory and Applications of Time Series Analysis* (pp. 443–458). Springer International Publishing. https://doi.org/10.1007/978-3-030-56219-9_29

Mohseni, M., & Jouzaryan, F. (2016). Examining the Effects of Inflation and Unemployment on Economic Growth in Iran (1996-2012). *Procedia Economics and Finance*, *36*, 381–389. https://doi.org/10.1016/S2212-5671(16)30050-8

Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, *4*(2), 101–119. https://doi.org/10.1016/j.jfds.2017.11.002

Otter, D. W., Medina, J. R., & Kalita, J. K. (2019). *A Survey of the Usages of Deep Learning in Natural Language Processing* (arXiv:1807.10854). arXiv. https://doi.org/10.48550/arXiv.1807.10854

Peng, B., Chersoni, E., Hsu, Y.-Y., & Huang, C.-R. (2021). *Is Domain Adaptation Worth Your Investment? Comparing BERT and FinBERT on Financial Tasks*. 37–44. https://doi.org/10.18653/v1/2021.econlp-1.5

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., & Benevenuto, F. (2016). SentiBench—A benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, *5*(1), Artigo 1. https://doi.org/10.1140/epjds/s13688-016-0085-1

Rosenblatt, F. (1958) The Perceptron A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65, 386. - References—Scientific Research Publishing. (sem data). Obtido 25 de novembro de 2023, de https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55.))/reference/referencespapers.aspx?referenceid=2067071

Sa'idu, B., & Muhammad, A. (2015). Do Unemployment and Inflation Substantially Affect Economic Growth? *Journal of Economics and Development Studies*, *3*. https://doi.org/10.15640/jeds.v3n2a13

Thorsrud, L. A. (2020). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics*, *38*(2), 393–409. https://doi.org/10.1080/07350015.2018.1506344

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tsang, J. (2024). *Johntwk/Diebold-Mariano-Test* [Python]. https://github.com/johntwk/Diebold-Mariano-Test (Trabalho original publicado 2017)

Urasawa, S. (2014). Real-time GDP forecasting for Japan: A dynamic factor model approach. *Journal of the Japanese and International Economies*, *34*, 116–134. https://doi.org/10.1016/j.jjie.2014.05.005

U.S. Bureau of Economic Analysis. (1946, janeiro 1). *Gross Domestic Product*. FRED, Federal Reserve Bank of St. Louis; FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/GDP

vladkens. (2023, julho 9). How to still scrape millions of tweets in 2023 using twscrape. *Medium*. https://medium.com/@vladkens/how-to-still-scrape-millions-of-tweets-in-2023-using-twscrape-97f5d3881434

Wang, H., & Raj, B. (2017). *On the Origin of Deep Learning*.

Werbos, P., & John, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences /*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. Em Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yucesan, M., Gul, M., & Celik, E. (2018). Performance Comparison between ARIMAX, ANN and ARIMAX-ANN Hybridization in Sales Forecasting for Furniture Industry. *Drvna Industrija*, *69*(4), 357–370. https://doi.org/10.5552/drind.2018.1770

APPENDIX

Annex 1

Dataset Counts after US Scope Filtering

Sources	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	All
BBC News (World)	58	118	113	83	67	47	40	17	14	13	6	7	3	59	27	672
Barron's	0	0	1	4	2	3	2	5	25	20	27	47	113	473	257	979
Bloomberg	0	0	19	68	60	62	97	111	151	200	297	705	1137	1952	1535	6394
CBS News	0	2	14	32	29	21	34	42	7	21	39	95	60	206	141	743
CNBC	0	0	17	87	75	89	100	90	96	172	220	480	573	1006	525	3530
CNN	0	12	11	11	7	2	3	2	7	42	56	139	243	530	206	1271
FOX Business	0	0	42	183	45	54	51	82	122	197	248	619	0	31	142	1816
Financial Times	0	0	73	50	79	80	118	154	195	190	169	220	175	89	79	1671
MarketWatch	28	33	29	57	101	140	236	313	223	277	447	655	571	987	1048	5145
NBC News	0	9	16	4	10	21	36	17	16	15	22	106	91	274	106	743
Reuters	14	63	104	44	54	81	66	49	44	101	179	458	806	1568	1135	4766
Seeking Alpha	0	0	0	42	276	141	159	173	259	339	210	26	278	181	2315	4399
Sky News	1	104	48	11	17	11	26	25	40	67	36	51	91	341	207	1076
The Economist	0	9	20	70	98	72	56	68	124	109	89	141	188	296	351	1691
The Wall Street Journal	0	16	19	63	58	103	147	115	153	209	181	279	282	618	413	2656
TheStreet	0	0	0	0	0	0	17	310	202	140	116	77	122	241	310	1535
All	101	366	526	809	978	927	1188	1573	1678	2112	2342	4105	4733	8852	8797	39087

Annex 2

dates	GDP	economy	inflation	recession	stock market	unemployment	All
2007	4	57	24	3	12	1	101
2008	7	194	51	71	23	20	366
2009	24	237	23	150	17	75	526
2010	75	350	122	73	78	111	809
2011	110	329	205	158	43	133	978
2012	110	414	110	103	62	128	927
2013	189	495	121	105	103	175	1188
2014	239	650	203	117	190	174	1573
2015	227	750	198	196	216	91	1678
2016	244	922	279	309	224	134	2112
2017	242	964	446	133	362	195	2342
2018	285	2059	503	226	708	324	4105
2019	283	2402	486	898	455	209	4733
2020	342	4560	355	857	1247	1491	8852
2021	187	3728	2261	154	1816	651	8797
All	2568	18111	5387	3553	5556	3912	39087

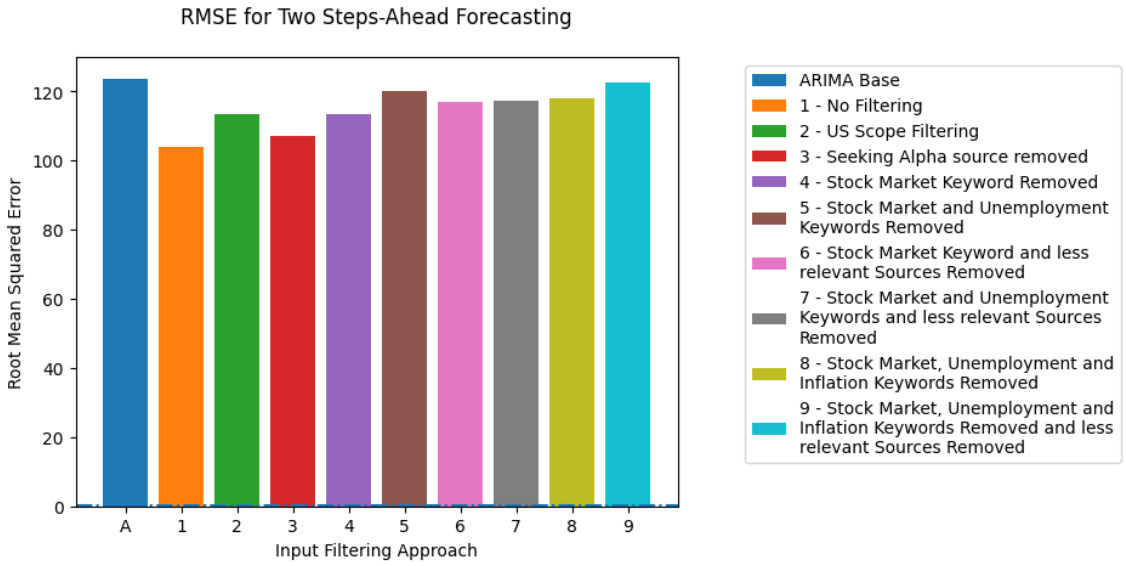
Annex 3

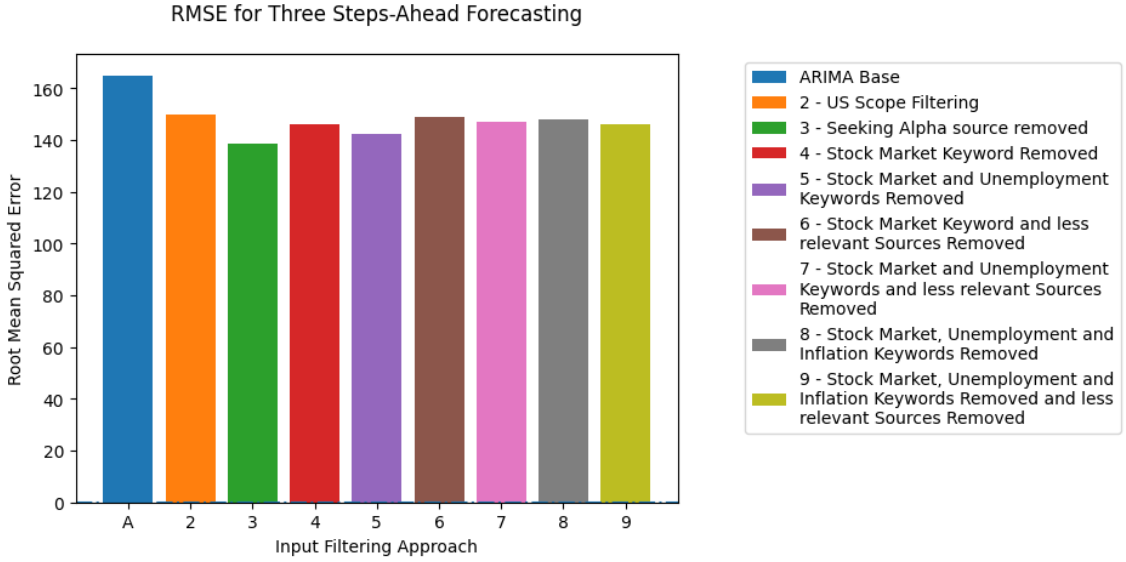
Full Description of Exogenous Regressors Used

Code	Meaning
neutral_percentage_tone	Percentage of news classified as neutral by the FinBERT-Tone model
neutral_percentage_prosus	Percentage of news classified as neutral by the FinBERT model
positive_percentage_tone	Percentage of news classified as positive by the FinBERT-Tone model
positive_percentage_prosus	Percentage of news classified as positive by the FinBERT model
negative_percentage_tone	Percentage of news classified as negative by the FinBERT-Tone model
negative_percentage_prosus	Percentage of news classified as negative by the FinBERT model
tone_negative_n-1 ... Tone_negative_n-5	Lags of negative_percentage_tone up to n-5
tone_positive_n-1 ... tone_positive_n-5	Lags of positive_percentage_tone up to n-5
prosus_negative_n-1 ... prosus_negative_n-5	Lags of negative_percentage_prosus up to n-5
prosus_positive_n-1 ... prosus_positive_n-5	Lags of positive_percentage_prosus up to n-5
tone_score	Sum of sentiment score based on FinBERT-Tone
prosus_score	Sum of sentiment score based on FinBERT
tone_score_n-1 ... tone_score_n-5	Lags of tone_score up to n-5
prosus_score_n-1 ... prosus_score_n-5	Lags of prosus_score up to n-5
ln_negative_percentage_tone	natual logarithm transformation of negative_percentage_tone
ln_negative_percentage_tone_n-1 ... ln_negative_percentage_tone_n-5	Lags of negative_percentage_tone up to n-5
diff_tone_negative_pct	Difference Between negative_percentage_tone and tone_negative_n-1
diff_tone_positive_pct	Difference Between positive_percentage_tone and tone_positive_n-1
diff_prosus_negative_pct	Difference Between negative_percentage_prosus and prosus_negative_n-1
diff_prosus_positive_pct	Difference Between prosus_percentage_tone and prosus_positive_n-1
var_tone_negative_n-1 ... var_tone_negative_n-5	Lags of diff_tone_negative up to n-5
var_tone_positive_n-1 ... var_tone_positive_n-5	Lags of diff_tone_positive up to n-5
var_prosus_negative_n-1 ... var_prosus_negative_n-5	Lags of diff_prosus_negative to n-5
var_prosus_positive_n-1 ... var_prosus_positive_n-5	Lags of diff_prosus_positive to n-5

Annex 4

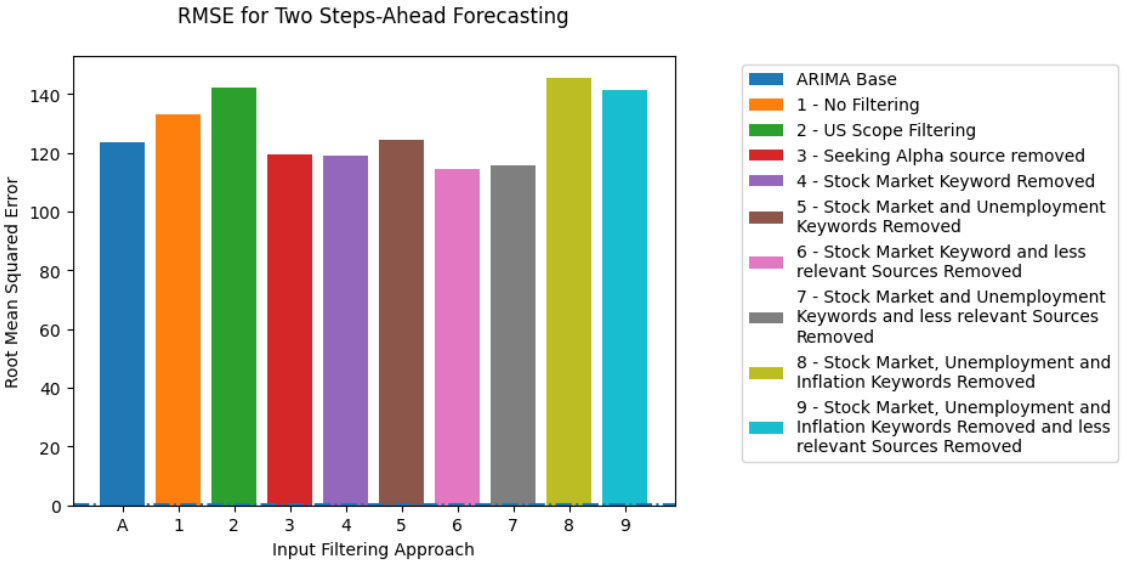
Plot of RMSE Metrics for Several Forecasting Horizons for the Single Exogenous Regressors Approach

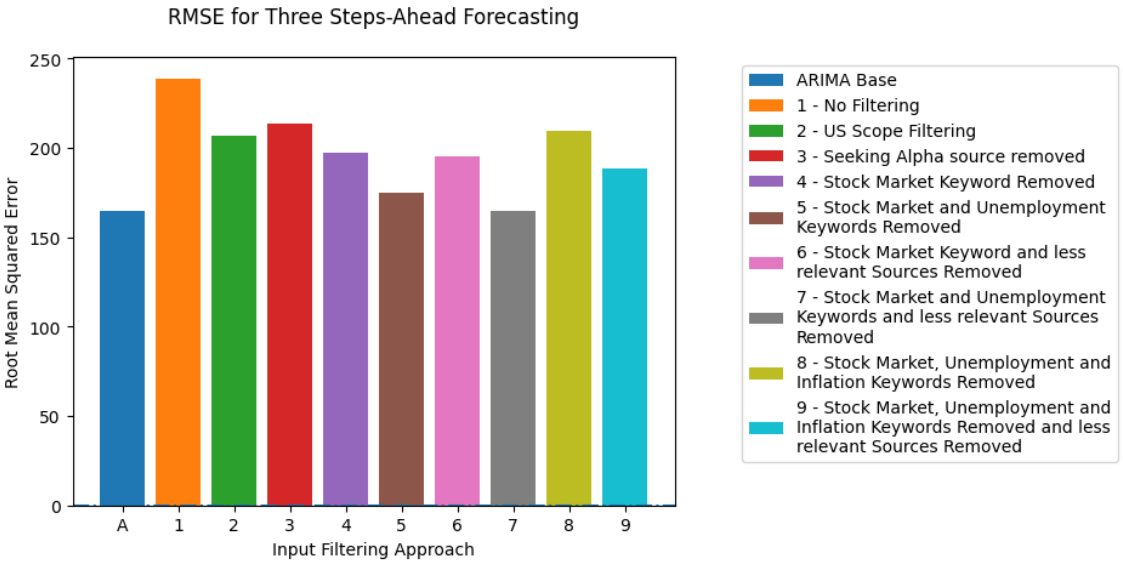


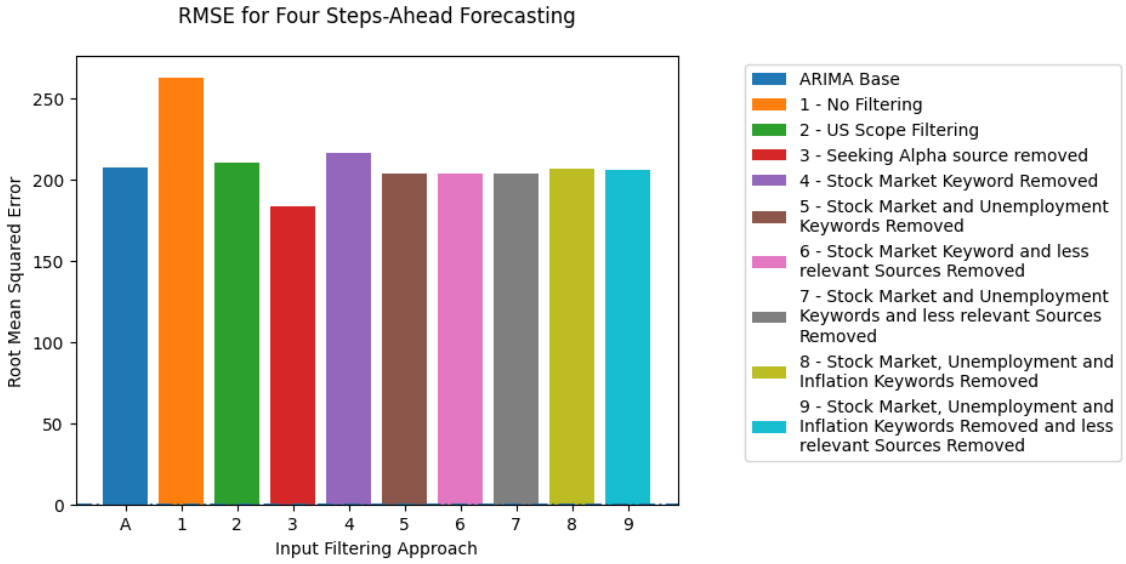


Annex 5

Plot of RMSE Metrics for Several Forecasting Horizons for the Pair of Exogenous Regressors Approach






RMSE for Four Steps-Ahead Forecasting
Root Mean Squared Error
0
50
100
150
200
250
A
1
2
3
4
5
6
7
8
9
Input Filtering Approach
ARIMA Base
1 - No Filtering
2 - US Scope Filtering
3 - Seeking Alpha source removed
4 - Stock Market Keyword Removed
5 - Stock Market and Unemployment Keywords Removed
6 - Stock Market Keyword and less relevant Sources Removed
7 - Stock Market and Unemployment Keywords and less relevant Sources Removed
8 - Stock Market, Unemployment and Inflation Keywords Removed
9 - Stock Market, Unemployment and Inflation Keywords Removed and less relevant Sources Removed

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa