

Anticipating The Stock Market Trend Using Natural Language Processing Techniques



By

Sana Sajid

Fall-2020-MS-CS 329039 SEECS

Supervisor

Dr. Seemab Latif

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of
Science in Computer Science (MS CS)

In

School of Electrical Engineering & Computer Science (SEECS),

National University of Sciences and Technology (NUST),

Islamabad, Pakistan.

(June 2024)

Dedication

This thesis is dedicated to my parents, Mr. Muhammad Sajid and Mrs. Shazia Sajid, as well as my spouse, Mr. Zaid Tariq. I trust that this accomplishment will fulfil the aspiration you held for me. Thank you for your continuous encouragement and support.

Approval

It is certified that the contents and form of the thesis entitled "Anticipating the Stock Market Trend using Natural Language Processing Techniques" submitted by Sana Sajid have been found satisfactory for the requirement of the degree

Advisor : Dr. Seemab Latif

Signature:  _____

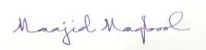
Date: 01-Jul-2024

Committee Member 1:Dr. Rabia Irfan

Signature:  _____

03-Jul-2024

Committee Member 2:Mr. Maajid Maqbool

Signature:  _____

Date: 11-Jul-2024

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Anticipating the Stock Market Trend using Natural Language Processing Techniques" written by Sana Sajid, (Registration No 00000329039), of SEecs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

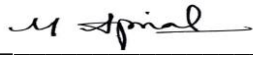
Signature: _____  _____

Name of Advisor: Dr. Seemab Latif

Date: 01-Jul-2024

HoD/Associate Dean: _____  _____

Date: 01-Jul-2024

Signature (Dean/Principal): _____  _____

Date: 01-Jul-2024

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at Department of Computing at School of Electrical Engineering Computer Science (SEECS) or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at School of Electrical Engineering & Computer Science (SEECS) or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Sana Sajid**

Signature: *Sana Sajid*

National University of Sciences & Technology

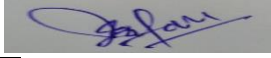
MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: (Student Name & Reg. #) Sana Sajid [00000329039]

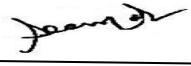
Titled: Anticipating the Stock Market Trend using Natural Language Processing Techniques

be accepted in partial fulfillment of the requirements for the award of Master of Science (Computer Science) degree.

Examination Committee Members

1. Name: Rabia Irfan Signature: 
15-Aug-2024 9:56 PM

2. Name: Maajid Maqbool Signature: 
15-Aug-2024 9:56 PM

Supervisor's name: Seemab Latif Signature: 
15-Aug-2024 9:59 PM



Muhammad Imran Malik
HoD / Associate Dean

19-August-2024

Date

COUNTERSIGNED

20-August-2024

Date



Muhammad Ajmal Khan
Principal

Acknowledgments

Commencing by expressing gratitude to the divine being, Almighty ALLAH (S.W.T), for bestowing upon me the courage, strength, and determination to successfully conclude my research tenure. I wish to convey my heartfelt appreciation to Dr. Seemab Latif for her unwavering assistance and mentorship during my research. I would also like to express my gratitude to her for consistently motivating me to venture outside my comfort zone and strive for improvement. The meetings and conversations were crucial in stimulating my ability to think creatively and critically from various angles, enabling me to develop a comprehensive and unbiased critique. In addition to expressing my gratitude to my adviser, I would like to extend my thanks to the other members of my thesis committee, Sir Maajid Maqbool, and Dr. Rabia Irfan, for their support and valuable input. The collaboration with the team at CPIs was an exceptional experience. In conclusion, I extend my sincere gratitude to the Higher Education Commission's National Research Program for Universities (NRPU) for their support and funding, which has been instrumental in advancing our research efforts under project number 20-15756 with the title "Generating Plausible Counterfactual Explanations using Transformers in Natural Language-based Financial Forecasting (NLFF)".

Sana Sajid

Abstract

Anticipating stock market fluctuations is a crucial routine that investors must engage in when participating in the stock trading market, making it an intriguing research area. The stock market is subjected to the impact of multiple factors, such as news events, economic data, and investor sentiments. Nevertheless, the intricate relationship between news and stock prices contains hidden trends contributing towards the trading recommendations. This study aims to anticipate stock market trends using Natural Language Processing (NLP) techniques, with a particular focus on the Pakistan Stock Market. By creating sequential snapshots of news along with financial data, and employing sentiment analysis to capture market sentiment, this research goes beyond traditional methodologies. Additionally, this research examines the correlation between stock market patterns and news events by employing features that are specifically tailored to analyse the distinct market dynamics of the Pakistan Stock Exchange. The results of our study reveal a prominent ratio of 1:2 between negative and positive news events, underscoring the substantial influence of negative events on market volatility. Consequently, this work represents progress in the direction of automating data-driven and well-informed trading recommendations.

Contents

1	Introduction and Motivation	1
1.1	Overview	1
1.2	Fintech and Stock Market.....	3
1.2.1	What is Fintech.....	3
1.2.2	Importance of Stock Market	4
1.2.3	Landscape of Pakistan’s Stock Market	5
1.2.4	Unique Dyanmics of PSX.....	7
1.2.5	Financial market prediction	7
1.3	Problem Statement	8
1.4	Relevance to National Needs	8
1.5	Research Objectives	10
1.6	Research Questions	10
1.7	Major Contributions	11
1.8	Upcoming Chapters	12
2	Literature Review	13
2.1	Financial News Sentiment Analysis	13
2.2	Financial News Named Entity Recognition.....	16
2.3	Financial News Event Extraction.....	18
2.4	Literature Gap	20

3	Methodology	21
3.1	Module1- Event-Driven Stock Movement Analysis	21
3.1.1	Standardized Dataset Creation	21
3.1.2	Financial Data Extraction and Preparation.....	26
3.1.3	Sector-Based Tagging of Extracted News.....	27
3.1.4	News Filtration	29
3.1.5	Comprehending Market Sentiment	29
3.1.6	News Keyword Extraction	31
3.1.7	News Named Entity Recognition.....	33
3.1.8	Embedding Generation for News Articles	34
3.1.9	Event-Stock Extraction and Mapping	35
3.2	Module2- Correlating Event Impact with Stock Trend Movement.....	39
3.2.1	News-Feature Formulation and Tagging Module.....	39
3.2.2	Feature Validation, Event Extraction and Visualization Module.....	40
4	Results and Discussion	42
4.1	Module-1	42
4.2	Module-2.....	50
5	Conclusion	55
6	Recommendations	56

List of Tables

3.1	Category Standardization of News Dataset	25
3.2	News Dataset.....	25
3.3	Active Stocks of Chosen Sectors	26
3.4	Financial Data.....	27
3.5	Sector's Keywords.....	28
3.6	Assigning Sector based Sub-Cat to the News Articles	28
3.7	News Tagged as Others	29
3.8	Configuration Parameters	31
3.9	News Articles with Keywords.....	32
3.10	News Articles with NER Information.....	34
3.11	Configuration Parameters	35
3.12	Processed_Financial_News_Dataset_Columns.....	36
4.1	Specifics of the experimental environment.....	43
4.2	News and Events Statistics	45
4.3	News Articles by Standard Categories	50

List of Figures

1.1	Share of the region’s population and GDP Growth	9
3.1	Methodology flowchart from data acquisition to visual analysis.....	22
3.2	Methodology flowchart from data acquisition to visual analysis.....	39
4.1	Distribution of News Articles by Source from 2020 to 2023.....	43
4.2	t-SNE visualization of RoBERTa embeddings for news categories clusters.....	44
4.3	A snapshot of Event Dataset.....	44
4.4	Word Cloud of Frequently Mentioned Terms in Cement Sector Events (2020-2023)	45
4.5	Word Cloud of Frequently Mentioned Terms in Oil Gas Sector Events (2020-2023)	46
4.6	Word Cloud of Frequently Mentioned Terms in Economic Sector Events (2020-2023)	47
4.7	Word Cloud of Frequently Mentioned Person Entities in News Events (2020-2023)	48
4.8	. Financial events trend with respective sentiment count on KSE 100.....	48
4.9	Financial events trend with respective sentiment count on Oil & Gas sector.....	49
4.10	. Visualization of key news events timeline aligned with KSE 100 bullish trend	49
4.11	Financial events trend with respective sentiment count on Oil & Gas sector.....	50
4.12	t-SNE Visualization of News Embeddings by Standardized Feature.....	52
4.13	Comparative Analysis of Top Features in the Petroleum Sector.....	53
4.14	Impact of “Rising Oil Prices" on OGDC Stock Price.....	53
4.15	Impact of “Oil Prices Fell" on OGDC Stock Price.....	54

List of Abbreviations and Symbols

Abbreviations

PSX	Pakistan Stock Exchange
NLP	Natural Language Processing
FinTech	Financial Technology
NLFF	Natural Language Processing for Financial Forecasting
GDP	gross domestic product
CAGR	Compound Annual Growth Rate
KSE	Karachi Stock Exchange
KATS	Karachi Automated Trading System
BATS	Bonds Automated Trading System
FinBERT	Financial Sentiment Analysis with Pre-trained Language Models
KITS	Internet- based order routing system
IT	Information Technology
NER	Named Entity Recognition
LLMs	Large Language Models
t-SNE	t-Distributed Stochastic Neighbor Embedding

CHAPTER 1

Introduction and Motivation

1.1 Overview

Financial markets are essential for promoting economic growth and encouraging innovation. They provide as a platform for the buying and selling of different assets, including bonds, stocks, and derivatives [58]. Financial analysts and investors must possess a thorough understanding of financial markets as it is essential for the operation of the economy [58]. The stock market is a financial market where companies raise funds by selling shares of stock, which is also referred to as equity, to investors [46]. Forecasting stock market fluctuations is an intricate dilemma that has captivated finance experts and professionals for a considerable period of time. In Pakistan, where ensuring economic stability and promoting prosperity are of utmost importance, the ability to predict market trends and investor sentiments becomes even more significant.

In 2023, the Pakistani economy had a nominal gross domestic product (GDP) of USD 339 billion [60]. It is characterised by a variety of economic sectors, including services, manufacturing, agriculture, and other industrial activities. Although the country has achieved significant accomplishments, such as maintaining an average real GDP growth rate of 3.9% over the past decade [60], it continues to struggle with converting its economic potential into concrete advantages for its people.

The stock market is a crucial component of Pakistan's financial ecosystem, functioning as an indicator of economic well-being and investor outlook. Nevertheless, the precise forecasting of stock market fluctuations continues to be challenging due to the intricate nature of financial markets, which encompass aspects such as volatility, uncertainty, and the intricate interplay of several elements, including geopolitical events and regulatory adjustments [42]. Researchers are

drawn to examine advanced strategies in order to improve prediction due to the unpredictable character of this phenomenon. Accurately forecasting stock market developments leads to substantial financial gains.

There are two primary methods used for predicting stock trends: fundamental analysis and technical analysis[57]. Technical analysis examines historical data and trading volumes of stock prices, whereas fundamental analysis takes into account not only stock information but also evaluates the performance of the industry, political events, and economic conditions[46]. Fundamental analysis is more pragmatic as it assesses the market in a wider context. Given this background, combining textual news sentiments with stock market data presents a promising method to improve the predictive skills of models in the Pakistani environment[46].

News stories that are based on text, and represent the combined knowledge and opinions of market players, provide vital information about investor sentiment and market trends. Textual data is a superior source of concealed information compared to numerical data since it enables the prediction of financial trends with supporting evidence[52]. For example, a news story on a company that includes terms such as "resignation" or "risk of default" enables investors to anticipate a decline in the company's stock values. Moreover, information on numerous unpredictable variables has the potential to influence fluctuations in the stock market[47]. Examples of such events include economic and political upheavals, armed conflicts, civil disturbances, acts of terrorism, and natural calamities. Consequently, there is a significant requirement for improved knowledge discovery processes to choose features from textual data.

The impact of everyday news and events on the development of market sentiment is undeniable[47]. The automation of this process is crucial because experienced traders require a thorough understanding of the correlation between these occurrences and market behaviour in order to make well-informed and smart judgements on buying and selling[51]. There has been a significant amount of study conducted on the correlation between news and stock market performance[51, 52, 55]. Many studies have explored how different elements can influence stock prices and returns[56, 58]. Lately, advancements in Natural Language Processing (NLP) have opened up new possibilities for tackling this specific difficulty. According to De Oliveira Carosia et al.[49], positive news had a more significant impact on stock returns than negative news in the Brazilian stock market [10]. Zhou et al.[9] examined the relationship between media attention and stock market performance, focusing specifically on the China-Pakistan Economic Corridor. He deduced that investors demonstrate a higher inclination to spend an additional amount for

the exact same stock when there is a sudden increase in positive news. As a result, the return of the stock market also increases. Rashid et al.[54] asserted that macroeconomic news exerts a significant influence on market results. The ongoing study of extracting events from news is a persistent area of research, with current investigations primarily centred around historical and news data. Furthermore, the absence of studies especially examining current achievements in a particular country, such as Pakistan, can be attributed to the varying market conditions in each country.

Historical stock price data is a crucial source of information for anticipating future patterns in the stock market. Stock market data offers numerical indicators of market activity and patterns[49]. The combination of these two distinct sources of information offers a distinct possibility to utilise machine learning algorithms and sophisticated analytical methodologies in predicting stock market movements. Through the examination of the association between the attitudes expressed in news articles and the fluctuations in stock values, researchers have the ability to discover concealed patterns and connections that may not be apparent through conventional analytical approaches.

1.2 Fintech and Stock Market

The financial sector has a long history of incorporating technological innovations. The current upsurge in financial technology (fintech) has fundamentally transformed the conventional banking and financial services industry.

1.2.1 What is Fintech

Financial Technology (**FinTech**) is the application of technology to enhance, streamline, and customise banking and financial services. FinTech is the convergence of digital technologies and finance, which has a substantial impact on developing countries[16]. It enhances the effectiveness of financial services by creating user-friendly advanced technical platforms and expanding the range and complexity of financial services. FinTech provides a wide array of financial goods, services, software, and distinctive communication platforms. FinTech advancement enables the progress of digital innovation, enhances conventional financial institutions, and reduces information asymmetry[16].

FinTech has significantly transformed the financial industry, including banking and investing.

Currently, it is also playing a crucial role in reshaping the methods through which investors engage in stock markets. The FinTech market is experiencing rapid growth. According to Statista, investments in FinTech companies between 2015 and 2019 amounted to \$428 billion. Additionally, it is estimated that there are approximately 20,000 active FinTech enterprises worldwide as of the end of 2020[59].

1.2.2 Importance of Stock Market

The functioning of a nation's financial market is a vital factor in determining its overall economic state, allowing economists and financial professionals to assess the nation's present economic well-being. Within the financial sector, the stock market is a prominent and influential industry. Due to its inherent volatility and potential for significant gains or losses, the stock market has consistently attracted the attention of investors[16]. Consequently, experts have devoted considerable attention to the study of stock forecasting.

The economic condition of a country has a direct or indirect influence on various industries, including finance, agriculture, metal, and investment banking, among others. The expansion of these industries depends on their instability, which adheres to the fundamental principle of supply and demand. The stock market is directly influenced by the demand for a specific sector[16]. When there is a rise in supply, traders and financial institutions are prompted to invest in that sector or stock, which leads to an increase in pricing. In addition, consistent dividend payments enhance the development of profits and yield on invested capital. Investors must accurately choose the optimal time to sell their shares in order to attain their intended profits. Financial markets consist of several sorts of markets, such as stock markets, derivatives markets, bond markets, and commodities markets. The stock market functions as a venue where investors can allocate capital to acquire and possess a share or fraction of a company[43]. As firms expand, they frequently need additional financial resources to sustain their future initiatives. Upon receiving approval from existing shareholders, firms have the option to sell newly issued shares to investors in order to raise capital, which may result in diluted ownership for the current shareholders. Positive outcomes lead to a rise in the stock market value of the shares.

Stocks traded on the stock market can be purchased for either short-term or long-term investment plans. Long-term investment entails retaining shares for an extended duration, whereas short-term investments involve purchasing and selling shares within shorter time periods, with investors seeking to generate profits within days or weeks[43]. Traders utilise a diverse array of

trading tactics, such as swing trading, day trading, position trading, and scalping.

The stock market plays a crucial role in a nation's economy by offering a platform for corporations and investors to get finance and make investments. Anticipating the future performance of the stock market not only offers investors investment guidance, but also aids enterprises in devising financing strategies, thereby fostering the sound growth of the economy[16].

The stock market exhibits high volatility due to the dynamic nature of stock prices, which fluctuate based on the trading activity and volume of shares in the market. The market is subject to the influence of national policies, global and regional economics, as well as psychological and human aspects. Consequently, external variables such as social media and financial news can either have a good or negative impact on stock prices[16]. Throughout history, the stock markets have been predominantly controlled by prominent investors and bankers, primarily due to their ability to get and utilise valuable information, enabling them to make well-informed decisions. For those who are inexperienced in the stock market, where having access to information is crucial, fintech has revolutionised the established order by making market insights and extensive data available to everyone through the use of technology[16].

1.2.3 Landscape of Pakistan's Stock Market

The Pakistan Stock Exchange was founded on September 18, 1947 and officially registered on March 10, 1949 as the 'Karachi Stock Exchange', operating as a limited liability company. A second stock exchange was established in Lahore in October 1970 to cater to the stock trading requirements of the provincial metropolis. The establishment of the Islamabad Stock Exchange took place in October 1989 with the aim of serving investors in the northern regions of the country. The Stock Exchanges (Corporatization, Demutualization and Integration) Act, 2012 was enacted by the Government of Pakistan to address the lack of coordination and shared structure among the three exchanges[54]. As a result, the exchanges merged their operations on January 11, 2016, forming the new entity called 'Pakistan Stock Exchange Limited' (PSX). This integration brought together the separate management, trading interfaces, and indices of the three exchanges.

Pakistan Stock Exchange (PSX) aims to become one of the top Exchanges globally, with a remarkable track record of being the highest-performing Stock Exchange in Asia. Recently, the PSX experienced Compound Annual Growth Rate (CAGR) of 26% across all indices from 2012 to 2017[54]. PSX strives to regain its previous standing and has implemented strategies to trans-

form into a resilient and highly efficient Stock Exchange that is on par with, or even surpasses, other global Stock Exchanges. It plays a leading role in Pakistan's economy being the sole significant institution facilitating capital formation in the country. The Pakistan Stock Exchange possesses operational capabilities that are on par with any international stock exchange. It has robust trading systems such as Karachi Automated Trading System (**KATS**), Internet- based order routing system (**KITS**), and Bonds Automated Trading System (**BATS**)[54]. Additionally, it has advanced Information Technology (**IT**) security systems and a progressive regulatory framework. Presently, there are 11 indexes that are listed on PSX, which are as follows:

- KSE 100 Index
- KSE 20 Index
- KSE All Share Index
- KMI 30 Index
- PSX KMI All Shares Index
- BKTi (Tradeable Banks Index)
- OGTI (Tradeable Oil & Gas Index)
- NITPG Index
- UPP9 Index
- MZNPI Index
- NBPPGI Index
- JSMF Index
- ACITEF Index
- PSX Dividend 20 Index

PSX has achieved notable progress throughout its existence, starting with a modest presence of 5 listed businesses and a total paid-up capital of Rs 37 Mn. In 1960, there were 81 firms with a market capitalization of Rs 1.8 billion. Currently, there are 522 companies registered on the stock exchange, including 2 on the GEM Board, with a market capitalization of PKR 7.2 trillion. The listed companies are divided across 37 areas or groups of industry[54].

1.2.4 Unique Dynamics of PSX

The Pakistan Stock Market encounters distinctive challenges that set it apart from the markets of other developed nations. Given its intrinsic volatility and dynamism, the market functions in an environment that is characterised by unpredictability and fast change[54]. The volatility is affected by various causes such as political instability, economic fluctuations, and external financial pressures, which can cause substantial and sudden shifts in market patterns. Moreover, the Pakistan Stock Exchange is influenced by unique features that differentiate it from other markets.

Various sociocultural factors have the potential to induce abrupt fluctuations in investor sentiment and market dynamics[24]. Occurrences such as political elections, alterations in government policy, and geopolitical conflicts in the region can swiftly and significantly influence the stock market.

The market is significantly influenced by macroeconomic factors, including inflation rates, foreign exchange reserves, and monetary policies established by the State Bank of Pakistan. Therefore, understanding the peculiarities of the Pakistan Stock Exchange is essential for both investors and active traders due to its intricate nature[24]. Gaining a comprehensive understanding of the distinctive forces in action is crucial for making well-informed trading recommendations

1.2.5 Financial market prediction

Stock data is a classic time series. Many researchers have used time series models for forecasting, such as ARIMA or GARCH models, but the assumptions of classic time series models are relatively high. For example, the series needs to be stable and linear. However, there are many factors that affect the stock price of stock data, which makes the stock data itself not stable and linear. Therefore, stock market data is not only enough for market prediction[19].

Financial market prediction based on news disclosures is attracting more and more attention in recent years, because financial news has great influence in the stock market[58]. Investors often rely on financial news information to make their decision of buying or selling. However, it is very difficult to predict the financial market accurately based on news disclosures, due to the complexity and ambiguity of natural languages used.

1.3 Problem Statement

Although there has been progress in predicting sentiment for news articles and stock movements, there is still a need to effectively combine technical indicators from the stock market with textual news sentiments to improve the predictive abilities of models[54]. This difficulty is especially noticeable within the distinct characteristics of Pakistan’s stock market. To address this deficiency, our research aims to meticulously create a well-organized dataset that includes news events and their accompanying stock market prices. We employ advanced sentiment analysis techniques to extract subtle and detailed insights from textual news articles. The primary goal is to understand the complex relationship between news sentiments and stock market movements. Our goal is to enhance predictive models by incorporating rich information provided by News events. This would enable the models to effectively analyse sequential information in time series data, leading to a more thorough knowledge and improved forecasting capacity in relation to Pakistan’s stock market. Our goal is to use natural language processing (NLP) to automatically detect and combine events in order to highlight important occurrences. These events can then be used to deliver knowledgeable trading recommendations to market players.

1.4 Relevance to National Needs

Pakistan’s economic landscape stands at a crucial juncture, where informed decision-making in financial markets plays a pivotal role in fostering economic stability, attracting investment, and driving sustainable growth. With a nominal GDP of USD 339 billion in 2023 as shown in Figure 1.1, Pakistan’s economic potential is evident, yet challenges persist in translating this potential into tangible outcomes for its citizens[60]. The stock market, as a barometer of economic health and investor sentiment, holds significant importance in this regard.

However, the accurate prediction of stock market movements remains a formidable challenge, fraught with uncertainties and complexities unique to Pakistan’s financial ecosystem. Despite an average real GDP growth of 3.9% over the last decade[stat], the volatility of Pakistan’s stock market often mirrors broader economic fluctuations, influenced by geopolitical events, regulatory changes, and economic indicators.

The accessibility and interpretation of timely information pose additional challenges for market participants, hindering their ability to make well-informed investment decisions. Despite private consumption accounting for 85% of GDP in 2021 and government consumption at 11%, fixed

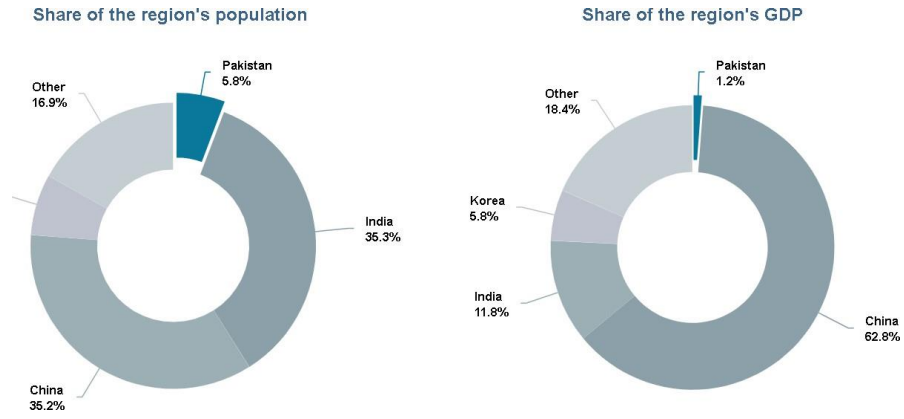


Figure 1.1: Share of the region’s population and GDP Growth

investment remains relatively low at 14%, signaling a need to bolster investor confidence and incentivize long-term capital allocation[60].

Against this backdrop, the fusion of textual news sentiments with technical indicators represents a promising avenue for enhancing the predictive capabilities of models in the Pakistani context. With services accounting for 58% of overall GDP in 2021, manufacturing at 12%, other industrial activity at 7%, and agriculture at 23%, there exists a diverse array of economic activities that can benefit from improved forecasting techniques *statista dp pakistan*.

By leveraging advanced sentiment analysis techniques and integrating key market indicators, our research seeks to address these challenges and provide valuable insights into the interplay between news events and stock market dynamics. In 2021, manufactured products made up 75% of total merchandise exports, indicating a reliance on a few key export categories that could benefit from enhanced market insights and diversification strategies[60].

The implications of our research extend beyond academia, with profound implications for national development goals and economic prosperity. By empowering investors with more accurate predictions and actionable insights, we aim to bolster investor confidence, attract domestic and foreign investment, and stimulate economic growth. Moreover, the adoption of data-driven decision-making processes within financial institutions and regulatory bodies can contribute to the efficiency and transparency of Pakistan’s financial markets, enhancing their resilience in the face of global economic uncertainties.

Furthermore, our research holds the potential to foster collaboration and knowledge-sharing among various stakeholders, including government agencies, financial institutions, and academic researchers. By facilitating dialogue and partnership, we can leverage collective expertise

and resources to address pressing challenges and seize emerging opportunities within Pakistan's financial landscape.

In conclusion, our research endeavors to bridge the gap between theory and practice, offering tangible solutions to enhance the predictive accuracy of stock market models and contribute to the broader objectives of economic development and prosperity in Pakistan. Through strategic partnerships and informed decision-making, we aspire to realize Pakistan's full economic potential and create lasting value for its citizens.

1.5 Research Objectives

The main research objectives that lead this thesis are:

- Extraction and formulation of an extensive financial news data relevant to the Pakistan Stock Exchange (PSX) from notable sources.
- Mapping the sequential snapshots of financial events to their corresponding stock market trends and identifying correlations between them.
- Identifying the features that are important to PSX (Pakistan Stock Exchange) because of the unique market dynamics in Pakistan, and validating the significance of these features with experts.
- Presenting a systematic approach for extracting events from news disclosures. This entails creating a chronological sequence of events and analysing how it aligns with the fluctuation of stock prices in order to generate trading suggestions.

In the context of identifying the correlation between stock prices and news events, some basic research questions are identified that direct/initiate this thesis work.

1.6 Research Questions

- Previous studies have utilized machine learning models to predict outcomes. How can attention-based models, specifically Financial Sentiment Analysis with Pre-trained Language Models ([FinBERT](#)), enhance the understanding of the semantics in financial news items and social media texts?

- How effective are the Natural Language Processing techniques for stock market prediction taking into account the market trend, news sentiments and new trends?
- What specific types of news (e.g., economic, political, corporate) have the most significant impact on stock market movements?
- What are the limitations and challenges in using NLP for correlating news with stock market movements?

1.7 Major Contributions

This thesis aims to fill a notable research gap by investigating the correlation between news and stock prices in the specific context of Pakistan’s Stock Exchange. This market is known for its distinctive characteristics and volatility as an emerging economy. The study’s significant contributions are diverse and provide significant progress in the field of financial analysis through the utilisation of natural language processing (NLP).

This research utilises sophisticated attention-based models, specifically Fin-BERT, to effectively capture the meaning and context of financial news articles. The unique implementation of Fin-BERT improves the capacity to predict fluctuations in the stock market by effectively deciphering the subtle language used in financial communications. This contribution is essential in a context such as Pakistan’s, where market responses to news can be particularly significant due to the distinct economic and political environment.

Furthermore, the study assesses the efficacy of different Natural Language Processing (NLP) methods in predicting stock market movements. This is achieved by employing market trend, news attitudes, and emerging trends. This showcases the pragmatic usefulness of NLP in predicting stock prices in an emerging market setting. This complete method facilitates the connection between theoretical models and practical implementations, offering a more resilient foundation for forecasting market behaviour.

An important contribution of this research is the recognition of distinct categories of news—economic, political, and corporate—that exert the greatest influence on stock market fluctuations in Pakistan. This distinction is crucial for investors and regulators, providing valuable information on which news categories should be regularly observed in order to predict market changes. This study emphasises the diverse levels of impact that different sorts of news have on stock prices, illustrating the distinct sensitivities of the Pakistani market.

Furthermore, this thesis examines the constraints and difficulties associated with utilising natural language processing (NLP) to establish a correlation between news and fluctuations in the stock market. The text identifies and examines the practical challenges faced, including data scarcity, linguistic subtleties, and the requirement for reliable data sources. By recognising these obstacles, the study offers a plan for future investigations to surmount these hindrances and enhance the precision and dependability of NLP-driven financial forecasts.

Overall, this thesis offers valuable contributions by utilising Fin-BERT to improve the analysis of financial news, assessing the efficacy of NLP techniques in predicting stock market trends, determining the most influential types of news on Pakistan's stock market, and addressing the obstacles associated with employing NLP in this particular domain. These contributions not only enhance academic comprehension but also provide practical perspectives for investors, analysts, and policymakers working in developing economies such as Pakistan.

1.8 Upcoming Chapters

The structure of this thesis is as follows. Section 2 offers a comprehensive summary of the relevant studies. Section 3 elaborates on the suggested framework and its implementation process. The evaluation of the strategy is carried out in Section 4. Section 5 concludes the findings of the study and Section 6 provides recommendations for future research efforts.

CHAPTER 2

Literature Review

As the equities market develops, more and more investors are realizing that there are a variety of lucrative investing options there [43]. Financial news is the first-hand information that the general public can access, and investors base their investment decisions on it [27]. Finding a way to forecast how the news will affect the stock market is therefore extremely important from a practical standpoint for investors. The development of various computer technologies has led to a considerable surge in research on the use of text mining and machine learning algorithms in the financial sector.

Numerous studies have explored various approaches to stock market prediction, including technical analysis, fundamental analysis, and machine learning algorithms. Technical indicators, derived from historical price data, are a popular tool among traders and analysts for making predictions about future stock price movements [28, 51]. This literature review has been segmented into three parts i.e. literature related to Sentiment analysis, Named Entity Recognition and Event Extraction.

2.1 Financial News Sentiment Analysis

Recent studies have shown that there is a significant relationship between articles about stock and changes in stock price.

The authors of [40] focused on the Brazilian electoral period of 2018 and examined sentiment from three perspectives: the absolute number of tweet sentiments, sentiments weighted by favorites, and sentiments weighted by retweets. The authors employ various Machine Learning techniques including Naive Bayes, Support Vector Machines, Maximum Entropy, and Multi-

layer Perceptron (MLP) for SA comparison. A thorough comparison of conventional ML approaches including Naive Bayes, Maximum Entropy, SVM, and MLP is given using the BoW embeddings. The dataset used for this study comprised tweets and news in Portuguese. The experiments show that the multilayer perceptron outperformed the rest of the models. The paper presents detailed insights into how predominant social media sentiment and stock market movement are correlated. This sets a research direction for our work too. Also, the scope of this research is limited to conventional ML approaches that are not capable of capturing the semantics of the text. In another study [50] making use of ML approaches, the authors tackled challenges in sentiment analysis for Lithuanian financial texts using supervised machine learning. The authors utilize multinomial Naive Bayes, LSTM, and SVM algorithms with hyperparameter optimization. The highest accuracy achieved is 71.1% with Naive Bayes (NB), followed by LSTM (71%) and SVM (70.4%). The SVM and NB model's assumption of feature independence compromises its ability to capture nuanced word relationships in sentiment analysis. Similarly, LSTM models excel at sequence learning but can face the vanishing gradient problem when the input sequence exceeds a certain length. Addressing these aspects would make the research more impactful.

Aiding further research, Wu et. al [57] introduced the SI-LSTM approach for accurate stock price prediction. It combines diverse data sources and sentiment analysis using a convolutional neural network. The Long Short-Term Memory (LSTM) network is employed for prediction. Experiments show improved accuracy, with a mean absolute error of 2.38, surpassing traditional methods. The approach's effectiveness is validated using real data from the China Shanghai A-share market. Another similar study [47] using sequence models introduces an innovative approach to enhance stock price prediction accuracy by integrating machine learning techniques and Long Short-Term Memory (LSTM) networks. The study leverages sentiments from news articles, alongside technical analysis, to improve predictions. The LSTM model is applied to historical stock data and sentiments extracted from the scrapped news headlines and OHLC dataset to create a more effective predictive model. The paper reported an accuracy of 88.73%. The only limitation of these two studies is that LSTMs struggle with capturing long-range dependencies in sequences due to vanishing gradient issues, limiting their ability to model complex relationships compared to newer architectures like Transformers [18].

The authors of [49], introduce a methodology using a Convolutional Neural Network (CNN) for sentiment analysis of Brazilian financial news. It establishes a significant correlation between daily news sentiment and stock market behavior, proposing profitable sentiment-based

investment strategies that outperform traditional methods with 86.5% accuracy. One challenge here is that using CNNs for textual analysis has some limitations. They struggle with complex semantics, varying sentence lengths, and hierarchical structure comprehension. Sequential context understanding is lacking, and they might not surpass specialized NLP models.

The authors in another study [55] aimed to predict stock market prices by combining PCA, EMD, and LSTM techniques alongside sentiment analysis. Experiments used stock market price data from Thailand (2018-2022) and financial news data of 12,667 news articles. Authors implemented FinBERT with Thai news fine-tuning, achieving an 82% accuracy outperforming other sentiment analysis models. Results showed the PCA-EMD-LSTM framework outperformed baseline methods in predicting stock prices. However, the application of news sentiment to EMD-LSTM did not consistently yield improvements in all components. One of the limitations of this study is that it used traditional or hybrid machine learning models, missing potential improvements from recently developed machine learning model combinations for time series forecasting.

Building on this notion of incorporating sentiment analysis, another paper aimed to enhance stock price prediction accuracy through a two-component model [56]. One component analyzed stock patterns using LSTM, while the other predicted sudden stock movements using sentiment analysis, achieving an average RMSE score of 0.33 and an accuracy of 84.9%. This study utilized Kaggle's dataset, which contained daily news headlines from 2000, merged and pre-processed for analysis. The integration of LSTM predictions and news sentiment enabled the model to effectively determine stock movement direction, offering more accurate forecasting. Nonetheless, this study encountered limitations in predicting sudden stock price changes with precision.

In a parallel vein, [52] employed deep learning, particularly LSTM, for stock price prediction, integrating news sentiment analysis through the FinBERT-LSTM model. Evaluations on NASDAQ-100 data and New York Times articles showcased the model's superior performance, underscoring the advantages of sentiment integration. The model suite encompassed a Multi-layer Perceptron (MLP), LSTM, and FinBERT-LSTM, with the latter showcasing a 20.2370% MAPE improvement and 0.3640% accuracy boost over MLP, as well as a 3.2424% MAPE enhancement and 0.0479% accuracy improvement over the vanilla LSTM approach. However, a limitation lay in the model's potential oversight of other external factors that could influence stock behavior, given its primary focus on news sentiment integration.

Furthermore, in the context of facilitating rapid stock market decision-making, [36] presented

the use of Bidirectional Encoder Representations from Transformers (BERT) for sentiment analysis of breaking news articles. The study involved fine-tuning a BERT BASE model using manually labeled sentiment data from 582 stock news articles, resulting in a commendable 72.5% F-score. Performance comparison with naive Bayes, SVM, and TextCNN revealed BERT's superior accuracy. Despite these promising sentiment analysis outcomes, the study acknowledged limitations stemming from a short data collection period and the complexity of interpreting the relationship between sentiment analysis and the Dow Jones index. Future research avenues were suggested, including individual company news analysis and the incorporation of accounting data for improved prediction accuracy.

In conclusion, recent research has highlighted the potential of sentiment analysis in enhancing stock market prediction accuracy. Various studies have utilized diverse techniques to combine sentiment information with market forecasts, yielding promising results. However, common limitations exist, such as the restricted capability of traditional machine learning models in capturing complex relationships and the potential oversight of external factors by models like LSTM. Newer approaches like Transformers offer solutions to these challenges. While sentiment integration through methods like BERT shows promise, challenges include interpreting sentiment-market relationships and addressing short data collection periods. These studies emphasize the significance of sentiment analysis, while also indicating the need for further advancements to overcome existing limitations and leverage evolving machine learning and NLP techniques.

2.2 Financial News Named Entity Recognition

Named Entity Recognition (NER) is a crucial natural language processing task that involves identifying and classifying named entities, such as names of individuals, organizations, dates, and more, within a text corpus.

Le et al [44] introduced the Maximum Entropy Named Entity (MENE) model, which aim to improve the accuracy of named entity recognition tagging by utilising a wide range of knowledge resources. Other maximum entropy models were proposed by Bender et al. [3] and Chieu and Ng [1]. McNamee and Mayfield [2] proposed an SVM (Support Vector Machines)-based model called SNOOD, requiring minimal linguistic knowledge and adaptable to various target languages. Their classifier was trained on 258 orthography and punctuation features and 1000 language-related features, rendering binary decisions for the current token's class. However,

SVMs lack consideration of neighboring words.

To address this limitation, Conditional Random Fields (CRFs) are employed. McCallum and Li [4] introduced a feature induction model for CRFs in named entity recognition, while Krishnan and Manning [7] presented a coupled CRF approach utilizing latent representations. CRF-based NER has found applications in multiple domains, including chemical text [11], tweets [9, 10], biomedical text [5, 45], and the legal domain, where Sharafat et al. [35] employed CRF models to extract named entities from civil court proceedings.

The authors of [14] explored the critical task of extracting named entities from financial press releases. By employing Conditional Random Fields (CRF) as the underlying model, the study addresses the challenge of recognizing entities such as companies, monetary amounts, and dates within the financial domain. The authors utilize a specialized dataset, and the results demonstrate the effectiveness of CRF in identifying entities in financial press releases, achieving a notable level of precision. However, the study's limitation lies in its focus on a specific domain, potentially hindering generalizability.

Huang et al. and colleagues [30] introduced a multi-tasking deep structural model, integrating "partially annotated" datasets to collectively recognize all potential entity types within training corpora. In a similar vein, Liu et al. [34] proposed diminishing the neural models' uncertainty through the incorporation of external gazetteers. Zie and Lu [32] advanced the field by presenting a dependency-guided LSTM-CRF model capable of encoding comprehensive dependency trees, capturing syntactic relationships, and deducing the presence of named entities in the context of the NER task.

Strubell and co-authors [17] introduced an expedited substitute for Bi-LSTMs known as Iterative Dilated Convolutional Neural Networks (ID-CNNs), displaying enhanced capability for encompassing substantial context and predictive structure, surpassing conventional CNNs. Ghaddar and Langlias [20] observed the tendency to disregard lexical features in neural network-driven NER methods and devised an approach to embed words and entity types into low-dimensional vector spaces, trainable from annotated data. They crafted feature vectors representing individual words and harnessed a basic recurrent neural network (RNN) for precise entity recognition. Batbaabar et. al [29] performed health-related entity identification in Twitter messages. Employing a deep learning framework with BiLSTM and CNN, they utilize CRF and the Viterbi algorithm for sequence labeling. The methodology leverages a healthcare ontology, achieving impressive precision, recall, and F1-scores: 93.99%, 73.31%, and 81.77% for disease/syndrome entities; 90.83%, 81.98%, and 87.52% for sign/symptom entities; and 94.85%, 73.47%, and

84.51% for pharmacologic substances. The study's manual annotation validates high-quality results despite complex medical terminology and tweet context limitations.

The study [53] provides a comprehensive survey of recent advancements in Chinese Named Entity Recognition (CNER). It covers both traditional and deep learning approaches, with a focus on the latter. The review outlines common datasets, tag schemes, evaluation metrics, and challenges specific to CNER. The deep learning perspective is explored through a three-layer architecture involving character representation, context encoding, and tag decoding, along with attention mechanisms and adversarial transfer learning. In another study [31] for Chinese NER, the authors proposed a model for named entity recognition (NER) in Chinese electronic medical records. By incorporating a BERT-BiLSTM-CRF architecture, the study aims to enhance word representation by utilizing BERT's pre-trained language model. The model's effectiveness is evaluated against baseline models using the CCKS 2017 datasets. Results indicate that the proposed BERT-BiLSTM-CRF model outperforms the baseline models with 88.45% F1-score in terms of NER performance. This approach capitalizes on the strengths of BERT's semantic representation and BiLSTM-CRF sequence labeling, showcasing its potential for improving NER tasks, particularly in specialized domains like medical records.

In conclusion, Named Entity Recognition (NER) plays a pivotal role in identifying entities within text. Previous studies have showcased diverse approaches such as MENE's integration of knowledge sources, SNOOD's SVM-based approach, and CRFs for context. These techniques find applications in specialized domains like legal and financial contexts. Innovations include ID-CNNs for context prediction, entity type embedding, and health-related NER. The field also benefits from comprehensive surveys in Chinese NER and the effectiveness of BERT-BiLSTM-CRF in medical records NER. These collective efforts underscore the continuous evolution of NER methods to address a wide range of challenges in information extraction and understanding.

2.3 Financial News Event Extraction

Another significant NLP application in the field of financial forecasting is event extraction. Event extraction focuses on identifying key event features within the text and structurally expressing these events, as opposed to the semantic method, which analyses whole texts, including sentences and paragraphs. Its main objective is to reduce the noise of unnecessary text by compressing important information by grouping related news stories. In this sense, an event is a

particular occurrence having a clear time, location, and involvement of one or more people, frequently denoting a change in status [39]. In order to filter out repeated information, the event extraction job entails recognising repetitive material and putting it into an event structure.

According to the research by [12], a news event may alter investor perceptions, prompt trade, and affect stock price movement. Event extraction is being used in a growing number of commercial and financial applications, including enabling businesses quickly identify market reactions, deriving signals for trading suggestions, risk analysis, and more [26].

Conventional methods for filtering pertinent text include both manual and automatic pattern discovery techniques [6]. Manual approaches require extensive knowledge bases and rule sets, limiting their adaptability to specific domains. Automatic methods encompass statistical, linguistic, and machine learning solutions [13], such as TF-IDF [23], which may underperform on specialized texts. More advanced strategies incorporate knowledge heuristics, with fuzzy logic showing promise but needing manual rule generation [15].

While supervised approaches are prevalent in Knowledge Extraction (KE), unsupervised methods offer practicality by eliminating the need for text tagging. Examples of supervised KE methods include Gottipati et al.'s ML-based course improvement solution [21], López-Úbeda et al.'s extraction of radiological report information [48], and Verneer et al.'s relevance detection from social media messages [37]. Among extraction methods for relevant topics from news articles, Jacobs et al. [22] employed supervised economic event extraction, Oncharoen et al. [25] used Open Information Extraction for tuple representation, Carta et al. [46] utilized real-time clustering for event extraction, and Harb et al. [8] introduced linguistic-based opinion extractions.

Considering the causal relationship between financial news and asset prices [41], both machine learning and deep learning have been explored for context-dependent stock market information gathering [58]. Atkins et al. [19] used Naive Bayes and LDA-derived feature vectors for volatility prediction, while Shilpa & Shambhavi [58] presented sentiment analysis-based prediction. Notably, prior approaches often overlooked the importance of temporality.

Despite the prevalence of KE solutions, particularly in financial NA, temporal analysis has often been neglected. Many systems rely on timestamps or verb tenses for temporal references. Recent advancements [38, 33] focus on enhancing predictive models. CapTE [33] employs a Capsule network based on Transformer to capture deep semantic and structural information from tweets. Conversely, Att-DCNN [42] introduces dilated causal convolution networks with attention for event knowledge embedding, considering direct, inverse, and financial indicator relationships

for S&P500 index prediction. Both models outperform baselines with 64.22% (CapTE) and 72.23% (Att-DCNN) accuracy on different datasets.

In conclusion, event extraction is a vital NLP application in financial forecasting, structurally representing key event features within texts. These events, marked by temporal and contextual attributes, impact investor perceptions and stock prices. Event extraction's reach extends to commercial and financial sectors, aiding quick market reaction identification, risk analysis, and trading signals. Traditional text filtering methods range from manual to automatic pattern discovery. While supervised methods dominate Knowledge Extraction, unsupervised approaches offer efficiency. Recent advances in predictive models like CapTE and Att-DCNN highlight deep learning's role in enhancing performance. This evolving landscape underscores NLP's transformative potential in financial analysis and prediction.

2.4 Literature Gap

The oversight of the crucial relationship or association between stock entities in contemporary stock price prediction research reveals a gap in understanding the intricate dynamics that influence stock prices [64]. Furthermore, with the varying market dynamics from country to country [10], there is a dearth of literature that examines the current developments in the context of a particular country such as Pakistan. Additionally, the distinct contextual features impacting the Pakistan Stock Exchange, separate from the advanced economic financial market, emphasize the need for tailored approaches in financial forecasting [10]. Despite the application of natural language processing in financial forecasting, there is still a need to explore forecasting for decision-making based on financial events' timelines, news classification based on NER, and market sentiment.

CHAPTER 3

Methodology

The methodology section outlines the systematic procedure employed to examine the correlation between news and stock market fluctuations utilising natural language processing (NLP) techniques. The methodology section is divided into two significant modules. Module 1 specifically concentrates on extracting events from textual news stories and utilising various Natural Language Processing (NLP) techniques to provide context to those news events and mapping of those events to stock market data. The second module primarily concentrates on the identification, extraction, and synthesis of news-based events to generate actionable trading recommendation specifically for the Pakistani stock market. Our attention in this module is specifically on the Petroleum sector, which has the highest market capitalization and trading activity in the PSX.

3.1 Module1- Event-Driven Stock Movement Analysis

The main object of this module is to extract, map, and visualize new events timeline which require a sequence of phases, presented in this module. Figure 3.1 depicts these phases of the proposed module pipeline.

3.1.1 Standardized Dataset Creation

The gathering of data is a crucial component of every research project. In order to anticipate stock market trends using Natural Language Processing (NLP) techniques, it was imperative to establish a standardized dataset of financial news articles.

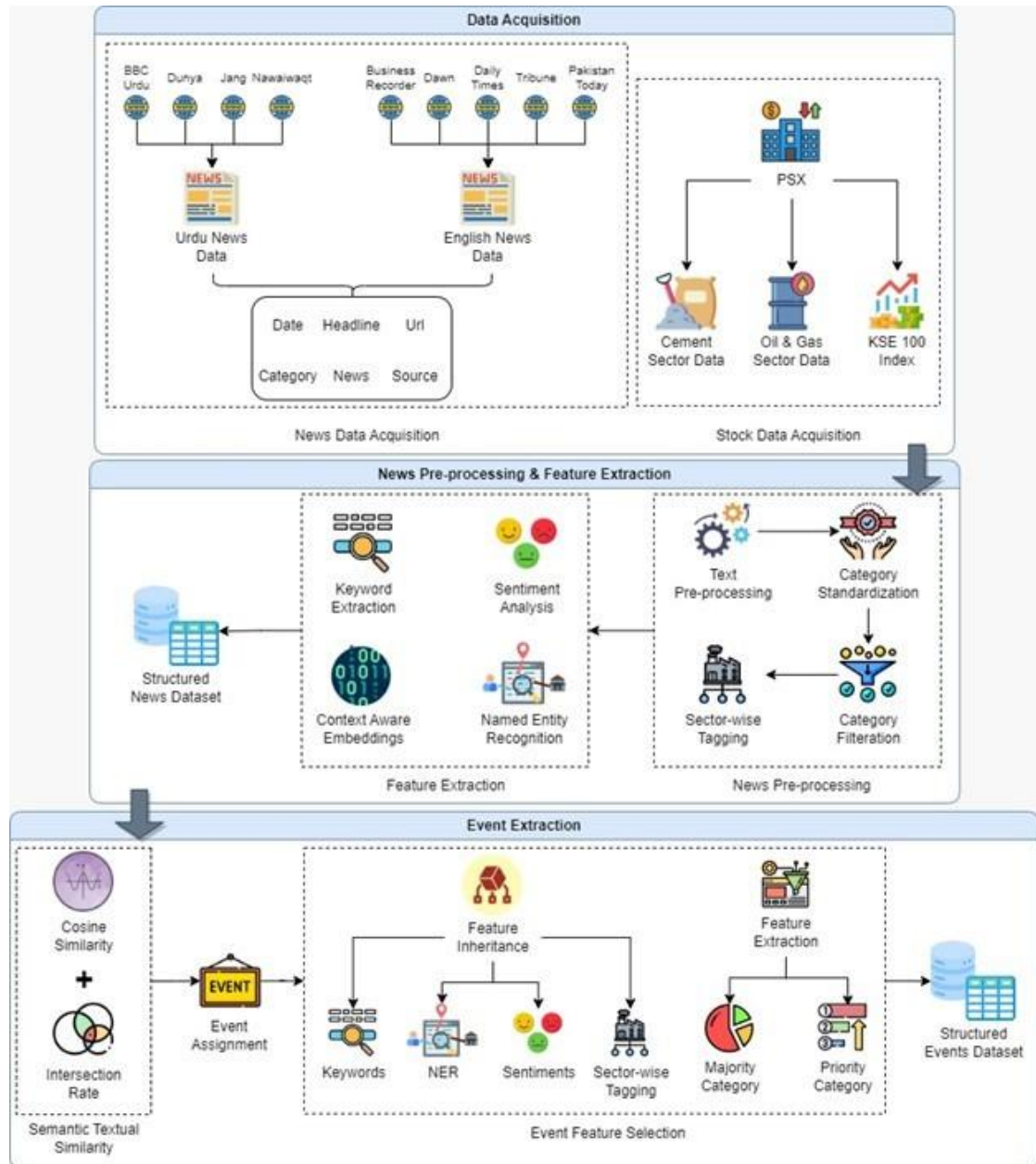


Figure 3.1: Methodology flowchart from data acquisition to visual analysis.

Data Gathering

The initial phase of dataset creation involved the systematic collection of financial news articles from five prominent Pakistani news websites: [Dawn¹, Business Recorder², Daily Times³, Tribune⁴ and Pakistan Today⁵]. These websites were selected based on their reputation for providing timely and comprehensive news coverage relevant to the Pakistani stock market.

Web scraping was employed as the primary method of data acquisition. A custom Python script was meticulously crafted to facilitate this process, utilizing widely recognized libraries such as BeautifulSoup, Requests, and Selenium. These libraries played a crucial role in seamlessly interacting with the target web pages and extracting essential components from the financial news articles. The script iterated through a predefined matrix of month-date combinations for all the years from 2020 till 2023. For each date, it sent a request to the website's archive page, retrieved the HTML content, and then used BeautifulSoup to parse the content. Additionally, for each article, it sent another request to the individual article link, retrieved the HTML content, and extracted the article's description. The extracted data was then processed and cleaned using various functions such as 'cleanhtml' and 'unicodedata.normalize'. Finally, the cleaned data was stored in a CSV file.

The script was designed to capture a comprehensive set of attributes for each article, ensuring the dataset's richness and relevance. These attributes included the headline, a precise timestamp indicating the publication date, the URL of the news article, the news source's identity, the news category and a concise description summarizing the news content. As a result of this data collection process, we assembled a well-organized news dataset covering three years (2020-2023). This dataset holds valuable information, ready for detailed analysis and exploration.

Data Preprocessing

Raw scraped data often contains noise and inconsistencies that can impede accurate analysis. To ensure the dataset's quality, a series of preprocessing steps were applied to the collected articles:

Text Cleaning One of the primary preprocessing steps was text cleaning, where the extracted text underwent a rigorous cleaning process. During this process, HTML tags, special characters,

¹<https://www.dawn.com/>

²<https://www.brecorder.com/>

³<https://dailytimes.com.pk/>

⁴<https://tribune.com.pk/>

⁵<https://www.pakistantoday.com.pk/>

digits, and unnecessary formatting artifacts were meticulously removed from the text. This crucial step aimed to eliminate potential sources of noise and distractions within the textual data, laying the foundation for more accurate analysis. **Deleting Unnecessary Information** Data cleanliness was maintained by removing any unnecessary or redundant columns. Columns with names starting with 'Unnamed:' were systematically deleted, and if an 'index' column existed, it was also removed. Rows containing null values in any of the columns were removed, as they could introduce biases or inaccuracies in the analysis. Ensuring the dataset's structural integrity was essential for coherent analysis. **Date Column Formatting** The temporal aspect of the data was considered. The 'date' column, initially in a varied format, was uniformly converted into a datetime format while retaining only the date component. Subsequently, the dataset was sorted in ascending order based on the date and the index was reset. This organization of data ensured that the temporal sequence was consistent and logical throughout the dataset. **Category Standardization** The significance of category standardization became evident due to the variation in terminologies used by different news websites to denote similar content classifications. For instance, while the website "Dawn" utilized the term 'Business' to represent its business category, the website "Business Recorder" opted for the term 'Business & Finance' for the same classification. This disparity in naming conventions underscored the necessity of harmonizing categories within the news dataset. The process of category standardization involved creating a comprehensive key mapping strategy to unify content classifications from various news websites. This strategy encompassed distinct category names employed by different sources, such as 'Business' and 'Business & Finance.', 'World' and 'International'. Through a meticulously designed category_mappings dictionary, each news source's categories were aligned to a standardized set. Subsequently, a mapping function was applied to the dataset, which referred to this dictionary for assigning a unified category label to each article. In cases where a direct mapping wasn't available, articles were categorized under 'Others.' This systematic approach ensured consistency and coherence in the dataset's content categories, enabling streamlined analysis and interpretation. Table 3.1 shows how the dataset looked like after performing category standardization on it.

Dataset Formulation

To offer a tangible illustration of the dataset's structure and contents, a representative sample is provided in Table 3.2, showcasing the integration of headline, timestamp, URL, news source, Category and description for each news article. This table serves as a glimpse into the dataset's

Table 3.1: Category Standardization of News Dataset

Headline	DateTime	URL	Source	Category	main_cat
Activities of Karachi Port and Port Qasim	12/27/2023	https://www.brecorder.com/news/40280631/..	Business Recorder	Markets	Business
Salaam Takaful, Syngenta Pakistan pilot crop insurance programme in Punjab	12/27/2023	https://www.brecorder.com/news/40280667..	Business Recorder	Business & Finance	Business
Turkey parliament committee approves Swedenâ€™s NATO bid	12/27/2023	https://www.brecorder.com/news/40280643/..	Business Recorder	World	International

composition and forms an integral component of the dataset creation process.

Table 3.2: News Dataset

Headline	DateTime	URL	Source	Category	News
IMF programme revival..	2023-05-01	https://tribune.com.pk/story/2414379/imf-..	Tribune	Business	Once done, imports should be relaxed,..
Kyrgyzstan eager ..	2023-05-01	https://dailytimes.com.pk..	Daily Times	Business	Ambassador of Kyrgyzstan..

This structured dataset, enriched with vital attributes, establishes a robust foundation for subsequent analysis and application of Natural Language Processing techniques to anticipate stock market trends effectively.

3.1.2 Financial Data Extraction and Preparation

The stock price time series data is comprised of two dimensions of transaction data: stock daily open, high, low, close price and volume, with closing price and volume being the most commonly used metrics. Investors frequently utilise technical indicators to assess market conditions. In the domain of financial data, we carefully gathered information from the official Pakistan Stock Exchange website. For this thesis work, 2 PSX sectors were selected along i.e Oil & gas and Cement with their active stocks. These key sectors play a big role in Pakistan's economy. We looked at data from 2020 to 2023, putting in a lot of effort to ensure accuracy. We examined prominent companies like MARI, PSO, PPL, OGDC in the oil & Gas sector, and ACPL, Fauji Cement Limited, BestWay Cement Limited, and Kohat Cement Limited in the cement sector. For Oil & Gas and Cement sector, the active stocks taken are mentioned in the Table 3.3.

Table 3.3: Active Stocks of Chosen Sectors

Sector	Company Name
Oil & Gas	OGDC (Oil & Gas Development Company Limited)
Oil & Gas	PPL (Pakistan Petroleum Limited)
Oil & Gas	POL (Pakistan Oilfields Limited)
Oil & Gas	MARI (Mari Petroleum Company Limited)
Cement	ACPL (Attock Cement Pakistan Limited)
Cement	BWCL (Bestway Cement Limited)
Cement	FCCL (Fauji Cement Company Limited)
Cement	LUCK (Lucky Cement Limited)

Our financial dataset includes various details like time intervals, opening and closing prices, high and low values, trading volumes, and symbols. To address missing values in the financial data, particularly during weekends and holidays when stock prices aren't available, we employed the forward fill method. This method involves filling in missing stock prices for empty days by carrying forward the last known price. This process is done so that for every day, news data as well as financial data is available. The Table 3.4 shows how our financial data is aligned

Table 3.4: Financial Data

(a) Stock Price Data of 2 Sectors							
Time	Open	High	Low	Close	Volume	Sector	Symbol
2020-10-01	71.5	74.66	70.5	74.66	20,000	Cement	ACPL
2020-10-02	76	76.9	74.62	75.76	3,000	Oil & Gas	PSO
2020-10-03	76	78.88	76	78.51	14,000	Oil & Gas	PPL

(b) KSE-100 Index Data						
Date	Open	High	Low	Close	Change	Volume
2020-01-02	11339.21	11339.21	11193.63	11282.01	-65.65	29301396
2020-01-03	11288.46	11412.97	11253.63	11402.04	120.03	53124168
2020-01-04	11434.31	11463.51	11334.85	11361.97	-40.07	39976120

3.1.3 Sector-Based Tagging of Extracted News

In the process of sub-category identification within our news filtration system, a methodical approach was adopted to refine and enhance the depth of news analysis. This method involved a meticulous examination of online news sources and the integration of insights gleaned from ChatGPT. Through this process, a predefined keyword list was curated to align with our specific analytical objectives. Some samples from our sector specific keywords list are depicted in Table 3.5. This list was meticulously crafted to encapsulate key themes and topics deemed critical within our domain of interest. Few keywords that are included in every sector list are given below. The selection of three distinct sub-categories—Economic, Oil & Gas, and Cement—was underpinned by their recognized potential to significantly influence and correlate with stock prices within the chosen sectors. Each sub-category represents a unique area of focus within the broader realm of financial and economic news, bearing significance for our analytical endeavors. By employing the predefined keyword list, a systematic matching process was conducted against the descriptions of news articles. This systematic approach facilitated the assignment of relevant sub-categories to individual news items based on their thematic content. The association of news articles with specific sub-categories yielded several notable benefits.

Firstly, the process enhanced the relevance of our analysis by ensuring that news articles were filtered and categorized with precision, thereby maintaining focus on topics directly pertinent

Table 3.5: Sector's Keywords

Sector Name	Keywords
Economic	State Bank of Pakistan, FBR, Taxation...
Oil & Gas	Pakistan Petroleum, Crude Oil, Ministry of Energy...
Cement	Cement, Concrete, Construction Materials...

to our analytical objectives. Additionally, the identification of sub-categories enabled a more granular analysis within specific thematic areas, thereby facilitating the discovery of nuanced insights and trends that might otherwise have been overlooked.

Moreover, the categorization of news articles according to relevant sub-categories such as Oil & Gas and Cement provided the groundwork for conducting correlation analysis between news events and stock price movements within these sectors. Such correlation analysis is integral for identifying potential causal relationships and informing investment decisions.

Furthermore, the systematic approach to sub-category identification streamlined the workflow of our news analysis, enabling efficient processing of large volumes of news data. The predefined keyword list can be continually refined and expanded to adapt to evolving market dynamics and analytical requirements. Table 3.6 shows how our data looked like after assigning sector specific Sub-Categories to the News Articles.

Table 3.6: Assigning Sector based Sub-Cat to the News Articles

Headline	Sub-Cat
FBR launches Automated Currency Declaration System	Economic
Petrol prices remain unchanged	Oil & Gas
Cement supply declines by 16.58pc in January	Cement

3.1.4 News Filtration

During this phase, the main objective was to identify news stories that directly influence the movements of the stock market. This is because the news covers a wide range of topics, including sports and entertainment, which are not important to our use case. The main goal was to identify events in the world, national, and corporate sectors that have a substantial impact on the stock market. In order to accomplish this, a filtration procedure was created to eliminate unnecessary news articles by analysing the main-cat column of the dataset. The Table 3.7 depicts the rows where 'Others' was assigned as a main-cat to the News articles.

The news articles were filtered by examining the values in main-cat column. Articles that were categorised under topics other than international, national, and business were considered irrelevant to the research subject and so eliminated. As a result, only news items that had the potential to affect stock market movements were kept for examination. The data cleaning approach entailed methodically removing articles from non-essential categories, thereby filtering the dataset to comprise just the most pertinent news for our unique use case. The filtration stage played a vital role in improving the precision and significance of the subsequent research, guaranteeing that the attention stayed on news events that were likely to influence stock market movements.

Table 3.7: News Tagged as Others

Headline	main-cat
ICC must speak out against social injustice, insists Sammy	Others
Zindagi Tamasha becomes first movie to premiere on TikTok	Others

3.1.5 Comprehending Market Sentiment

Sentiment analysis plays a pivotal role in stock price prediction by capturing the emotional undercurrents and perceptions that influence market behavior. Understanding market sentiment provides valuable insights into investor attitudes, enabling more informed decision-making. In this context, sentiments serve as a lens through which one can gauge market reactions to events, news, and trends, thereby enhancing the predictive accuracy of stock price movements.

In pursuit of comprehending market sentiment, I employed a sophisticated approach using the FinBERT model⁶. FinBERT is a specialized Natural Language Processing (NLP) model, rooted in BERT's transformative architecture, tailored for financial text analysis. By fine-tuning FinBERT using a Financial News sentiment analysis dataset from Kaggle⁷, which features sentiments categorized as negative, neutral, or positive, I harnessed its capabilities for sentiment analysis within the financial domain. This dataset was selected to ensure that the model could accurately capture the linguistic and contextual nuances of financial news in Pakistan, thereby enhancing its relevance and accuracy in this specific domain.

The fine-tuning process involved several key steps. First, we prepared the dataset by ensuring it included appropriately labeled sentiments reflecting the financial context. This preparation was crucial for training the model to predict the sentiment of Pakistani financial news accurately. The ProsusAI/FinBERT model was then fine-tuned using the AdamW optimizer, known for its efficiency in handling large-scale data and its ability to improve the model's convergence rate. The learning rate was set to $2e-5$, a value chosen to balance the speed and accuracy of the model's training process. Table 3.8 depicts the Configuration parameters used for fine tuning our model.

The loss function used during the training was CrossEntropyLoss, which is well-suited for multi-class classification problems such as sentiment analysis. The batch size was set to 16, ensuring that the model could process a reasonable amount of data in each iteration without overwhelming the computational resources. The training process was conducted over four epochs, which was sufficient to achieve a robust model performance without overfitting.

The model was trained to classify news articles into three sentiment categories: positive, negative, and neutral. The fine-tuning process yielded an impressive F1 score of 0.88, indicating the model's high accuracy and reliability in sentiment classification.

The fine-tuned model was then applied to the filtered unlabelled dataset of news articles, providing sentiment scores that were crucial for subsequent analysis. Given the variability in description lengths, I tactfully segmented descriptions into individual sentences. For each sentence, the model generated sentiment predictions. To determine the overall sentiment of a news article, I adopted a majority voting mechanism. If the majority of sentence sentiments for a given description were positive, the overall sentiment was classified as positive. Similarly, if positive and negative sentence sentiments were balanced, the overall sentiment was assigned as negative.

This process facilitated the assignment of sentiment labels to previously unlabelled news ar-

⁶<https://huggingface.co/ProsusAI/finbert>

⁷<https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news>

Table 3.8: Configuration Parameters

Configuration Parameter	Value
Model	ProsusAI/FinBERT
Optimizer	AdamW
Learning Rate	2e-5
Loss Function	CrossEntropyLoss
Batch Size	16
Number of Labels	3
Epochs	4

ticles, adding a layer of nuanced understanding to the dataset. By harnessing the power of FinBERT and implementing a systematic sentiment aggregation strategy, I successfully transformed unstructured news data into quantifiable sentiment indicators. This approach enables a more informed analysis of market sentiment’s impact, contributing to a more comprehensive framework for anticipating stock market trend.

3.1.6 News Keyword Extraction

Keyword extraction is a crucial step in the data pipeline, aimed at identifying the most important terms and concepts within news articles. This process is enriched by the integration of various techniques such as TextRank, TF-IDF, KeyBERT, and spaCy, each of which contributes uniquely to the extraction of valuable keywords. The importance of keyword extraction lies in its ability to highlight key themes and topics from news articles, which are essential for correlating news with stock market trends.

TextRank, as a graph-based algorithm, delves into the text to reveal contextually relevant keywords by analyzing the relationships between words. This technique effectively uncovers the significance of terms within the document, making it valuable for understanding the contextual relevance of various news articles. TF-IDF, a classic yet effective method, evaluates the importance of words based on their frequency within the document relative to a broader corpus of texts. This approach helps in identifying distinctive keywords specific to each document, providing a comparative measure of term significance across multiple texts.

KeyBERT leverages transformer-based models, particularly BERT, to capture nuanced and context-

aware keywords. By using contextual embeddings, KeyBERT excels in modern NLP applications, making it an excellent choice for extracting keywords that reflect the subtleties and complexities of financial news. Although spaCy is not exclusively designed for keyword extraction, it contributes significantly by providing linguistic annotations, part-of-speech tagging, and named entity recognition. These features enhance text analysis, leading to more accurate keyword extraction by offering a deeper understanding of the text's structure and entities. Once

Table 3.9: News Articles with Keywords

Headline	URL	Keywords
Cement supply declines by 16.58pc in January	https://profit.pakistantoday.com.pk/2022/02/02/cement-supply-declines-by-16-58pc-in-january/	['domestic', 'tons', 'decrease', 'cement', 'mills', 'exports', 'period', 'cent', 'markets', 'south', 'supply', 'january', 'declines', 'mainly', 'million', 'months', 'year', 'despatches', 'north', 'local', 'total', 'fiscal', 'reduction', 'massive', 'month']
SBP is now 'State Bank of IMF': PML-N	https://dailytimes.com.pk/862450/sbp-is-now-state-bank-of-imf-pml-n/	['oil', 'supplies', 'marketing', 'company', 'profit', 'pso', 'receivables', 'utilities', 'gas', 'winter', 'major', 'current', 'producers', 'government', 'high', 'residential', 'domestic', 'account', 'year', 'lng', 'pakistan', 'power', 'billion', 'defaulters', 'consumers']

the keywords are extracted using these techniques, they are combined to form a comprehensive keyword list. This list is then streamlined by converting it into a set, eliminating duplicate keywords as shown in Table 3.9. This holistic approach ensures that the most relevant and distinctive financial themes are identified and categorized from the news articles. By highlighting key terms and concepts, keyword extraction provides valuable insights for subsequent analysis, enabling a more precise correlation between news events and stock market movements. This

step enhances the understanding of how specific news topics influence market trends, thereby contributing to more informed and effective market analysis.

3.1.7 News Named Entity Recognition

The extraction of named entities from news articles is a critical step in linking important entities with stock market movements. This process involves identifying and extracting entities such as people, organizations, and locations from textual news data, providing valuable context and insights into market-related events. For this study, the primary focus was on identifying names and organizations within the Pakistani context, as these entities are pivotal in assessing the impact of news on the stock market.

To accomplish this, various NER models, including NLTK, Spacy, Stanford Core NLP, Allen NLP, and Polyglot, were evaluated. However, Flair⁸, an advanced NLP library built on PyTorch, was found to be the most effective for this purpose. Flair is renowned for its exceptional performance in a wide range of NLP tasks, including Named Entity Recognition (NER), POS tagging, and text classification. It is user-friendly and supports state-of-the-art embeddings like BERT and ELMO, which significantly enhance its ability to capture contextual information and identify entities accurately.

Using Flair, the news articles were processed to extract entities categorized as persons, organizations, and others (such as locations and miscellaneous entities). This categorization facilitates a detailed analysis of how different types of entities mentioned in the news correlate with stock market movements. Flair's NER model was applied to the textual data, generating three columns in the dataset: NER_Persons, NER_Organizations, and NER_Others as shown in Table 3.10. These columns captured the specific named entity types within the news articles, enhancing the depth of analysis and understanding of the news content.

This systematic extraction and categorization of entities enable a more comprehensive assessment of the potential impacts of news events on the stock market, providing valuable insights into how specific entities mentioned in the news influence market movements.

⁸<https://huggingface.co/flair>

Table 3.10: News Articles with NER Information

Headline	Category	Source	NER_Persons	NER_Org	NER_Others
SBP is now 'State Bank of IMF': PML-N	Business	Daily Times	['miftah ismail', 'musaddik malik']	['SBP', 'PML-N']	['pakistan', 'islamabad']
Fawad Chaudhry admits need to 'reduce bitterness'	Pakistan	Dawn	['fawad chaudhry']	[]	['pakistani', 'pakistan', 'islamabad']

3.1.8 Embedding Generation for News Articles

To transform news articles into numerical representations, we utilized the RoBERTa-base model from the Hugging Face Transformers library⁹. The RoBERTa-base model is a robust transformer-based language model known for its excellence in capturing contextual information and semantic meaning from text. As a variant of BERT, RoBERTa enhances the training procedure by removing the next sentence prediction objective, which leads to improved performance in understanding the nuances of language. This specific model, RoBERTa-base, has been pretrained on extensive datasets including Wikipedia and BookCorpus, providing it with a comprehensive understanding of both general language patterns and nuanced literary structures.

The embedding generation process for this study involved several key steps. First, the text data contained in the 'CC_wo_Stem_Lema_Punc' column of the input DataFrame was prepared for processing. Text exceeding the model's maximum sequence length of 512 tokens was truncated to ensure compatibility. The RoBERTa tokenizer then converted the text into tokens suitable for model processing. This tokenized text was fed into the RoBERTa-base model, which processed these tokens and extracted the last hidden state output.

To obtain a fixed-size vector representation for each input, the last hidden state output was averaged along the sequence dimension. This step ensured that the resulting embedding vector captured the overall semantic information of the entire text. The resulting embedding vectors were

⁹https://huggingface.co/docs/transformers/en/model_doc/roberta

then stored in the 'Embedding_Vector' column of the DataFrame for further analysis. These embeddings are crucial for various downstream tasks such as similarity comparison, clustering, and any application requiring fixed-size vector representations of text. The fine-tuned RoBERTa-

Table 3.11: Configuration Parameters

Configuration Parameter	Value
Model	RoBERTa-base
Architecture	RobertaForMaskedLM
Attention Heads	12
Hidden Layers	12
Hidden Size	768
Maximum Position Embeddings	514
Training Data	Wikipedia, BookCorpus
Tokenizer	RoBERTa Tokenizer
Output	Last Hidden State
Embedding Vector Dimension	Averaged across sequence

base model, although not specifically adjusted for downstream tasks in this instance, leveraged its pretrained capabilities to generate high-quality embeddings for the news articles. This approach allowed us to utilize the robust language understanding inherent in the RoBERTa-base model to create accurate and reliable numerical representations of the news articles, facilitating deeper analysis and more precise insights into the impact of news on stock market movements. Our one processed Financial News dataset is generated here containing the following columns as shown in Table 3.12

3.1.9 Event-Stock Extraction and Mapping

The event extraction method is divided into two time periods: "1-Day Event Extraction" and "5-Day Event Extraction". The previous method included recognising occurrences in news items that have the same publication date and utilising the similarities between these articles on the same day as a financial event occurs. On the other hand, the "5-Day Event Extraction" approach use a longer time frame to extract events that occur across multiple days.

Let $D = [d_0, d_0 + 1, d_0 + 2, \dots, d_L]$ be a list of all the unique consecutive dates of the structured

Table 3.12: Processed_Financial_News_Dataset_Columns

Column Name
Headline
DateTime
Url
Source
Category
News Combined_Cleaned
Sentiment
NER_Persons
NER_Orga
NER_Other
Main_Cat
Keywords_CC
Sub_Cat

news dataset SN , where d_0 is the oldest date and d_L is the latest date.

Let $W = [w_1, w_2, \dots, w_n]$ be the list of 5 day rolling window sizes for D , where $w_1 = [d_0, d_0 + 1, d_0 + 2, d_0 + 3, d_0 + 4]$, $w_2 = [d_0 + 1, d_0 + 2, d_0 + 3, d_0 + 4, d_0 + 5]$, $w_n = [d_{L-4}, d_{L-3}, d_{L-2}, d_{L-1}, d_L]$, and n is the number of windows created based on the total number of consecutive dates in D .

$$n = |D| - (\text{Window Size}) + 1 \quad \text{Eq (1)}$$

where $|D|$ is the length of D and Window Size = 5. If $|D| = 31$ we will have 27 windows.

$\forall w \in W$ let $df(w)$ be the subset of SN where the dataset df contains the information of the news articles only for the window w .

$$df(w) = SN('Date' == w) \quad \text{Eq (2)}$$

Let $E(df)$ define the function for event extraction for the dataset df .

$$SE_2 = \sum_{i=w_1}^{w_n} E(df(i)) \quad \text{Eq (3)}$$

where SE_2 is the structured event dataset for 5-Day Event Extraction.

Semantic Similarity Calculation

The initial technique entails computing the cosine similarity score for the embeddings of the news items in the dataset, which allows for the assessment of their semantic similarity. Cosine similarity quantifies the cosine of the angle θ formed by two nonzero vectors in a multi-dimensional space. The mathematical expression for cosine similarity between two vectors A and B can be represented by Equation 4 as follows:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \|B\|} \quad \text{Eq (4)}$$

where: $A \cdot B$ represents the dot product of vectors A and B. $\|A\|$ represents the magnitude (Euclidean norm) of vector A. $\|B\|$ represents the magnitude (Euclidean norm) of vector B. Cosine similarity values increase as the similarity between vectors increases. A value of 1 indicates that the vectors are identical, while a value of 0 indicates that the vectors are orthogonal.

The second mechanism includes using the intersection rate approach to measure the degree of overlap in content between articles. The rate is calculated by comparing the number of words that are common to both articles, compared to the total number of words in the smaller article. The mathematical equation for determining the intersection rate can be expressed as Equation 5:

$$\text{Intersection Rate} = \frac{|w_1 \cap w_2|}{|w_{L_S}|} \quad \text{Eq (5)}$$

where:

$$w_{L_S} = \min(w_1, w_2)$$

w_1 and w_2 are the lists of words of two articles and $|w_1 \cap w_2|$ represents the cardinality (number of elements) in the intersection of the two lists, which is the count of common elements between them. Whereas $|w_{L_S}|$ represents the cardinality of the smaller of the two lists, ensuring that the intersection rate is calculated relative to the size of the smaller list.

Threshold optimization for similarity score

A combined similarity score is calculated by considering the relative significance of the intersection rate and the cosine similarity. Multiple trials are conducted to determine the most effective threshold for the ultimate similarity score used in event assignment. The optimised threshold

condition is met when the similarity score between any two news articles is more than 0.8 and the intersection rate exceeds 0.5. Ultimately, the articles that meet these criteria are accurately recognised as belonging to the same event. Equation 6 represents the optimised threshold values for each individual.

$$\text{Similarity Score} = (\text{Cosine Similarity} \times 0.8) + (\text{Intersection Rate} \times 0.2) \quad \text{Eq (6)}$$

Events feature extraction

The event assignment method is further improved to enhance event selection in the production of the events timeline. This entails incorporating the inheritance of features from the group of news items that are important to the events, in order to construct the contextual basis of the identified occurrences. In addition, to provide in-depth analysis of the connected news events in the stock market and its sectors, the selected events are mapped on a timeline and categorised based on the most common category found among the articles. The factors determining this classification include the Majority Category, Priority Category, and their respective levels of purity. The Majority Category represents the category that encompasses the most number of news items related to a specific occurrence. The event category is determined based on priority levels allocated to each category, with the following category priorities: Business (1), Pakistan (2), and World (3). In addition, purity calculations provide a measurable measure that enables us to evaluate the level of dominance displayed by the chosen category in a certain event. A purity score of 1.0 signifies that the category completely dominates event-related news. In our instance, a threshold of ≥ 0.5 is deemed appropriate to ensure a more accurate depiction of the category in event-related news stories. This technique not only facilitates comprehension of the prevailing topics within the news dataset but also offers valuable perspectives from a variety of sources. Therefore, the comprehensive approach enables us to explore the complex connection between news content, event extraction, and how they can potentially affect stock sectors. The research pipeline we offer allows for the visualisation of the relationship between news and stock data by utilising the persistent features that we have calculated and preserved during the analysis.

3.2 Module2- Correlating Event Impact with Stock Trend Movement

To cater the unique dynamics of a Pakistan’s volatile market, we have proposed a novel Petro-Market Insight Framework (PMIF). It identifies and synthesizes news-based events to produce actionable trading recommendations for the Pakistani stock market. Specifically, we focused on the Petroleum sector, which dominates the market’s capitalization and trading activity. The PMIF framework consists of three sub-modules as depicted in Figure 3.2.

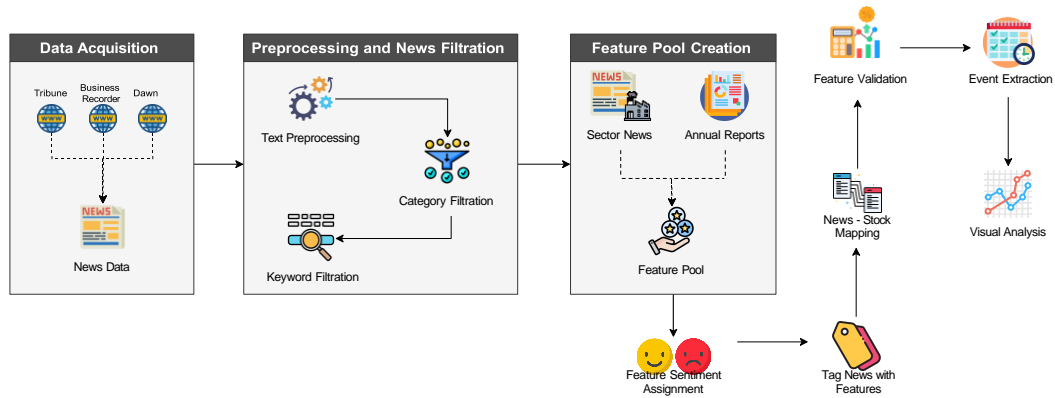


Figure 3.2: Methodology flowchart from data acquisition to visual analysis.

3.2.1 News-Feature Formulation and Tagging Module

This module is tasked with the preparation of Petroleum sector-specific features along with their sentiments and tagging them with our filtered news dataset obtained from the previous module. A dedicated feature pool for our sector is formulated by assessing the sector-specific news and the Annual Reports issued by the companies representing the Petroleum sector. A Retrieval Augmented Generation (RAG) based approach utilizing Large Language Models (LLM) was employed to query the Annual Reports for feature extraction specific to the sector. RAG is an approach in NLP that merges the advantages of retrieval-based and generative-based AI models. Subsequently, the filtered news dataset was tagged with features within the feature pool using a hybrid approach. The approach utilized the Exact Match and Fuzzywuzzy algorithms with a threshold of 0.9. Fuzzywuzzy is a string-matching library that works based on the principle of Levenshtein distance. This process facilitated the retention of a granular set of news relevant to our sector and also served as a foundation for evaluating the impact of these features on trading

recommendations.

3.2.2 Feature Validation, Event Extraction and Visualization Module

A two-staged approach of feature validation and reduction to extract news-based events that have a significant impact on the stock movement of the petroleum sector is used. Recent news from the filtered news corpus were mapped to sector data (i.e., change percentage value and close prices) where each feature was assigned the change percentage value to represent the impact of the news on the sector. The impact value was determined by comparing each feature's sentiment with the change percentage polarity. An aggregated feature impact value for each feature based on its frequency is then calculated. This incorporates the information regarding feature representation along the market trend. Finally, the impact of each feature is scaled using Equation 1 and Equation 2 based on its representation in either of the positive or negative subset of the feature pool.

For positive feature scaling:

$$F_{s+} = \frac{f_{imp+}}{H_{imp+}} \times 5 + \frac{f_{occ+}}{H_{occ+}} \times 5 \quad (3.2.1)$$

For negative feature scaling:

$$F_{s-} = \frac{f_{imp-}}{H_{imp-}} \times 5 + \frac{f_{occ-}}{H_{occ-}} \times 5 \times (-1) \quad (3.2.2)$$

where:

- H_{occ-} and H_{occ+} are the maximum negative and positive feature occurrences, respectively.
- H_{imp-} and H_{imp+} are the highest negative and positive impact percentages, respectively.

This scaling mechanism takes into account not only the occurrence of the feature in the news dataset but also the highest percentage of impact. To evaluate feature accuracy within the respective subset of the pool following rules were formulated:

$$C(F_s) = \begin{cases} 1 & \text{if } F_s = F_s^+ \text{ and } \Delta P_t > 0 \\ 1 & \text{if } F_s = F_s^- \text{ and } \Delta P_t < 0 \\ 0 & \text{otherwise} \end{cases}$$

where:

CHAPTER 3: METHODOLOGY

- $C(F_s)$ is the correctness of the feature F_s .
- ΔP_t is the change percentage value at time t .

Features with accuracy below 60% were dropped and news-based events were extracted with the qualified features. An event impact on the sector's trend is calculated by averaging the scaling values of features representing the recent events, which occurred in close proximity to the trend. The events generated were then visualized on the stock trend of the petroleum sectors for providing trading recommendations to the active traders. The events generated were then visualized on the stock trend of the petroleum sectors to provide trading recommendations to active traders.

Results and Discussion

This Results and discussion section has been divided into two major sections. The initial segment presents the findings of Module 1, in which events are collected from news items and subsequently mapped to the corresponding stock trend. The second section consolidates the outcomes of the module, where a correlation has been established between news events and the trend of the stock market in order to provide informed trading recommendations.

4.1 Module-1

In our experimentation, optimization, and fine-tuning of the models involved conducting necessary experiments on an RTX 3060 GPU with a memory size of 12GB. For news scrapping, Python libraries including Selenium, Requests, and BeautifulSoup are utilized. The English news dataset consists of **414,578** news articles. Analyzing the distribution of these articles among different sources from the year 2020 to 2023 reveals significant variations. The largest contributor is Business Recorder, which accounts for a substantial 198,944 articles. Tribune follows with 80,262 articles, indicating its prominent role in news reporting. Daily Times and Dawn contribute 59,659 and 53,659 articles, respectively, showcasing their considerable presence in the news landscape. Pakistan Today has the smallest share, with 22,054 articles. The combined total of these sources sums up to 414,578 articles, illustrating the extensive coverage and diversity of the news sources in the dataset. The Figure 4.1 visually emphasizes the dominance of Business Recorder and the notable contributions from other sources, providing a clear overview of the news article distribution across these platforms.

Due to the complexity and scale of the data, the news articles were split into monthly datasets.

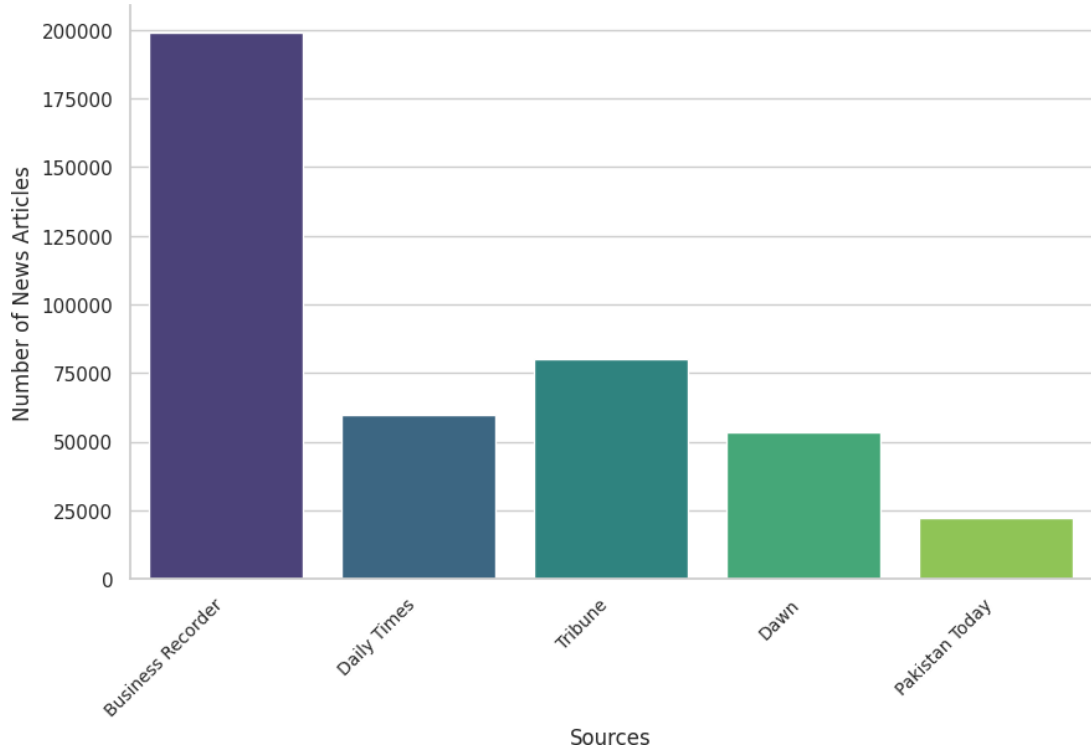


Figure 4.1: Distribution of News Articles by Source from 2020 to 2023

Task	Model	Hyperparameters	F1 Score
Sentiment Analysis	ProsusAI/finbert	optimizer: AdamW learning rate: 2e-5 loss function: CrossEntropyLoss batch size: 16 num_of_labels: 3 epochs: 4	0.88
Embeddings	RoBERTa-base	max length: 512 output vector length: 768	-

Table 4.1: Specifics of the experimental environment

Each monthly dataset contains aggregated news from all sources, facilitating efficient processing for subsequent tasks in the proposed pipeline. Various NLP tasks are executed using the Python version (3.10.12) and advanced libraries built on Keras and PyTorch. Table 4.1 outlines the specifics of the experimental settings utilized in fine-tuning the models for essential NLP tasks along with the evaluation metric where applicable.

The visualization results of the news embeddings demonstrate the capacity and versatility of the RoBERTa model in encoding rich contextual information for both languages. Figure 4.2 represents the effective semantic representation of the news embeddings concerning the categories from our dataset for May 2023. The proximity of the embeddings within their respective color sections highlights the model’s effectiveness in capturing semantic similarities and grouping similar articles. The prevalence of news distribution in ‘Pakistan’ as a main category is prominent in the t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization, suggesting that a significant number of events captured in the news articles are centered around or impact Pakistan. This insight is valuable for understanding the regional dynamics in Pakistan, providing a comprehensive overview of the news landscape.

The labeling of the financial news dataset, encompassing subcategories from economic and

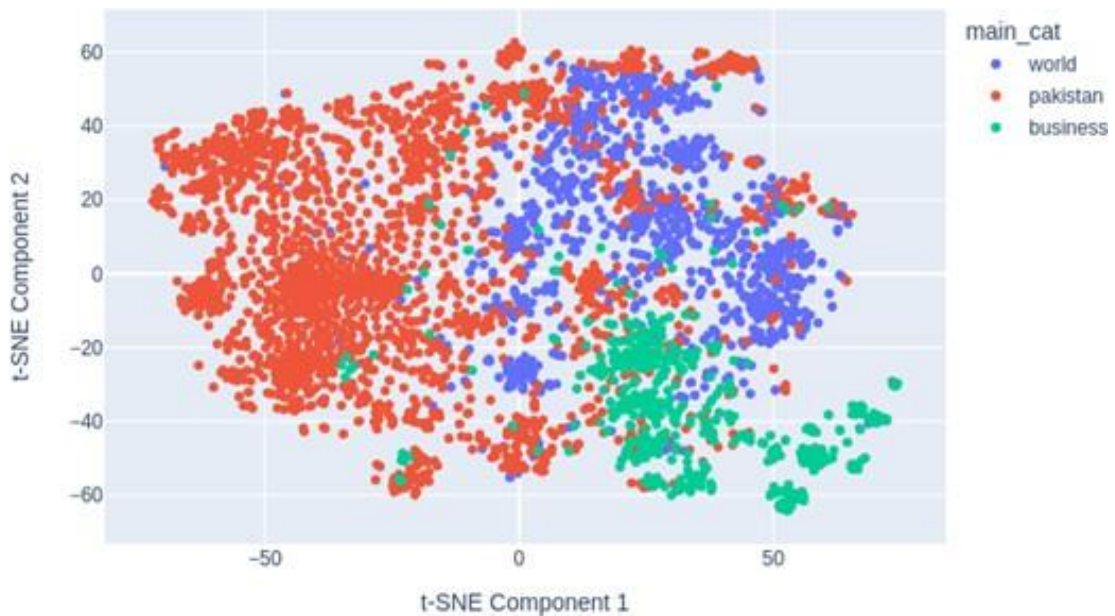


Figure 4.2: t-SNE visualization of RoBERTa embeddings for news categories clusters

stock market sectors, provides useful insights into the sources. Each news event is meticulously cataloged, detailing its headline, sources, URLs, the number of related articles and other things as shown in Figure 4.3. The correlation of the news events with the dynamics of the Pakistan

Event	Sources	Urls	Article Start Date	End Date	Headlines Description	Categories	Sub_Categor	Major_Cat	Priority_cat	Priority_Pri	Maj_Purit	Sentiment	Keywords	Similarity	NER_Pers	
Oil edges higher or	['Tribune']	['https://']	5	2/1/2022	2/1/2022	['Oil edge: LONDON:']	['Business', []]	Business	Business	Business	1	1	2	['europe', [0.904003	['putin', 'ji	
Pakistan raises \$1b	['Dawn',]	['https://']	3	2/2/2022	2/2/2022	['Pakistan KARACHI:']	['Business', ['Economic']]	Business	Business	Business	1	1	2	['europe', [0.886326	['asad riz	
PM's China visit	['hi']	['Daily Tin']	2	2/2/2022	2/2/2022	['PM's Chi Interior M']	['National', []]	National	National	National	1	1	1	['saud', 'hi']	[0.909708	['rashid h.
Industries losing R:	['Dawn',]	['https://']	2	2/3/2022	2/3/2022	['Industrie KARACHI:']	['Business', ['Oil & Gas']]	Business	Business	Business	0.5	0.5	0	['chairmar']	[0.936334	['muhamr
Attack Group of Cc	['Business']	['https://']	5	2/12/2022	2/12/2022	['Attock G TEXT:']	['Business', ['Oil & Gas']]	Business	Business	Business	1	1	2	['pioneer', [0.896935	['adikhatt	
Wait-and-see moo	['Daily Tin']	['https://']	4	4/6/2022	4/6/2022	['Wait-anc Pakistans']	['Business', ['Cement', 'E Business']]	Business	Business	Business	1	1	1	['cotton', '']	[0.890874	['rogers vi

Figure 4.3: A snapshot of Event Dataset

Stock Exchange is shown in Table 4.2. A prevalent focus on the economic sector in most articles is observed as compared to specific stock industries. Extracting and filtering events from each labeled sector for representation in the financial event timeline reveals that the business community in Pakistan exhibits a notable interest in English news sources. It can be inferred from the statistics that this preference may stem from the perceived quality and trustworthiness associated with such sources.

Table 4.2: News and Events Statistics

Sector	No. of news articles	No. of events extracted
Oil & Gas	45024	12512
Cement	2455	979
Economic	80453	18059

Furthermore, the analysis of news events related to the stock market has been enhanced through Named Entity Recognition (NER) and sector-specific list tagging to extract important person and sector-specific entities. The word clouds below for the Cement, Oil & Gas, and Economic sectors events provide a visual summary of the most frequently mentioned terms within these sectors, revealing key topics and trends that are likely to influence stock performance. The Ce-

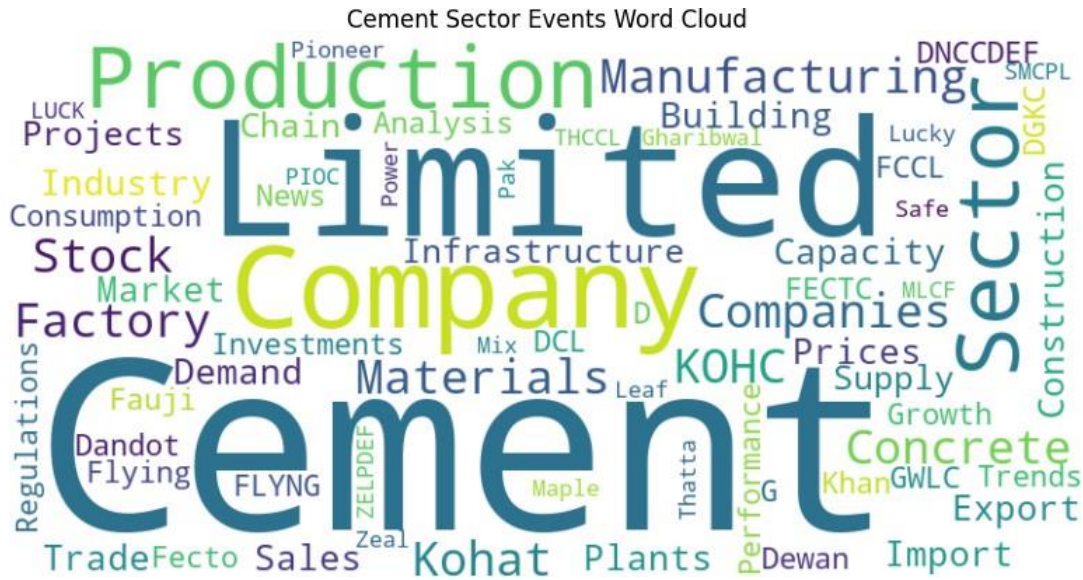


Figure 4.4: Word Cloud of Frequently Mentioned Terms in Cement Sector Events (2020-2023)

ment Sector Events Word Cloud highlights as shown in Figure 4.4 significant terms such as "Cement", "Limited", "Company", "Production", and "Sector", which dominate the discussion. Other notable terms include "Manufacturing", "Factory", "Prices", and "Infrastructure".

This word cloud encapsulates various aspects of the cement industry, including production levels, company activities, market dynamics, and infrastructural developments. By understanding the frequency and context of these terms, investors can gain insights into the factors driving stock price movements within the cement sector. The Oil & Gas Sector Events Word Cloud as

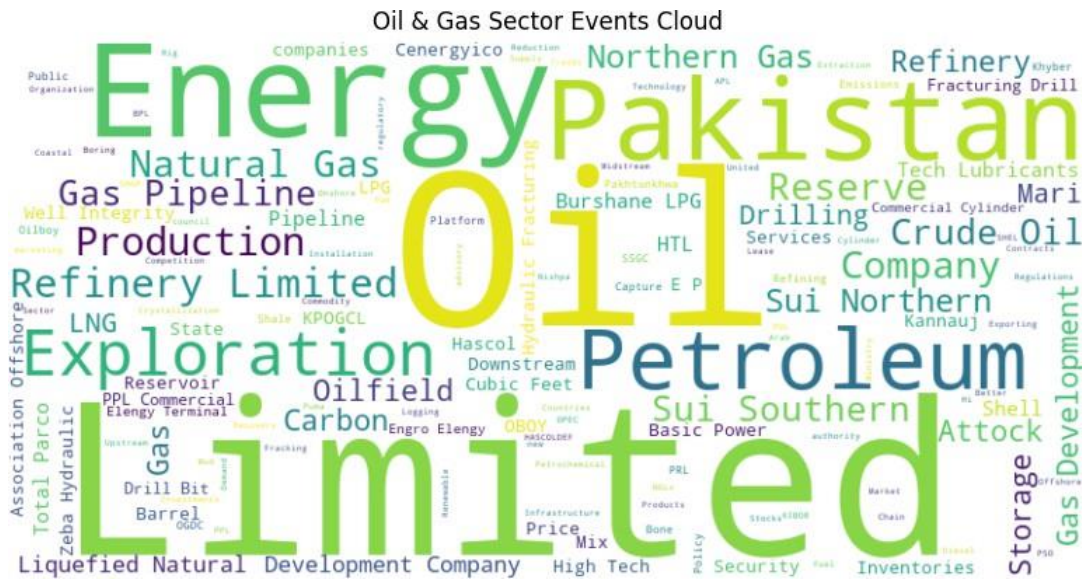


Figure 4.5: Word Cloud of Frequently Mentioned Terms in Oil Gas Sector Events (2020-2023)

shown in Figure 4.5 prominently features terms like "Oil", "Energy", "Pakistan", "Exploration", "Limited", and "Petroleum". Key concepts such as "Natural Gas", "Pipeline", "Refinery", and "Production" also appear frequently. This visualization sheds light on the critical elements of the oil and gas industry, including energy production, exploration activities, infrastructure, and market trends. Recognizing the importance of these terms helps in understanding the sector's impact on stock performance and identifying potential investment opportunities. The Economic Sector Events Word Cloud as shown in Figure 4.6 displays significant terms like "Pakistan", "Economic", "National", "Bank", and "Investment". Other essential terms include "Policy", "Market", "Trade", "Revenue", and "International". This word cloud captures the broad range of topics that influence the economic landscape, such as fiscal policies, market conditions, trade activities, and financial regulations. By analyzing the frequency and context of these terms, stakeholders can better understand the economic factors affecting stock market trends and make informed investment decisions. The word clouds serve as a valuable tool for correlating news coverage with stock market fluctuations. By identifying the most frequently mentioned entities and terms within each sector, investors and analysts can pinpoint the news events and trends that are likely to impact stock prices. For instance, frequent mentions of production changes, regula-

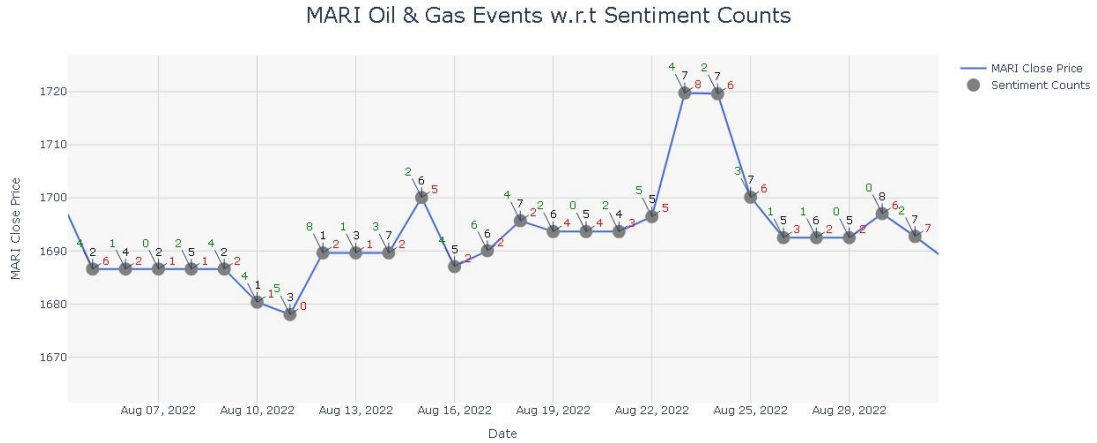


Figure 4.9: Financial events trend with respective sentiment count on Oil & Gas sector

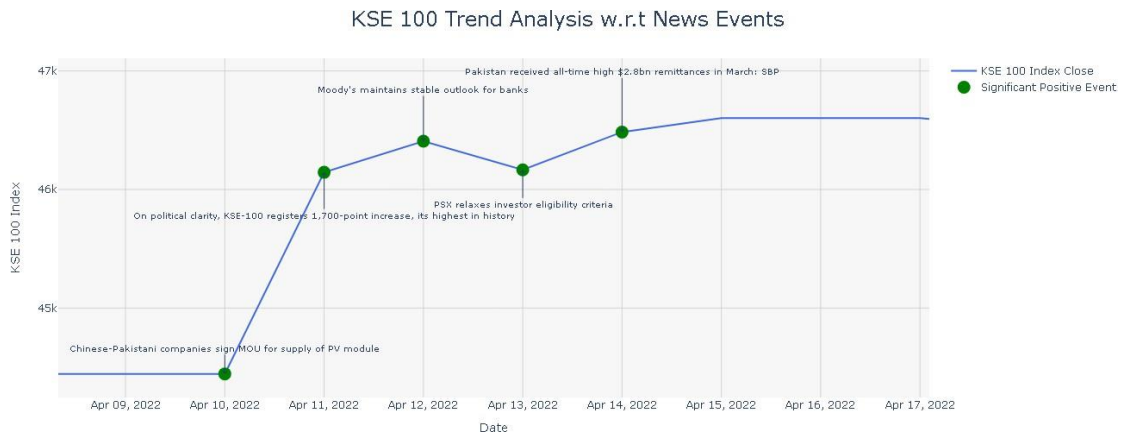


Figure 4.10: . Visualization of key news events timeline aligned with KSE 100 bullish trend

investment news across several business sectors, relaxation in eligibility conditions for stock investors, and an upsurge in foreign remittance. These variables contributed to a substantial surge in stock momentum, leading to a favourable market trend. Contrarily, the prevailing sentiment of most events, as seen in Figure 4.11, is negative, which is causing a significant decline in the KSE 100 index. The trend is characterised by notable occurrences such as a substantial fiscal deficit, a decrease in State Bank reserves, a rise in interest rates, and challenges in reaching settlements with the IMF for loan packages. The rapid succession of these events within a few timeframe had a significant adverse effect on the future trajectory of the stock market. Our suggested model has the capability to identify important events and their connection to stock movement. This allows us to highlight crucial information and gain valuable insights from large amounts of unstructured business-related news data.

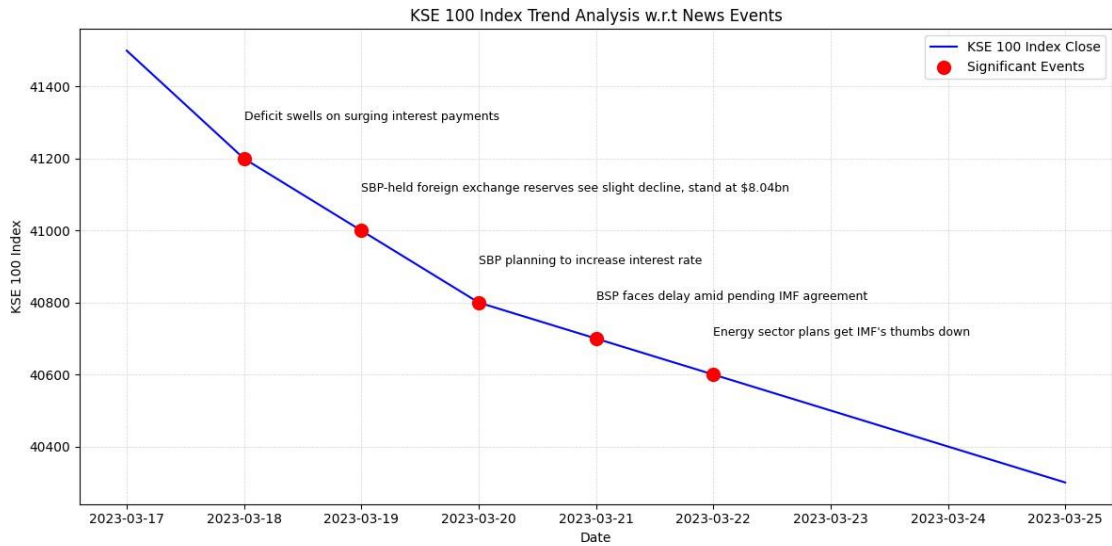


Figure 4.11: Financial events trend with respective sentiment count on Oil & Gas sector

4.2 Module-2

The evaluation of the acquired news articles produced several significant observations concerning the influence and dependability of different features, as well as the distribution of themes and sentiments throughout the dataset. The studies were conducted using Python 3.12.0 and utilised an RTX 3060 GPU with 12GB of memory. We employed Python modules such as Selenium, Requests, and BeautifulSoup for the purpose of news scraping. By leveraging the power of GPU acceleration, we successfully performed embedding calculations and visualisations, enabling us to process the data fast and create complete visual tools that unveil significant patterns and trends.

Table 4.3: News Articles by Standard Categories

STANDARDIZED CATEGORY	COUNT
Business	1,460
International	33
Pakistan	20
Other	46

Our extensive financial news data on the PSX includes a wide variety of articles pertaining to the Petroleum industry. The dataset has been classified into different standardised categories, showcasing the frequency and distribution of articles within each category, as shown in Table 4.3.

The articles are distributed throughout multiple areas, with a significant focus on the "Business" and "International" sections. The Business category news largely focuses on giving coverage of news related to the domain, while also incorporating Pakistan-specific news that takes into consideration local challenges that impact the oil sector. Moreover, international news offers a holistic outlook by incorporating other viewpoints from the area of news known as the Other. This enables us to precisely identify and incorporate crucial elements that have a substantial impact on the Petroleum business. The provided context is crucial for comprehending the feature analysis, since it illustrates the wider thematic scope of the news stories.

After analysing three years of news data, which included a total of 656 events, we found that there was a ratio of 1:2 between positive and negative events. In order to gain a deeper understanding of the data, we utilised the t-distributed Stochastic Neighbour Embedding (t-SNE) method to visually represent complex news embeddings in a two-dimensional plane. By leveraging GPU acceleration, we achieved efficient processing and visualisation of the embeddings. Each data point in the plot, depicted in Figure 4.12, corresponds to a news story and is visually differentiated by its standardised feature. The visualisation displays clear clusters that correspond to different features, such as "rising oil prices" and "oil prices fell". The presence of these clusters suggests that articles with common characteristics are arranged in close proximity, showcasing the framework's proficiency in accurately capturing the semantic resemblances among news items. The aggregation of news stories based on their characteristics in Figure 4.12 implies that market responses are affected by comparable categories of news. It is worth mentioning that unfavourable factors such as "oil prices fell" often gather closely together, suggesting a persistent market sentiment and the possibility of affecting stock price developments. These visualisations enhanced comprehension of the data by revealing intricate links and patterns.

Feature mapping of these events to the stock's sector price reveals that International events like Brent oil price fluctuations (e.g., falls due to global oversupply) and OPEC's (Organization of the Petroleum Exporting Countries) decisions (e.g., decrease in oil production) dominate with unfavorable news for the Petroleum sector of Pakistan ¹. Conversely, positive news encompassed occurrences such as like supportive government policies, new oil reserve discoveries by OGDCL (Oil and Gas Development Company Limited), and Brent oil price rises due to global demand ². It has been observed that the sentiment associated with these events has a substantial impact

¹<https://www.dawn.com/news/1728612/gas-reserves-discovered-in-di-khan>

²<https://tribune.com.pk/story/2211584/oil-slips-26-weak-demand-supply-glut-weigh>

on stock price movements. Positive events tend to increase prices, while negative events tend to cause losses. This correlation emphasises the crucial significance of event sentiment in offering trading recommendations to the stakeholders who require to be aware of changing market conditions and feelings. In order to determine the primary elements that impact trading recom-

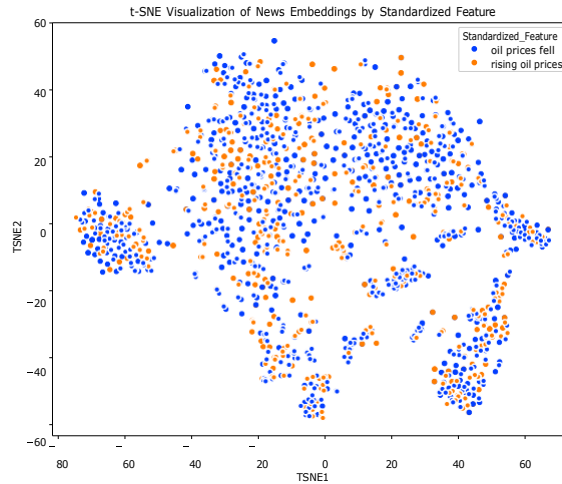


Figure 4.12: t-SNE Visualization of News Embeddings by Standardized Feature

mendations, we conduct an analysis of the top five features extracted from the events dataset to obtain a thorough comprehension. Figure 4.13 depicts the assessment criteria, highlighting the notable occurrence of events related to the oil prices fell. This occurrence has a substantial influence on the trading recommendations scale. Nevertheless, these events exhibit only a moderate level of precision and impact. On the other hand, the phrase "oil prices jumped" clearly indicates a substantial and accurate effect, implying its strong reliability when describing the trend in the petroleum industry. The prioritisation of features is determined by these insights, which encompass their reliability, impact, and frequency in recommendations.

In addition, we conducted an analysis on how particular news features affect stock prices by utilising line graphs. Figures 4.14 and 4.15 depict the graphical representation of events affecting the sector's overall direction and the corresponding trading suggestions. The latest news-driven events extracted using our technique effectively capture both bullish and bearish patterns. The production cuts announced by OPEC+ resulted in a substantial rise in oil prices, as indicated by the extracted characteristic of "rising oil prices" and the subsequent upward trend in the petroleum industry, as shown in Figure 4.14. However, the worries about tightening monetary policies and excessive crude oil supplies resulted in a negative event feature extraction, specifically the decline in oil prices. This accurately represented the subsequent downward movement

in the price trend of the petroleum sector, as shown in Figure 4.15.

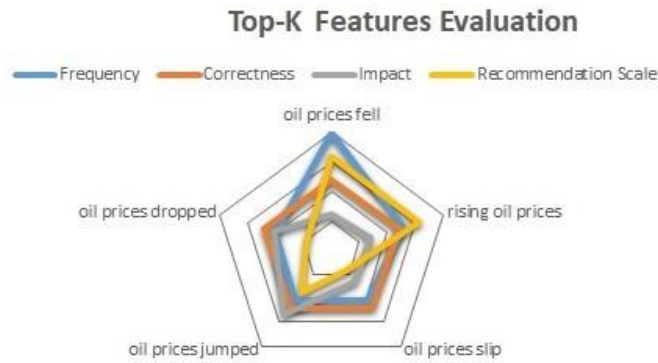


Figure 4.13: Comparative Analysis of Top Features in the Petroleum Sector

These examples showcase the resilience and flexibility of our methodology in extracting events and effectively responding to a wide range of market conditions. As a result, market participants in the petroleum sector are given the ability to make better-informed choices. The model’s accuracy is 61% when utilising the top-K features derived from the events data. This suggests that the chosen features possess the capacity to identify the patterns that impact the trajectory of the petroleum sector inside the intricate dynamics of the PSX. Future developments could prioritise investigating complex relationships among characteristics using advanced modelling approaches to increase suggestions for market participants.



Figure 4.14: Impact of “Rising Oil Prices” on OGDC Stock Price



Figure 4.15: Impact of "Oil Prices Fell" on OGDC Stock Price.

CHAPTER 5

Conclusion

This study presented a methodical Natural Language Processing for Financial Forecasting (NLFF) pipeline that utilises important indicators in the domain. The markers utilise semantic similarity, sentiment analysis, and named entity identification to create a chronology of financial events taken from news sources. We have compiled a well-organized dataset of financial news in English, taking into account the economic situation in Pakistan. By utilising news article embeddings and implementing a strategy to optimise the similarity threshold, we were able to effectively identify and extract events that are potential candidates for generating a financial timeline. Financial timeline visualisations were created to forecast the trajectory of the stock market and aid investors in making informed decisions about future stock purchases. We have achieved a considerable level of correlation between stock price fluctuations and our news sentiments.

Furthermore, the actions of investors in the shallow-plate market in Pakistan are directly impacted by publicly accessible information that has an effect on the overall investment climate. The primary achievement of this study was the automated identification and synthesis of events using natural language processing (NLP) to emphasise important occurrences that can be used to offer knowledgeable trading suggestions to market participants. This study entailed the aggregation of substantial volumes of financial news, with a specific emphasis on the PSX. One of its major contributions was to identify crucial elements exclusive to the PSX (Pakistan Stock Exchange) by considering the unique market dynamics in Pakistan. These identified traits were then validated with specialists. By correlating the events with the company's sector data, it becomes apparent that news factors have a substantial influence on inducing volatility in the stock market.

Recommendations

There are other study directions that can be explored as the future of this research endeavour. It is clear that stock prices are not affected by all news connected to business. Therefore, it would be worthwhile to investigate and identify the more significant events that do have an impact on stock prices. An avenue worth exploring is to examine the correlation between news sentiments and stock prices by analysing the potential influence of positive or negative sentiments in news items and events. Moreover, there is potential for enhancement in English sentiment analysis, particularly in accurately differentiating negatives that are often incorrectly categorised as neutral. Accurate event sentiment assignment and effective stock trend analysis from English news sources heavily rely on this.

Utilising a knowledge graph can be a valuable approach to analyse the correlation between stock entities and events in order to predict future stock trends. Knowledge graphs enable financial analysts to effectively identify patterns and trends within intricate datasets, hence offering valuable data insights and capturing crucial information. Moreover, in the majority of cases, there was a noticeable decrease in the proportion of news events related to specific stock sectors compared to economic developments. This provides opportunity for additional investigation into analysing sector-specific patterns for distinct possibilities within particular industries. There are other potential measures that can be implemented to enhance and broaden the scope of this study. To improve the precision of trading suggestions, it is feasible to integrate advanced Natural Language Processing (NLP) techniques, such as Large Language Models (LLMs). In addition, social media platforms and expert opinions can reveal different perspectives on market sentiment and improve understanding of market dynamics. Considering the effects of several external factors on news sentiment, such as economic indicators, geopolitical events, and world-

wide market patterns as well as incorporating market technical indicators, can provide a more thorough understanding of the intricate factors that influence stock market behaviour.

Our objective is to develop a sophisticated financial forecasting system that combines technical indications obtained from previous stock data with a wide array of economic factors. This approach will not only enhance the overall forecasting outcomes but also enable a comprehensive assessment of the financial event timeline produced by our study. Consequently, this allows for the prediction of market trends and the creation of customised investment strategies tailored to the tastes of individual investors.

Bibliography

- [1] Hai Leong Chieu and Hwee Tou Ng. “Named entity recognition: a maximum entropy approach using global information”. In: *COLING 2002: The 19th International Conference on Computational Linguistics*. 2002.
- [2] Paul McNamee and James Mayfield. “Entity extraction without language-specific resources”. In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002.
- [3] Oliver Bender, Franz Josef Och, and Hermann Ney. “Maximum entropy models for named entity recognition”. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. 2003, pp. 148–151.
- [4] Andrew McCallum and Wei Li. “Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons”. In: (2003).
- [5] Burr Settles. “Biomedical named entity recognition using conditional random fields and rich feature sets”. In: *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*. 2004, pp. 107–110.
- [6] Katharina Kaiser and Silvia Miksch. “Information extraction”. In: *A survey, Institute of Software Technology & Interactive Systems, Vienna University of Technology* (2005).
- [7] Vijay Krishnan and Christopher D Manning. “An effective two-stage model for exploiting non-local dependencies in named entity recognition”. In: *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*. 2006, pp. 1121–1128.
- [8] Ali Harb et al. “Web Opinion Mining: How to extract opinions from blogs?” In: *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. 2008, pp. 211–217.

BIBLIOGRAPHY

- [9] Xiaohua Liu et al. “Recognizing named entities in tweets”. In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011, pp. 359–367.
- [10] Alan Ritter, Sam Clark, Oren Etzioni, et al. “Named entity recognition in tweets: an experimental study”. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011, pp. 1524–1534.
- [11] Tim Rocktäschel, Michael Weidlich, and Ulf Leser. “ChemSpot: a hybrid system for chemical named entity recognition”. In: *Bioinformatics* 28.12 (2012), pp. 1633–1640.
- [12] Xiao Ding et al. “Using structured events to predict stock price movement: An empirical investigation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1415–1425.
- [13] Slobodan Beliga, Ana Meštrović, and Sanda Martincić-Ipšić. “An overview of graph-based keyword extraction methods and approaches”. In: *Journal of information and organizational sciences* 39.1 (2015), pp. 1–20.
- [14] Rebecca J Passonneau, Tifara Ramelson, and Boyi Xie. “Named Entity Recognition from Financial Press Releases”. In: *Knowledge Discovery, Knowledge Engineering and Knowledge Management: 6th International Joint Conference, IC3K 2014, Rome, Italy, October 21-24, 2014, Revised Selected Papers 6*. Springer. 2015, pp. 240–254.
- [15] Muhammad Azhari and Yogan Jaya Kumar. “Improving text summarization using neuro-fuzzy approach”. In: *Journal of Information and telecommunication* 1.4 (2017), pp. 367–379.
- [16] Thomas Puschmann. “Fintech”. In: *Business & Information Systems Engineering* 59 (2017), pp. 69–76.
- [17] Emma Strubell et al. “Fast and accurate entity recognition with iterated dilated convolutions”. In: *arXiv preprint arXiv:1702.02098* (2017).
- [18] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [19] Adam Atkins, Mahesan Niranjan, and Enrico Gerding. “Financial news predicts stock market volatility better than close price”. In: *The Journal of Finance and Data Science* 4.2 (2018), pp. 120–137.

BIBLIOGRAPHY

- [20] Abbas Ghaddar and Philippe Langlais. “Robust lexical features for improved neural network named-entity recognition”. In: *arXiv preprint arXiv:1806.03489* (2018).
- [21] Swapna Gottipati, Venky Shankararaman, and Jeff Rongsheng Lin. “Text analytics approach to extract course improvement suggestions from students’ feedback”. In: *Research and Practice in Technology Enhanced Learning* 13 (2018), pp. 1–19.
- [22] Gilles Jacobs, Els Lefever, and Véronique Hoste. “Economic event detection in company-specific news text”. In: *1st Workshop on Economics and Natural Language Processing (ECONLP) at Meeting of the Association-for-Computational-Linguistics (ACL)*. Association for Computational Linguistics (ACL). 2018, pp. 1–10.
- [23] Changzhou Li et al. “LDA meets Word2Vec: a novel model for academic abstract clustering”. In: *Companion proceedings of the the web conference 2018*. 2018, pp. 1699–1706.
- [24] Adeel Mumtaz, Tahir Saeed, and M Ramzan. “Factors affecting investment decision-making in Pakistan stock exchange”. In: *International Journal of Financial Engineering* 5.04 (2018), p. 1850033.
- [25] Pisut Oncharoen and Peerapon Vateekul. “Deep learning for stock market prediction using event embedding and technical indicators”. In: *2018 5th international conference on advanced informatics: concept theory and applications (ICAICTA)*. IEEE. 2018, pp. 19–24.
- [26] Hang Yang et al. “Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data”. In: *Proceedings of ACL 2018, System Demonstrations*. 2018, pp. 50–55.
- [27] Syed Aliya Zahera and Rohit Bansal. “Do investors exhibit behavioral biases in investment decision making? A systematic review”. In: *Qualitative Research in Financial Markets* 10.2 (2018), pp. 210–251.
- [28] Manish Agrawal, Asif Ullah Khan, and Piyush Kumar Shukla. “Stock price prediction using technical indicators: a predictive model using optimal deep learning”. In: *Learning* 6.2 (2019), p. 7.
- [29] Erdenebileg Batbaatar and Keun Ho Ryu. “Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach”. In: *International journal of environmental research and public health* 16.19 (2019), p. 3628.

BIBLIOGRAPHY

- [30] Xiao Huang et al. “Learning a unified named entity tagger from multiple partially annotated corpora for efficient adaptation”. In: *arXiv preprint arXiv:1909.11535* (2019).
- [31] Shaohua Jiang et al. “A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition”. In: *2019 12th international conference on intelligent computation technology and automation (ICICTA)*. IEEE. 2019, pp. 166–169.
- [32] Zhanming Jie and Wei Lu. “Dependency-guided LSTM-CRF for named entity recognition”. In: *arXiv preprint arXiv:1909.10148* (2019).
- [33] Jintao Liu et al. “Transformer-based capsule network for stock movement prediction”. In: *Proceedings of the first workshop on financial technology and natural language processing*. 2019, pp. 66–73.
- [34] Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. “Towards improving neural named entity recognition with gazetteers”. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 5301–5307.
- [35] Shahmin Sharafat, Zara Nasar, and Syed Waqar Jaffry. “Legal data mining from civil judgments”. In: *Intelligent Technologies and Applications: First International Conference, INTAP 2018, Bahawalpur, Pakistan, October 23-25, 2018, Revised Selected Papers 1*. Springer. 2019, pp. 426–436.
- [36] Matheus Gomes Sousa et al. “BERT for stock market sentiment analysis”. In: *2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI)*. IEEE. 2019, pp. 1597–1601.
- [37] Susan AM Vermeer et al. “Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media”. In: *International Journal of Research in Marketing* 36.3 (2019), pp. 492–508.
- [38] Yaowei Wang et al. “EAN: Event attention network for stock price trend prediction based on sentimental embedding”. In: *Proceedings of the 10th ACM conference on web science*. 2019, pp. 311–320.
- [39] Wei Xiang and Bang Wang. “A survey of event extraction from text”. In: *IEEE Access* 7 (2019), pp. 173111–173137.
- [40] AEO Carosia, Guilherme Palermo Coelho, and AEA Silva. “Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media”. In: *Applied Artificial Intelligence* 34.1 (2020), pp. 1–19.

BIBLIOGRAPHY

- [41] Cosmin-Octavian Cepoi. “Asymmetric dependence between stock market returns and news during COVID-19 financial turmoil”. In: *Finance research letters* 36 (2020), p. 101658.
- [42] Divyanshu Daiya, Min-Sheng Wu, and Che Lin. “Stock movement prediction that integrates heterogeneous data sources using dilated causal convolution networks with attention”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 8359–8363.
- [43] Audeliano Wolian Li and Guilherme Sousa Bastos. “Stock market forecasting using deep learning and technical analysis: a systematic review”. In: *IEEE access* 8 (2020), pp. 185232–185242.
- [44] Jing Li et al. “A survey on deep learning for named entity recognition”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (2020), pp. 50–70.
- [45] Shifeng Liu et al. “HAMNER: Headword amplified multi-span distantly supervised method for domain specific named entity recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 8401–8408.
- [46] Salvatore Carta et al. “Event detection in finance using hierarchical clustering algorithms on news and tweets”. In: *PeerJ Computer Science* 7 (2021), e438.
- [47] Shilpa Gite et al. “Explainable stock prices prediction from financial news articles using sentiment analysis”. In: *PeerJ Computer Science* 7 (2021), e340.
- [48] Pilar López-Úbeda et al. “Pre-trained language models to extract information from radiological reports.” In: *CLEF (Working Notes)*. 2021, pp. 794–803.
- [49] Arthur Emanuel de Oliveira Carosia, Guilherme Palermo Coelho, and Ana Estela Antunes da Silva. “Investment strategies applied to the Brazilian stock market: a methodology based on sentiment analysis with deep learning”. In: *Expert Systems with Applications* 184 (2021), p. 115470.
- [50] Rokas Štrimaitis et al. “Financial context news sentiment analysis for the Lithuanian language”. In: *Applied Sciences* 11.10 (2021), p. 4443.
- [51] Manish Agrawal et al. “Stock Prediction Based on Technical Indicators Using Deep Learning Model.” In: *Computers, Materials & Continua* 70.1 (2022).
- [52] Shayan Halder. “FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis”. In: *arXiv preprint arXiv:2211.07392* (2022).

BIBLIOGRAPHY

- [53] Pan Liu et al. “Chinese named entity recognition: The state of the art”. In: *Neurocomputing* 473 (2022), pp. 37–53.
- [54] Abdul Rashid et al. “Time-Varying Impacts of Macroeconomic Variables on Stock Market Returns and Volatility: Evidence from Pakistan”. In: *Journal for Economic Forecasting* 3 (2022), pp. 144–66.
- [55] Krittakom Srijiranon, Yoskorn Lertratanakham, and Tanatorn Tanantong. “A hybrid Framework Using PCA, EMD and LSTM methods for stock market price prediction with sentiment analysis”. In: *Applied Sciences* 12.21 (2022), p. 10823.
- [56] Swati Srivastava et al. “Stock price prediction using LSTM and news sentiment analysis”. In: *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE. 2022, pp. 1660–1663.
- [57] Shengting Wu et al. “S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis”. In: *Connection Science* 34.1 (2022), pp. 44–62.
- [58] Shilpa BL and Shambhavi BR. “Combined deep learning classifiers for stock market prediction: integrating stock price and news sentiments”. In: *Kybernetes* 52.3 (2023), pp. 748–773.
- [59] Statista. *Fintech - Statistics & Facts*. Accessed: 2024-06-13. 2023. URL: <https://www.statista.com/topics/2404/fintech/>.
- [60] Statista. *Gross Domestic Product (GDP) in Pakistan 2023*. Accessed: 2024-06-13. 2023. URL: [https://www.statista.com/statistics/383739/gross-domestic-product-gdp-in-pakistan/#:~:text=Gross%20domestic%20product%20\(GDP\)%20in%20Pakistan%202023&text=The%20gross%20domestic%20product%20\(GDP,GDP%20than%20the%20preceding%20years..](https://www.statista.com/statistics/383739/gross-domestic-product-gdp-in-pakistan/#:~:text=Gross%20domestic%20product%20(GDP)%20in%20Pakistan%202023&text=The%20gross%20domestic%20product%20(GDP,GDP%20than%20the%20preceding%20years..)