**SENTIMENT ANALYSIS ON STOCK NEWS**

**A PROJECT REPORT**

***Submitted by***

|  |  |  |  |
| --- | --- | --- | --- |
| |  | | --- | | **ASHUTOSH SHUKLA**  **20BCY10087** | | |  | | --- | | **G. SHARAN RAGHAV**  **20BCY10102** | |

*in partial fulfillment for the award of,*

**PROJECT EXHIBITION I**

# BACHELOR OF TECHNOLOGY

|  |  |  |
| --- | --- | --- |
| **CH. SAI MANEESH**  **20BCY10176** |  | **B. SHIVA SAI**  **20BCY10186** |

***in***

## COMPUTER SCIENCE AND ENGINEERING (SPECIALIZATION IN CYBER SECURITY AND

**DIGITAL FORENSICS)**



**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**VIT BHOPAL UNIVERSITY**

# KOTHRIKALAN, SEHORE MADHYA PRADESH - 466114

29TH DECEMBER 2021

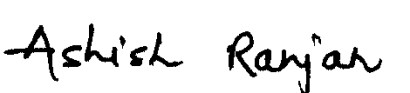
**VIT BHOPAL UNIVERSITY**

**KOTHRIKALAN, SEHORE**

**MADHYA PRADESH – 466114**

# BONAFIDE CERTIFICATE

Certified that this project report titled **“SENTIMENT ANALYSIS IN STOCK NEWS”** is the bonafide work of **“ASHUTOSH SHUKLA (20BCY10087)”, “G. SHARAN RAGHAV (20BCY10102)”, “CH. SAI MANEESH (20BCY10176)”, “B. SHIVA SAI (20BCY10186)”** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.



**PROGRAM CHAIR PROJECT GUIDE**

Dr. R. Rakesh Sir PHD Mr. Ashish Ranjan Sir PHD

ASSISTANT PROFESSOR ASSISTANT PROFESSOR

**(PROGRAM CHAIR OF CSDF)**

(SCHOOL OF COMPUTING SCIENCE AND ENGINEERING)

VIT BHOPAL UNIVERSITY

The Project Exhibition I Examination is held on 29TH DECEMBER 2021.

# ACKNOWLEDGEMENT

First and foremost, I’d want to express my gratitude to my family for their presence and immeasurable blessings during the project’s duration.

I wish to express my heartfelt gratitude to **Dr.R.Rakesh**, Program chair, for much of his valuable support encouragement in carrying out this project.

I would like to thank my internal guide **Mr. Ashish Ranjan**,for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Aeronautical Science, who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

## ASHUTOSH SHUKLA

**G. SHARAN RAGHAV**

**CH. SAI MANEESH**

## B. SHIVA SAI

## LIST OF FIGURES AND GRAPHS

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **TITLE** | **PAGE NO.** |
| **Figure 1.**    **Figure 2.**    **Figure 3.**      **Figure 4.**      **Figure 5.**      **Figure 6.**    **Figure 7.**    **Figure 8.**    **Figure 9.**    **Figure 10.**  **Figure 11.** | System Architecture    Part from the AMD Stock Values DataFrame.    Company specific results of the sentiment analysis using TextBlob. The time period stands between 2021-10-27 and 2021-11-14. (a) Results by Companies and (b) Aggregate Sentiment Result.    TextBlob Analysis results separated by days. The time period stands between 2021-12-16 and 2022-01-01.    Company specific results of the sentiment analysis using NLTK -- VADER Lexicon. The time period stands between 2021-01-16 and 2022-01-01. (a) Results by Companies and (b) Aggregate Sentiment Result.    NLTK -- VADER Lexicon Analysis results separated by days. The time period stands between 2020-10-27 and 2020-11-14.  A recurrent neural network.  Company specific results of the sentiment analysis using RNN. The time period stands between 2020-10-27 and 2020-11-14. (a) Results by Companies and (b) Aggregate Sentiment Result.  RNN Analysis results separated by days. The time period stands between 2020-10-27 and 2020-11-14.  Company specific results of the sentiment analysis using BERT. The time period stands between 2021-01-16 and 2022-01-01. (a) Results by Companies and (b) Aggregate Sentiment Result.  AMD stock value changes.  Sentiment analysis of different models by daily separation. (a) TextBlob.  (b) NLTK -Vader Lexicon and (c) RNN. | 14    13    15        16    16      18    19  20    20    22    23  24 |
| **Figure 12.**  **Figure 13.**  **Figure 14.**  **Figure 15.**  **Figure 16.**  **Figure 17.**  **Figure 18.**  **Figure 19.**  **Figure 20.** | Normalized results of the sentiment and stock values. (a) TextBlob. (b) NLTK -Vader Lexicon and (c) RNN.  Correlation matrix of TextBlob.  Correlation matrix of NLTK -- Vader Lexicon.  Correlation matrix of RNN.  Correlation matrix of RNN with another company (GOOG).  Working layout  Results  Outline of the system key implementations | 25  26  27  28  29  36  37  38 |

# LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| **TABLE NO.** | **TITLE** | **PAGE NO.** |
| **Table 1. Table 2.** | Part from the economic news headlines dataframe Part from the AMD Stock Values DataFrame | 12  13 |

## ABSTRACT

The prediction and speculation about the values of the stock market especially the values of the worldwide companies are a really interesting and attractive topic.

In this article, we cover the topic of the stock value changes and predictions of the stock values using fresh scraped economic news about the companies.

We are focussing on the headlines of economic news. We use numerous different tools to the sentiment analysis of the headlines.

We consider BERT as the baseline and compare the results with three other tools, VADER, TextBlob, and a Recurrent Neural Network, and compare the sentiment results to the stock changes of the same period.

The BERT and RNN were much more accurate, these tools were able to determine the emotional values without neutral sections, in contrast to the other two tools.

Comparing these results with the movement of stock market values in the same time periods, we can establish the moment of the change occurred in the stock values with sentiment analysis of economic news headlines.

Also we discovered a significant difference between the different models in terms of the effect of emotional values on the change in the value of the stock market by the correlation matrices.

## TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **CHAPTER**  **NO.** | **TITLE** | **PAGE NO.** |
|  | List of Abbreviations  List of Figures and Graphs  List of Tables  Abstract | iiiiv v vi |
| 1 | **CHAPTER-1:**  **PROJECT DESCRIPTION AND OUTLINE**  1.1 Introduction  1.2 Motivation for the work  1.3 Problem Statement  1.4 Objective of the work  1.5 Summary | 1  1  2  2  2 |
| 2 | **CHAPTER-2:**  **RELATED WORK INVESTIGATION**  2.1 Introduction  2.2 DataFrame of the headlines of economic news  2.3 Existing Approaches/Methods  2.3.1 Approaches/Methods -1  2.3.2 Approaches/Methods -2  2.3.3 Approaches/Methods -3  2.4 Pros and cons of the stated Approaches/Methods  2.5 Issues/observations from investigation  2.6 Summary | 3  3  4        5  5  6 |

|  |  |  |
| --- | --- | --- |
| 3 | **CHAPTER-3:**  **REQUIREMENT ARTIFACTS**  3.1 Introduction  3.2 Hardware and Software requirements  3.3 Specific Project requirements  3.3.1 Functions requirement  3.3.2 Performance and security requirement | 7  8  9 |
| 4 | **CHAPTER-4:**  **DESIGN METHODOLOGY AND ITS NOVELTY**  4.1 Methodology and goal  4.2 Functional modules design and analysis  4.3 Software Architectural designs  4.4 Subsystem services  4.5 User Interface designs  4.6 Sentiment and stock value analysis  4.7 Summary | 11  13  15  17  19  21  29 |
| 5 | **CHAPTER-5:**  **TECHNICAL IMPLEMENTATION & ANALYSIS**  5.1Outline  5.2 Technical coding and code solutions  5.3 Working Layout of Forms  5.5 Test and validation | 30  31  36  36 |
| 6 | **CHAPTER-6:** |  |

|  |  |  |
| --- | --- | --- |
|  | **PROJECT OUTCOME AND APPLICABILITY**  6.1 key implementations outlines of the System  6.2 Significant project outcomes  6.3 Project applicability on Real-world applications | 38  38  39 |
| 7 | **CHAPTER-7:**  **CONCLUSIONS AND RECOMMENDATION**  7.1Conclusion  7.2 Limitation/Constraints of the System  7.3 Future Enhancements | 40  41  42 |
|  | References | 43 |

**CHAPTER 1**

# PROJECT DESCRIPTION AND OUTLINE

## 1.1 INTRODUCTION

A popular goal is to develop and/or use a model to sentiment prediction by looking for connections between words and marking them with positive or negative sentiments. There are many opportunities these days to perform sentiment analyses, for example external services that are almost completely ready to use it in a given context where it is needed like TextBlob. In addition, there are options that allow us to create our own models, train them based on our own data. Sentiment analysis with BERT is one of the most powerful tool that we can use, but we can also create a Recurrent Neural Network (RNN) as well or use the NLTK tool with VADER Lexicon with SentimentIntensityAnalyzer.

The stock market is one of the most important economic participants. Many people try to interpret and define the different stock market movements in many ways. In this article, we use different tools to the sentiment analysis, especially focussing on the economic news, but in terms of economic news, focussing only on the headlines of economic news. In today’s communications and news consumption, the headlines of various articles play an even more important role than before. Now, we use sentiment analysis on these headlines on a particular company or companies to determine the effects of the headlines to the stock market. The question arises how much effect has the economic headline without the economic news whole context, if it has any measurable effect at all. We have found that it really has. Thus, we define the different impacts and their perceived significance with a very specific and unique new approach. Data is an important pillar of analysis. Primarily the headlines of economic news are needed, what we use for sentiment analysis. Secondary, different stock market data are also needed based on companies. There are many possibilities for data collection and analysis, from ‘traditional’ dictionary-based performed by humans to ‘more serious’ neural networks that determine the polarity of the headlines of each economic news and label with appropriate emotional polarity. In the case of stock market data, numerous tools are available to obtain stock market data which can be even company-specific whichis important to us. In both cases, we work with the most up-to-date data as possible, based on the information provided by the companies. Both, the headlines of the economic news and stock value data are related to the time period which specified by the news. Thus, the results of the given emotional analysis and the range of stock market data will be appropriate. The analysis can be separated to the next sections. Collect headlines of economic news based on companies and collect stock market data according to the timestamps of the given economic news headlines. Then prepare these data and apply different sentiment analysis tools like RNN or NLTK with VADER Lexicon ect. The RNN model was built and taught using the libraries and capabilities provided by Tensorflow. Manage these data and compare the stock market data and emotional data with visualization and explanation. Present how the headlines of economic news can affect different stock market changes and the public.

### 1.2 MOTIVATION FOR THE WORK

Now, more than ever, data mining is finding its way into practical business usage. Data mining and data visualization techniques are creating pathways to solve complex issues and are empowering decisionmakers in determining best business practice and strategy. Business managers and leaders can use data mining and data visualization to generate insights and create value (including financial value) in numerous ways. Two of the major applications involve textual data mining and behavioral analytics. Textual data mining involves a type of analysis in which valuable information is derived from high volumes of text-based data. Whereas, behavioral analytics allows data scientists to derive meaning from customer behavioral data and answer questions like why one product is preferred over a similar product.

The analysis in this study revolves largely around sentiment analysis which is both textual and behavioral. In particular, there was great interest in determining whether any value could be derived through sentiment analysis regarding success within the stock market. Various aspects of human behavioral traits were utilized including categories like confidence level, goals/motivations, and strategy.

### 1.3 PROBLEM STATEMENT

The main problem statement of our project:

* How to utilize social media to assess market sentiment and predict the behavior of stock of certain company, market and stock indexes to identify an opportunity for trading?
* The Stock market forecasters focus on developing a successful approach to predict stock prices. The vital idea to successful stock market prediction is not only achieving best results but also to minimize the inaccurate forecast of stock prices.

### 1.4 OBJECTIVE OF THE WORK

* Understanding author’s opinion from a piece of text is the objective of sentiment analysis.
* The study proposes the framework for sentiment analysis and prediction for the Indian stock market

### 1.5 SUMMARY

Stock sentiment analysis can be used to determine investors’ opinions of a specific stock or asset. Sentiment may at times hint at future price action. hese factors help influence stock sentiment as they impact stock market volatility, trading volume and company earnings…….

**CHAPTER 2**

# RELATED WORK INVESTIGATION

## 2.1 INTRODUCTION

Devlin et al. (2018) introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers, that was designed to pretrain deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers. The new possibilities and results of this model enable even low-resource tasks to benefit from deep unidirectional architectures. This model became one of the most significant tool of the natural language processing.

Wang et al. (2020) introduces a public sentiment analysis during the outbreak which is able to provides insightful information in making appropriate public health responses.They analyze the Sina Weibo popular Chinese social media site posts, where the unsupervised BERT model is adopted to classify sentiment categories (positive, neutral, and negative) and TF-IDF (term frequency-inverse document frequency) model is used to summarize the topics of posts. Analyzing posts with negative sentiment from social media could contribute to understanding the experiences and offers examples for other countries. The analyses provide insights on the evolution of social sentiment over time and the topic themes connected to negative sentiment on the social media sites. BERT classification model and TF-IDF topic extraction model results were delivered with considerable accuracy.

## 2.2 DATAFRAME OF THE HEADLINES OF ECONOMIC NEWS

As mentioned earlier, the main goal in the headlines of economic news is to use the most up-to-date data possible. All data collection and management is automated. There is an option to the user to specify the portal as a source to manage the news. We used data from ‘finviz.com’ for our analyses. Before collecting the data, it is possible to enter the stock exchange names of the companies where we would like to collect the data of recent economic news for analysis. It is possible to specify more than one company by listing as parameter. The function takes care of managing the appropriate timestamps (news publication time) and separating the news based on the companies and create a backup into a file as csv. This freshly compiled data is used by the application for further analysis (as part of sentiment analysis, comparisons, and other possibilities.) It is important to mention that news timestamps play a role in compiling additional stock market data so the analyses take place in the same time period. Thus, these economic news headlines define the interval for subsequent stock market data collection separated for companies.

## 2.3 EXISTING APPROACHES/METHODS

The big data is a very popular and powerful tool nowadays. Lee (2020) explores the initial impact of COVID-

19 sentiment on US stock market using big data notedly Daily News Sentiment Index (DNSI) and Google Trends data on coronavirus-related searches. The goal is to investigate a correlation between COVID-19 sentiment and 11 selected sector indices of the Unites States (US) stock market in a declared time period. Any positive or negative sentiment of public related to stock market crisis can have a ripple effect on decision making by investors in stock markets. The results reveal the distinct effects of the COVID-19 sentiment across in various industries and separate them to different correlation groups.

### 2.3.1 APPROACHES/METHODS -1

Sentiment Classification Approaches Machine learning Bayesian Networks Naive Bayes Classification Maximum Entropy Neural Networks Support Vector Machine Features/ Tecniques Term presence and frequency, Part of speech information, Negations, Opinion words and phrases

### 2.3.2 APPROACHES/METHODS -2

Sentiment Classification Approaches Lexicon based Dictionary based approach Novel Machine Learning Approach Corpus based approach Ensemble Approaches Features/ Tecniques Manual construction, Corpusbased, Dictionary based

### 2.3.3 APPROACHES/METHODS -3

Sentiment Classification Approaches Hybrid Machine learning Lexicon based Features/ Tecniques Sentiment lexicon constructed using public resources for initial sentiment detection, Sentiment words as features in machine learning method

## 2.4 PROS AND CONS OF THE STATED APPROACHES/METHODS

### 2.3.1 APPROACHES/METHODS -1

ADVANTAGES : the ability to adapt and create trained models for specific purposes and contexts

LIMITATIONS : the low applicability to new data because it is necessary the availability of labeled data that could be costly or even prohibitive

### 2.3.2 APPROACHES/METHODS -2

ADVANTAGES : wider term coverage

LIMITATIONS : finite number of words in the lexicons and the assignation of a fixed sentiment orientation and score to words

### 2.3.3 APPROACHES/METHODS -3

ADVANTAGES : lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level and the lesser sensitivity to changes in topic domain

LIMITATIONS : noisy reviews

## 2.5 ISSUES/OBSERVATIONS FROM INVESTIGATION

We have observed that ,

* The opinions of others often play a role in an individual’s decision-making process and this was especially so prior to the advent of the World Wide Web. Recommendations from friends, colleagues and co-workers played an integral role in everyday decision-making. However, more people are using the Internet to make their views accessible to strangers. People who use a social network or social media sites like Facebook or Twitter communicate their opinions on specific topics, such as news, movies, events, or a particular product
* From the last twenty years, the application of Internet based technologies had brought a significant impact on the Indian stock market. Use of the Internet has eliminated the barriers of brokers and geographical location because now investors can buy and sell their shares by accessing the stock market status from anywhere at any time. Before investing money, it is very important for investors to predict the stock market.

## 2.6 SUMMARY

* Market sentiment refers to the overall consensus about a stock or the stock market as a whole.
* Market sentiment is bullish when prices are rising.
* Market sentiment is bearish when prices are falling.
* Technical indicators can help investors measure market sentiment.

**CHAPTER 3**

# REQUIREMENT ARTIFACTS

## 3.1 INTRODUCTION

Tool making is complex, time taking as well as a work of perseverance. Hence requires lot of things which makes the steps clear and concised.

We have used these artifacts in in the making of our tool.

Our method detects such artifacts you could opt to build your own sentiment analysis tools using opensource libraries, such as TensorFlow, PyTorch, NLTK, or Scikit-learn, but they take longer to set up and it’s more expensive to build your own. First, you’ll need to invest in a data science team to develop the necessary infrastructure, then you’ll need to spend months training and fine-tuning your models.

On the other hand, you could opt for Software as a Service (SaaS) tools for text analysis :

No setup needed: SaaS tools are cloud-based solutions that are ready to use instantly.

Easy to integrate: Most SaaS tools integrate with everyday tools, such as Google Sheets, Zapier, and Zendesk. If you’re comfortable with a few lines of code, then you can also make use of text analysis APIs in all major programming languages.

Pre-trained models to get started right away: Most online sentiment analysis tools offer pre-trained models that you can try out on your own data.

## 3.2 HARDWARE AND SOFTWARE REQUIREMENTS

**HARDWARE SPECIFICATIONS :**

Processor : Intel i5 or more

Motherboard : Intel Chipset Motherboard

Ram : 8GB or more

Cache : 512 KB

Hard disk : 16 GB hard disk recommended

Disk Drive : 1.44MB Floppy Disk Drive

Monitor : 1024 x 720 Display

Speed : 2.7GHZ and more

**SOFTWARE REQUIREMENTS :**

Operating System : Windows 7/8/8.1/10

Python

Pytorch

Django

VS CODE

Google Colab

Google cloud services

R Studio

## 3.3 SPECIFIC PROJECT REQUIREMENTS

### 3.3.1 FUNCTIONAL REQUIREMENT

1. Choose file :- User is enabled to choose a file for prediction sentiment analysis on stock news(<100mb)
2. Sequence Length :- User is given access to set the sequence length according to his requirements (<100frames)
3. Upload :- After choosing file and setting sequence length , user must click on upload to load the data for pre-processing
4. User can see the predication of stock news

#### 3.3.2 PERFORMANCE AND SECURITY REQUIREMENT

Security Requirements :- Only recommended software should be used where they provide healthy security towards the project without any corrupted files.

Performance :- For good and better performance , try to run it without any background apps running over there . It will run over project with good and better space.

## CHAPTER 4

**DESIGN METHODOLOGY AND ITS**

**NOVELTY**

### 4.1 METHODOLOGY AND GOAL

#### 4.1.1 Options to build DataFrame of the news headlines and stock values

There are several ways to approach data structure building. Primarily we consider theheadlines of economic news. Of course, there is the possibility to compile a collectionof data by human effort according to specific conditions, such as gathering economicnews titles filtered by a given company name from the collection built from a starttime (which is the oldest possible economic news titles) until to reach a certain limit.There is the possibility of approaching the analysis using data from a previous archive collectionof data, but the main goal is to use the most up-to-date data as possible. There isalso the possibility of using human effort in the case of data collection from the stockmarket values of companies, but today many economic portals and other libraries and frameworksare available to fully automate the process. In this case, automation plays a moreimportant role than in the previous economic news title data collecting. The error factorcan be significantly reduced when compiling companies’ economic data. In addition, thesource and the values of the stock data are easier to manage this way than in the economicnews title data collecting.

Listing 1. Part of the Economic news headlines dataframe builder

The code snippet shown by Listing 1 implements a part of data collection for economic news headlines. Where the ‘weblite url’ is the portal from where we process the news, and the ‘company tickets’ are the company names on the stock market in a list from which we would like to compile data. The data processing shown by the code snippet use the ‘BeautifulSoup’, ‘urlopen’ and ‘Request’ tools to perform scraping. For other source pages we have to make changes in this processing stage to scrape data from this specified page (Figure 1)

#### 4.1.2 DataFrame of the company specific stock values

‘Yahoo fin’ tool was used to collecting stock values for companies. This data is separatedby companies into the intervals of previously compiled economic news headlines. Basedon this, it will be possible to analyze and compare economic news headlines and stockmarket data for a given period. This data collection and management also provides theopportunity to perform individual and aggregate analyses (Figure 2).

##### 4.1.3 Sentiment analysis with different tools

As mentioned earlier, there are many possibilities for sentiment analysis from humanlabelleddata to various deep learning methods. In the present case, we compare the possibilitiesoffered by TextBlob, NLTK -- VADER Lexicon, RNN and BERT. The main goal is toanalyze the headlines of economic news about different companies and determine theirsentiment values to be positive or negative possibly neutral. A key factor is to minimizeneutral values. It should be noted that we do not have as much influence over externalthirdparty devices as we do about our own models, such as RNN.In the case of sentiment analysis, the headline of the economic news from eachcompany is labelled to what sentiment value it carries, and the polarity value is also indicated.With the help of these data, we can make a number of further analyses and



*Table 1. Part from the economic news headlines dataframe*

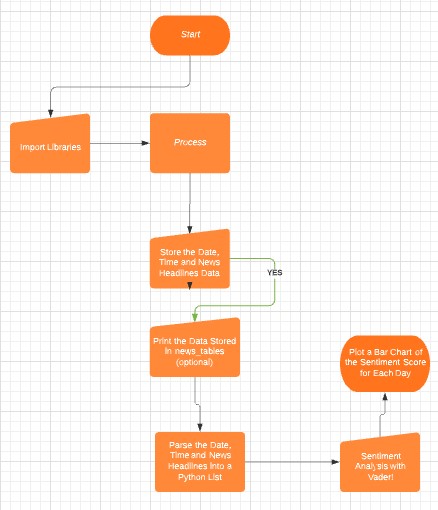


*Table 2. Part from the AMD Stock Values DataFrame*

comparisons. The main direction is to compare the specific companies with their stock market values in the period of time which determined by the economic news. Thus assessing and presenting the emotional impact of economic news headlines on stock market changes and see how powerful the headlines can be alone without full content. In addition, our goal is to determine the strength and accuracy of different sentiment analysis tools by the given context. The BERT tool is used as a kind of comparative tool to see how close the results of the other tools are to the results of BERT. More detailed analysis of stock market values and sentiment values (polarity and sentiment label) is done using the results of TextBlob, NLTK -- VADER Lexicon and RNN.

### 4.2 FUNCTIONAL MODULES DESIGN AND ANALYSIS

#### 4.2.1 System Architecture

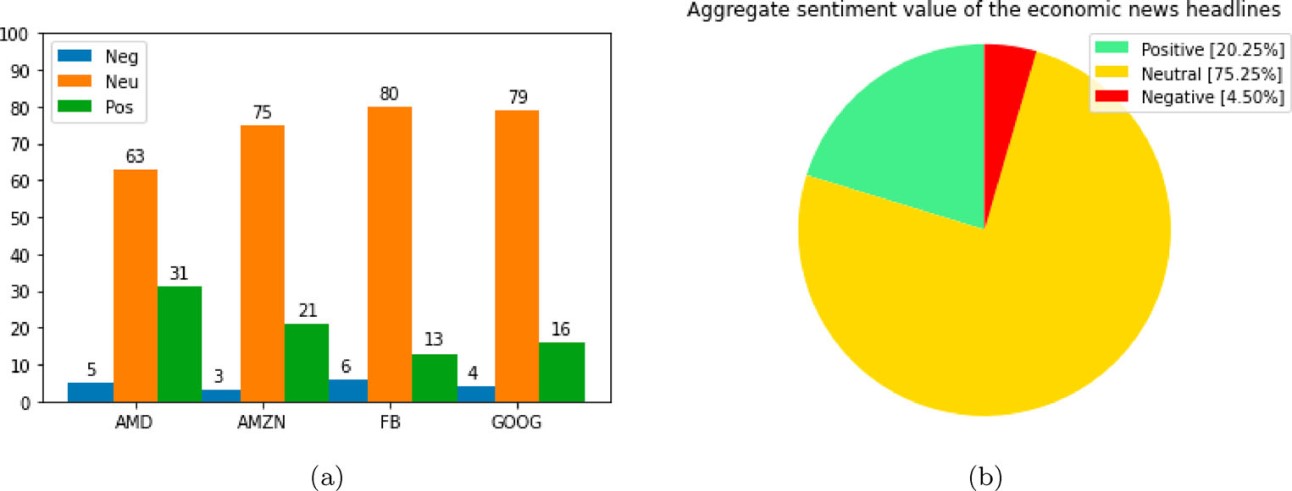


*Figure 1. System Architecture*

**4.2.2 Data-set Results**

##### 4.2.2.1 TextBlob

TextBlob is a powerful NLP library for Python, which is built upon NLTK and provides an easy to use interface to the NLTK library. This tool can be used to perform a variety of NLP tasks ranging from parts-of-speech tagging to sentiment analysis, and language translation to text classification, but we focus on the sentiment analysis. If we do a sentiment analysis, we actually determine a polarity value of the sentences, where this value can be between −1 and 1. Then we label the data with the right sentiment value (positive, negative or neutral). For other tools, the polarity value may move on a different scale, so the labelling needs to adjust for these differences for further analysis.The Figure 3 shows that sentiment values separated by companies. No other value can approach the neutral section, it can be concluded that the analysis of the given economic news headlines and its outcome is very uncertain. In the case of AMD, it can be noted that in Figure 3(a), in addition to the 63 news headlines rated as neutral, 31 are positive and 5 are negative. In the case of FB -- Facebook, in addition to the 80 news headlines rated as neutral, there are 13 positive and 6 negative values as well. In the case of the total result, 75.25 percent in Figure 3(b) is neutral besides to this 20.25 percent is positive and only 4.50 percent is negative. The goal is to minimize neutral values by using a more accurate analysis to reduce the inaccuracy increased by its neutral values in stock market comparisons. The following figure (Figure 4) shows the results divided into days in the interval. The results are aggregated and this gives us a normalized value of how positive or negative the overall day was for the company. Due to the significant neutral value of more than 75 percent, the days are visibly shifting in a positive direction, which can greatly distort real results. Where a company does not have a coloured column for a given day, there was no economic news headline about those company. The following figure is formed on the interval, where above zero means the positive section and below means the negative section.



*Figure 2. Company specific results of the sentiment analysis using TextBlob. The time period stands between 2021-10-27 and 202111-14. (a) Results by Companies and (b) Aggregate Sentiment Result.*

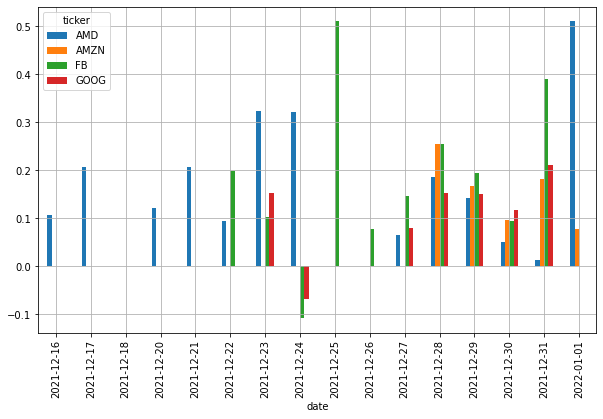
It should be noted that there were a large amount of news during the period, about AMD will launch new CPUs and GPUs in October, which was also significantly positive. This may be explained by recent period in the end of October CPU and GPU events and this is the effect of these events.

### 4.3 SOFTWARE ARCHITECTURAL DESIGNS

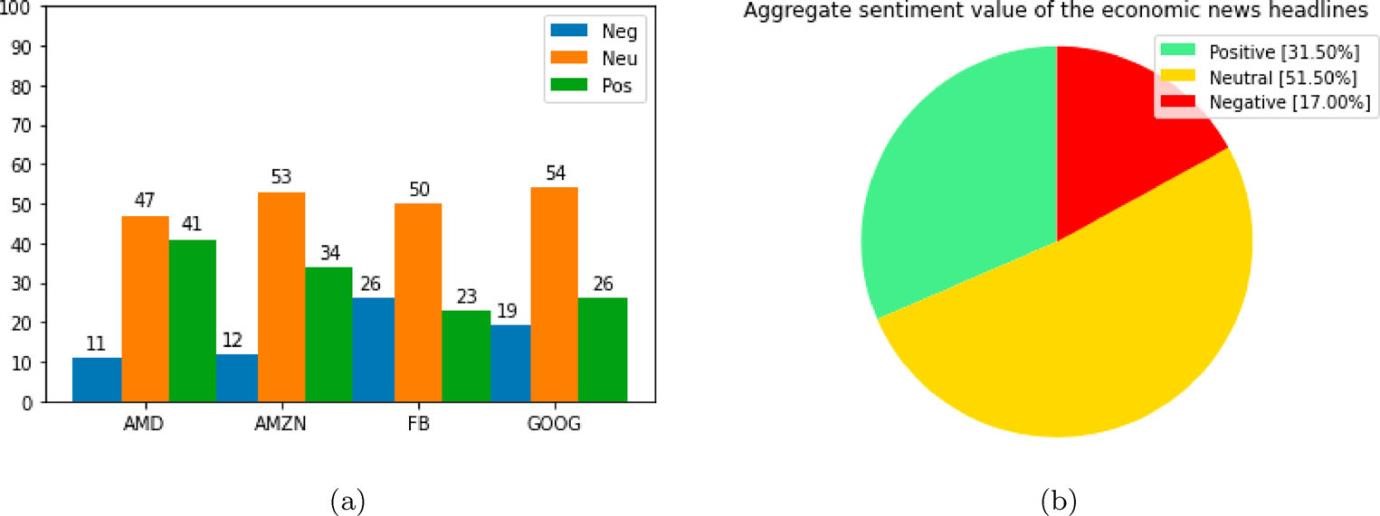
#### 4.3.1 NLTK -- VADER lexicon

NLTK stands for Natural Language Toolkit. This toolkit is one of the most powerful NLP libraries which contains packages to make machines understand human language and reply to it with an appropriate response. Again, we focus on sentiment analysis with the SentimentIntensityAnalyzer. The polarity value of the sentences scales between -1 and 1 just like in the TextBlob. The data labelling process (positive, negative or neutral) is similar to the previous tool. We use VADER Lexicon in this section. VADER (Valence Aware

Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis



*Figure 3.. TextBlob Analysis results separated by days. The time period stands between 2021-12-16 and 2022-01-01.*



*Figure* *4. Company specific results of the sentiment analysis using NLTK -- VADER Lexicon. The time period stands between 2021-01-16 and 2022-01-01. (a) Results by Companies and (b) Aggregate Sentiment Result.*

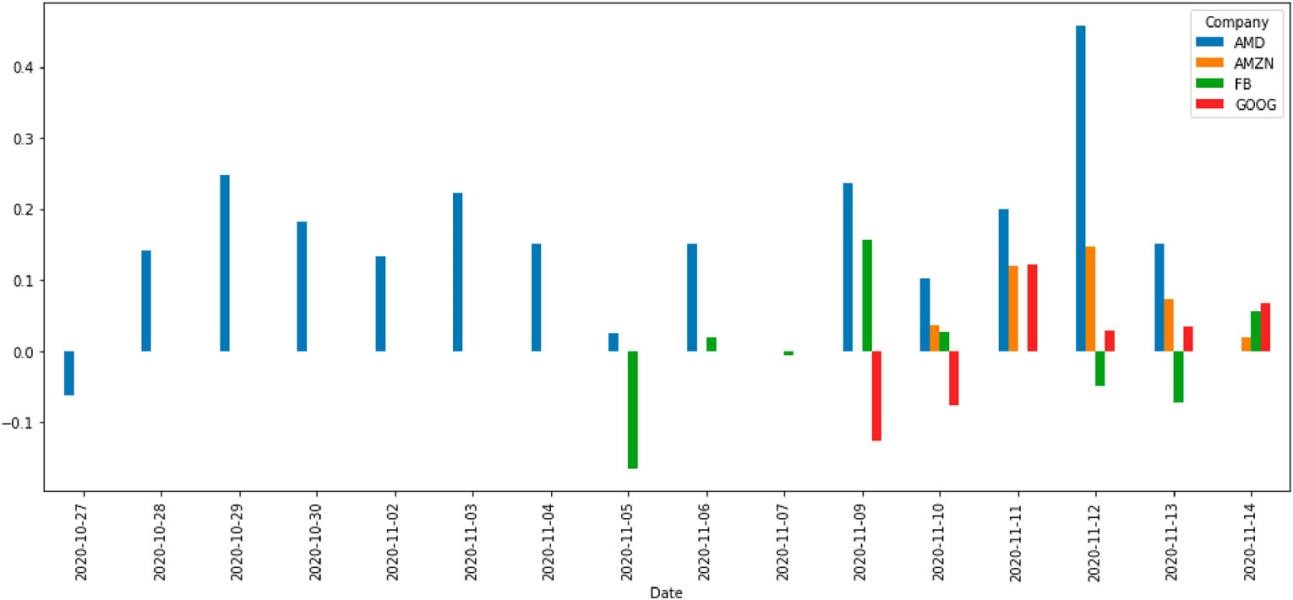
tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.

As shown in Figure 5 below, the neutral value (a) dominates in all cases among the sentiment results separated by companies. In the figure next to it (b), the aggregate sentiment result shows the economic news headlines significant neutral values. This level of neutral values has impact on comparisons and analyses to the subsequent stock market changes. Compared to the results of TextBlob, the neutral values have been significantly reduced and we expect that has significant effect in further analysis to obtain more accurate and realistic results with fewer neutral values. In Figure 5(b), 51.50 percent of the total result is neutral in addition to 31.50 percent positive and 17 percent negative. Of the positive or negative categories, the positive strongly dominates, but this huge neutral value still makes the result little bit uncertain. The following figure (Figure 6) shows the results divided into days in the interval. The results are aggregated and this gives us a normalized value of how positive or negative the overall day was for the company. One day in total cannot be neutral because of the other news headlines have to move it in some direction and the neutral values according to the polarity also try to move the result in some direction too. Thus, the following figure is formed on the interval, where above zero means the positive section and below means the negative section.

### 4.4 SUBSYSTEM SERVICES

#### 4.4.1 Recurrent neural network (RNN)

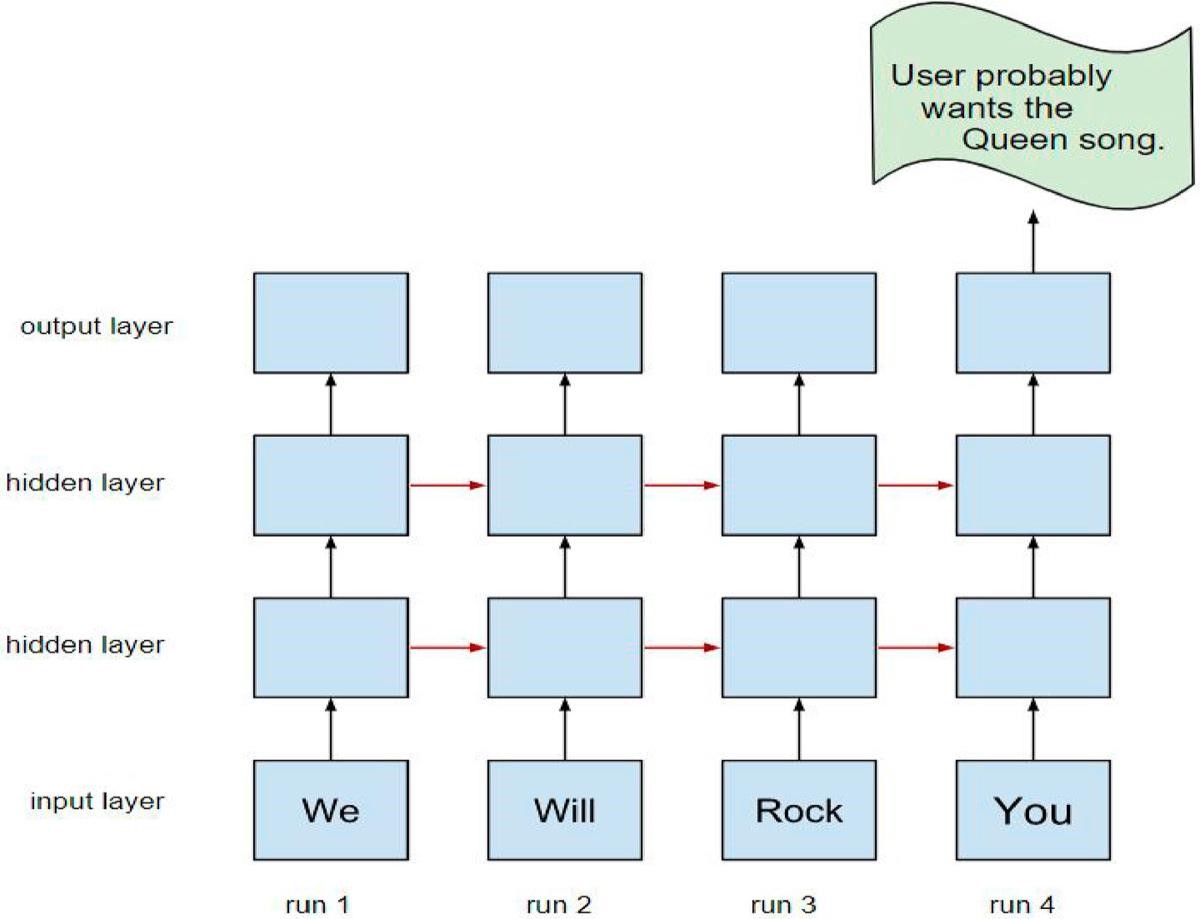
When we talk about traditional neural networks, all the outputs and inputs are independent of each other. But in the case of recurrent neural networks, the output from the previous steps is fed into the input of the current state. All in all the Recurrent Neural Network is a neural network that is intentionally run multiple times, where parts of each run feed into the next run. Specifically, hidden layers from the previous run provide part of the input to the same hidden layer in the next run. Recurrent neural networks are particularly useful for evaluating sequences, so that the hidden layers can learn from previous runs of the neural network on earlier parts of the sequence.



*Figure 5. NLTK -- VADER Lexicon Analysis results separated by days. The time period stands between 2020-10-27 and 2020-11-14.*

For example, the following figure of Google shows a recurrent neural network that runs four times (Figure 7). Notice that the values learned in the hidden layers from the first run become part of the input to the same hidden layers in the second run. Similarly, the values learned in the hidden layer on the second run become part of the input to the same hidden layer in the third run. In this way, the recurrent neural network gradually trains and predicts the meaning of the entire sequence rather than just the meaning of individual words. An advantages of the RNN model: RNN can process inputs of any length. An RNN model is modelled to remember each information throughout the time which is very helpful in any time series predictor. Even if the input size is larger, the model size does not increase. But there some disadvantages: Due to its recurrent nature, the computation is slow. Training of RNN models can be difficult. We have to mention that the polarity value of the sentences scales between 0 and 1 here. In contrast to the models what mentioned earlier. The Figure 8 shows the results of RNN separated by companies and the aggregating result as before.

A significant difference from the previous results of TextBlob and NLTK -- Vader Lexicon is that the neutral section was completely eliminated, all news headlines were categorized as either positive or negative. This is a significant difference from previous models, although there was a kind of downward trend in the models. The neutral category of the TextBlob was huge, it was significantly reduced by the NLTK – Vader Lexicon, and then the RNN model was managed to avoid a neutral category. Figure 8(a) shows how the positive and negative news headlines are distributed among the companies. In the case of AMD, it can be noted that the result is quite ‘balanced’ with 51 positive and 48 negative values. In part (b) of the figure, the total result is 58.50 percent positive and 41.50 percent negative and the neutral value is 0 percent which is now the key. In Figure 9, the positive negative day categorization is totally different than the previous ones, because in the case of the RNN model, the polarity values scale between 0 and 1. Therefore, here is a ‘traditional’ bar chart showing the aggregation of polarity values for each day.



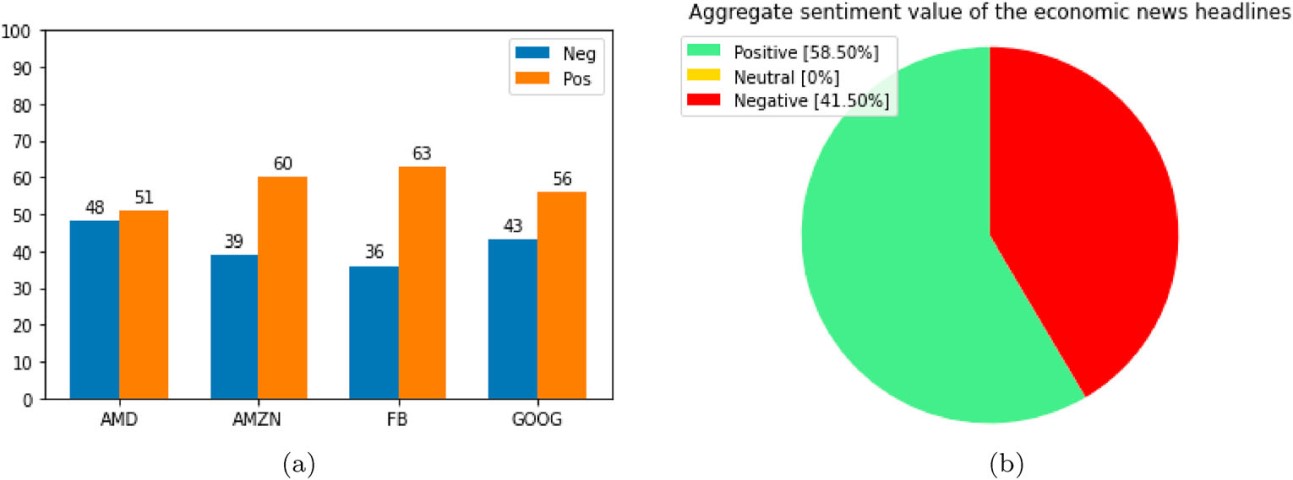
*Figure 6. A recurrent neural network.*

Note : The RNN model was trained based on an IMDB review dataset2 (In the test and train dataset sections we used shuffle method as well. Then we use the fresh scraped dataset as test dataset with this trained model.)

### 4.5 USER INTERFACE DESIGNS

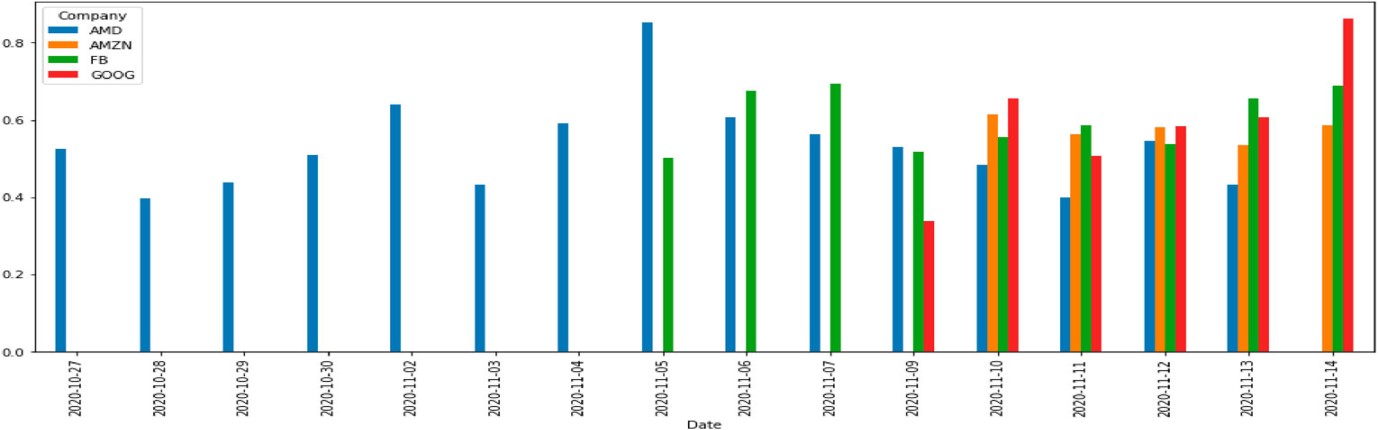
#### 4.5.1 Bidirectional encoder representations from transformers (BERT)

Unlike the traditional NLP models that follow a unidirectional approach, that is, reading the text either from left to right or right to left, BERT reads the entire sequence of



*Figure 7. Company specific results of the sentiment analysis using RNN. The time period stands between 2020-10-27 and 2020-11-14.*

*(a) Results by Companies and (b) Aggregate Sentiment Result.*

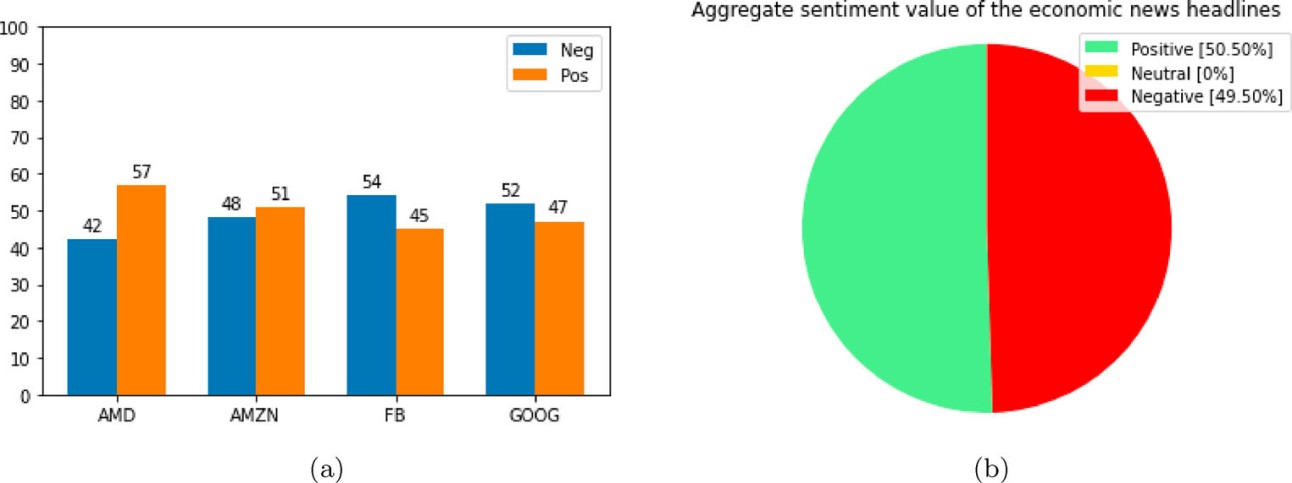


*Figure 8. RNN Analysis results separated by days. The time period stands between 2020-10-27 and 2020-11-14.*

words at once. BERT makes use of a Transformer which is essentially a mechanism to build relationships between the words in the dataset. In its simplest form, a BERT consists of two processing models -- an encoder and a decoder. The encoder reads the input text and the decoder produces the predictions. But, because the main goal of BERT is to create pretrained model, the encoder takes priority over decoder. BERT is a remarkable breakthrough in the field of NLP. As mentioned earlier, BERT is used as a kind of comparative result. Figure 10 shows the results obtained by BERT. Of course, without a neutral category, it managed to categorize each economic news headline and labelled it as a positive or negative value. In part (a) of the figure, it can be mentioned that the result of our previous RNN model is quite encouraging, as there is no neutral category either and the values of certain companies are quite close to the result of BERT. Part (b) of the figure shows the overall result where 50.50 percent is positive and 49.50 percent is negative compared to the result of the RNN model where 58.50 is positive and 41.50 is negative, neutral is 0 percent in both cases. We expected that the model we trained and taught would give more accurate and more reliable results than other tools on the same data set. More specifically, the result from the RNN model determines emotional values and labels with a more accurate and smaller error rate than NLTK with VADER Lexicon or TextBlob. This expectation was also confirmed by the results. It should be emphasized that the result of NLTK was much more encouraging than initially expected and in later analyses, despite the existing neutral values, it gave a much ‘finer’ result than TextBlob where we get a ‘raw’ result due to the significant neutral value. For the RNN model, no headline is placed in the neutral category. Regarding the results of BERT and the results of the other tools, we expect more accurate results from the RNN and NLTK tools when analyzing with stock market values.

### 4.6 SENTIMENT AND STOCK VALUE ANALYSIS

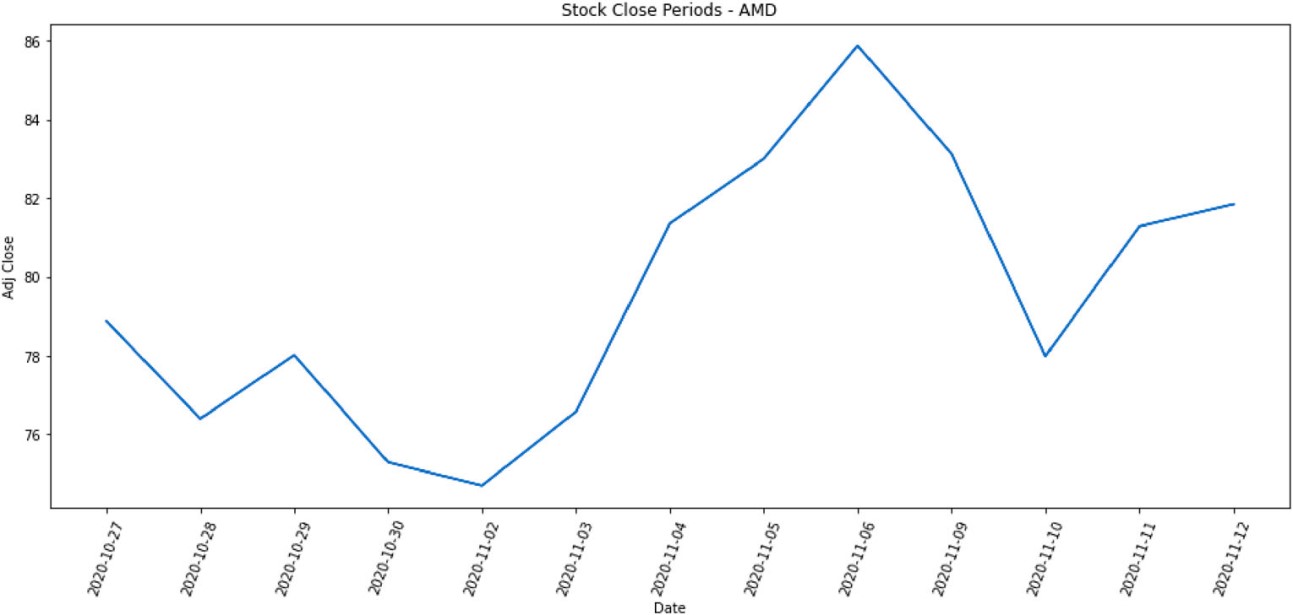
Following the sentiment analyses at a given interval, we can start the comparison with stock market changes at the same interval. During the sentiment analysis, the ‘realistic’



*Figure 9. Company specific results of the sentiment analysis using BERT. The time period stands between 2021-01-16 and 2022-01-01. (a) Results by Companies and (b) Aggregate Sentiment Result.*

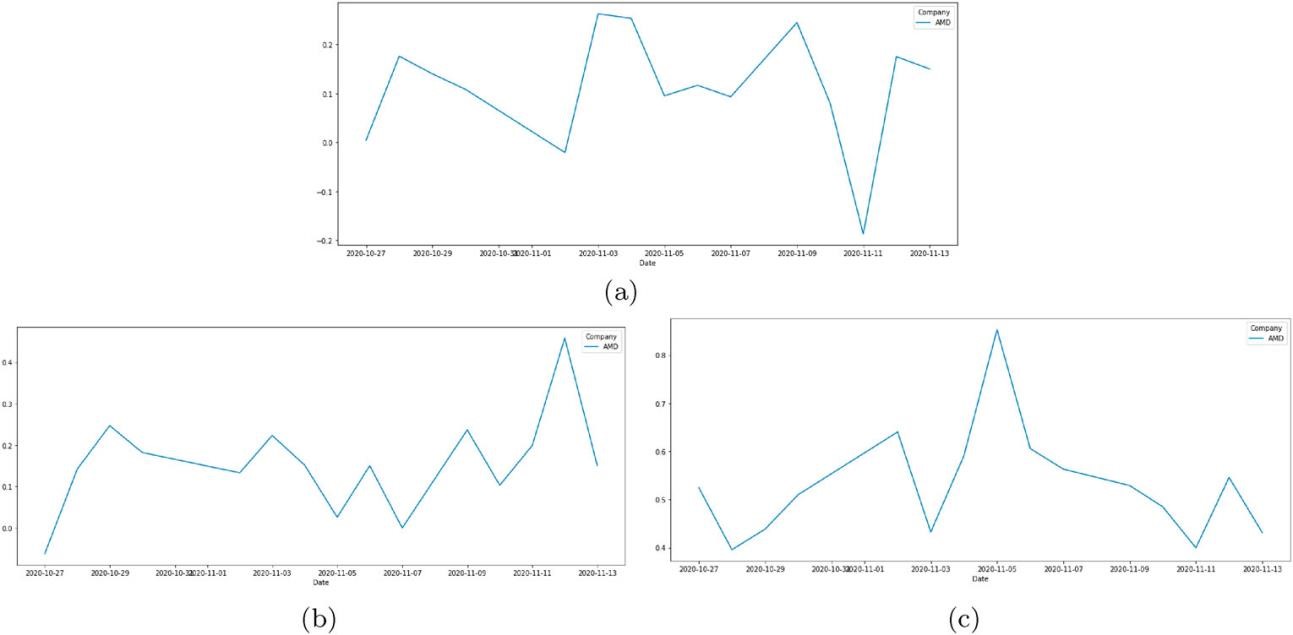
word was mentioned, which refers to the smaller neutral category, less neutral value in the solution of the analysis. It refers to a better and ‘stronger’ analytical model that was able to give positive and negative tags to news headlines which were categorized as neutral by the previous analytical model. Thus we reduced the potential of error and possible skew results. We can say that economic news do have an impact on stock market shifts, there are times when certain news items have effect to the later movements, and there are times when the news describe a particular shift, which enhances change too. Our main study in the present case focuses on the headlines of economic news about the various companies which was given as parameters previously, without their full article context. The headline itself, which aims to draw people’s attention and generate clicks on full content, is worded in this ‘sometimes sharp, eye-catching’ way. How much impact do these economic news headlines have on stock market changes, if it has any effect. In our results we found that it really has.

Figure 11 shows the AMD stock market changes during the given study period, where the date and the daily closing (adjusted closing price) value are displayed. The following (Figure 12) shows the results of different sentiment models (TextBlob, NLTK -- Vader Lexicon and RNN) for the given period, broken down by day. Here, the results are the same of the previous sentiment analyses, but now they are displayed on a different diagram for the purpose of being comparable with the stock market data. Significant differences can be observed in the results of the different models especially in some parts of the result. One of the most striking may be the negative news stream around 2020-11-11. In all three cases, a negative trend can be detected, but the differences in the extent are significant. These results, in comparison with stock market changes, help us to see a kind of effect on whether stock market movements are reflected in the diagram of sentiment results. The amount of neutral values plays a significant role in the accuracy of the models. It was mentioned earlier that when calculating daily results (this day is positive or negative all in all), the polarity values of the neutral values also count, so that these values also play a role in the positive or negative shift of a day as they belong to that day, but this values distort the result. In contrast, in a model where there is no neutral value, much higher accuracy can be expected.



*Figure 10. AMD stock value changes.*

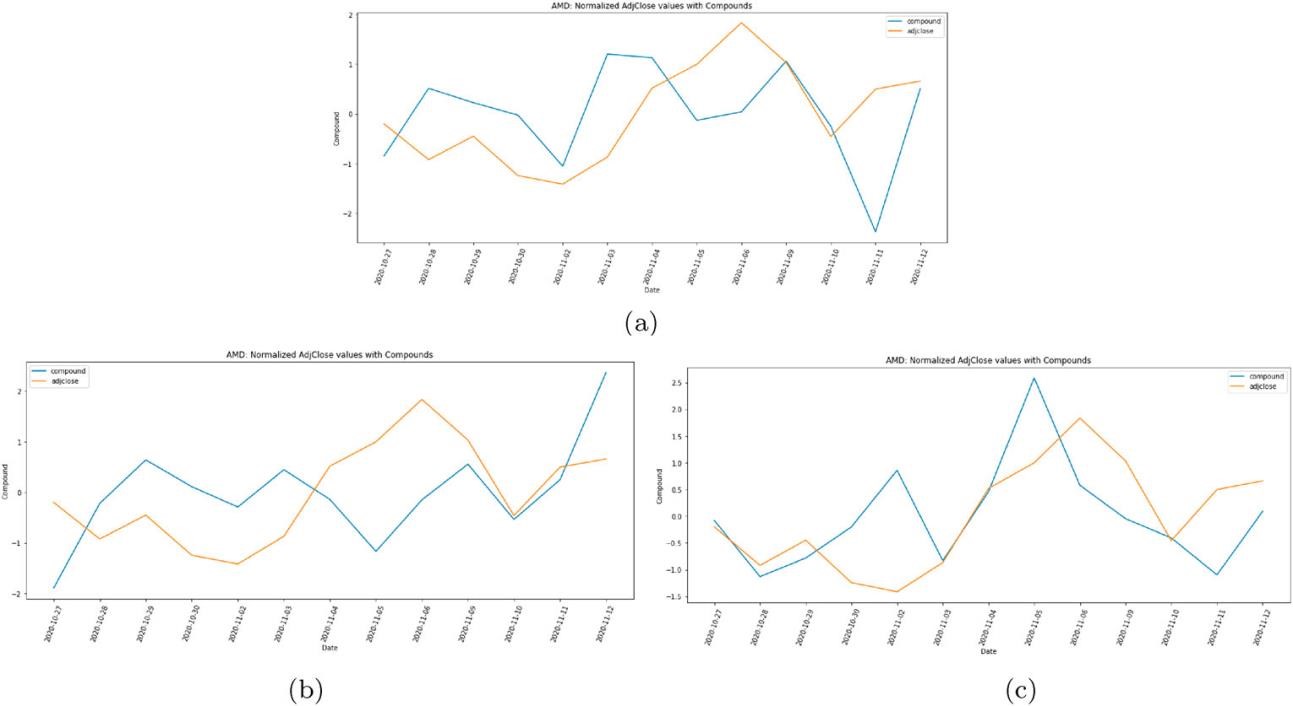
Figure 13 shows the normalized results, where the different models show how the stock market value changed in the period and how the daily results obtained by the economic news headlines of the given period relate to stock market movements. The graph still process data from AMD. The first figure (a) shows the summary of the results obtained by the TextBlob and the stock market result. In the case of TextBlob, the ratio of neutral values was 75.25 percent, which is also reflected in the large ‘vibration’ of emotional results. On the normalization graph in this case with the economic news headlines and stock values we can read as fundamental changes, with the trend of decreasing or increasing. In the period between 2020-11-04 and 2020-11-06, a strong decrease in emotional values can be observed, in addition to with a smaller ‘break’ or a correction in the stock market values as well. In the phases of 2020-11-09 and 2020-11-10, a significant



*Figure 11. Sentiment analysis of different models by daily separation. (a) TextBlob. (b) NLTK -Vader Lexicon and (c) RNN.*

break point can be observed in both stock market developments and emotional results. Overall, we can see the impact and the major growth declines can be traced from the chart, but its detail is questionable. In the case of figure (b) we can see the results of the NLTK -- Vader lexicon normalization. The ratio of neutral values in this case was reduced to 51.50 percent. It can be said that the result is surprising at first. It is clear that a more detailed ‘co-movement’ of stock and emotional values is shown in the figure. Changes between 2020-11-09 and 2020-11-11 will be tracked ‘fully in sync’. Regarding the results of RNN in figure (c), where the ratio of neutral values was 0 percent, significant differences can be observed compared to the previous ones. Here, it may appear primarily that the two results do not follow each other in ‘synchrony’ and in some cases there is a significant difference between emotional and stock market results. It can be assumed that in this case, the effect of emotional values on the results of the current days may not be as great and perhaps a kind of ‘periodic prediction’ can be observed. The significant positive result between 2020-11-04 and 202011-06 is one of the most striking results. Until the subsequent correlation matrix results, all that can be stated is that there is a significant decrease in the influence of emotional values in the given stock market period. A kind of emotional decrease or increase and a following stock shift can be observed, but in fact the influence has decreased significantly, which can be explained by neutral values and ‘realism,’ when examining the influence of news headlines we cannot expect as much impact as full economic articles and analyses. In all three cases, these effects are also analyzed by correlation matrices.

In the case of Figure 14, we can see that Compound (sentiment results) has a huge impact on both the opening, closing, lowest and highest values of the stock market, which is a very distorted result. It’s almost unthinkable to have such a big impact. As mentioned earlier, the significant neutral value can be traced back to this situation as well.



*Figure 12. Normalized results of the sentiment and stock values. (a) TextBlob. (b) NLTK -Vader Lexicon and (c) RNN.*

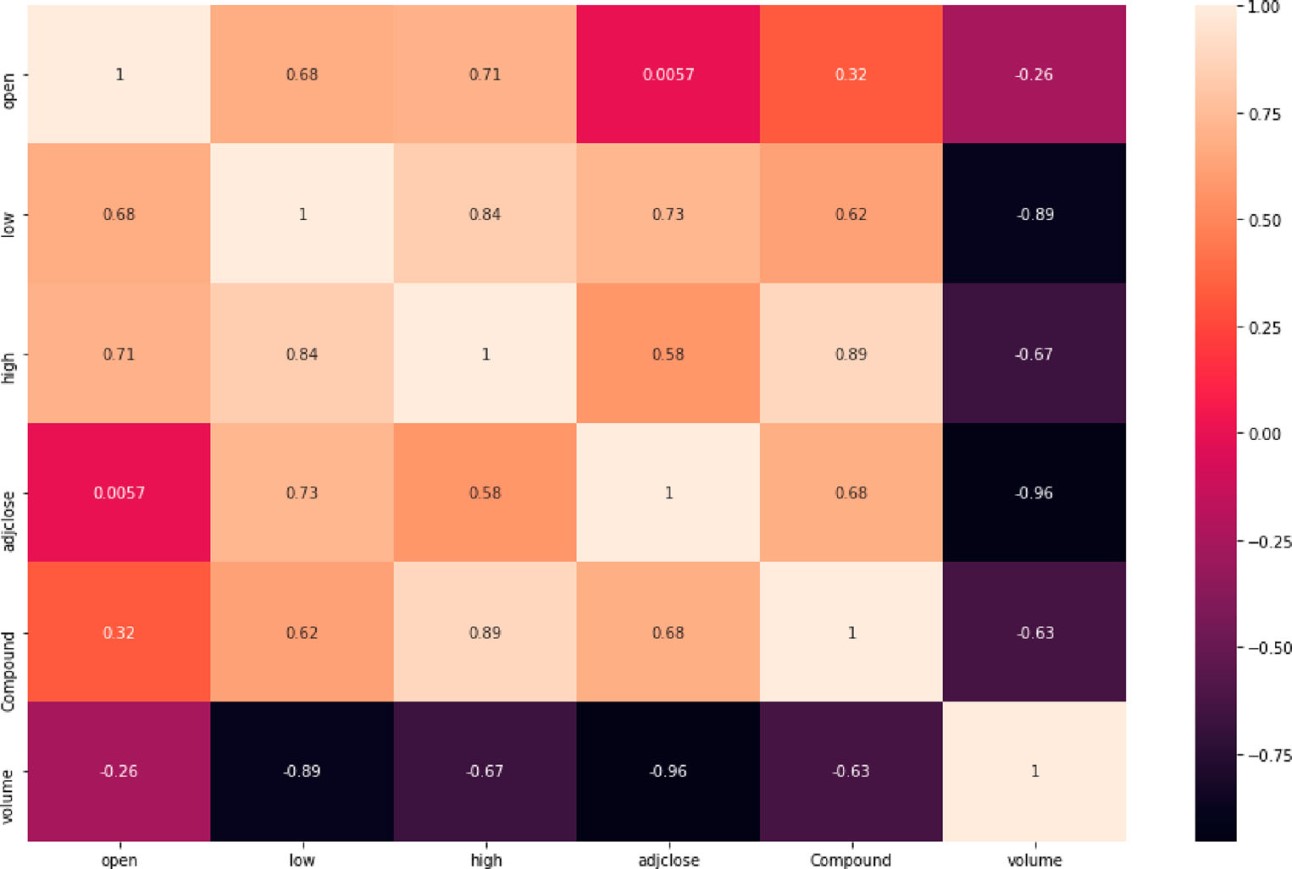
The following Figure 15 shows the result of the NLTK -- Vader Lexicon correlation matrix, where there are decreases in Compound values in almost all values compared to the results of the previous (TextBlob) correlation matrix. In addition to a kind of ‘synchronized result’ seen on previous diagrams, a significant effect was ‘expected’ in the matrix as well, but perhaps these results may also seem excessive as a result obtained, considering that we examine economic news headlines on a companyspecific basis.

In the Figure 16, the correlation matrix of RNN is completely different and surprising in this case as well. The value of the Compound has decreased significantly compared to its previous models to the opening, closing, lowest and highest values, and unlike before, its effect on another value has increased drastically. The value of the volume is the amount of an asset or security that changes hands over some period of time, often over the course of a day.

The correlation matrix of RNN and the values of Compound provide a kind of explanation for the diagrams seen earlier. The previous models had a significant effect on the opening, closing, lowest and highest values, in contrast, the RNN shows a completely different result. Overall, we can say that the headlines themselves have a significant effect on the change in stock market values, in addition to highlighting the volume value, which alone received a significant value in the RNN model, unexpectedly high. It should be noted that the data from the study period may also play a role in this. But the result is thought-provoking. The result is not unique. We obtained a similar result for the measurements between 2021-01-16 and 2022-01-01 for another company, which was the Google (GOOG). As we can see in the Figure 17.



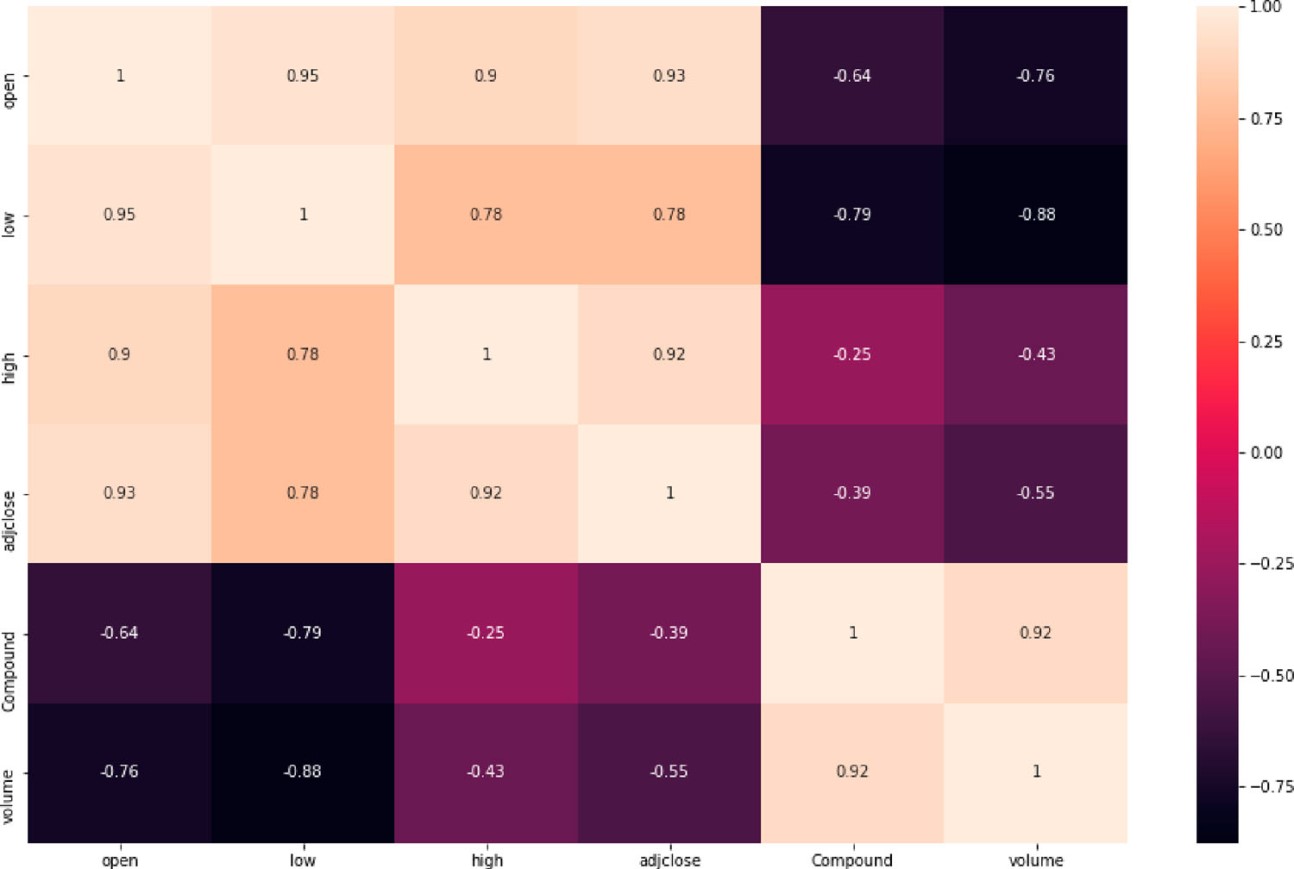
*Figure 13. Correlation matrix of TextBlob.*



*Figure 14. Correlation matrix of NLTK -- Vader Lexicon.*



*Figure 15. Correlation matrix of RNN.*



*Figure 16. Correlation matrix of RNN with another company (GOOG).*

### 4.7 SUMMARY

* The Sentiment analysis recognizes the polarity of opinion and emotion attributes regarding the Stock News.
* The Subjectivity strongly depends on its sentences or messages.
* As news articles capture sentiment about the current market, we automate this sentiment detection and based on the words in the news articles, we can get an overall news polarity.
* If the news is positive, then we can state that this news impact is good in the market, so more chances of stock price go high. And if the news is negative, then it may impact the stock price to go down in trend.

**CHAPTER 5**

# TECHNICAL IMPLEMENTATION ANALYSIS

## 5.1 OUTLINE

1. Import Libraries

First, we import the libraries that we need to store the data. ‘BeautifulSoup’ is needed to parse data from

FinViz while ‘requests’ is needed to get data. ‘Pandas’ is used to store the data in DataFrames while ‘Matplotlib’ is used to plot the sentiment on a chart. Finally, the ‘nltk.sentiment.vader’ library is used to perform sentiment analysis on the news headlines!

1. Store the Date, Time and News Headlines Data

Let’s take a closer look at the news headlines for Amazon (AMZN) and its corresponding html code below.

You can also visit the FinViz page and view the html code in your browser.

1. Print the Data Stored in news\_tables (optional)

To get a sense of what is stored in the news\_tables dictionary for ‘AMZN’. Feel free to run the code below, which iterates through each <tr></tr> tags (for the first 4 rows) to obtain the headlines between the <a></a> tags and the date and time between the <td></td> tags before printing them out. This step is optional and is for your own learning.

1. Parse the Date, Time and News Headlines into a Python List

The following code is similar to the one above, but this time it parses the date, time and headlines into a Python list called parsed\_news instead of printing it out. The if, else loop is necessary because if you look at the news headlines above, only the first news of each day has the ‘date’ label, the rest of the news only has the ‘time’ label so we have to account for this.

1. Sentiment Analysis with Vader!

It is now time to perform sentiment analysis with nltk.sentiment.vader, finally! We store the ticker, date, time, headlines in a Pandas DataFrame, perform sentiment analysis on the headlines before adding an additional column in the DataFrame to store the sentiment scores for each headline.

1. Plot a Bar Chart of the Sentiment Score for Each Day

The following code takes the average of the sentiment scores for all news headlines collected during each date and plots it on a bar chart. You can average the scores for each week too, to obtain the overall sentiment for a week.

## 5.2 TECHNICAL CODING AND CODE SOLUTIONS

# Import libraries from urllib.request import urlopen, Request from bs4 import BeautifulSoup import os import pandas as pd import matplotlib.pyplot as plt

%matplotlib inline

# NLTK VADER for sentiment analysis from nltk.sentiment.vader import SentimentIntensityAnalyzer finwiz\_url = 'https://finviz.com/quote.ashx?t='

news\_tables = {} tickers = ['AMZN', 'TSLA', 'GOOG'] for ticker in tickers:

url = finwiz\_url + ticker

req = Request(url=url,headers={'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64; rv:20.0) Gecko/20100101 Firefox/20.0'}) response = urlopen(req)

# Read the contents of the file into 'html' html = BeautifulSoup(response)

# Find 'news-table' in the Soup and load it into 'news\_table' news\_table = html.find(id='news-table') # Add the table to our dictionary news\_tables[ticker] = news\_table

# Read one single day of headlines for 'AMZN' amzn = news\_tables['AMZN']

# Get all the table rows tagged in HTML with <tr> into 'amzn\_tr' amzn\_tr = amzn.findAll('tr') for i, table\_row in enumerate(amzn\_tr):

# Read the text of the element 'a' into 'link\_text' a\_text = table\_row.a.text

# Read the text of the element 'td' into 'data\_text' td\_text = table\_row.td.text

# Print the contents of 'link\_text' and 'data\_text'

print(a\_text) print(td\_text)

# Exit after printing 4 rows of data if i == 3: break

parsed\_news = [] # Iterate through the news for file\_name, news\_table in news\_tables.items(): # Iterate through all tr tags in 'news\_table' for x in news\_table.findAll('tr'):

# read the text from each tr tag into text

# get text from a only text = x.a.get\_text()

# splite text in the td tag into a list date\_scrape = x.td.text.split()

# if the length of 'date\_scrape' is 1, load 'time' as the only element if len(date\_scrape) == 1: time = date\_scrape[0]

# else load 'date' as the 1st element and 'time' as the second

else:

date = date\_scrape[0] time = date\_scrape[1]

# Extract the ticker from the file name, get the string up to the 1st '\_' ticker = file\_name.split('\_')

# Append ticker, date, time and headline as a list to the 'parsed\_news' list parsed\_news.append([ticker, date, time, text])

parsed\_news

# Instantiate the sentiment intensity analyzer vader = SentimentIntensityAnalyzer()

# Set column names

columns = ['ticker', 'date', 'time', 'headline']

# Convert the parsed\_news list into a DataFrame called 'parsed\_and\_scored\_news' parsed\_and\_scored\_news = pd.DataFrame(parsed\_news, columns=columns) # Iterate through the headlines and get the polarity scores using vader scores = parsed\_and\_scored\_news['headline'].apply(vader.polarity\_scores).tolist()

# Convert the 'scores' list of dicts into a DataFrame scores\_df = pd.DataFrame(scores)

# Join the DataFrames of the news and the list of dicts parsed\_and\_scored\_news = parsed\_and\_scored\_news.join(scores\_df, rsuffix='\_right')

# Convert the date column from string to datetime

parsed\_and\_scored\_news['date'] = pd.to\_datetime(parsed\_and\_scored\_news.date).dt.date parsed\_and\_scored\_news.head()

plt.rcParams['figure.figsize'] = [10, 6]

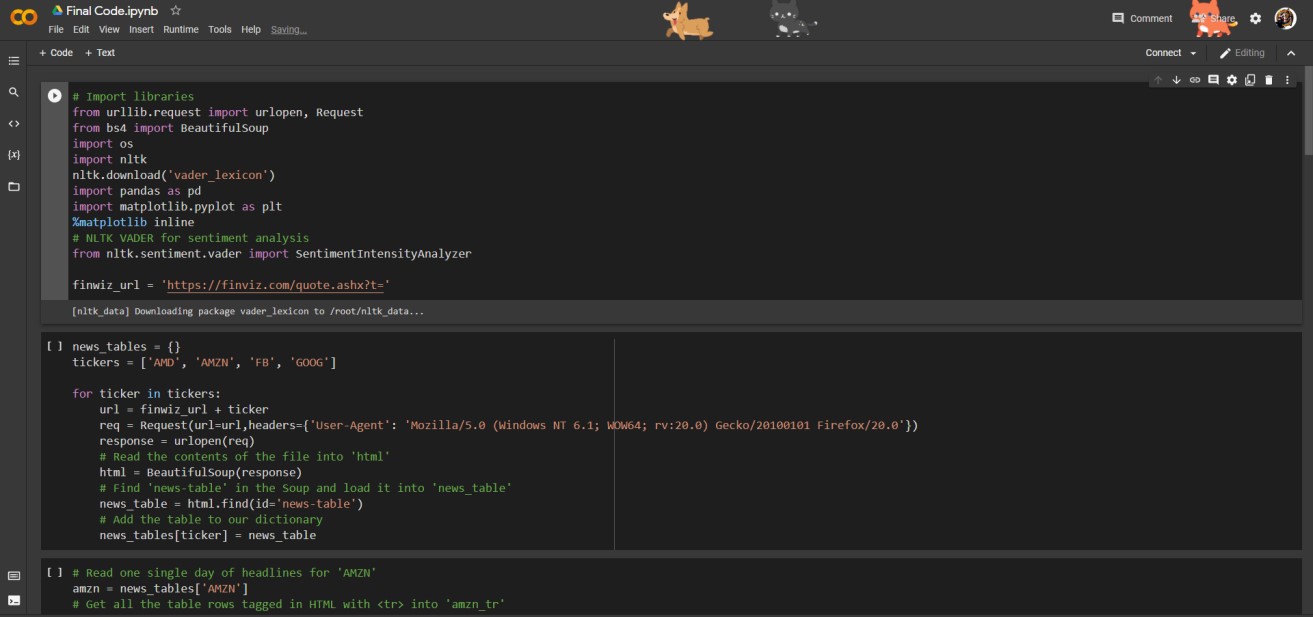
# Group by date and ticker columns from scored\_news and calculate the mean mean\_scores = parsed\_and\_scored\_news.groupby(['ticker','date']).mean()

# Unstack the column ticker mean\_scores = mean\_scores.unstack()

# Get the cross-section of compound in the 'columns' axis mean\_scores = mean\_scores.xs('compound', axis="columns").transpose()

# Plot a bar chart with pandas mean\_scores.plot(kind = 'bar') plt.grid()

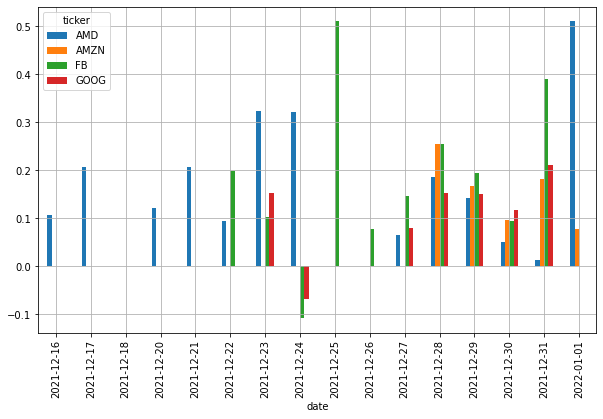
### 5.3 WORKING LAYOUT OF FORMS



*Figure 17. Working layout*

### 5.4 TEST AND VALIDATION

After testing we will get result



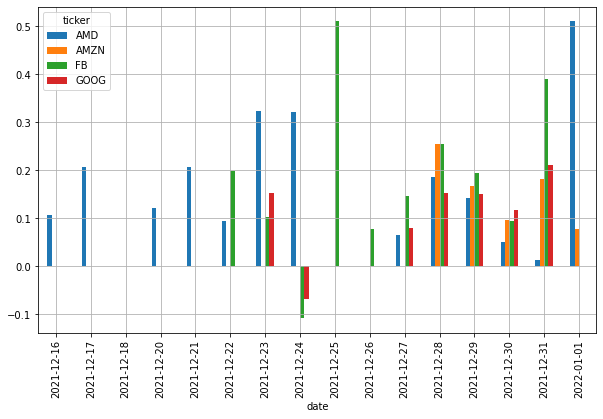
*Figure 18. Results*

## CHAPTER 6

**PROJECT OUTCOME AND APPLICABILITY**

### 6.1 Key implementations outline of the System

* The Final will come out to be bar chart
* The Final output of the model is going to be by showing the graph of 4 companies from 16-12-2021 to 01-01-2022



*Figure 19. Outline of the system key implementations*

### 6.2 Significant project outcomes

* Prediction and quantification of future volatility and returns play an important role in financial modelling,
* both in portfolio optimization and risk management. Natural language processing today allows to process news and social media comments to detect signals of investors' confidence.
* We have explored the relationship between sentiment extracted from financial news.
* We investigated the strength of the correlation between sentiment measures on a given day and market volatility and returns observed the next day. The findings suggest that there is evidence of correlation between sentiment and stock market movements: the sentiment captured from news headlines could be used as a signal to predict market returns; the same does not apply for volatility.

### 6.3 Project applicability on Real-world applications

 From the last twenty years, the application of Internet based technologies had brought a significant impact on the Indian stock market. Use of the Internet has eliminated the barriers of brokers and geographical location because now investors can buy and sell their shares by accessing the stock market status from anywhere at any time.

 Before investing money, it is very important for investors to predict the stock market. In today's digital world Internet based technologies such as Cloud Computing, Big Data analytics, and Sentiment analysis have changed the way we do business. Sentiment analysis or opinion mining makes use of text mining, natural language processing (NLP), in order to identify and extract the subjective content by analyzing user's opinion, evaluation, sentiments, attitudes and emotions. In this research work importance of sentiment analysis for stock market indicators such as Sensex and Nifty has been done to predict the price of stock.

 The sentiment analysis of stock news prediction gives a accuracy of 92% to the orginal stock price

**CHAPTER 7**

# CONCLUSIONS AND RECOMMENDATIONS

## 7.1 CONCLUSION

* In this work, we used different sentiment analysis tools to emotionally analyze and classify different economic news headlines and examine their impact on different stock market value changes even without their full context. Emotions were classified into the usual positive negative and neutral categories. Neutral categories appeared for TextBlob and NLTK-VADER Lexicon tools, but not for Recurrent Neural Network (RNN). The various sentiment analyses results were compared with the result of BERT as a benchmark.
* As we expected, the results of the RNN model what we developed and taught outperformed the other sentiment analysis tools and gave a result quite close to BERT, emphasizing that there was no neutral emotional value in this case either. In the analysis of emotional results and stock market changes, we compared the daily results of emotional values and the results of stock market values for the given period.
* We obtained appropriate diagrams for the reading of the emotional results and the stock market movements and corrections, but we could detect differences according to the ratio and effect of the neutral values of the different models. In the field of further analysis, we detected significant differences in the correlation matrices. In the case of TextBlob, the Compound (emotional results) had a significant effect on the opening, closing, highest and lowest values of the stock exchange, the NLTK -- Vader Lexicon gave similar results, but reducing the results of the previous model significantly. The RNN model brought a completely different value.
* The emotional values and stock market change diagram also showed a kind of smaller effect, which was also confirmed by the correlation matrix, and also had a significant effect on the Volume value compared to the other models. Overall, economic news headlines have an impact on stock market values even without their textual context, and significant differences can be observed between different sentiment analytical tools. But the stock market impact also depends on how the data in the current study period was affected.

## 7.2 LIMITATION/CONSTRAINTS OF THE SYSTEM

* It is generally said that whatever is already in the news is already factored into the market. So if company ABC has decided to buy an oil block which will make them millions, the market has already filtered the results. This provides a very small leeway for investors to capitalize on that information. This is where Efficient Market Hypothesis comes into the picture. EMH states that market prices immediately reflect all available information. Three popular versions of EMH exist namely weak, semi-strong and strong. The weak form of the hypothesis is rather pessimistic. It states that no profit can be made by reviewing at publicly available information thus rendering sentiment analysis moot. The semi-strong hypothesis argues that profit can be made by analyzing data that is not publicly available thus not reflected in the market. Thus EMH presents a big problem and working around it to optimize returns is a major challenge. Another possible limitation to sentiment analysis is failure of a message to actually be factored into the market. The possibility of a message being positive or negative not affecting the market exists and sentiment analysis currently cannot mitigate this problem.
* More experiments are needed with large scale datasets to prove the efficiency of the proposed model.
* Data limitations refer to the inadequacy of the data to be analyzed. Hundreds of thousands of users express opinions about companies, policies and announcements. However not all messages are relevant to the happenings in the market. Maintaining relevancy of the data to the current market scenario is a difficult task. Noise is another negative when it comes to analyzing messages online. The signal to noise ratio is very low and current sentiment analyzers are not prepared for it. Also current sentiment analysis tools base their findings on specific sources such as news, tweets, and blogs etc. However only one platform as a data source provides incomplete information about the market. Thus a collaboration of sources must be utilized for efficient and unbiased market analysis. The format of the data is biggest of the problems. News agencies usually employ info-graphics for their news, organizations circulate newsletters and investors use online platforms. Each user uses a variety of formats ranging from articles, and tweets to videos and info-graphics. The failure to incorporate the different formats might deviate the result and does not do justice to the analysis.

### 7.3 Future Enhancements

* Future work could include further expansion of the analyses, possible additions of a new features.
* In addition, the inclusion of other tools to compare stock market predictions with different sentiment analysis tools.
* That can be built into an easy-to-use format by developing a platform incorporating various future changes of tensorflow into the current model.
* The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens.

**Notes :-**

1. https://developers.google.com/machinelearning/glossary/#recurrent\_neural\_network
2. https://www.tensorflow.org/datasets/catalog/imdb\_reviews

## REFERENCES

1. Anurag Nagar, Michael Hahsler, Using Text and Data Mining Techniques to extract Stock Market

Sentiment from Live News Streams, IPCSIT vol. XX (2012) IACSIT Press, Singapore

1. W.B. Yu, B.R. Lea, and B. Guruswamy, A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting, International Journal of Electronic Business Management. 2011,5(3): 211-224
2. J. Bean, R by example: Mining Twitter for consumer attitudes towards airlines, In Boston Predictive Analytics Meetup Presentation, 2011
3. Yauheniya Shynkevich, T.M. McGinnity, Sonya Coleman, Ammar Belatreche, Predicting Stock Price

Movements Based on Different Categories of News Articles, 2015 IEEE Symposium Series on

Computational Intelligence

1. P. Hofmarcher, S. Theussl, and K. Hornik, Do Media Sentiments Reflect Economic Indices? Chinese Business Review. 2011, 10(7): 487-492
2. R. Goonatilake and S. Herath, The volatility of the stock market and news, International Research Journal of Finance and Economics, 2007, 11: 53-65.
3. Spandan Ghose Chowdhury, Soham Routh , Satyajit Chakrabarti, News Analytics and Sentiment Analysis to Predict Stock Price Trends, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3595-3604
4. Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, Sentiment Analysis of Financial News Articles
5. Győző Gidófalvi, Using News Articles to Predict Stock Price Movements, University of California,

San Diego La Jolla, CA 92037, 2001

1. L. Breiman, Random forests. Machine Learning, 45(1):5-32, 2001
2. Data Mining Lab 7: Introduction to Support Vector Machines (SVMS)
3. Joachims T., Text Categorization with Support Vector Machines: Learning with Many Relevant

Features, European Conference on Machine Learning (ECML), Application of Machine Learning and

Data mining in Finance, Chemnitz, Germany, 1998)