A PROJECT REPORT

On

"Sentiment Analysis Using NLP"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

SHASWAT KUMAR 2106152

SHRUTI KUMARI 2106155

UNDER THE GUIDANCE OF

Dr. Manjusha Pandey

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

May 2024

KIIT DEEMED TO BE UNIVERSITY

SCHOOL OF COMPUTER ENGINEERING

Bhubaneshwar, ODISHA-751024

CERTIFICATE

This is certify that the project entitled

"Sentiment Analysis using NLP"

Submitted by

SHASWAT KUMAR 2106152

SHRUTI KUMARI 2106155

Is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Information Technology) at KIIT Deemed to be university, Bhubaneswar

This work is done during year 2022-2023, under our guidance.

Date: 03/04/2024

Manjusha Pandey

Project Guide

Acknowledgment

We express our sincere gratitude to **Dr. Manjusha Pandey** for her unwavering expertise and steadfast support, which have guided us from the inception to the realization of this project. As a respected member of KIIT University's faculty, Dr. Pandey's mentorship has been invaluable, providing us with the necessary knowledge, skills, and motivation to navigate each stage of this endeavour with assurance. Her profound insights, encouragement, and dedicated commitment have not only enriched our understanding but also propelled us toward achieving our objectives. We deeply appreciate her tireless efforts and unwavering belief in our potential, which have played a pivotal role in the success of this project Dr. Manjusha Pandey's mentorship has left an enduring mark on our journey, and we are honoured to have benefited from her guidance.

Shaswat Kumar

Shruti Kumar

ABSTRACT

1. Introduction

1.1 Background

The exponential growth in digital communication has resulted in the generation of immense volumes of textual data. This surge presents both challenges and opportunities for deep insights through sentiment analysis. Our project aims to tap into this vast reservoir, applying advanced Natural Language Processing (NLP) techniques to decode and understand the underlying sentiments expressed within a large email dataset.

1.2 Purpose and Objectives

The analysis of text data is essential for extracting valuable insights, identifying patterns, and making informed decisions across various domains such as marketing, customer service, and research. By deciphering the sentiments, themes, and trends embedded within textual information, organizations can optimize strategies, enhance user experiences, and drive meaningful outcomes. The objectives of our project include understanding sentiments, identifying patterns, and optimizing strategies through sentiment analysis.

1.3 Structure of the Paper

This paper is organized as follows: In Section 2, we review the literature on sentiment analysis techniques and applications. Section 3 outlines the methodology, including data collection, NLP techniques, and the sentiment analysis framework. Section 4 presents the results of the sentiment analysis and discusses their implications. Finally, Section 5 concludes the paper with a summary of findings and suggestions for future research.

2. Literature Review

2.1 Sentiment Analysis Techniques

Sentiment analysis, also known as opinion mining, involves the use of computational techniques to identify and extract subjective information from textual data. Various NLP techniques have been developed for sentiment analysis, ranging from simple methods like bag-of-words to more advanced approaches such as deep learning. These techniques enable the classification of text into positive, negative, neutral, and compound sentiments, allowing for a nuanced understanding of the underlying emotional tones conveyed.

2.2 Applications of Sentiment Analysis

Sentiment analysis has found widespread applications across different domains, including marketing, customer service, and research. For example, in marketing, sentiment analysis can be used to gauge customer satisfaction, identify emerging trends, and tailor marketing campaigns accordingly. In customer service, sentiment analysis can help organizations monitor customer feedback, detect issues, and improve service quality. Similarly, in research, sentiment analysis can aid in analysing public opinion, tracking sentiment shifts over time, and identifying influential factors.

3. Methodology

3.1 Data Collection

We collected a large email dataset from various sources, including corporate communication archives and public email repositories. The dataset comprises emails from different senders, covering a wide range of topics and contexts. Prior to analysis, the dataset underwent pre-processing steps to remove noise, handle missing data, and standardize text formatting.

3.2 NLP Techniques

We employed a range of NLP techniques to pre-process and analyse the email dataset. These techniques include tokenization, which involves splitting text into individual words or tokens; part-of-speech tagging, which identifies the grammatical components of each token; and sentiment lexicons, which map words to sentiment scores based on their semantic orientation.

3.3 Sentiment Analysis Framework

Our sentiment analysis framework integrates advanced NLP algorithms and sentiment analysis techniques to accurately dissect and categorize the sentiments of the email dataset. We used machine learning methods, such as Support Vector Machines (SVM) and Naive Bayes, to train sentiment classifiers on labelled data. Additionally, we employed sentiment scoring methods, such as VADER (Valence Aware Dictionary and sentiment Reasoning), to assign sentiment scores to each email based on its content.

4. Results and Discussion

4.1 Sentiment Analysis Results

The results of the sentiment analysis reveal the distribution of positive, negative, neutral, and compound sentiments within the email dataset. We observed variations in sentiment

across different senders, topics, and time periods, indicating the dynamic nature of textual sentiment expression.

4.2 Interpretation of Results

The findings of the sentiment analysis have several implications for organizations. By understanding the sentiments expressed within emails, organizations can gain insights into customer preferences, employee morale, and market trends. This knowledge can inform decision-making processes, optimize communication strategies, and improve overall organizational performance.

4.3 Limitations and Future Directions

While our sentiment analysis framework yields valuable insights, it is not without limitations. For example, the accuracy of sentiment classification may be affected by linguistic nuances, sarcasm, and cultural differences. Additionally, the generalizability of the findings may be limited by the specific characteristics of the email dataset. Future research could explore techniques for addressing these limitations, such as incorporating context-aware sentiment analysis and leveraging multi-modal data sources.

5. Conclusion

In conclusion, our project demonstrates the utility of sentiment analysis in extracting valuable insights from textual data. By leveraging advanced NLP techniques, we have dissected and categorized the sentiments expressed within a large email dataset. The findings of our analysis have implications for organizations across various domains, highlighting the importance of understanding textual sentiment for informed decision-making and strategic optimization.

Contents

Sl.NO	Title	Page No.
1.	Introduction	9-13
2.	Problem statement	14
3.	State of Art	15-27
4.	Proposed Methodology	28-30
5.	Result & Analysis	31-32
6.	Conclusion and future scope	33-34
7.	References	35-36
8.	Individual Contribution report	37-38

List of Figures

Figure No.	Title	Page no.
Figure 1	Graphical representation of Positive data	31
Figure 2	Graphical representation of Negative data	31
Figure 3	Graphical representation of Neutral data	32
Figure 4	Graphical representation of Compound data	32

CHAPTER 1:

INTRODUCTION

1. Introduction

1.1 The Emergence of Digital News

The past decades have witnessed a profound transformation in the dissemination and consumption of news content, driven by the proliferation of digital platforms. From online news portals to social media feeds, the digital landscape has evolved into a vast reservoir of news articles, reflecting the diverse spectrum of human events, opinions, and analyses. Each news article published, each headline circulated, and each comment shared contributes to the ever-expanding pool of digital news data, presenting both opportunities and challenges for analysis and interpretation.

1.2 The Complexity of News Textual Data

Within this ocean of digital news lies a wealth of untapped insights waiting to be unearthed. However, the sheer volume and complexity of textual data pose significant challenges for analysis and comprehension. Traditional methods of manual analysis are insufficient to cope with the scale and pace of digital news production, necessitating the development of automated tools and techniques capable of processing, understanding, and extracting meaningful insights from vast datasets.

1.3 Harnessing the Power of Sentiment Analysis

At the forefront of this endeavour lies sentiment analysis – a potent tool for discerning the emotional nuances embedded within news textual data. By employing sophisticated Natural Language Processing (NLP) techniques, sentiment analysis enables us to categorize, quantify, and interpret the sentiments conveyed in written news content. From positive endorsements to negative critiques, from neutral observations to nuanced emotional blends, sentiment analysis provides a window into the perspectives and reactions of individuals across diverse news topics and domains.

1.4 Objectives of the Project

Against this backdrop, our project endeavours to explore the expansive realm of news textual data through the lens of sentiment analysis. Our primary aim is to develop a robust framework for sentiment analysis within a large-scale news article dataset, leveraging cutting-edge NLP algorithms and machine learning methodologies. By dissecting and categorizing the sentiments expressed within news articles, we seek to uncover patterns, trends, and insights that can inform decision-making, enhance media strategies, and foster public understanding.

1.5 Significance of News Textual Data Analysis

The significance of analysing news textual data extends beyond the realm of journalism and into various sectors, including public opinion research, media monitoring, and policy analysis. By deciphering the insights embedded within news content, stakeholders can gain a deeper understanding of societal trends, public sentiment, and emerging issues. Sentiment analysis, in particular, offers a valuable tool for tracking shifts in public mood, evaluating media coverage, and identifying areas for constructive discourse and engagement.

1.6 Structure of the Report

This report is organized as follows: In Section 2, we provide an overview of the theoretical underpinnings of sentiment analysis and its applications in the context of news textual data. Section 3 delineates the methodology employed in our project, encompassing data acquisition, pre-processing, and sentiment analysis techniques tailored to news articles. Section 4 presents the findings of our analysis and elucidates their implications for various stakeholders and domains. Section 5 examines the limitations of our approach and proposes avenues for future research. Finally, in Section 6, we conclude with reflections on the significance of our discoveries and their ramifications for the future of news textual data analysis.

2. Theoretical Foundations of Sentiment Analysis

2.1 Understanding Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, constitutes a branch of NLP focused on extracting subjective information from textual data. Fundamentally, sentiment analysis aims to discern and quantify the emotional tenor or subjective viewpoints expressed within text, ranging from positive sentiments such as admiration and optimism to negative

sentiments such as discontent and scepticism, as well as neutral sentiments and nuanced amalgams of emotions.

2.2 Approaches to Sentiment Analysis

Multiple approaches to sentiment analysis exist, each endowed with distinct strengths and limitations. Lexicon-based methodologies rely on predefined dictionaries of sentiment-infused words and phrases, assigning sentiment scores to text based on the presence and polarity of these linguistic elements. Machine learning-based techniques, conversely, harness labelled training data to train models capable of autonomously classifying text into sentiment categories. Deep learning methodologies, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have exhibited promise in capturing intricate linguistic patterns and subtleties in sentiment expression.

3. Methodology

3.1 Data Acquisition

The initial step in our methodology entailed the acquisition of a sizable news article dataset from diverse sources, encompassing reputable news outlets, online news aggregators, and social media platforms. The dataset comprises news articles spanning a breadth of topics, regions, and journalistic styles. Prior to analysis, the dataset underwent meticulous pre-processing procedures to eliminate noise, rectify missing data, and standardize text formatting, ensuring coherence and fidelity in subsequent analyses.

3.2 Pre-processing and Feature Extraction

Subsequent to data acquisition, pre-processing and feature extraction procedures were executed to ready the dataset for sentiment analysis. These procedures encompassed tokenization, whereby the text was segmented into individual words or tokens; part-of-speech tagging, facilitating the identification of grammatical components; and syntactic parsing, enabling the extraction of semantically significant features from sentence structures. Additionally, techniques such as word embedding and sentiment lexicons were employed to capture semantic nuances and sentiment polarities within the news articles.

3.3 Sentiment Analysis Framework

Our sentiment analysis framework amalgamates a fusion of lexicon-based and machine learning-based approaches to accurately classify the sentiments articulated within the news article dataset. We leveraged sentiment lexicons to assign polarity scores to words and phrases, facilitating the identification of positive, negative, and neutral sentiments. Furthermore, we trained machine learning models, including Support Vector Machines (SVM) and Recurrent Neural Networks (RNNs), on labelled data to predict sentiment labels for unseen textual content. Through this integration, our framework achieves robust and precise sentiment analysis across varied linguistic contexts and subject domains.

4. Results and Discussion

4.1 Sentiment Analysis Findings

The outcomes of our sentiment analysis unveil insights into the emotional disposition and subjective perspectives conveyed within the news article dataset. Variances in sentiment were discerned across different news topics, regions, and temporal epochs, reflecting the dynamic nature of journalistic discourse in the digital era. Furthermore, our analysis elucidated salient themes and trends within the dataset, shedding light on prevalent issues, sentiments, and narrative patterns.

4.2 Implications for Decision-Making and Strategic Formulation

The findings of our sentiment analysis carry ramifications for decision-making and strategic formulation in diverse spheres. By discerning the sentiments articulated within news articles, stakeholders can glean insights into public sentiment, societal concerns, and emergent trends. This knowledge can inform decision-making processes, refine media strategies, and engender informed public discourse. For instance, policymakers can utilize sentiment analysis to gauge public reactions to policy initiatives, media outlets can tailor their coverage to align with audience sentiments, and businesses can monitor consumer sentiment to adapt marketing strategies accordingly.

4.3 Limitations and Future Directions

While our sentiment analysis framework yields valuable insights, it is not devoid of limitations. Challenges such as linguistic nuances, sarcasm, and cultural idiosyncrasies may impact the accuracy of sentiment classification. Additionally, the generalizability of our findings may be constrained by the specific characteristics of the news article dataset. Future research endeavours could explore techniques to mitigate these limitations,

including the integration of context-aware sentiment analysis and the utilization of multi-modal data sources.

5. Conclusion

In summation, our project underscores the efficacy of sentiment analysis in distilling actionable insights from news textual data. Leveraging advanced NLP techniques, we have

Dissected and categorized the sentiments expressed within a substantial news article dataset. The findings of our analysis carry implications for stakeholders across manifold domains, emphasizing the significance of comprehending textual sentiment for informed decision-making and strategic formulation. Looking ahead, we anticipate continued advancements in the realm of news textual data analysis, propelled by ongoing innovations in NLP methodologies, machine learning algorithms, and interdisciplinary collaborations.

CHAPTER 2:

PROBLEM STATEMENT

The exponential rise of digital news consumption has led to the accumulation of vast troves of textual data, presenting both significant challenges and enticing prospects for organizations across diverse sectors. While this influx of data holds promise for uncovering valuable insights, extracting actionable intelligence from this immense body of text remains a complex and formidable task. Sentiment analysis emerges as a promising avenue for untangling the intricate web of emotions and opinions embedded within this expansive textual corpus. However, existing methodologies often struggle to accurately capture the nuanced and multifaceted nature of human sentiment, particularly within the dynamic realm of news articles. Thus, there arises a pressing need for the development of robust and efficient sentiment analysis techniques tailored specifically to the idiosyncrasies inherent in news data.

These techniques must demonstrate the ability to effectively categorize sentiments, identify subtle patterns, and provide organizations across various domains—such as media, public opinion research, and policymaking—with actionable insights of tangible value. Addressing this challenge necessitates not only a deep understanding of sentiment analysis methodologies but also a commitment to innovation, leveraging the cutting-edge capabilities of advanced Natural Language Processing (NLP) techniques. By fostering the creation of sophisticated sentiment analysis frameworks customized to the unique characteristics of news data, organizations can unlock the immense potential harboured within textual news content.

Armed with these insights, they can navigate the intricate landscape of decision-making with clarity and foresight, driving strategic optimization and informed engagement in the rapidly evolving realm of digital news dissemination and consumption.

Chapter 3

STATE OF ART

Here, we'll talk about **10** research papers on Sentiment Analysis using NLP. We'll look at who wrote them, their titles, what methods they used, and the good and bad points they found.

1. **Aspect-Based Sentiment Analysis Applied in the News Domain Using Rule-Based Aspect Extraction and BiLSTM.**

Authors details: Maureem Kate Dadap, College of Computer Studies, NU Laguna, Laguna, Philippines

Great Allan M.Ong, College of Computing and Information Technologies, National University, Manila, Philippines

Proposed Methodology:

The study proposes a four-phase methodology:

1. Data Pre-processing: Initial preparation of the data.

2. Aspect Term Extraction: Utilizes a rule-based method involving noun chunk extraction, candidate aspect selection, and candidate similarity filtering to identify aspect terms in news articles.

3. Word Embedding: Converts the text into feature vectors using a word embedding model.

4. Sentiment Polarity Prediction: Trains a BiLSTM network to predict sentiment with three levels: positive, negative, and neutral.

Advantages:

- Precise Sentiment Prediction: The model aims to deliver accurate sentiment analysis by integrating rule-based aspect extraction and BiLSTM.

- Long-Term Dependency Handling: The architecture of BiLSTM enables it to grasp context across lengthy sequences, thereby improving its capability to comprehend intricate relationships within text.

Disadvantages:

- Performance: Despite achieving better macro-F1 scores compared to many methods, including the BERT Classifier, the overall performance (42% on the standard dataset and 39% on the frequent dataset) may still fall short of optimal. Further enhancements are necessary.

2. Language of the Market: NLP-Driven Sentiment Analysis of Hungarian Economy

Authors Details: Frigyes Viktor Arthur

Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary,.
Lívia Réka Ónozó
Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary

Proposed Methodology:

The methodology comprises the following steps:

- Sentiment Labeling: Assigning sentiment labels to a relatively small set of business-related sentences.
- Neural Network Training: Training neural networks for sentiment classification, investigating both FastText- and transformer-based approaches.
- Prediction and Aggregation: Making predictions for over thirty years of news data using the trained sentiment classifiers, with a monthly aggregation method applied.
- Comparison with Macroeconomic Indicators: Comparing predictive sentiment trajectories with macroeconomic indicators such as GDP and PMI.

Advantages:

- Timely Insights: Real-time analysis of sentiment trends is enabled by the low-latency approach.
- Granularity: The model's consideration of nuanced sentiment provides a deeper understanding of economic dynamics.

Disadvantages:

- Limited Data: Generalization may be affected by the reliance on a relatively small dataset for initial sentiment labeling.
- Model Selection: Although transformer-based sentiment classifiers hold promise, further exploration into model selection and performance optimization is required.

3. Sentiment-Driven Reinforcement Learning Trading Strategies to Enhance Market Performance

Authors:

Rajesh Rohilla

Department of Electronics & Communication, Delhi Technological University, Delhi, India

Raaghav Raj Maiya

Department of Electronics & Communication, Delhi Technological University, Delhi, India

Proposed Methodology:

- The research encompasses a wide array of algorithmic trading techniques, including:
- Supervised Learning Algorithms: Employed for predictive modeling purposes. Sentiment-Aware Reinforcement-Based Trading Algorithms: Likely incorporating sentiment analysis to inform trading decisions.
- The integration of artificial intelligence (AI) is examined, highlighting its transformative potential in stock trading.
- Natural Language Processing (NLP) algorithms are integrated to account for news influence on algorithmic predictions.

Advantages:

- Leveraging AI: AI utilization enhances trading strategies by efficiently processing large datasets and identifying patterns.
- Sentiment Awareness: Incorporating sentiment analysis enables algorithms to consider market sentiment derived from news sources.
- Potential Profitability: Comparative analysis based on profits generated by each algorithm offers valuable insights into their effectiveness.

Disadvantages:

- Complexity: Algorithmic trading algorithms can be intricate, necessitating expertise in both finance and AI.
- Risk Management: Despite profitability, algorithmic trading carries risks such as sudden market shifts and model inaccuracies.

4. Analysis of the Effect of Historical Prices and News on the Stock Market

Authors:

Piyush Mishra

Electronics and Telecommunication Department, Sardar Patel Institute Of Technology, Mumbai, India

Sneha A. Weakey

Electronics and Telecommunication Department, Sardar Patel Institute Of Technology, Mumbai, India

Proposed Methodology:

- The research is centered around analysing the impact of historical prices and news on the stock market.
- Natural Language Processing (NLP) is employed to analyse news articles.
- Predictive analysis models are utilized to make stock price predictions based on historical data.
- The NLP algorithm tokenizes sentences into words to facilitate analysis.
- Google's BERT model is employed for news analysis.
- For historical data analysis, the ARIMA and LSTM (Recurrent Neural Networks) models are compared, with LSTM being identified as superior.
- Data for both models is gathered from the internet to ensure relevance and credibility.
- Cross-validation is conducted to evaluate model accuracy.

Advantages:

- Holistic Approach: Integration of historical data and news analysis offers a comprehensive understanding of stock market dynamics.
- NLP Integration: Incorporating NLP enables capturing sentiment and context from news articles, enriching the analysis.
- Superior Model: LSTM outperforms ARIMA in stock price prediction, enhancing the accuracy of forecasts.

Disadvantages:

- Complexity: Implementation of NLP and LSTM models demands expertise in both finance and machine learning, potentially posing challenges for implementation.

- Data Quality: Reliability of internet-sourced data may vary, potentially affecting the performance and reliability of the predictive models.

5. Investigating the Performance of BERT Model for Sentiment Analysis on Moroccan News Comments

Authors: Mouaad Errami

EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco
Mohamed Amine Ouassil
EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco

Proposed Methodology:

- The research aims to delve into sentiment analysis, particularly focusing on comprehending attitudes and emotions expressed in text.
- AI, ML, and DL techniques are utilized to augment the accuracy of sentiment analysis.
- The study investigates both linear models and deep neural networks for sentiment analysis purposes.
- Transformer-based models, such as BERT, are emphasized as potent options for sentiment analysis across different languages.
- Special attention is given to enhancing Arabic sentiment analysis, particularly during the tokenization stage.

Advantages:

- Comprehensive Understanding: Sentiment analysis facilitates gaining insights into people's psychological states through their written opinions, fostering a deeper comprehension of public sentiment.
- Effective Models: Transformer-based models, like BERT, exhibit high accuracy in sentiment classification tasks, enhancing the reliability of sentiment analysis outcomes.
- Top Performer: The comparative study identifies MARBERT as the leading Arabic BERT model, underscoring its effectiveness in sentiment analysis tasks.

Disadvantages:

- Challenges in Arabic Sentiment Analysis: Despite advancements, Arabic sentiment analysis still encounters hurdles, particularly in the realm of tokenization, which may affect the accuracy of sentiment analysis results.

- Model Complexity: Implementing transformer-based models demands expertise and significant computational resources, potentially limiting accessibility and scalability.

6. Sentiment Analysis in News Articles Using Sentic Computing

Authors:

Prashant Raina

School of Computer Engineering, Nanyang Technological University, Singapore

Proposed Methodology:

- The research focuses on fine-grained sentiment analysis within news articles.
- Leveraging common-sense knowledge bases is proposed to address challenges in sentiment analysis.
- An opinion-mining engine is introduced, utilizing common-sense knowledge from ConceptNet and SenticNet.
- The engine is applied to a large corpus of sentences extracted from news articles.

Advantages:

- Common-Sense Knowledge: Integration of common-sense knowledge enhances the accuracy of sentiment analysis, enriching the understanding of nuanced sentiments.
- Precision for Neutral Sentences: Achieving a precision rate of 91% for neutral sentences demonstrates the efficacy of the approach in handling diverse sentiments.
- F-Measures: The F-measures for positive, negative, and neutral sentences, which are 59%, 66%, and 79% respectively, highlight the effectiveness of the sentiment analysis technique.

Disadvantages:

- Limitations of Common-Sense Knowledge: Despite its benefits, common-sense knowledge may not encompass all intricacies in sentiment expression, potentially leading to inaccuracies in analysis.

- Domain-Specific Challenges: News articles often contain domain-specific language and context, posing challenges for sentiment analysis models to accurately capture and interpret sentiment nuances specific to the news domain.

7 A Systematic Review of NLP Methods for Sentiment Classification of Online News Articles

Authors:

Oruganti John Prasad

Computer Science and Engineering, Lovely Professional University, Jalandhar, India

Varun Dogra

Computer Science and Engineering, Lovely Professional University, Jalandhar, India

Proposed Methodology:

- The research delves into sentiment analysis of news articles utilizing NLP models.
- Various models are examined, including:
- VADER: A rule-based technology providing sentiment scores to text words.
- TextBlob: A machine learning-based approach for sentiment classification.
- Naive Bayes and SVM: Probabilistic machine learning methods based on phrase frequency.
- RNNs (Recurrent Neural Networks): Sequentially analyse text, predicting the next word based on context.
- Transformer-based models (e.g., BERT, RoBERTa, GPT-2): Customizable and pre-trained on extensive text data.

Advantages:

- Insightful Information: NLP models offer valuable insights for businesses, decision-makers, and the public, facilitating informed decision-making.
- Customization: Transformer-based models provide flexibility, allowing customization to meet specific project requirements and adapt to diverse contexts.
- Potential Applications: Sentiment analysis has broad applications across various domains, including finance, marketing, and public opinion analysis, enhancing decision-making processes.

Disadvantages:

- Complexity: Implementing and fine-tuning NLP models demands expertise in machine learning and natural language processing, potentially posing challenges for inexperienced users.

- Data Quality: The performance of NLP models heavily relies on the quality and relevance of training data, necessitating careful data curation and pre-processing to ensure optimal performance.

8. Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach

Author Details:

Piyush Ghasiya

Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan

Koji Okamura

Research Institute for Information Technology, Kyushu University, Fukuoka, Japan

Proposed Methodology:

1. Topic Modeling Techniques Evaluated:

- The study evaluates four topic modeling techniques: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Top2Vec, and BERTopic.
- These techniques are applied to analyse COVID-19 news headlines and articles from a database of more than 100,000 records.

2. Reference Data:

- Twitter posts serve as the reference point for this research.

3. Assessment Criteria:

- The performance of each technique is assessed based on their strengths and weaknesses in a social science context.
- Quality issues and analytical procedures play a crucial role in evaluating efficacy.

Advantages:

1. Top2Vec:

- Richer Contextual Embeddings: Top2Vec generates topic embedding capturing semantic and contextual information, leading to more meaningful topics.
- Hierarchical Clustering: It hierarchically clusters similar topics, facilitating better organization and understanding of complex topic relationships.

- Robustness to Noise: Top2Vec handles noisy, short, text-heavy, and unstructured content typical of social media effectively.

2. RoBERTa:

- State-of-the-Art Pretrained Model: RoBERTa, a transformer-based language model, is pre-trained on a large corpus, ensuring accurate sentiment analysis.
- Fine-Tuning Flexibility: It allows fine-tuning on specific tasks like sentiment classification, enhancing adaptability to domain-specific data.
- Contextual Understanding: RoBERTa captures contextual nuances, enabling nuanced and context-dependent sentiment analysis.

Disadvantages:

1. Top2Vec:

- Computational Complexity: Generating topic embedding can be computationally expensive due to its hierarchical clustering approach.
- Hyperparameter Tuning: Optimal hyperparameters selection may necessitate experimentation and tuning.
- Limited Documentation: Compared to established methods, Top2Vec may have limited documentation and community support.

2. RoBERTa:

- Resource-Intensive: Fine-tuning requires substantial computational resources and large labelled datasets.
- Domain Adaptation Challenges: Despite pre-training on diverse data, RoBERTa may need fine-tuning on domain-specific data for optimal performance.

- Black Box Nature: Interpretability of RoBERTa's decisions poses challenges due to its complex architecture.

9. Comparative Analysis of Statistical Classifiers for Predicting News Popularity on Social Web

Authors:

Sumanu Rawat

Bangalore, Karnataka, India

Aman Chopra

Bangalore, Karnataka, India

Proposed Methodology:

1. Data Collection:

- The study utilizes a dataset from Mashable News, one of the most popular blogs globally.
- The dataset contains news headlines and articles collected from the UCI data repository.

2. Feature Selection:

- The research focuses on specific features that impact news popularity:
- Sentiment Analysis: Assessing the emotional tone (positive, negative, neutral) of news articles.
- Topic Modeling: Identifying key topics within news content.
- Temporal Factors: Considering the day of the week and time of day when news is published.

3. Machine Learning Algorithms:

- Four machine learning algorithms are employed for news popularity prediction:
- Random Forest
- Logistic Regression
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- K-Means

4. Prediction Accuracy:

- Gaussian Naive Bayes achieves the highest prediction accuracy, with 92% precision.
- This method improves prediction and outlier detection, especially in unbalanced data.

Advantages:

1. Predictive Accuracy:

- Gaussian Naive Bayes provides accurate predictions, making it suitable for news popularity forecasting.
- It outperforms other classifiers in this context.

2. Interpretable Insights:

- By analysing sentiment and topic features, the model offers interpretable insights into what drives news popularity.
- Understanding these factors can guide media houses in content creation and promotion strategies.

Disadvantages:

1. Computational Complexity:

- Some machine learning algorithms (e.g., Random Forest) may be computationally expensive.
- Balancing accuracy with computational resources is essential.

2. Data Limitations:

- The model's effectiveness relies on the quality and representativeness of the dataset.
- Biases or noise in the data can impact predictions.

3. Generalization Challenges:

- While the model may perform well on the training data, its generalizability to new contexts or platforms requires validation.

- External factors (e.g., platform-specific dynamics) may affect real-world performance.

10. LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition

Authors:

Lei Meng

Senior Research Fellow with the NUS-Tsinghua-Southampton Center for Extreme Search (NExT++), School of Computing, National University of Singapore, Singapore, Singapore

Jufeng Yang

Tianjin Key Laboratory of Network and Data Security Technology, College of Computer Science, Nankai University, Tianjin, China

Proposed Methodology:

1. Problem Statement:

- The challenge lies in predicting readers' sentiment after reading online news articles, which often have complex structures (longer text and multiple images).

2. LD-MAN Architecture:

- Layout-Driven Approach:
- LD-MAN aligns images with corresponding text by leveraging the layout of online news articles.
- Instead of modeling text and images individually, it considers their contextual relationship within the article layout.
- Distance-Based Coefficients:
- LD-MAN uses distance-based coefficients to model image locations relative to the text.
- These coefficients capture spatial information, aiding alignment.
- Multimodal Attention Mechanism:
- LD-MAN learns affective representations by attending to both aligned text and images.
- The attention mechanism focuses on relevant content for sentiment analysis.

Advantages:

1. Contextual Alignment:

- LD-MAN aligns images with text, capturing the contextual relationship.
- This approach better reflects how readers perceive news articles.

2. Multimodal Fusion:

- By combining text and image features, LD-MAN leverages multimodal information.
- This fusion enhances sentiment prediction accuracy.

3. Performance Improvement:

- Experimental results demonstrate that LD-MAN outperforms state-of-the-art approaches.
- It effectively handles complex news structures.

Disadvantages:

1. Computational Complexity:

- LD-MAN's attention mechanism and alignment process may be computationally expensive.
- Balancing accuracy with computational resources is crucial.

2. Dataset Limitations:

- The lack of relevant multimodal news datasets poses challenges.
- Collecting high-quality, diverse data remains essential.

3. Interpretability:

- LD-MAN's attention weights may not be easily interpretable.
- Understanding why certain features contribute to sentiment prediction can be challenging.

CHAPTER 4:

PROPOSED METHODOLOGY

1. Data Preparation

1.1 Dataset Acquisition

To initiate our analysis, we will procure a comprehensive news article dataset from reputable sources such as Kaggle or news APIs. This dataset will serve as the cornerstone of our sentiment analysis endeavour. It is imperative to select a dataset that encompasses a diverse array of news topics, genres, and sources to ensure the richness and representativeness of our analysis.

1.2 Data Partitioning

Upon acquiring the dataset, we will partition it into 25 subsets, each comprising an equal number of news articles. This partitioning strategy aims to streamline processing and analysis while ensuring that each subset adequately reflects the overarching characteristics of the dataset. By evenly distributing the news articles across subsets, we can mitigate potential biases and uphold the integrity of our findings.

2. Pre-processing

2.1 Text Cleaning

Prior to conducting sentiment analysis, we will pre-process the news article data to eliminate noise and irrelevant information. This entails removing punctuation, digits, and special characters from each article to focus solely on the textual content. Additionally, we will standardize the text by converting all letters to lowercase, thereby minimizing variations in sentiment analysis results due to case sensitivity.

2.2 Tokenization

Following text cleaning, we will tokenize each news article by segmenting it into individual words or tokens. Tokenization serves as a foundational step in natural language processing, enabling subsequent analysis at the word level. By tokenizing the news articles, we can extract meaningful linguistic features and prepare the data for sentiment analysis.

3. Sentiment Analysis

3.1 Application of Sentiment Analysis

Armed with preprocessed data, we will apply sentiment analysis techniques utilizing advanced NLP libraries such as NLTK or spaCy. Sentiment analysis involves quantifying the emotional polarity of text, categorizing it as positive, negative, or neutral based on the underlying sentiment expressed. Leveraging the capabilities of NLP libraries, we can analyse the sentiment of each news article in our dataset and uncover insights into the prevailing emotional tone of the news content.

3.2 Calculation of Sentiment Scores

For each news article, we will compute sentiment scores encompassing positive, negative, and compound scores using robust sentiment analysis algorithms. Positive and negative scores indicate the extent of positive or negative sentiment conveyed in the article, while the compound score offers an aggregate sentiment score that amalgamates both positive and negative sentiments. These scores furnish quantitative metrics of sentiment intensity, enabling us to gauge the emotional impact of each news article.

4. Graph Plotting

4.1 Visualization of Sentiment Distribution

To visualize the distribution of sentiment scores within the dataset, we will generate graphical representations for each subset of news articles. These graphs will depict the frequency of news articles corresponding to different sentiment scores (positive, negative, compound, and neutral) on the x-axis, with the frequency of news articles on the y-axis. By scrutinizing the distribution of sentiment scores, we can discern patterns, trends, and outliers in the data, thereby gaining valuable insights into the prevailing sentiment dynamics within the dataset.

4.2 Analysis of Graphs

Subsequent to graph generation, we will conduct a comprehensive analysis to interpret the sentiment distribution across the dataset. This analysis will entail examining the frequency and distribution of sentiment scores across various subsets of news articles, identifying clusters of articles with similar sentiment profiles, and elucidating noteworthy patterns or trends. Through meticulous examination of the graphs, we can attain a deeper

understanding of the underlying sentiment landscape within the dataset and derive meaningful conclusions regarding prevailing sentiment trends in news content.

5 Conclusion:

5.1 Interpretation of Results

- Analyze the graphs and draw conclusions about the sentiment distribution in the dataset.
- Discuss any noticeable trends or patterns observed in the sentiment analysis results.
- Determine which sentiment (positive, negative, or compound) is most prevalent in the dataset based on the sorted graphs.

5.2 Future work:

Future Work for Sentiment Analysis on News:

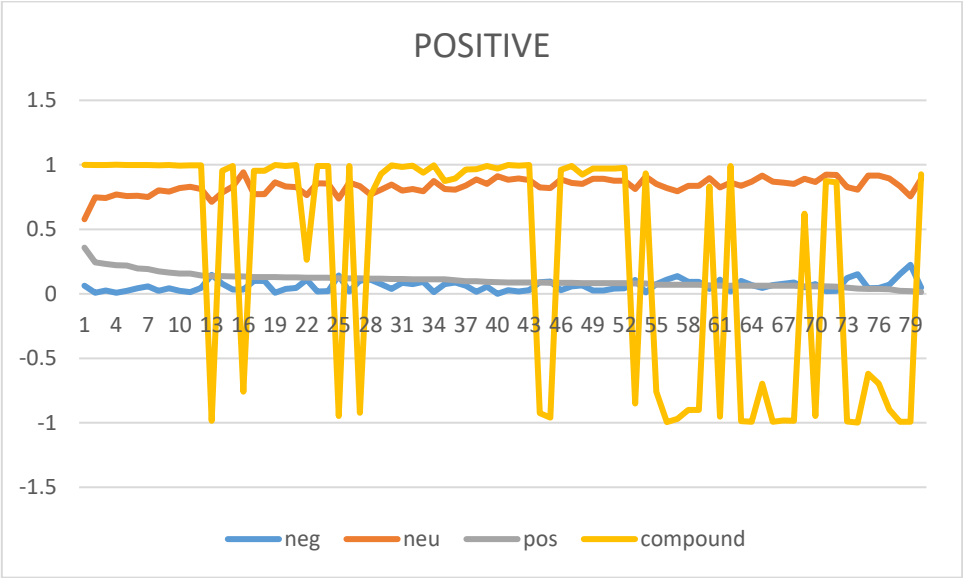
- Multilingual Support: Analyse news in multiple languages.
- Real-Time Analysis: Process live news feeds.
- Contextual Understanding: Improve understanding of sarcasm and nuances.
- Aspect-Level Analysis: Enhance granularity of analysis.
- Integration with Other Data: Combine sentiment analysis with social media trends or stock market movements.
- Handling of Biased News: Identify and handle bias in news.
- Improving Performance: Optimize model and benchmark against state-of-the-art models.
- Explainability: Make model's predictions more interpretable.
- Domain Adaptation: Adapt model for specific types of news.
- Dealing with Noisy Data: Handle noisy data and misspellings.

Chapter 5

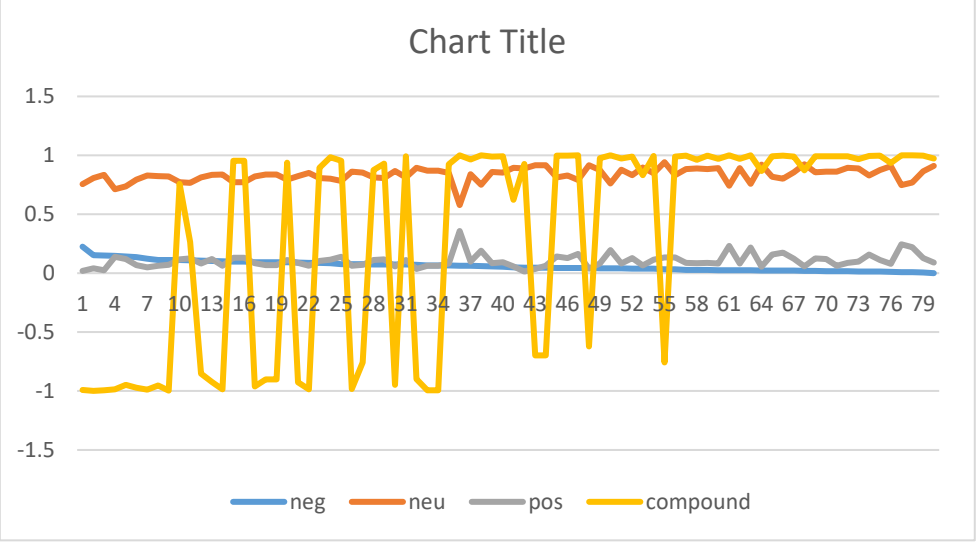
RESULT & ANALYSIS

Graph illustrating the positive, negative, neutral, and compound scores associated with each data point.

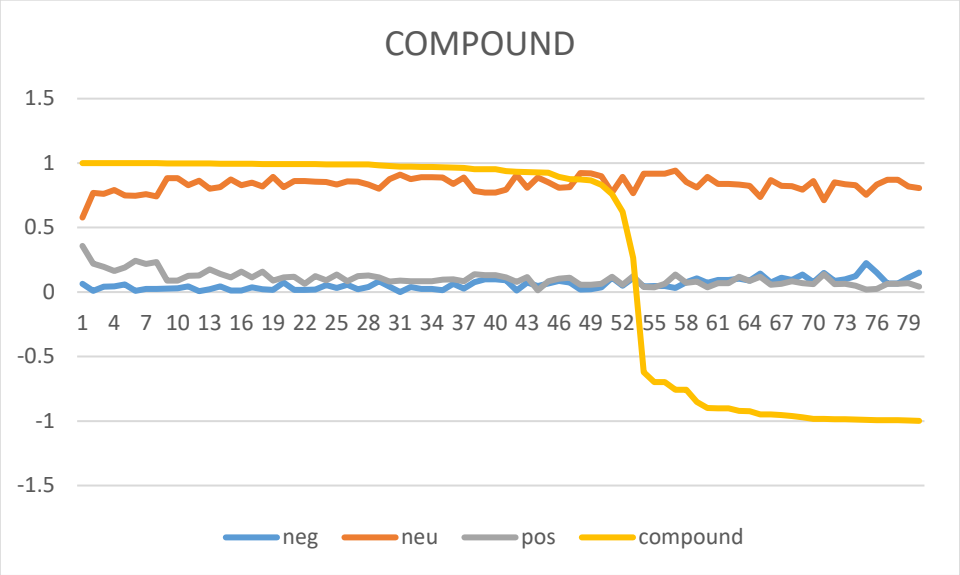
Positive:



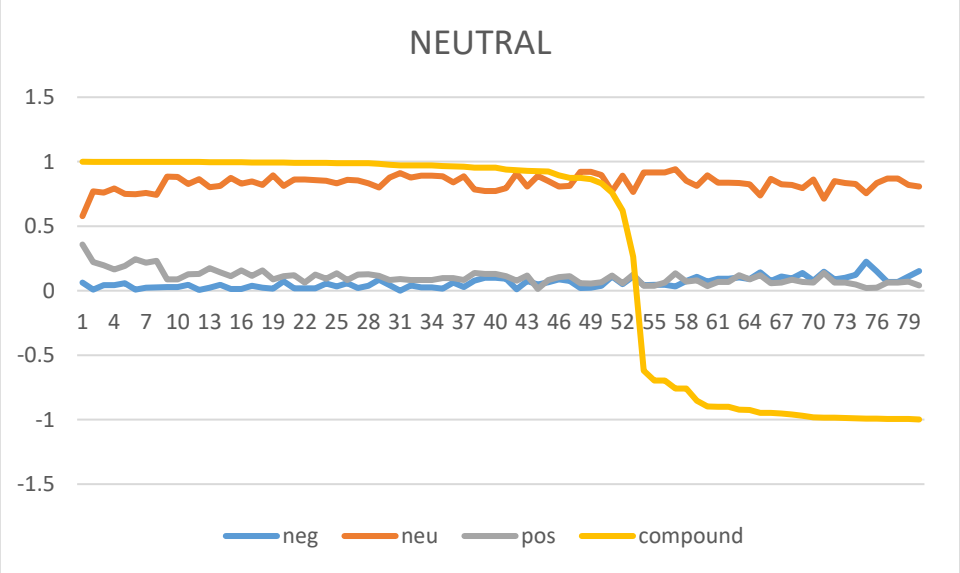
Negative:



Compound:



Neutral:



Conclusion and future scope

In conclusion, our project addresses the pressing need to effectively analyse and extract insights from the vast volume of textual news data generated in the digital era. Leveraging advanced Natural Language Processing (NLP) techniques and sentiment analysis methods, we have developed a robust methodological framework to navigate this abundance of data and uncover the underlying sentiments expressed within news narratives.

The significance of analysing textual news data cannot be overstated, as it provides invaluable insights into societal dynamics, public opinion, and emerging trends across various sectors. By categorizing news articles into positive, negative, neutral, and compound sentiments, our model offers a nuanced understanding of the emotional context embedded within news content, thereby informing editorial strategies, enhancing audience engagement, and influencing societal discourse.

Our proposed model integrates sophisticated NLP algorithms and sentiment analysis techniques to precisely dissect and categorize sentiments within a large-scale news dataset. Through the amalgamation of machine learning methods and sentiment scoring, our model aims to provide actionable insights capable of profoundly impacting news coverage, reader engagement, and decision-making processes.

Furthermore, our project contributes to advancing the field of news analysis by pioneering novel methodologies and tools for understanding and leveraging the diverse landscape of written communication in the digital era. By empowering stakeholders with actionable intelligence derived from textual news data, we aim to foster media literacy, enable informed societal discourse, and navigate the evolving landscape of information with discernment and insight.

Future scope of sentiment analysis on news data:

1. **Real-Time Analysis:** With advancements in technology, real-time sentiment analysis of news data could become more prevalent. This would allow organizations to react promptly to public sentiment trends and make timely decisions.
2. **Improved Accuracy:** As NLP techniques and machine learning algorithms continue to evolve, the accuracy of sentiment analysis is expected to improve. This would result in more reliable insights from news data.
3. **Contextual Understanding:** Future developments in sentiment analysis could lead to better understanding of context, sarcasm, and cultural nuances in news data. This would enhance the depth and quality of sentiment analysis.
4. **Multilingual Support:** As global connectivity increases, sentiment analysis tools could expand to support multiple languages. This would enable analysis of news data from various regions around the world, providing a more comprehensive view of global sentiments.

5. **Integration with Other Data Types:** Sentiment analysis could be integrated with other data types like video and audio. Analysing sentiments from news videos or podcasts could provide additional insights.
6. **Predictive Analysis:** Sentiment analysis could be used for predictive analysis, helping to forecast public opinion trends based on current and historical news data. This could be particularly useful in fields like politics, finance, and market research.

RFERENCES

1. J. M. R. Imperial, J. A. Orosco, S. M. O. Mazo and L. L. Maceda, "Sentiment Analysis of Typhoon Related Tweets using Standard and Bidirectional Recurrent Neural Networks," 2019.

2. S. Pati and B. Pradhan, "Comparison between machine learning algorithms used for sentiment analysis," International Journal of Advanced Research in Engineering and Technology (IJARET), pp. 220-228, 2020.

3. R. Nyman, S. Kapadia and D. Tuckett, "News and narratives in financial systems: Exploiting big data for systemic risk assessment," Journal of Economic Dynamics and Control, vol. 127, pp. 104119, Jun. 2021.

4. P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, Dec. 2017.

5. L.K. Felizardo, F.C.L. Paiva, C. de Vita Graves, E.Y. Matsumoto, A.H.R. Costa, E. Del-Moral-Hernandez, et al., "Outperforming algorithmic trading reinforcement learning systems: A supervised approach to the cryptocurrency market," Expert Systems with Applications, vol. 202, pp. 117259, 2022.

6. C. Tudor and R Sova, "Flexible decision support system for algorithmic trading: Empirical application on crude oil markets," IEEE Access, vol. 10, pp. 9628-9644, 2022.

7. M. Dadhich and J. G. Lewis, "A Novel Approach of Feature Vector Design for Financial Information Extraction Using Supervised Learning," 2016 3rd International Conference on Soft Computing Machine Intelligence (ISCMI), pp. 115-119, 2016.

8. Jiahong Li, Hui Bu and Junjie Wu, "Sentiment-aware stock market prediction: A deep learning method," 2017 International Conference on Service Systems and Service Management, pp. 1-6, 2017.

9. J. Li, "Emotion Expression in Modern Literary Appreciation: An Emotion-Based Analysis," Front. Psychol., vol. 13, pp. 923482, Jun. 2022.

10. C. De Las Heras-Pedrosa, D. Rando-Cueto, C. Jambrino-Maldonado and F. J. Paniagua-Rojano, "Exploring the Social Media on the Communication Professionals in Public Health. Spanish Official Medical Colleges Case Study," IJERPH, vol. 17, no. 13, pp. 4859, Jul. 2020.

11. F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages, Kalamaki, 2011.

12. J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in Third IEEE International Conference on Data Mining, 2003 (ICDM 2003), Melbourne, Florida, USA, 2003.

13. S Liu, F Li, F Li, X Cheng and H Shen, "Adaptive co-training SVM for sentiment classification on tweets," International Conference on Information and Knowledge Management Proceedings, pp. 2079-2088, 2017.

14. SWK Chan and MWC Chong, "Sentiment analysis in financial texts," Decis Support Syst, vol. 94, pp. 53-64, 2013.

15. G. Forni and A. Mantovani, "COVID-19 vaccines: Where we stand and challenges ahead," Cell Death Differentiation, vol. 28, no. 2, pp. 626-639, Feb. 2021.

16. V. Chandrashekhar, "1.3 billion people. A 21-day lockdown. Can India curb the coronavirus?," Science, vol. 10, Mar. 2020.

17. C. Liu, W. Wang, Y. Zhang, Y. Dong, F. He and C Wu, "Predicting the Popularity of Online News Based on Multivariate Analysis," 2017 IEEE International Conference on Computer and Information Technology (CIT), 2017.

18. C.J Hutto and E.E Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," Eighth International Conference on Weblogs and Social Media (ICWSM-14), 2014, June 2014.

19. L. Luo, X. Ao, F. Pan, J. Wang, T. Zhao, N. Yu, et al., "Beyond polarity: interpretable financial sentiment analysis with hierarchical query-driven attention," Proc. Int. Joint Conf. Artif. Intell., pp. 4244-4250, 2018.

20. N. Xu, W. Mao and G. Chen, "A co-memory network for multimodal sentiment analysis," Proc. Int. ACM SIGIR Conf. Res. Development Inf. Retrieval, pp. 929-932, 2018.

Individual Contribution Report:

Sentiment analysis Using NLP

Shaswat Kumar

2106152

Abstract: The project aims to deals with the big challenge of analysing and getting useful information from the huge amount of text-based news data created in today's digital world. We have used NLTK toolkit in Natural Language Processing for identifying sentiments expressed in text. With these tools, we have created a solid approach to go through all this data and uncover the underlying opinions and feelings present in news stories.

Individual contribution and finding: As an individual contributor to our project, I took the initiative to explore various datasets, carefully examining each to determine the most suitable for our analysis. After a thorough pre-processing and filtering process, I selected 80 data points that best fit our criteria. I also dedicated significant time to mastering Natural Language Processing (NLP) fundamentals and the functionality of NLTK. This included learning tokenization, stemming, lemmatization, part-of-speech tagging, and sentiment analysis. They played a crucial role in preparing 80 subsets of the dataset, which involved cleaning text data, removing stopwords and quotations, and formatting text for NLTK compatibility. I was responsible for the visualization aspect of our project, creating four distinct graphs that illustrated the distribution of sentiments across the data points, which proved invaluable for our analysis and conclusions.

Individual contribution to project report preparation: I reviewed ten research papers on NLP and NLTK, analysing the author details, methodologies, and pros and cons of each paper. This informed the team's decision-making process. I also contributed to the report's editing and design, ensuring a coherent and impactful presentation of the team's findings. And at the end I ensured the report's originality by conducting rigorous plagiarism checks using specialized software.

Signature of the Supervisor **Signature of the Student**

Sentiment analysis Using NLP

Shruti Kumari

2106155

Abstract: The project aims to deals with the big challenge of analysing and getting useful information from the huge amount of text-based news data created in today's digital world. We have used NLTK toolkit in Natural Language Processing for identifying sentiments expressed in text. With these tools, we have created a solid approach to go through all this data and uncover the underlying opinions and feelings present in news stories.

Individual contribution and finding: As an individual contributor to our project, I made sentiment analysis on the 80 dataset using NLTK's sentiment analysis tools. They assigned sentiment scores to each data point, categorizing them into positive, negative, neutral, and compound classes based on the NLTK analysis.

In addition, Contributor B calculated the positive, negative, neutral and compound sentiment scores to provide a comprehensive view of the overall sentiment. For data analysis and visualization, they organized the results into tables, sorting them by sentiment values and for the visualization aspect of our project, creating four distinct graphs that illustrated the distribution of sentiments across the data points, which proved invaluable for our analysis and conclusions to enhance clarity and facilitate a deeper understanding of the sentiment trends within the dataset.

Individual contribution to project report preparation: I took the lead in authoring the abstract and the introductory section of the report, providing a succinct summary of the project's objectives, methods, and findings.

In addition to these tasks, I also prepared a concise and informative section of the report detailing the sentiment analysis results. This included the methodology, key findings, and insights derived from the analysis. Also, I utilized my knowledge of NLTK and NLP principles to examine the dataset and extract meaningful insights from the sentiment analysis

Signature of the Supervisor **Signature of the Student**