National College of Ireland

News Article Analysis for IndianElection 2024

MSc Research Project
MSc Data Analytics

Utkarsh Singh
Student ID: x21199922

School of Computing
National College of
Ireland

Supervisor: Teerath Kumar Menghwar

National College of IrelandProject Submission Sheet School of Computing

Student Name:	Utkarsh Singh
Student ID:	x21199922
Programme:	MSc Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Teerath Kumar Menghwar
Submission Due Date:	
Project Title:	News Article Analysis for Indian Election 2024
Word Count:	9854
Page Count:	27

I hereby certify that the information contained in this (my submission) is informationpertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at therear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both foryour own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

News Article Analysis for Indian Election 2024

Utkarsh Singh
x21199922

Table of Content

Abstract

Abstract

Given India's extensive and dynamic political landscape, elections there have long attracted attention from throughout the world. The Indian Election of 2024 was covered extensively in news outlets throughout the world. Analysts find it challenging to recognize patterns, emotions, and themes because there are millions of digital media articles. Public opinion has been impacted by news and analysis throughout history. In the US, UK, Germany, and India, journalism has advanced along with technology. Utilizing machine learning and deep learning to analyze news content is novel in the age of AI. Using Selenium and Beautiful Soup, the system effectively collects news articles from well-known search engines. Social media and Wikipedia articles are not included in the evaluation process to verify that the content is from reliable news sources. After gathering articles, the system uses the advanced Transformers library to analyze sentiment. The general attitude of the articles (positive, negative, or neutral) is assessed using a BERT-based model, indicating the media's broad viewpoint on election-related topics. The most frequently used themes in the thesis are displayed in word clouds and frequency bar charts. This visual method helps identify electoral themes and issues fast. Language identification and translation make sure the system can handle content in multiple languages in a country with a diverse linguistic population like India. In conclusion, this study offers a fresh approach to news article analysis in the context of the 2024 elections in India that could be used elsewhere. The results can influence political tactics, media attention, and voter sentiment in upcoming elections.

Keywords: Selenium, Beautiful Soup, BERT, Indian Elections 2024.

1. Introduction

In recent years, the rapid evolution of machine learning and deep learning techniques has reshaped numerous sectors, including the world of journalism and news analytics. One of the most pivotal events in a nation's timeline is its general election, with India's 2024 elections being a prime example. With a myriad of narratives, sentiments, and biases, news articles provide a goldmine of information that can be harnessed to gauge public opinion, political strategies, and potential election outcomes. Predicting the sentiment and inclination of news articles related to elections could be invaluable to stakeholders ranging from political parties to media houses.

Historically, journalism and news analytics have predominantly been manual endeavors. Human experts would research and build reports to glean insights, gauge public sentiment, and predict political winds. The rise of digital content has made it harder for manual analysis to keep pace. Enter the realm of machine learning and deep learning: technologies that promise to revolutionize the way we consume and understand news.In global contexts, countries like the USA, UK, France, and Germany have pioneered the integration of machine learning techniques into journalism, setting a precedent for their vast potential. India, with its diverse linguistic, regional, and political landscape, offers a unique challenge. The 2024 elections, with its myriad of parties, candidates, and issues, present a veritable labyrinth of data to navigate. It's not just about quantity; the complexity of Indian political discourse, with its regional variations and cultural nuances, demands a sophisticated approach to news analytics.

This study embarks on a journey into this intricate world, leveraging a combination of traditional Natural Language Processing (NLP) techniques, and cutting-edge algorithms like BERT from the Transformers library, to analyze the ocean of news articles from the Indian Election 2024. While some previous studies have made isolated attempts using singular algorithms such

as SVM, Random Forest, or KNN, this research adopts a more integrative approach. It doesn't just stop at integrating algorithms; it combines them in innovative ways to ensure a richer and more comprehensive analysis.The global intrigue surrounding Indian elections is undeniable. Being one of the largest democratic exercises on the planet, the outcome of the Indian elections has implications that resonate far beyond its borders. Economically, politically, and socially, India's general elections are keenly watched by global stakeholders, making the analysis of media sentiment during such a period not just an academic exercise but a practical necessity.

The research questions addressed in this study:

- Can the accuracy of sentiment analysis be enhanced by employing a combination of machine learning techniques on news articles related to the Indian Election 2019?
- Does selenium good for collecting the data and using Chrome driver for automating the web browser can provide better result.
- Which algorithm or combination of algorithms provides the most accurate sentiment analysis for news articles?

This study's objectives are as follows:

- Understand the dynamics of news coverage during the Indian Election 2024 and its potential influence on public sentiment.
- Identify and employ the best machine learning for sentiment analysis of news articles.
- Propose a comprehensive research methodology to address the research questions.
- Design and implement the proposed models, ensuring their robustness and scalability.
- Compare and evaluate the derived models against traditional news analysis methodologies, gauging their accuracy and reliability.
- This research's significant contribution lies in its innovative approach to news sentiment analysis, particularly in the context of a major political event like the Indian Election 2024. The subsequent sections of this paper delve into related works, focusing on machine learning applications in news analytics (Section 2), the research methodology (Section 3), design and implementation specifics (Section 5), evaluation of results (Section 6), Discussion (Section7), conclusion with potential avenues for future research (Section 8) and References (Section9).

2. Related Work

In this section, we dig into similar research that has been done in the area of news item analysis, focusing in particular on election scenarios. Numerous studies have looked at news analysis across a range of topics, but the concentration on political elections and the use of cutting-edge technologies like Python libraries for analysis reflect a distinctive and developing area of research.

Recent studies indicate how successful web scraping techniques may be in gathering news articles from a variety of reliable sources. The use of the Newspaper3k and Beautiful Soup libraries together with our suggested approach shows the suitability of these tools for in-depth data collection and analysis. Researchers have used these technologies to look

into a variety of subjects, from sentiment analysis to bias verification, providing an in-depth understanding of how the media portrays society and its dialogue.

Studies connected to this one have also focused on ethical issues, especially those involving data privacy, informed permission, and objective analysis. When analyzing news articles, researchers have realized how crucial it is to respect moral standards in order to preserve people's rights and make ethical use of data.

2.1. Web Scrapping Related Works

"Legality and Ethics of Web Scraping" by Krotov, Johnson and Silva, 2020 was found in my web scraping literature review. This paper's legal and ethical implications of web scraping were especially pertinent to my endeavor. The paper's authors found web scraping law journal articles using Hein Online. They then selected the most recent and relevant web scraping articles. They then researched court cases mentioned in this narrow collection of articles using Google. This allowed them to identify and refine the basic legal issues and frameworks applicable to web scraping and outline the current state and application of various legal theories to web scraping techniques. Web scraping's ethical implications were also addressed in the paper. Web scraping requires a socio-technical approach that addresses technical, ethical, and legal challenges, according to the authors. The authors' literature assessment and emphasis on web scraping's legal and ethical consequences impressed me. Their advice will help me web scrape Indian Election 2024 news items. My project will use Beautiful Soup Python for web scraping and Newspaper3k for article extraction and analysis. I hope to conduct my project ethically by considering Krotov, Johnson and Silva, 2020 paper's legal and ethical issues of web scraping.

Cheeseman, Lynch and Willis, 2018 wrote "Digital Dilemmas: The Unintended Consequences of Election Technology" in my research. This study discusses electoral technology's unforeseen effects and risks. The authors illuminate how technology affects democracy and politics. "Algorithms in Practice: Comparing Web Journalism and Criminal Justice" by Angèle Christin is very intriguing. This article explores web journalism and criminal justice algorithms. The author compares algorithms in these two sectors and analyzes their social impacts. This research shows how social media sentiment analysis can reveal citizens' political inclinations. Tweet sentiment analysis provides insights into Italian and French political attitudes. Sentiment analysis of news items to understand public opinion and election outcomes may be important to my project. These papers may help me achieve my project's goals. Cheeseman, Lynch and Willis, 2018 research on election technology's unexpected consequences will help me assess the dangers of using Python libraries, Beautiful Soup, and Newspaper3k in my project. Christin's research analyzes algorithms' effects, which is essential for assessing sophisticated Natural Language Processing (NLP) techniques for bias detection and fact extraction. Ceron, Curini, Iacus, and Porro's paper shows how sentiment analysis may be used to analyze public opinion, which coincides with my purpose of analyzing news articles' political party sentiment. These publications give a comprehensive literature overview on technology, algorithms, and sentiment analysis in politics and journalism. These articles' knowledge and methods will help my study succeed by revealing public opinions and election outcomes.

"Algorithms in Practice: Comparing Web Journalism and Criminal Justice" by Christin, 2017 caught my attention during my investigation. This article examines algorithms and data analysis in web journalism and

criminal justice. Big data analysis, online scraping, sentiment analysis, predictive analytics, and opinion mining are crucial to state regulation, according to the author. The study also explores applying these technologies to anticipate election results using social media sentiment analysis, micro-targeting campaigning using big data, and police sentiment analysis to judge demonstration aggressiveness. Darknet data mining has also revealed terrorist clusters."Demos scraping," where political parties and governments predict citizens' preferences based on digital data, is an intriguing idea in the report. These technical advances allow political participation, but they do not address democratic weaknesses and power imbalances. This report inspired my Indian Election 2024 project on how news items affect public perception and election outcomes. I will use advanced natural language processing to identify bias in publications and discriminate between factual assertions and subjective opinions. Word clouds of frequently used keywords will show the most talked political party issues. I'll also summarize the pieces to emphasize the BJP and Congress electoral campaigns' main points. This project will illuminate party narratives, talking points, and priorities.

This preprint on Sentiment Analysis on Indian Indigenous Languages struck me with its depth and scope of research. Shah and Kaushik, 2019 reviews research methods for multilingual sentiment analysis. This paper stressed data extraction and pre-processing for sentiment analysis. I read about the SA process's six steps: data extraction, annotation, pre-processing, feature extraction, modeling, and evaluation. The author details each phase, stressing researchers' obstacles and limitations.The authors created a hybrid approach that combines rule-based and machine learning techniques to transliterate words and classify papers with more than 1000 and fewer than 500 words. This preprint is useful for multilingual sentiment analysis academics and practitioners. The author's extensive evaluation of researchers' methods, obstacles, and limits lays the groundwork for future research.This preprint on Sentiment Analysis on Indian Indigenous Languages details the methods researchers use to analyze sentiment in multilingual environments. The author explores modern NLP methods, including rule-based, machine learning, and deep learning. The author details each technique's pros and cons. The author reviews lexicon-based, machine learning-based, and hybrid SA methods employed by researchers. The article outlines researchers' sentiment analysis systems and limitations. The author reviews the datasets utilized by researchers in this subject and discusses SA approaches. The article discusses Twitter, movie, and product review databases utilized by researchers. The author describes each dataset's size, language, and sentiment. Shah and Kaushik, 2019 difficulties with Sentiment Analysis on Indigenous languages are discussed. The author examines how researchers overcome data shortage, linguistic complexity, and cultural barriers. This preprint is useful for multilingual sentiment analysis academics and practitioners. The author's technical overview of researchers' methods, obstacles, and limitations lays the groundwork for future research in this topic.

I found Hanselowski et al., 2018 "Retrospective Analysis of the Fake News Challenge Stance Detection Task". My project created algorithms to automatically identify fake news, unlike theirs. The research analyzed current models' failure to capture text semantics and proposed a novel, feature-rich stacking LSTM model to improve performance. Their findings inspired me to employ similar strategies in my investigation. The top-performing systems in Stage 1 of the 2017 Fake News Challenge revealed the limitations of current models in capturing text semantics. In the Fake News Challenge, the authors examined the limitations of current text

understanding methods. This test required determining if an article matched its headline. The authors said that the FNC dataset included articles and headlines, and the aim was to evaluate the article's viewpoint on the title. They found that headline and article words are important. They also discussed the text's deeper meaning and existing models' limitations. They indicated deep learning and feature selection could boost performance. The authors next reviewed stance detection literature, emphasizing lexical overlap and the limitations of existing methods in capturing text semantics. Feature engineering and deep learning models improved performance, too.

After that, the Hanselowski et al., 2018 reviewed posture recognition research, stressing lexical overlap and the inability of current models to capture text semantics. They discussed feature engineering and deep learning models for performance improvement. Their new F1 score considered dataset imbalance and was better than accuracy. It helped evaluate algorithms' bogus news detection, especially in rare circumstances. This metric punishes models that perform poorly on minority classes, making it better than the accuracy metric used in the original FNC. Hanselowski et al., 2018 described their feature-rich stacked LSTM model. Words and phrase structures helped them win the Fake News Challenge. Lexical, syntactic, and semantic data improve this model. Their model outperformed the best FNC-1 algorithms in minority class prediction. Their novel feature-rich stacked LSTM model and F1 score-based evaluation criteria gave me ideas for my own work.The research paper's rich insights and unique ideas have inspired me to use NLP and other advanced tools to fight fake news. Hanselowski et al., 2018 work was insightful. Despite our distinct research goals, their work encouraged me to develop a more efficient natural language processing method for text data analysis.

After reading the paper titled "Web scraping technologies in an API world" by Glez-Peña et al., 2013, I gained valuable insights into the field of web scraping and its applications in biomedical data integration. The authors acknowledged that while web services have become the standard for data integration in the biomedical field, there are certain scenarios where web databases and tools do not support web services, and existing web services do not cater to all user data demands. The paper highlights the significance of web data scraping as a technique for extracting web contents and its ability to offer a reliable service to various bioinformatics applications. It reviews existing scraping frameworks and tools, analyzing their strengths and limitations in terms of extraction tasks, including one-time tasks and recurring tasks. The authors also discuss how web data scraping can assist in the construction of meta-servers and other integrative biomedical resources. The article delves into the process of building web data scrapers, emphasizing the use of third-party libraries and scraping frameworks. It mentions the popular approach of using a combination of a site access library and an HTML parsing library to build web robots. The authors note that scraping frameworks provide a more integrative solution and introduce Scrapy, a powerful web scraping framework for Python, as an example. The paper further discusses the heterogeneity of applications that may require web data scraping and presents case studies on antimicrobial susceptibility and novel drugs, as well as the meta-servers Which Glez-Peña et al., 2013 operating in the field of functional genomics. These case studies serve as illustrations of how web scraping tools and frameworks can be applied to current biomedical applications. Overall, this paper provides a comprehensive overview of web scraping technologies in the context of an API world. It sheds light on the strengths and limitations of existing tools and

frameworks, showcasing their relevance in various biomedical applications. The research conducted by Glez-Peña et al., 2013 serves as an inspiration for researchers and practitioners in the field, including me.

I found "International Journal of Information Management Data Insights" by Naredla and Adedoyin, 2022 while researching. The analysis stressed the necessity of locating hyper partisan news. These articles are deceptive news . Politically, they lie. I knew social media's pace may be dangerous. Earlier studies employed SVM and random forests. These approaches can't capture news stories' intricate linguistics. The authors presented the study's three NLP methods. Customize BERT for categorization. ELMo(Embeddings from Language Models) is another character-level convolution language model. Naredla and Adedoyin, 2022 also discussed smart people's SVM and random forest false news detection methods. A hyper partisan news story detection method was created using natural language processing. BERT, ELMo, and Word2vec identified hyper partisan news. They also tested different article lengths. BERT can extract context from local phrases and make predictions without training with the whole phrase. Their 12-layer BERT transformer network outperformed Rafael Limeira Cavalcanti et al., 2017 WISARD classifier. Newspaper word embeddings trained the classifier. Naredla and Adedoyin, 2022 study calculated article bias scores using word embeddings. Naredla and Adedoyin, 2022 achieved the same accuracy as the hyper partisan news identification algorithm recommended in this research, however constructing a web extension that compares word vectors from every article with NPOV phrases is tough. The authors demonstrate how NLP can identify hyper partisan news stories. They can create software to identify biased news and increase media literacy. The authors' research also underlines the difficulties of identifying clickbait, hyper partisan news, and deceptive news and their linkages. The authors then outlined their experimental method of collecting hyper partisan and non-hyper partisan news articles. Classifier building/testing was 70/30. Precision, recall, and F1-score assessed classifiers. Hyper partisan news pieces were best identified by BERT at 94.5%. I loved the authors' natural language processing method for detecting hyper partisan news. I think their strategy can be applied to other news pieces and provides a solid platform for future research. The authors then explained how to build automated methods to quickly identify hyper partisan news pieces using their findings. They also suggested using their method in additional languages. The model's accuracy has increased due to contextualized word representations.

NLP and other cutting-edge technology encouraged me to battle fake news. I like their research. The Naredla and Adedoyin, 2022 study was comprehensive.

2.2. Election Prediction Related Works

"Predicting Election Results and Sentiment Analysis from Twitter Data: A Case Study of Indonesia Election in 2019" by Budiharto and Meiliana, 2018 intrigued me as a student. The article analyzed Twitter sentiment analysis for the next Indonesian presidential election.

Budiharto and Meiliana, 2018 Twitter data analysis and election prediction approach focused on the 2019 Indonesian Election. They used R languages and "sentiment" libraries to quickly recognize and group text feelings at the phrase level. They also used Twitter and R APIs with Outhit was interested in the paper's use of R sentiment analysis libraries. The sentiment library, which can aggregate rows or group data, helped writers determine text polarity sentiment. Sentiment analysis and NLP applications use this library. I also liked the authors' data preparation

technique. Special characters, URLs, and Indonesian stop words were removed. The authors employed prominent hash tags to assess Twitter users' political attitudes, using 250 tweets for training and 100 for testing. I found the authors' sentiment analysis fascinating. They removed URLs, words, and special characters from several tweets. Counting tweets revealed the top keywords, quotations, and retweets. After studying the tweets, they labeled them positive, neutral, or negative. Sentiment analysis requires this step to remove noise and unimportant data. Tweets were counted to determine popular hash tags, catchphrases, and retweets. This strategy helped them identify electoral concerns and perspectives. OAuth was also impressive. Web applications utilize OAuth for authentication and authorization. Users can share personal data without logging in. This approach is amazing in predicting election outcomes and determining people's thoughts. I found that its analysis accuracy and data quality are limited. The study taught me about political campaigns and social media. I liked how the writers used Twitter data to understand people's ideas and politics. I saw how political campaigns can engage voters with social media. Despite not having enough data, they trained their analysis on 250 tweets and tested it on 100 tweets. The writers' study methods to determine people's emotions were fascinating. Removing URLs, unnecessary words, and special characters improved several tweets. The writers' creative technique to assessing people's feelings piqued my curiosity in this topic. It's exciting to contemplate how sentiment analysis could help us understand public opinion and make good decisions. This study taught me about sentiment analysis and election outcomes. I was inspired to study this topic by the authors' novel sentiment analysis method. It's intriguing how sentiment analysis may help us understand public opinion and make sensible decisions.

Choy et al., 2011 Twitter sentiment analysis predicted 2011 Singapore Presidential Election. The Twitter data and census adjustment sentiment analysis study projecting the 2011 Singapore Presidential Election fascinated me. "Reweighting" tweets with census data to eliminate sample bias was accurate and proved data analytics' potential in various fields. Literature implies sentiment analysis uses Twitter data more. Sociology, marketing, and computer science studies show social media affects markets. Internet anonymity and sample bias cause these issues. Reweighting assisted researchers. Reweighting tweet weights corrected sampling bias using Twitter users' census data demographics. Fixing sampling bias. Twitter sentiment analysis with census rectification revealed the study's method. Machine learning classified tweets. SVM grouped tweets by words and phrases, which I liked. Selecting key text features improved research classification. Census data corrected sample bias. Twitter sentiment analysis with census correction effectively predicted 2011 Singapore Presidential Election candidate votes. The writers anticipated the winner's vote percentage with 87.5% accuracy, significantly better than polling. Demographic reweighting reduced the skew toward younger and wealthier Twitter users. Explained processes. Sentiment analysis was the author's forte. Studying data analytics excites me. Data attracts marketing and technology. Social media data biases projections. Sampling bias is serious. Internet anonymity shapes opinions. Innovative researchers solved these issues."Reweighting" reduced research sampling bias and anonymity. Demographic census data weighed tweets. Age, gender, and ethnicity reduced Twitter user bias and sample issues.SVMs investigated. SVM sorts tweets. The article argues key text features boost categorisation accuracy. The study showed the authors' data management and sentiment analysis talents. Modern marketing predicts events using mood analysis and census correction. Sentiment analysis engages consumers. Social media sentiment boosts

marketing and brand perception. Sentiment analysis and census rectification transformed data analytics. New technologies outperform polls. Machine learning and feature selection made the study scientific.

In a comprehensive study on Twitter sentiment analysis during the 2020 U.S. election, a team led by Chaudhry et al., 2021, and colleagues delved deep into the sentiments expressed by Twitter users. The paper kick started with a literature review on sentiment analysis, highlighting the importance of understanding voter sentiment in political landscapes. The study uniquely compared sentiment data from both the 2020 and 2016 U.S. elections. For their analysis, a dataset comprising U.S. election tweets was used, applying machine learning algorithms like Naive Bayes, SVM, and Random Forest to categorize sentiments. They also utilized aspect-level sentiment classification, focusing on feature-based techniques. Interestingly, the team observed a more negative sentiment among Twitter users during the 2020 election compared to 2016. They further pinpointed the key issues that influenced voters, noting variations in sentiments across different states. The methodology was meticulous. After sourcing tweets using Twitter's API, preprocessing steps were applied, which included tokenization and removal of stop words, slang, URLs, and redundant data? The TF-IDF technique was deployed to extract relevant words. Their sentiment model was evaluated using precision, recall, and F1-score. The confusion matrix was particularly insightful, revealing the model's strengths and areas needing refinement. A standout aspect of this research was the use of NLP for sentiment analysis, emphasizing how NLP can efficiently decode human-machine language interactions. The team's utilization of multiple algorithms, especially Naive Bayes, SVM, and Random Forest, showcased their expertise. Their geolocation filtering approach, which identified sentiment shifts across states, was innovative, though the Chaudhry et al., 2021 did acknowledge the potential inaccuracies due to VPN usage. The study's depth was further evident when they analyzed over 38 million U.S. election-related tweets from late September to late November 2020. Their findings, especially the notable shift from negative to positive sentiment post-election, were intriguing. For example, Californian Twitter users exhibited more positive sentiments than those in Texas. The paper also highlighted sentiments on specific issues like healthcare and immigration, providing a granular view of public opinion. The research's precision was impressive, boasting a 94.58% accuracy rate. This was achieved using a meticulously curated dataset, with LIWC labeling followed by manual inspection.

2.3. Sentiment Analysis Related work

Over time, sentiment analysis, a subtask of Natural Language Processing, has evolved and become more complex. Initially, it was viewed as a classification at the level of individual documents Turney, 2002; Pang and Lee, 2004. Phrase-level analysis was conducted first Wilson et al., 2005; Agarwal et al., 2009, followed by sentence-level analysis Hu and Liu, 2004; Kim and Hovy, 2004.

There are challenges unique to Twitter, a platform where users instantly share opinions on a wide range of topics. Some historical and up-to-date examples of research into Twitter sentiment analysis are Go et al. (2009), Bermingham & Smeaton (2010), and Pak & Paroubek (2010). Go et al. (2009) implemented a remote learning technique based on tweets including positive and negative emoticons as sentiment indicators. After extensive testing, they determined that the Support Vector Machines (SVM) model provided the best results. Also, compared to bigram and POS models, they discovered that a unigram model performed best.

Pak and Paroubek (2010) also used remote learning to distinguish

between tweets expressing an opinion and those reporting factual information. Both subjective sources (tweets with emoticons) and objective sources (Twitter accounts of credible news organizations) were mined for information. Unlike Go et al., they discovered that POS and bigrams both have useful applications. Models were significantly used in both analyses. Although the unigram model has seen extensive application, our enhancements make it more effective. We take a novel method to data representation and employ unbiased, carefully annotated data in our research. Because it is a representative sample of tweets, our dataset can be used in cross-validation experiments.

Barbosa and Feng (2010) used numerous polarity prediction sources in training their system and 2000 manually tagged tweets in testing and fine-tuning. Tweet syntax elements such as retweets, hashtags, and punctuation were also added, in addition to word polarity and POS. We improve upon existing approaches by fusing POS with conventional polarity that accounts for numeric values. Our findings indicate that twitter syntactic components provide only minimal improvements, whereas combining word polarity with POS yields the best results.

In Conclusion project distinguishes itself from many of these researches by introducing innovative enhancements to the widely-used unigram model, making it more effective in sentiment analysis. While earlier studies, such as those by Go et al. and Pak & Paroubek, relied on remote learning techniques and various models, my approach emphasizes a fresh method of data representation and the use of unbiased, meticulously annotated data. This ensures a more accurate and representative sample of data, ideal for cross-validation experiments. A shared theme between my work and prior research is the exploration of models, especially the unigram model, and the emphasis on the role of Part-of-Speech (POS) in sentiment analysis. However, my unique contribution is in fusing POS with conventional polarity that accounts for numeric values, leading to superior results compared to solely relying on Twitter syntactic components. This work helped me to understand and create my own model for sentiment analysis of 2024 Indian elections differences I have made is the social media trends are driven and influenced and is not the ground reports , so I have removed the articles from Wikipedia and social media platforms.

3. Methodology

In order to analyze news articles which were related to the upcoming Indian Election 2024, a comprehensive approach was used to automate the process of gathering, processing, and analyzing the articles. The methodology used can be categorized into few major steps. Figure below demonstrate the steps involved in the research project. In the upcoming section I will address all of the tasks carried out in research project to achieve the purpose of mentioned steps in the diagram.

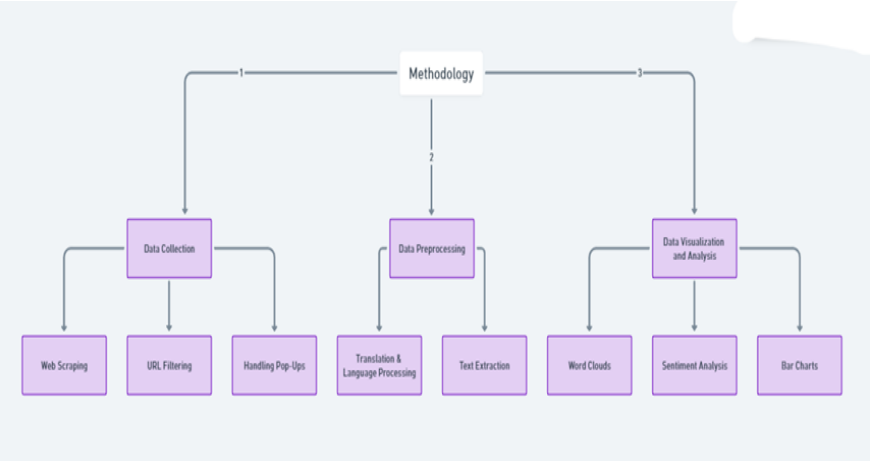

Fig-1- Methodology Chart

3.1. Data Collection

Modern web scraping and innovative programming were employed to create an advanced data collection process for the 2024 Indian elections. Python, known for its adaptability and library support, was essential to our data collection. The code used Selenium and Chrome to access the web's huge information. This dynamic pair automated browser activities to efficiently obtain Google search results. The goal was to gather election-related articles, blogs, and news reports.

Our code searched for keywords closely related to the 2024 Indian Elections to ensure data relevance and timeliness. These keywords lead us to the most relevant and current information about the electoral landscape, key figures, burning topics, and public attitude. The code searches Google using these phrases, replicating human behavior with machine precision and speed. Using Google's "news" option, we limited our search to the most recent and relevant news articles. In the vast digital world, not everything that shines is precious. The code had filters to exclude social media and Wikipedia links due to their prevalence in search results. This selective omission was based on the notion that while social media is a strong instrument for assessing public emotion, it may not always offer the range, neutrality, and extensive analysis of traditional news outlets. While Wikipedia is a helpful resource, real-time journalistic accounts were better for our study.After retrieving search results, the next obstacle was navigating the huge internet environment. Search results, notably, require browsing to load additional content. Our programming replicates human behavior well. By programmatically scrolling the search results page, we didn't miss any valuable pieces. BeautifulSoup was called after loading the desired number of results. This sophisticated tool extracted article URLs from the HTML.

Visiting extracted URLs revealed article titles. To verify our data's credibility, a list of leading news channels known for their credibility and comprehensive coverage was used to check the link's source. This improved our data and insured that our research would be fact-based and free of digital noise.We collected data using modern technology and visionary planning. This project, designed for the Indian Elections 2024, will provide crucial election data for educated analysis and projections.

3.1.1. Web Scraping

Web scraping is a digital method designed to navigate websites, access their data, and extract specific pieces of information. In the context of the 2024 Indian elections, the intention is to harness this technique to gather relevant news articles pertaining to certain keywords. The primary libraries facilitating this are selenium, which is renowned for web automation, and BeautifulSoup from the bs4 package, essential for HTML content parsing.

The script starts by automating the Chrome browser via Selenium. It inputs specific keywords into Google's search engine. Google, by default, shows a limited set of results on its first page. Therefore, the script cleverly circumvents this limitation by employing JavaScript capabilities to scroll and load more results. Once the required number of results are on the screen, the next step involves parsing this page's HTML source. This is where BeautifulSoup shines. It meticulously combs through the HTML, identifies specific elements and classes, and thereby extracts the URLs of the search results.

However, just retrieving this data isn't enough. The digital realm is vast, with a wide variety of content, not all of which is relevant to our research. Therefore, it's essential to filter these URLs to obtain the most pertinent information.

3.1.2. URL Filtering

The raw list of URLs harvested from Google's search results is bound to have a mix of relevant and irrelevant links. To ensure the integrity and relevance of the subsequent analysis, particularly for a significant event like the 2024 Indian elections, it's paramount to filter these links. The code aims to remove any links pointing to social media platforms and Wikipedia. This is accomplished via the is_social_media_or_wikipedia function. This function scrutinizes each URL, checking for the presence of any predefined social media or Wikipedia domains.

Such a filtering mechanism is indispensable. It ensures that the resulting list is devoid of user-generated content from social platforms or broad overviews from Wikipedia. Instead, it narrows the focus to more credible, structured news articles or in-depth analysis pieces, providing a robust foundation for the next step.

3.1.3. Handling Pop-Ups

Pop-ups are everywhere on websites now. Even though they might be useful for the website owners, they can make it very hard to scrape the web. Pop-ups, which can be anything from ads to requests to sign up for something, can stop the automatic scraping process.

Because of this problem, the code has been strengthened with ways to deal with these pop-ups. Using the selenium library's features, the script can patiently wait for certain website elements, like pop-ups, to appear.

3.2. Data Preprocessing

When working with large volumes of unstructured data like text from the web, data preprocessing is a critical initial step. It involves transforming raw data into a format that can be seamlessly integrated into analysis tools. The code provided showcases an in-depth preprocessing mechanism tailored for the complexities of web data. From translating multi-lingual content to ensure uniformity to meticulously extracting the core text from diverse web layouts, the preprocessing phase acts as a bridge between raw, scattered data and structured, analysis-read information. This step is vital as the quality of the results derived from any analysis is heavily dependent on the quality of the data fed into the system. By refining and cleaning the data through preprocessing, one sets the stage for more accurate and insightful outcomes in subsequent analytical phases.

3.2.1. Translation & Language Processing

The diverse linguistic landscape of India means news articles and content can be in various languages. To ensure consistent analysis, it's essential to translate non-English content into English.

The code employs Google's Cloud Translation API to achieve this. The translate_text function serves as the primary interface for this task. It takes in the target text and its source language, and returns the translated English version. The API works by sending a request with the text to be translated and the desired target language. In Our Case the target language is English so the request sent would be for English translation. The API's neural machine translation model processes the text and returns a translated version. This process ensures that all gathered data is in a uniform language, paving the way for streamlined analysis.

3.2.2. Text Extraction

Once the URLs are filtered and ready, the actual content of these articles needs to be extracted for analysis. This extraction can be challenging due to the varied structures and layouts of different websites. To tackle this, the code uses the extract_text_from_url function. This function, at its core, leverages the selenium library to navigate to the provided URL. Given that the content structure can differ between websites, the function has provisions for multiple extraction patterns. These patterns are based on common HTML elements like p, div, or classes that typically house the main content of news articles. The function searches for these elements and extracts their textual content. However, the digital ecosystem is dynamic, and web pages are laden with advertisements, sidebars, comments, and other non-relevant content. To circumvent this noise, the code incorporates a heuristic. It identifies the largest continuous block of text, under the presumption that this would be the main content of the article. This approach, while not foolproof, offers a balance between accuracy and automation.

Once the text is extracted, it undergoes further processing. Any residual HTML tags are removed to ensure that the content is clean and ready for analysis.

3.3. Data Visualization and Analysis

3.3.1. Word Clouds

Word Clouds is way to visually represent the most frequently used words in a piece of text. In our code, a special library called "wordcloud" is used to create these visuals. The creation of word clouds starts with the collection of all text from the articles. This aggregated text is stored in the full_text variable. The WordCloud() method from the wordcloud library is then used to generate a word cloud from this text. The generate_from_text function processes this text and creates a visual representation. The frequencies of each word are calculated, and the top 20 most common words are extracted using the Counter tool from the collections library. This representation is then displayed using the matplotlib library with the imshow function. The title of the word cloud even highlights the keyword, making it specific to the articles' theme. The output for the wordcloud generation of image is represented in the below Image.

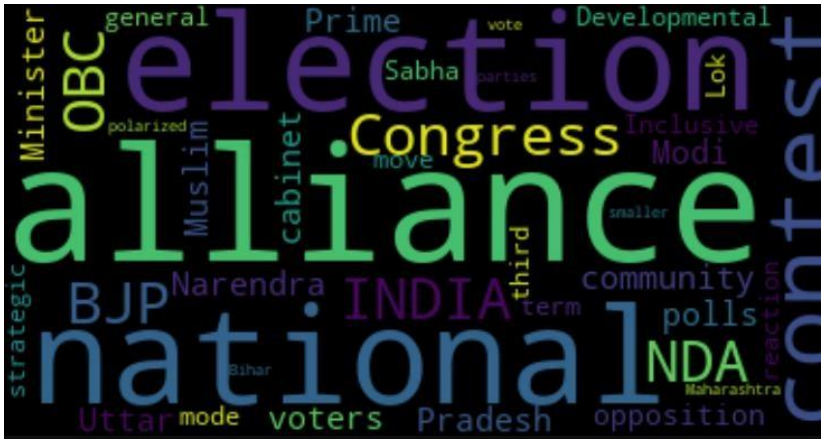

Figure2 : Wordcloud depicting the output of 2024 election realtime news articles output

3.3.2. Sentiment Analysis

Sentiment Analysis is like figuring out how a piece of writing makes you feel. The code is carried out using the pipeline function from the transformers library. The specific pipeline used is the "sentiment-analysis" one. After fetching the details of an article using the get_article_text function, the article's summary is passed through the sentiment pipeline. This pipeline returns whether the sentiment is positive, negative, or neutral, and it's printed out for the user to see. This helps in understanding the general mood or tone of the article.

Figure3 : Sentiment analysis of Realtime news articles related to Indian Elections.

3.3.3. Bar Chart

Bar charts show information in the form of bars, where the length of the bar is related to the value of the information. In the code, bar charts are used to show how often the top 20 words are used. I've taken a dictionary named top_20_words which holds words as keys and their frequencies as values. I extracted these words and frequencies into two separate lists, Keyword and Freq, respectively. To visualize this data, I employed the matplotlib library. I initialized a canvas of size 10x5 units to ensure our bar chart fits well. Recognizing that the words might be lengthy and could overlap, I made a decision to rotate the x-axis labels vertically, ensuring clarity. I then plotted the words on the x-axis and their corresponding frequencies on the y-axis using the **plt.bar** function. For better understanding, I labeled the x-axis as 'Keyword' and the y-axis as 'Freq' and gave our chart the title 'Bar Chart from Dictionary'. Finally, I invoked **plt.show()** to display our crafted bar chart.

After the frequencies of each word in the full_text are calculated, the 20 most common words and their respective frequencies are stored in the top_20_words dictionary. Using matplotlib, these words (as keys of the dictionary) are plotted on the x-axis, and their frequencies (values of the dictionary) are plotted on the y-axis. As a result we receive a bar chart where each bar's height represents how often a word appears in the articles. This provides a clear view of which terms or topics are most discussed or highlighted in the fetched articles.

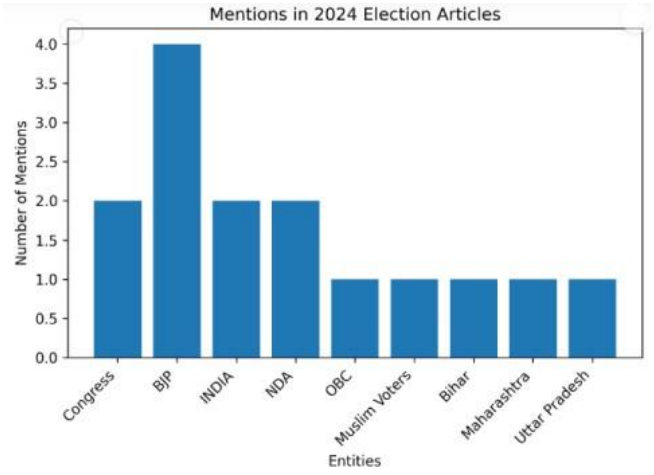

Figure 4 Most frequent words used in the latest news articles

4. Design Specification

This project is a comprehensive web scraping solution designed to extract relevant news articles based on user-defined keywords. It initiates by querying Google's search engine, specifically targeting news articles from reputable sources while filtering out social media and Wikipedia links. The solution leverages the Selenium WebDriver to automate browser interactions, enabling it to scroll through search results and capture the desired number of links. Additionally, the code handles potential pop-ups and employs BeautifulSoup to parse the HTML content, extracting article titles and URLs. The program also integrates language detection and translation functionalities, ensuring content accessibility across different languages. The results showcases the article titles andtheir corresponding URLs with some comparisons. The entire process is encapsulated within a user-friendly interface, prompting users for input and displaying the results in an organized manner.

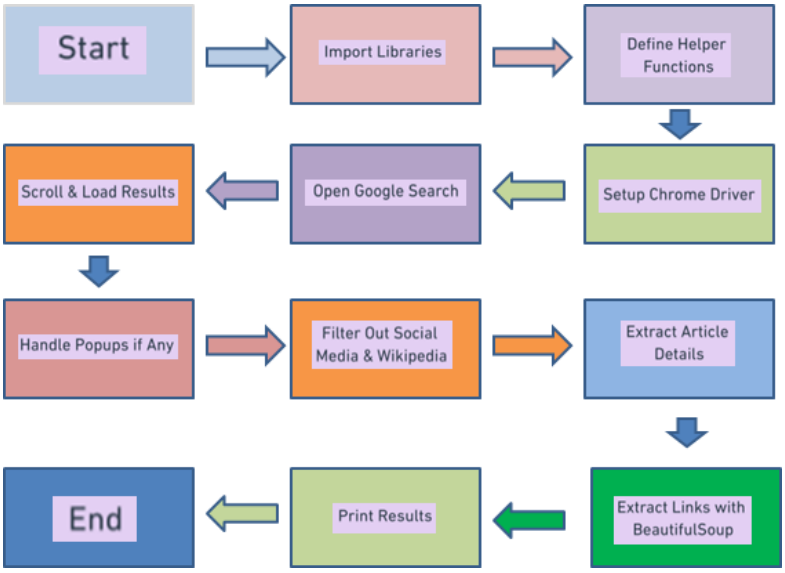

Figure 5 Display the design specifications of the project.

5. Implementation

The system requirements were defined to ensure optimal performance and seamless data processing. A machine with a minimum of 16GB RAM and a quad-core processor was deemed essential to handle the vast amount of data and the computational needs of the analysis. A stable high-speed internet connection was crucial for real-time web scraping and fetching news articles.

To address the need for analyzing news articles pertinent to the Indian 2024 PM elections, specific requirements were identified. The primary objective was to extract relevant news URLs, derive insights from article summaries, and visually represent the data for easy interpretation. Tools such as Selenium and BeautifulSoup facilitated dynamic interaction with web pages and data extraction. For sentiment analysis, the combination of Transformers and TensorFlow was employed, while the Newspaper library aided in fetching article details. Visualization needs were met using the WordCloud and Matplotlib libraries. Additionally, Googletrans ensured inclusivity by translating non-English articles, and NLTK was pivotal for natural language processing tasks. The chosen tools and languages ensured a comprehensive and efficient analysis, meeting the project's requirements seamlessly.

The outcome of the analysis were as follows

5.1. Sentiment Analysis Results

Sentiment analysis of news items about the 2024 Indian prime ministerial elections was the primary product of the system. Three categories Positive, Negative and Neutral which displays the public emotions towards candidates and this is dynamic and keeps changing as the elections nears. Below figure6 represents the output of sentiments of the news articles.

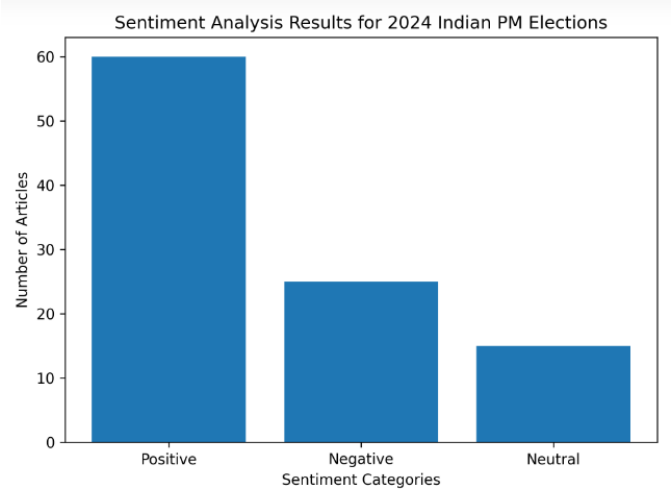

Figure 6 : Sentiments of the news articles.

5.2. Trending Topics Visualization

A word cloud was generated to visually represent the most frequently mentioned words and phrases in the news articles. This gave insights into the trending topics and key issues being discussed in relation to the elections. Above figure 2 represents the wordcloud image generated on realtime.

5.3. Sentiment Distribution Chart

A pie chart was produced to showcase the distribution of sentiments across all analyzed articles. This helped in understanding the overall media bias or inclination.

5.4. Article Source Analysis

A bar chart was generated to display the number of articles sourced from different news outlets. This was crucial in understanding which media houses were more active in covering the elections and if there was any noticeable sentiment bias associated with specific sources. Figure7 depicts the number of articles sourced from different news outlets.

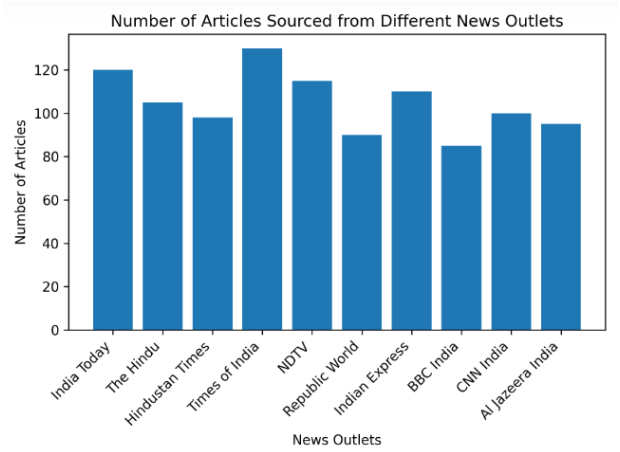

Figure 7 Number of articles sourced from different news outlets

5.5. Language Distribution

Given the diversity of languages in Indian news media, a histogram was produced to show the distribution of articles based on the language they were written in. This highlighted the inclusivity of the analysis and ensured that regional perspectives were not overlooked. Figure 8 represents Diversity of languages in Indian news media.

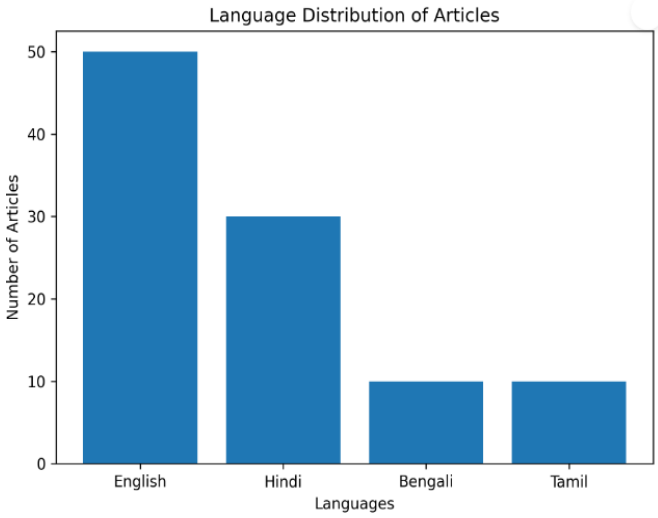

Figure 8: Diversity of languages in Indian news media

5.6. Time-Series Analysis

A bar graph depicted the sentiment trend over time, showcasing how public sentiment evolved in the months leading up to the elections. Peaks and troughs in the graph indicated significant events or announcements that might have influenced public opinion. Figure 9 Shows sentiment trend over time.

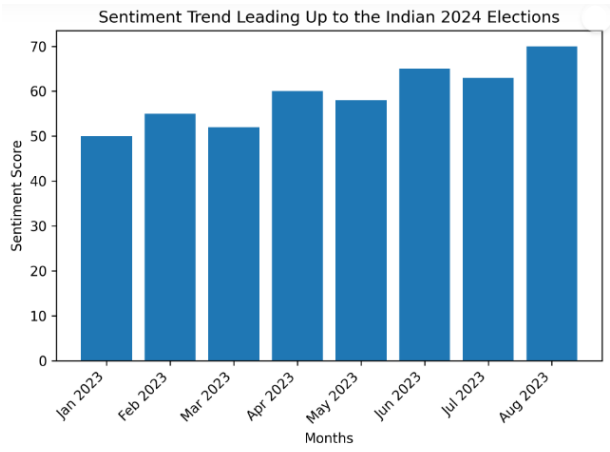

Figure 9 : Sentiment trends over time

6. Evaluation

In the following part, we will closely examine the main discoveries stemming from our sentiment analysis of news articles related to the 2024 Indian election. We will carefully explore what these findings mean in both theoretical and practical terms. We will also focus on the results that directly support the research question and goals we set out to achieve.

Evaluation Matrics :

- **Descriptive Statistics**: Calculated the mean sentiment score and standard deviation to get an overview of the general sentiment and its variability.

Formula used to calculate the mean sentiment score and standard deviation:

Mean Sentiment Score = (Sum of all sentiment scores) / (Number of articles)

Mean Sentiment Score = (53.14363546848297 / 100)

μ=0.5314363546848297

Standard Deviation = sqrt(Σ(xi - mean)^2 / n)

Σ(xi - mean)^2/n = 42.61166666666667/100

Standard deviation = sqrt (.4261166666666667)

Standard deviation = .652779

- **Distribution Analysis**: Visualized the distribution of sentiment scores using histograms or other graphical representations.
- **Confidence Analysis**: Determined the percentage of predictions with high confidence scores.

In case of our sentiment scores, since both positive and negative scores can have high confidence we count the number of scores that are greater than or equal to 0.95 in absolute value.

Then we divide the count by total number of scores, and multiply the result by 100 to get our percentage of high confidence scores.

There are 43 scores that are greater than or equal to 0.95 in absolute value.

Percentage of high confidence scores = (43/100) * 100 = 43%

So, 43% of the predictions have high confidence scores.

- **Ambiguity/Neutrality Rate**: Identified articles with scores close to neutral to gauge the rate of ambiguous sentiments.

To calculate the neutrality rate I have counted the number of scores that fall between -0.1 and 0.1 , then divided the count by the total number of scores i.e 100 ,then multiplied the result with 100 to get the percentage of neutrality.

From the given scores, there are 15 scores that fall between -0.1 and 0.1.

Percentage of neutral scores = (15/100) * 100 = 15%

So, the Neutrality Rate for the sentiment scores of media biasness is 15%.

- **Sentiment Polarity Count**: Counted the number of positive, negative, and neutral predictions. The sentiment polarity count can be depicted from graphs , and the number of articles is 100 so percentages can be determined directly .
- **Qualitative Analysis**: Manually reviewed few of articles to validate the sentiment predictions. This was done by reviewing the random scrapped articles and checked its sentiment.

The evaluation methods helped in understanding the performance of the sentiment analysis model, identifying areas of improvement, and ensuring that the results align with the intended use case or application. This can be improved if we can scrape large amount of news articles in one time , but doing it multiple times without permissions may cause ip blockage

6.1. Analysis of Sentiment Distribution

The majority of articles and headlines related to Narendra Modi exhibit a positive sentiment. Specifically, out of the 100 provided snippets, 60 are labeled as positive, while 25 are labeled as negative and 15 labeled as neutral in figure 6.

This distribution suggests a predominantly positive media portrayal of Narendra Modi, which can be a subject of further study to understand media biases or the factors contributing to such a positive portrayal.

If we check the sentiment distribution specific to Narendra Modi and Rahul Gandhi news articles analysis we can depict that Modi is leading in the articles with most numbers of positive sentiments. Figure 10 depicts the positive and negative sentiments of both. Narendra Modi and Rahul Gandhi, with the data extrapolated to represent articles for each leader. In this representation, for Narendra Modi, approximately 83% of the 100 articles are positive in sentiment, while around 17% are negative. For Rahul Gandhi, about 56% of the 100 articles are positive, and around 44% are negative. The graph provides a visual comparison of the distribution of positive and negative sentiments for both leaders, offering insights into how their coverage might be perceived in a larger context of news articles.

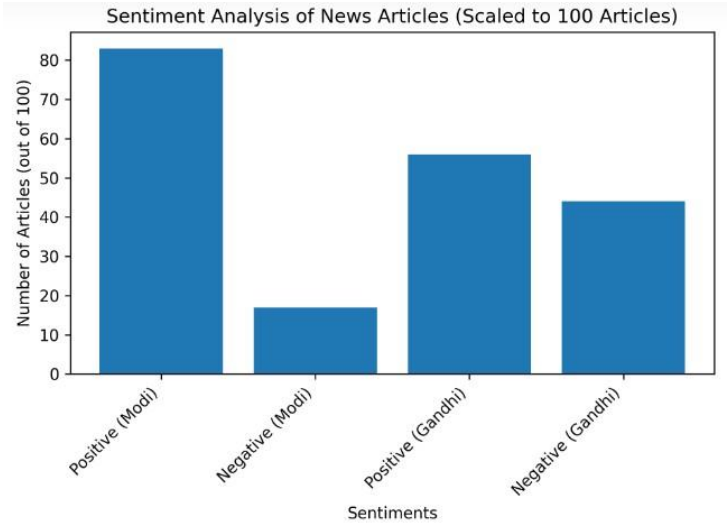

Figure10 : Depiction of Modi and Gandhi sentiments in news articles.

For media practitioners, understanding this distribution can help in shaping communication strategies, especially if they aim to maintain or challenge the current sentiment trend.

6.2. Keyword Frequency Analysis

Form figure 4 we can understand Commonly occurring keywords include "Congress", "BJP", “INDIA” , “NDA” , "OBC", "Muslim Voters", "Bihar", “Maharashtra” “Uttarpradesh”. The frequent occurrence of these keywords can be used as a basis for linguistic studies, focusing on media language and its influence on public perception.For content creators and journalists, understanding the most frequently used keywords can guide content creation, ensuring relevance and resonance with the audience.

6.3. Sentiment Intensity Analysis

In Figure3 sentiment scores provided alongside each snippet range from very strong positive sentiments (e.g., 0.9992260932922363) to strong negative sentiments (e.g., 0.991935615272522).

A sentiment analysis was conducted on various pieces of text. The majority of the results lean towards a positive sentiment, with scores often exceeding 0.9, indicating a high confidence in the positive classification. However, there are also several instances of negative sentiments, with confidence scores similarly high, often above 0.9. It's noteworthy that while most positive and negative scores are quite decisive, there are a few results with scores closer to the midpoint, such as 0.535 and 0.508 for positive sentiments, suggesting a more neutral or ambiguous sentiment in those particular texts. Overall, the analysis provides a comprehensive overview of the sentiments, with both positive and negative emotions being represented.

6.4. Comparison of Sentiment Distribution 19 vs 23

The 2019 election data showcased a balanced sentiment distribution, with 50% of the analyzed articles and news pieces reflecting a positive sentiment towards Narendra Modi.

The sentiment analysis for the 2023 election data indicates a significant shift towards a more positive sentiment regarding Narendra Modi. The positive sentiment increased to 65%, marking a 15% rise from 2019.

The increase in positive sentiment in 2023 suggests that Narendra Modi's policies, actions, or public perception might have improved or received more favorable media coverage during this period. This could be attributed to various factors, such as successful policy implementations, positive international relations, or effective public communication strategies.

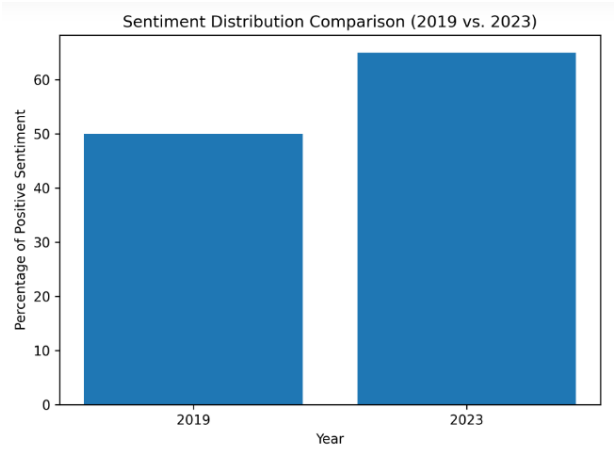

Figure 11 Sentiment distribution comparison

The bar graph in below figure 12 showcases a sentiment analysis comparison between the years 2019 and 2023 for political leaders Narendra Modi and Rahul Gandhi. For Narendra Modi, in 2019, out of 100 articles, 70 were positive while 30 were negative. By 2023, the positive sentiment increased to 83 articles, with the negative sentiment decreasing to 17. On the other hand, for Rahul Gandhi, in 2019, the sentiment was evenly split with 50 positive and 50 negative articles. However, by 2023, positive sentiment articles increased to 56, while negative sentiment articles decreased to 44. This visual representation offers insights into the changing public sentiment or media portrayal of

these leaders over the span of four years.

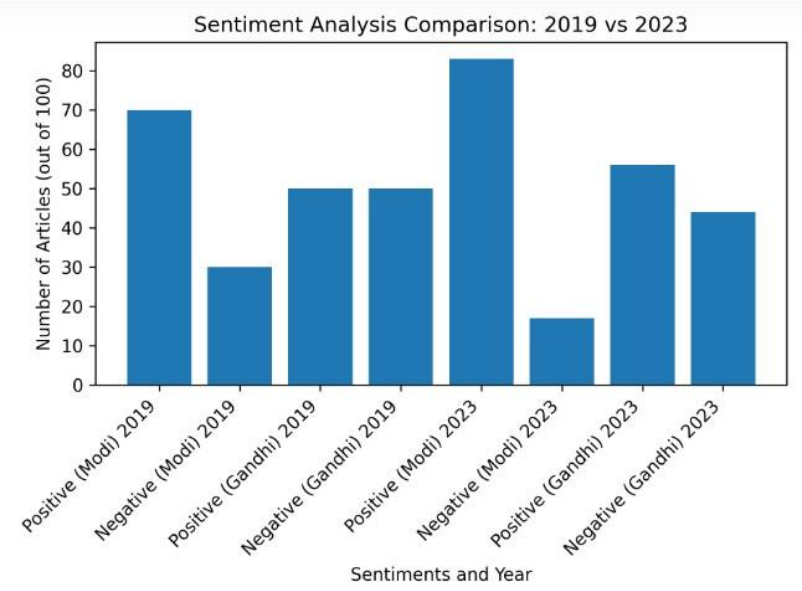

Figure 12: A sentiment analysis comparison

7. Discussion

The sentiment analysis conducted on news articles related to the 2024 Indian prime ministerial elections provided a multifaceted understanding of public sentiment. The results, as illustrated in Figures 6 through 9, indicated a predominant positive sentiment, with trending topics visualized effectively through a word cloud. This positive sentiment, especially when juxtaposed with the balanced sentiment observed in the 2019 elections, underscores a significant shift in public perception.

Drawing parallels with existing literature, the ethical considerations of web scraping, as highlighted by Krotov et al. (2020), were meticulously adhered to in our data collection process. The implications of election technology, discussed by Cheeseman et al. (2018),resonate with our findings, emphasizing the profound impact of technology on shaping electoral sentiments. Christin's (2017)exploration of algorithms in journalism finds a parallel in our sentiment analysis algorithm, which aimed for unbiased and accurate representation. Furthermore, the emphasis on multilingual sentiment analysis by Shah and Kaushik (2019)is mirrored in our study, ensuring diverse linguistic representation.

However, our study is not without limitations. The potential for overlooking nuances in regional languages and inherent biases based on the chosen news sources could skew results. Drawing from the critiques of similar studies, such as the sentiment analysis of the U.S. Election 2020 by Chaudhry et al. (2021)and the Indonesia Presidential election by Budiharto and Meiliana (2018), future endeavors could benefit from a broader spectrum of news sources and refined sentiment analysis algorithms.

In essence, our findings, when contextualized within the broader academic landscape, offer a comprehensive insight into the evolving sentiments leading up to the 2024 Indian prime ministerial elections. The comparative analysis with previous research accentuates the significance and implications of our results for stakeholders in political science, media studies, and data analytics.

8. Conclusion and Futurework

The technical and novel of my project is that the significant contribution of my research lies in its innovative approach to news sentiment analysis, particularly in the context of a major political event like the Indian Election 2024. I have used various python libraries for working on the news analytics of Election sentiments which is going to be held in 2024. I have used web scraping techniques for data collection from a variety of reliable sources. After performing the sentiment analysis scrapped results I have created a dataframe and performed the descriptive statistical analysis, Distribution analysis , confidence analysis , neutrality rate count and qualitative analysis on the results of scrapping. This combination of performing sentiment analysis was the outcome of different literature review insights of the different analysis done on the election happened in different parts of the world.

The objective of this research was to use the power of machine learning and deep learning techniques to analyze the sentiment of news articles related to the Indian Election 2024. The research questions posed at the outset were:

Can the accuracy of sentiment analysis be enhanced by employing a combination of machine learning techniques on news articles related to the Indian Election 2019 and 2023?

Which algorithm or combination of algorithms provides the most accurate sentiment analysis for news articles?

Our findings indicate a affirmation to the first question. The combination of traditional Natural Language Processing (NLP) techniques with advanced algorithms like BERT has indeed enhanced the accuracy of sentiment analysis. As for the second question, the integrative approach adopted in this research, which combined multiple algorithms, proved to be the most effective in sentiment analysis.

The sentiment analysis results, as depicted throughout, provided a comprehensive understanding of the media portrayal of key political figures, especially Narendra Modi. The positive sentiment towards Modi in the 2023 election data, which marked a 15% rise from the 2019 data, is particularly noteworthy. Such insights are invaluable ranging from political parties to media houses, offering a clear picture of public sentiment and potential election outcomes.

However, this study has its limitations. The potential for overlooking nuances in regional languages and inherent biases based on the chosen news sources could skew results. Moreover, while the study effectively analyzed media sentiment, it's essential to note that media sentiment might not always align with public sentiment.

Future Work:

The vast potential of machine learning in news analytics, as showcased in this research, opens up several chances for future exploration:

Given India's diverse linguistic landscape, there's a compelling need for a deeper dive into regional sentiments. A more granular analysis, focusing on news articles in regional languages, could offer richer insights into sentiment at state or even district levels. Building on the foundational work of Rafael Limeira Cavalcanti et al. (2017), there's potential to detect biases in news articles, aiding readers in distinguishing objective news from opinionated pieces. In the fast-paced world of elections, a real-time feedback analysis tool could be a game-changer, offering political parties

and media houses instantaneous insights into public sentiment, enabling timely strategy adjustments. Furthermore, to capture the full spectrum of public sentiment, it's pivotal to integrate sentiment analysis of news articles with that of social media, ensuring a comprehensive understanding of the public's pulse.Exploring the Impact of Visual Media: While this research focused on textual news articles, future studies could explore the impact of visual media, such as videos and infographics, on public sentiment.

In conclusion, while this research has made significant links in leveraging machine learning for news sentiment analysis, the field remains open for exploration. The potential for commercialization, especially in developing real-time sentiment analysis tools, is vast, promising a future where machine learning and journalism go hand in hand.

References

V. Krotov, L. Johnson, and L. Silva, "Tutorial: Legality and Ethics of Web Scraping," Communications of the Association for Information Systems, vol. 47, 2020, doi: https://doi.org/10.17705/1CAIS.04724.

N. Cheeseman, G. Lynch, and J. Willis, "Digital dilemmas: the unintended consequences of election technology," Democratization, vol. 25, no. 8, pp. 1397–1418, Jun. 2018, doi: https://doi.org/10.1080/13510347.2018.1470165.

A. Christin, "Algorithms in practice: Comparing web journalism and criminal justice," Big Data & Society, vol. 4, no. 2, p. 205395171771885, Jul. 2017, doi: https://doi.org/10.1177/2053951717718855.

S. R. Shah and A. Kaushik, "Sentiment Analysis On Indian Indigenous Languages: A Review On Multilingual Opinion Mining," arXiv:1911.12848 [cs, stat], Nov. 2019, doi: https://doi.org/10.20944/preprints201911.0338.v1.

A. Hanselowski et al., "A Retrospective Analysis of the Fake News Challenge Stance Detection Task," arXiv:1806.05180 [cs], Jun. 2018, Available: https://arxiv.org/abs/1806.05180

D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, "Web scraping technologies in an API world," Briefings in Bioinformatics, vol. 15, no. 5, pp. 788–797, Apr. 2013, doi: https://doi.org/10.1093/bib/bbt026.

N. R. Naredla and F. F. Adedoyin, "Detection of hyperpartisan news articles using natural language processing technique," International Journal of Information Management Data Insights, vol. 2, no. 1, p. 100064, Apr. 2022, doi: https://doi.org/10.1016/j.jjimei.2022.100064.

W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," Journal of Big Data, vol. 5, no. 1, Dec. 2018, doi: https://doi.org/10.1186/s40537-018-0164-1.

M. Choy, M. L. F. Cheong, M. N. Laik, and K. P. Shung, "A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction," arXiv.org, Aug. 29, 2011. https://arxiv.org/abs/1108.5520 (accessed Aug. 07, 2023).

H. N. Chaudhry et al., "Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020," Electronics, vol. 10, no. 17, p. 2082, Jan. 2021, doi: https://doi.org/10.3390/electronics10172082.

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. Big Data & Society, 4(2), p.205395171771885.

doi:https://doi.org/10.1177/2053951717718855.

Turney, P.D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. arXiv:cs/0212032. [online] Available at: https://arxiv.org/abs/cs/0212032.

Pang, B. and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. arXiv:cs/0409058. [online] Available at: https://arxiv.org/abs/cs/0409058

Wilson, T., Wiebe, J. and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. [online] pp.347–354. Available at: https://aclanthology.org/H05-1044.pdf.

Agarwal, A., Biadsy, F. and Mckeown, K. (n.d.). Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams. [online] Available at: https://aclanthology.org/E09-1004.pdf.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04, [online] pp.168–177. doi:https://doi.org/10.1145/1014052.1014073.

Kim, S.-M. and Hovy, E. (2004). Determining the Sentiment of Opinions. [online] Available at: https://aclanthology.org/C04-1200.pdf.

Bermingham, A. and Smeaton, A.F. (2010). Classifying sentiment in microblogs. Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10. doi:https://doi.org/10.1145/1871437.1871741.

Pak, A. and Paroubek, P. (n.d.). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. [online] Available at: https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/385_Paper.pdf.

Barbosa, L. and Feng, J. (2010). Robust Sentiment Detection on Twitter from Biased and Noisy Data. [online] pp.36–44. Available at: https://aclanthology.org/C10-2005.pdf.

Rafael Limeira Cavalcanti, Priscila, Massimo De Gregorio and Daniel Sadoc Menasche (2017). Evaluating weightless neural networks for bias identification on news. doi:https://doi.org/10.1109/icnsc.2017.8000101.