Sentiment Analysis for News Covered in Print Media

Twelve Month Training Report

Of

Sabudh Internship

Submitted by: Team

Members

Chinmai Thota

Sri Vyshnavi Perla

Meghana Gunnada

Asritha Boorlagadda

D-228, Third Floor, Sector 74,Sahibzada Ajit Singh Nagar,Punjab- 140307

CANDIDATE'S DECLARATION

We hereby certify that we have undergone twelve months industrial training at SABUDH FOUNDATION and worked on project entitled, "Sentiment Analysis for News Covered in Print Media", is an authentic record of our own work carried out during a period from July, 2023 to August, 2024 under the supervision of Shafila Bansal.

Team Members
Chinmai Thota
Meghana Gunnada
Sri Vyshnavi Perla
Asritha Boorlagadda

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Shafila Bansal

ACKNOWLEDGEMENT

I am highly grateful to Shafila Bansal and Sanchit Umate, Sabudh Foundation, Mohali, for providing an opportunity to carry out twelve months of training at Sabudh Foundation from July 2023-August 2024.

Shafila Bansal and Sanchit Umate have provided great help in carrying out my work and is acknowledged with reverential thanks. Without wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank Shafila Bansal and Sanchit Umate, for stimulating me from time to time. I would also like to thank the entire team of the Sabudh Foundation. I would also thank my friends who devoted their valuable time and helped me in all possible ways towards the successful completion of this training.

ABSTRACT

Sentiment Analysis for News Covered in Print Media," aims to analyze the sentiment of Indian political news articles. By leveraging Natural Language Processing (NLP) techniques, we assess whether the content is positive, negative, or neutral. We also identify which political party is the focus of the article and determine the sentiment towards that party

Keywords - *NLTK, VADER SENTIMENT*

LIST OF ABBREVIATIONS

Abbreviation	Full Form
NLTK	Natural Language Toolkit
NLP	Natural Language Processing

Contents

Chapter 1

Introduction to Organisation

Sabudh Foundation - An NGO that applies data science for social good. Sabudh Foundation is formed by the leading data scientists in the industry with the objective to bring together data and young data scientists to work on focused, collaborative projects for social benefit. Sabudh foundation is working on solving the real and high impact problems in areas such as education, governance, healthcare, and agriculture using Artificial Intelligence and Machine Learning techniques.

Data science can be used across a number of industries in order to be beneficial for the society. For example in agriculture, there are now Agro-bots and drones being used to gauge the health of the harvest that can help farmers improve their crop yield and reduce costs. With the help of advanced technologies, we're able to save 90% of the spraying costs. These technologies can help states like Punjab which has always been the food basket of India to rehabilitate food security while improving crop health.

The foundation has taken steps to involve Colleges, Universities, and Industry from the region for the social cause. Particularly, the foundation has signed academic and researchbased MoUs with Punjab University, Chandigarh, GNDEC, Ludhiana, BML Munjal University,
Punjab Government (Punjab Police), Punjabi University, Patiala, and Punjab Engineering College,
Chandigarh.

Chapter 2

Introduction to Project

2.1 Literature Review

- Antony Samuels, John Mcgonical 29 Nov, 2021

  News Sentiment Analysis
  https://drive.google.com/file/d/1-OQtyrlj5G2UTjVTWI_LvM4vClag3-7n/view?usp=drivesdk

The research paper focuses on the lexicon-based approach to sentiment analysis in news articles. It explores several key studies that have contributed to the field:

- **Lexicon-Based Sentiment Analysis**: The lexicon-based method relies on pre-defined dictionaries of words annotated with sentiment scores. Early studies, such as those by Turney (2002), emphasize the effectiveness of this approach in text classification, particularly in detecting sentiment polarity. Another key contribution is the work by Godbole et al. (2007), who implemented sentiment analysis on large-scale news and blog datasets using a lexicon-based approach, showcasing its scalability and effectiveness.
- **Challenges in Sentiment Analysis**: The literature identifies several challenges in sentiment analysis, including the ambiguity of words, context dependency, and the dynamic nature of language. These challenges often result in sentiment misclassification, which is a significant limitation of lexicon-based approaches.
- **Integration with Machine Learning**: While this paper primarily discusses lexicon-based methods, it acknowledges the growing trend of integrating these methods with machine learning algorithms to enhance accuracy. The combination of machine learning models with lexicon-based features is a notable direction in recent studies, aiming to leverage the strengths of both approaches.

- Jeelani Ahmed and Muqeem Ahmed , 6 Apr 2020
  A Framework for Sentiment Analysis of Online News Articles
  https://drive.google.com/file/d/10pe0lVqU78Ds95ms6TvzDvc-eyzld4yw/view?usp=sharing

This document delves deeper into machine learning methodologies applied to sentiment analysis. It reviews the evolution from traditional lexicon-based methods to more advanced machine learning models:

- **Supervised Learning Models**: The literature review highlights the use of supervised learning models, such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees, which have been widely adopted in sentiment analysis. These models are trained on labeled datasets to classify text into sentiment categories. Pang et al. (2002) were among the first to apply machine learning techniques to sentiment classification, showing significant improvement over rule-based methods.
- **Deep Learning Techniques**: The paper further explores the rise of deep learning in sentiment analysis. Neural networks, especially Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have shown promising results due to their ability to capture complex patterns and context in text data. The use of pre-trained models like Word2Vec and GloVe for embedding words into vectors has also been instrumental in improving the performance of sentiment analysis models.
- **Hybrid Models**: Recent studies have proposed hybrid models that combine lexicon-based approaches with machine learning algorithms. These models aim to improve sentiment classification by utilizing lexicon-based sentiment features as input to machine learning models. This hybrid approach addresses some limitations of purely lexicon-based methods, such as handling negations and capturing contextual information more effectively.
- **Challenges and Future Directions**: The literature also discusses the ongoing challenges in sentiment analysis using machine learning, such as handling sarcasm, detecting implicit sentiment, and the need for large labeled datasets. Future research directions include the development of more robust models that can handle multilingual sentiment analysis and the integration of external knowledge sources to improve model performance.

- Prateek Majumder , 29 Nov, 2021
  https://www.analyticsvidhya.com/blog/2021/11/web-scraping-a-news-article-and-performing-sentiment-analysis-using-nlp/

Chapter 3

Data-Set Description

3.1 Study Area

The study area for this project focuses on analyzing the sentiment of news articles related to the 2024 Indian elections, as covered in English-language print media. This analysis aims to identify the sentiment expressed towards various political parties and understand the overall tone of the news coverage. The study encompasses a broad range of print media sources available through the NewsAPI, which aggregates content from multiple reputable news outlets. The analysis will consider the textual content of these articles, focusing on how different political parties are portrayed in the media.

3.2 Explain Data Collection

Data collection for this project involved several key steps:

1. **API Integration**:
    - The NewsApiClient was used to access the NewsAPI, which aggregates news articles from various sources. The query was specifically tailored to fetch articles related to the 2024 Indian elections, filtered by language (English) and limited to the most relevant 100 articles using the page_size parameter.

2. **Fetching Article Content**:

    - The URLs of the articles fetched from the NewsAPI were processed using the requests library and BeautifulSoup to extract the full text of each article. This extraction focused on retrieving the main body content, usually found within <p> tags.

3. **Filtering Bad URLs**:

    - Some URLs may lead to inaccessible content or pages that do not contain the expected article text. A function was employed to test each URL, and those that resulted in errors or did not contain substantial content were removed from the dataset.

4. **Textual Data Processing**:

    - The nltk library was used for natural language processing tasks, such as tokenization, removing stopwords, and lemmatization, to prepare the text data for sentiment analysis. The VaderSentiment library was utilized to analyze the sentiment of each article, categorizing it as positive, negative, or neutral.

5. **Party Identification**:

    - A dictionary of keywords was created to identify mentions of political parties within the text. The spaCy NLP library was used to recognize these entities, ensuring that each article's sentiment analysis was correctly associated with the relevant political parties.

6. **Sentiment Analysis**:

    - Sentiment scores were calculated using the VaderSentiment library. The analysis was enhanced by weighting the sentiment scores according to the frequency and context of party mentions within each article.

7. **Data Export**:

    - The processed data, including sentiment scores, identified parties, and full article content, was saved in both Excel and CSV formats for further analysis and reporting. The openpyxl and pandas libraries facilitated this data export.

Chapter 4 Exploratory Data

Analysis

1.Summary of DataFrame Structure

Purpose: To understand the structure and data types of the DataFrame.

Description: The DataFrame summary shows the number of entries, columns, non-null counts, and data types of each column.

2. Checking Missing Values

Purpose: To identify missing values in the dataset.
Description: A heatmap is used to visualize where missing values exist in the DataFrame. data.isnull().sum() is used to calculate the number of missing values in each column.

3. Numerical and Categorical Summary

Purpose: To get statistical summaries of both numerical and categorical features.
Description: data.describe() provides statistical summaries (e.g., mean, count, std) for numerical columns. data.describe(include=['object']) provides summaries of categorical columns (e.g., count, unique, top, freq).

4. Visualizing the Distribution of the Target Variable

Purpose: To visualize the distribution of sentiment labels. **Description:** A bar plot is created using sns.countplot() to display the counts of different sentiment labels ('Senti'). The count of each label is shown on the respective bars.

5. Visualizing the Distribution of Key Features

Purpose: To analyze the distribution of sentiment scores.
Description: A histogram with a KDE (Kernel Density Estimate) overlay is plotted using sns.histplot() to visualize the distribution of sentiment_score.

6. Checking for Class Imbalance in the Target Variable

Purpose: To check for imbalance in the sentiment classes.
Description: A pie chart is created to visualize the distribution of sentiment classes ('Positive', 'Negative', 'Neutral'). The percentage of each class is shown on the pie slices. Additionally, the mean sentiment score for each sentiment class is calculated using data.groupby(['Senti'])['sentiment_score'].mean().

7. Handling Missing Values

Purpose: To handle missing values in the DataFrame.

Description: Several columns are filled with default values (e.g., "NA" for categorical columns). Missing sentiment_score values are replaced with the mean of the column, and missing values in

the Senti and party columns are filled with the mode of the respective columns. Rows with missing full_content values are dropped.

Chapter 5

Methodology

Initialization and Setup:

- Installed necessary Python libraries such as newsapi-python, pandas, beautifulsoup4, spacy, and vaderSentiment to handle API requests, data manipulation, web scraping, natural language processing, and sentiment analysis.

- Loaded the spaCy model (en_core_web_sm) for entity recognition and initialized the VADER sentiment analyzer.

- Configured the NewsApiClient with the provided API key to fetch news articles.

Fetching News Articles:

- Used the News API to retrieve articles related to the query "Indian elections 2024" in English. The response includes metadata such as the title, source, URL, and description of each article.

Web Scraping for Full Article Content:

- Developed a function fetch_article_content() to scrape the full content of each article using BeautifulSoup. The function extracts text from all <p> (paragraph) tags, ensuring the complete content is retrieved.

- Created a filtering mechanism to discard URLs that do not return valid content or trigger an error during the scraping process.

Filtering and Data Preparation:

- Filtered the DataFrame to remove rows where article content could not be retrieved (marked as "removed").

- Compiled a list of working URLs and bad requests for further reference.

Sentiment Analysis:

- Performed sentiment analysis on the full content of the articles using the VADER sentiment analyzer.

- The sentiment scores (compound scores) were calculated for each article and stored in the DataFrame.

- Categorized the sentiment as positive, negative, or neutral based on the calculated score.

Entity Recognition and Party Identification:

- Created a dictionary of political party keywords, including variations and aliases, for several major Indian political parties.

- Developed a function extract_parties() to identify political parties mentioned in the text using spaCy's named entity recognition (NER) and matching the detected organizations with the predefined party keywords.

Weighted Sentiment Scoring:

- Created a function determine_main_party_weighted() that calculates the sentiment for each sentence mentioning a political party. It assigns weights based on sentiment scores and the frequency of mentions to determine the most prominently discussed party in the article.

- The function returns the party with the highest weighted sentiment or the most frequently mentioned party if sentiment scores are neutral.

Data Filtering and Output:

- Filtered the DataFrame to exclude articles where no political party could be identified.

- Saved the filtered DataFrame, which includes full content, sentiment scores, identified parties, and sentiment labels, to an Excel file and a CSV file.

- Also saved the bad request URLs to a separate Excel file for reference.

Verification and Download:

- Verified that the generated Excel and CSV files exist and allowed users to download them.

- Previewed the first few rows of the saved CSV file to ensure correctness.

Final Output:

- The processed and analyzed news articles were saved to Excel and CSV files, containing information such as full content, sentiment analysis, identified political parties, and sentiment labels.

- A separate Excel file listed URLs that failed to fetch content.

Chapter 6

Conclusion and Future Scope

Total 214 News Articles are extracted from the News API Key and sentiment is predicted for all these 214 articles.

By the analysis of all these articles we have total 213 articles in which 180 are positive, 22 are negative and 11 articles are neutral.

The future scope of this project can be significantly expanded by incorporating more advanced natural language processing (NLP) techniques and real-time monitoring capabilities. One potential enhancement is the integration of deep learning models, such as BERT or GPT-based transformers, to improve the accuracy of sentiment analysis. The project could also be expanded to support multilingual sentiment analysis, especially for regional Indian languages, which would make it more inclusive and relevant to the diverse linguistic landscape of India. Additionally,

implementing real-time news monitoring and sentiment tracking could provide live insights into public opinion during critical political events, allowing stakeholders to react swiftly.

Beyond sentiment analysis, the project can delve into predictive analytics to forecast election outcomes based on historical trends in sentiment and political mentions. Integrating social media data, such as Twitter or Facebook posts, would allow for a broader analysis of public sentiment and the virality of political topics. Enhanced data visualization tools could create interactive dashboards for exploring trends, while entity co-occurrence analysis could provide insights into political alliances and issue centrality. Furthermore, bias detection across various media sources could be a valuable addition, helping to uncover potential influences in political reporting. These developments would transform the project into a comprehensive political sentiment analysis platform with applications in election forecasting, media analysis, and political strategy planning.