Feature Selection And Classification For Sentiment Analysis Of Newspaper Articles On Election Reviews

Ms. Divya Jaiswal [1*], Dr. Smita Suresh Daniel [2],

Dept of Computer Science, St. Thomas College Bhilai, C.G India

Abstract— The aim of this study is to explore the potential of sentiment analysis on election news article titles using machine learning techniques and to determine the most effective methods for text representation in this context. Traditionally, sentiment analysis has relied on part-of-speech tagging and word polarity counts, which work well in broad domains and when large labeled datasets are unavailable.

I. INTRODUCTION

Sentiment analysis is an emerging research area within the field of text mining. It plays a vital role in decision-making by extracting and identifying opinions, particularly from product reviews. However, the increasing volume of data necessitates the use of automated data analysis techniques. Information extraction focuses on identifying the writer's sentiments expressed through positive or negative comments. By analyzing large collections of documents, sentiment analysis seeks to determine the opinions or emotions associated with specific subjects. It leverages natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiments from typically unstructured text

This study explores the potential of sentiment analysis on election news article titles using an artificial neural network (ANN) and a support vector machine (SVM). The objective was to compare the results obtained from different feature sets and identify which features contribute most to improving classification accuracy, in line with the research problem. The results were evaluated using both accuracy and weighted metrics, as the dataset was imbalanced. An election news title comprises a wide range of features, such as individual words, their sequence, punctuation, quotation marks, total word count, and the number of words belonging to specific grammatical categories. This thesis aims to determine whether certain features have a greater impact on the classifier's performance. Sentiment-analyzed election news article titles can serve as valuable information for use in, for example, public opinion analysis or media monitoring.

.

However, in more specific domains with pre-labeled data, supervised learning methods are more suitable. This thesis evaluates the performance of a Convolutional Neural Network (CNN) and a Support Vector Machine (SVM) on various datasets, which were designed to capture different linguistic features.

The proposed system is implemented using Python 3.13.

The summary of this paper is as given.

- The extract election data are preprocessed using Natural Language Toolkit techniques.
- Extract the important news of the election
- In order to select the best features, frequency distribution is used. Based on these results, we implemented a multi- label classification algorithm to categorize them.
- The sentiments of each category was evaluated using polarity algorithm.

II. RELATED WORK

Sentiment analysis is a critical research area with a wide range of applications. It is particularly effective for analyzing text-based content, whether sourced from textual reviews or embedded in graphical data, to extract meaningful insights. This technique is extensively applied across various fields, including online platforms, email spam detection, political opinion analysis, and numerous other domains that require the interpretation of subjective information.

III. METHODOLOGY

This section outlines the methodology of the proposed study, as depicted in Figure 1. Initially, review documents were collected and preprocessed using basic natural language processing (NLP) techniques. The research method is explained in detail in this section. To achieve the objectives, a significant number of documents were gathered and analyzed. Relevant sections were then created to facilitate efficient sentiment analysis and to interpret specific index graphs. Subsequently, the Naïve Bayes Classifier was trained, and sentiment-based classification was performed. The goal of sentiment analysis was to classify the reviews based on their sentiment.

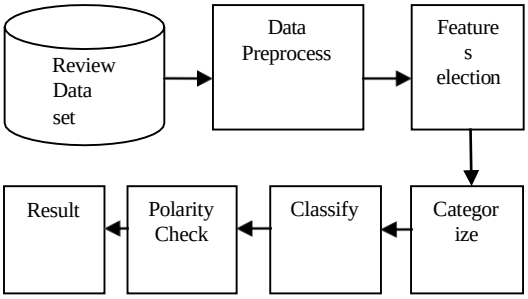

Figure1.Flow Diagram Of The System

IV. DATAPREPROCESSING

A. Data Acquisition

For the purpose of this study, we obtained our dataset from an election news website, focusing on reviews of the Kindle product. The dataset is in text file format and includes columns containing information such as the URL ID, URL link, positive score, negative score, polarity score, and more. However, to simplify the analysis, we selected only the review text for further processing. The dataset exhibited significant irregularities and diversity in the language used, along with considerable noise, making it a challenging task to utilize without thorough content analysis.

B. Data Pré-Processing

A. Each document contained a significant amount of informal language, sarcasm, acronyms, and misspellings, making the meaning often ambiguous and open to human interpretation. Faulty assumptions may arise if automatic algorithms are applied without a qualitative examination of the data. Text preprocessing refers to the process of refining the review text by removing unwanted words or phrases that contribute no meaningful information to the overall sentence structure.

From each of the review texts all punctuation except periods, apostrophes, and hyphens were removed. Users sometimes repeat letters in words so that to emphasize the words, for example, ‘greeeeat” toogooood”. Besides, common stop words such as “a, an, and, of, he, she, it”, non- letter symbols, and punctuation also bring noise to the text. So we pre-processed the texts before training the classifier:

1. We removed all the unwanted text from the reviews like repeated letters, stop words, non-letter symbols from the text texts.
2. Negative words are useful for detecting negative emotion and issues. So we substituted word sending within” and

other common negative words (e.g. Not, no, nothing, never, none, cannot) with“negk ok”.

3. Were moved all words that contain non-letter symbols and punctuation. This included there movalof@andhttplinks.
4. For repeating letters in words, our strategy was that when we detected two identic callettersrepeating, wekeptbothof them. If we detected more than two identical letters repeating, we replaced them with one letter. Therefore, “SOOO” were corrected to “so” .“much” was kept as “much” . Originally correct words such as “too” and “sleep” were kept as they were.

C. Tokenization

This process splits the text of a document into sequence of tokens. The splitting points are defined using all non letter characters. This results in tokens consisting of one single word (unigrams).

D. Pruning

There view data set was pruned to ignore the too frequent and too infrequent words. Absolute pruning scheme was used for the task. Two parameters were used for the pruning task namely, *prune below* and *prune above*. The value of these parameters was set as: *pruned below=5andprunedabove =200* i.e. ignoring the words that appear in less than 5 documents and in more than 200 documents.

E. Filtering tokens

Length based filtration scheme was applied for reducing the generated token set. The parameters used to filter out the tokens are the minimum length and maximum length. The parameters define the range for selecting the tokens. In the proposed model the minimum length was set to 4 characters and maximum length to 25 characters i.e. tokens with less than 4 characters and more than 25 characters were discarded.

F. Stemming

Stemming defines a technique that is used to find the root or stemofa word. The filtered token set undergoes stemming to reduce the length of words until a minimum length is reached. This resulted in reducing the different grammatical forms of a word to a single term. We used potter stemmer for the data set .

The general rules for dropping the endings from work.To optimize sentiment analysis for election reviews, we perform a series of preprocessing steps such as stopping, stemming, and lemmatization. These techniques are vital for enhancing the accuracy and effectiveness of text classification models. Here are the preprocessing rules applied to the election review dataset:

1. **If a word ends with "s", drop the "s".** *(Example: "votes" → "vote")*
2. **If a word ends with "Ing", delete "Ing" unless the remaining word consists of a single letter or ends in "the".**
   (Example: "voting" → "vote", but "thing" remains unchanged)
3. **If a word ends in a consonant (excluding "s") and is followed by "s", delete the "s".** *(Example: "issues" → "issue")*

Table1. Example of Preprocessed Election Review

Words and their stem.			
Review	**After Stopping**	**After Stemming**	**After Lemmatization**
Election results are important for our country's future.	Election result important country future.	Elect result import country future.	Election result importa nt country future.

Table 1 shows that stemming techniques reduce words to their root forms by removing suffixes such as "-ed," "-Ing," and "-s." This not only enhances the efficiency of sentiment classification but also contributes to a higher recall rate. By matching similar words, such as "debates" and "debate," or "voters" and "voter," stemming ensures that variations of the same word are treated as equivalent. This is especially beneficial when analyzing newspaper articles on election reviews, where similar sentiments may be expressed using different grammatical forms. Furthermore, stemming reduces the indexing size by approximately 40–50%, which is crucial when managing large volumes of text data, as commonly encountered in election reviews**.**

Po*larity Check*

The classifier achieved an accuracy of 76%. The pre-processed datasets were connected to the polarizer method, where each review was collected and split into sentences. Each sentence was tokenized and stored in a list, with each word classified as either positive or negative by the Naïve Bayes classifier. If the word is classified as positive, the variable **post** is incremented by 1; if negative, the variable **neg** is incremented by 1.

The sentence is then tested for the presence of the word "negtok". If this word exists in the list, everything after the word "not" is examined to determine whether the following words are positive or negative. This process considers the polarity of consecutive words, ultimately changing the sentiment orientation. For example, "screen is not good" is classified as negative, as the word "good" is negated by the word "not".

Thus, the polarity of each word is determined by the classifier, and the collective polarity is considered for each sentence. If the review contains more than one sentence, an average score is calculated by summing the total score and dividing it by the number of sentences in the review.

.

V. FEATURE EXTRACTION

In sentiment analysis of election reviews, features (mostly **nouns**) are extracted to identify key topics discussed. By analysing the frequency of these nouns, we gain insights into important election attributes. To improve the model's efficiency, we perform **feature selection** by removing irrelevant features with low frequency.

Common feature selection methods include:

1. **DF (Document Frequency)**: Measures how frequently a feature appears in documents. Higher frequency means higher relevance.
2. **IG (Information Gain)**: Assesses the ability of each feature to distinguish between sentiment categories (positive, negative, neutral).
3. **CHI (Chi-Square Test)**: Identifies features most strongly associated with sentiment categories.
4. **GR (Gini Index)**: Measures feature impurity, with lower values indicating more relevant features.
5. **Relief-F**: Evaluates the relevance of features based on how well they differentiate between sentiment categories.

A. Feature Selection Methods

- ***Document Frequency (DF)****: Measures how frequently a term appears across documents. Features appearing in too many or too few documents are considered non-informative.*
- ***TF-IDF****: Combines term frequency and inverse document frequency to give more weight to words that are important within individual documents but rare across the entire dataset.*

B. Selected Features

After applying feature selection, the following important features were identified for classification: "screen," "battery," "price," "weight," "library

A. AlgorithmforFeatureExtraction

B.

1. Input Dataset.
2. Generate New Dataset using preprocessing (removal of stopwords, punctuation, and noise).
3. Use POS tagger to extract only noun words into a word vector.
4. Construct a token set using extracted noun words.
5. Calculate frequency distribution using TF-IDF and identify top 10 most frequent nouns.
6. Filter the required features from this list.
7. Categorize reviews based on selected features.

VI. CLASSIFICATION

Once feature selection is done, the next step is **classification** of the sentiment in newspaper articles (positive, negative, or neutral).

Here are some **commonly used classifiers** for sentiment analysis:

1. **Naive Bayes Classifier:**
   - As mentioned earlier, Naive Bayes works well for text classification tasks. It is simple and fast, often yielding reasonable results in text classification, especially when features are independent of each other.
2. **Support Vector Machine (SVM):**
   - SVM is a popular choice for text classification problems, especially with high-dimensional data. It works by finding the optimal hyperplane that best separates different classes.
3. **Logistic Regression:**
   - This is a robust linear classifier, often used for binary sentiment classification (positive vs. negative), and can be extended to multiclass classification.
4. **Deep Learning (LSTM, CNNs):**
   - Long Short-Term Memory (LSTM) networks or Convolutional Neural Networks (CNNs) can be very effective for text classification tasks, especially when there is a need to capture sequential or contextual information in the text.

VII. RESULTS AND DISCUSSION

After preprocessing the test data file of newspaper articles and passing it through the **Feature Extraction** method, we identified the **10 most frequently used nouns** and their respective frequencies across various categories. The table below presents these findings

S.No.	Category	Words
1	Politics	election, government, policy, party, leader, vote, campaign, reform, opposition, mandate
2	Economy	market, growth, inflation, investment, economy, GDP, recession, trade, policy, sector
3	Health	disease, treatment, health, patient, hospital, doctor, vaccine, care, outbreak, research
4	Environment	climate, pollution, environment, carbon, emission, energy, policy, sustainability, green, ecosystem

andtheirtotalpolarity of each category was quantified whichisasshowninthe graph. we only focused on the above five attributes of the product kindle. The total polarity of each attribute was stored in variables. Then we add all positive score and all negative score and graphical representation of the result is shown below

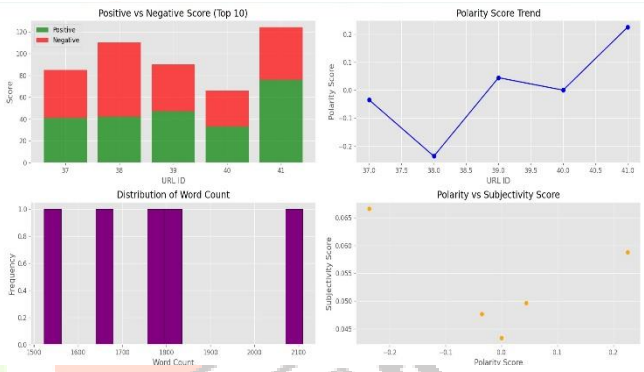

Figure2.Result of Graphical Analysis of polarity of each category

	URL_ID	URL	POSITIVE SCORE	NEGATIVE SCORE	POLARITY SCORE	SUBJECTIVITY SCORE	AVG SENTENCE	PERCENTAGE OF COMPLE	FOG INDEX	AVG NUMBER OF	COMPLEX WORD COUNT	WORD COUNT	SYLLABLE PER WORD	PERSONAL PRONOU	AVG WORD LENGTH
0	37	https://ti	41	44	-0.03529	0.047699	93.78947	8.080808	40.74811	93.78947	144	1782	1.544772	54	6.079125
1	38	https://ti	42	68	-0.28636	0.066626	75.04545	7.571169	33.04665	75.04545	125	1651	1.467378	53	6.00848
2	39	https://ti	47	43	0.044444	0.049614	90.7	7.22161	39.16864	90.7	131	1814	1.407085	59	5.992834
3	40	https://ti	33	33	0	0.043336	89.58824	7.879186	38.98697	89.58824	120	1523	1.448814	44	6.010506
4	41	https://ti	76	48	0.225806	0.058796	84.36	8.629682	37.19587	84.36	182	2109	1.463596	90	5.872926

Figure3 Result of Data Produced Analysis of polarity of each category

: Table2

From the above we can interpret the positivity and negativity about each of the features of the product kindle where we find that the battery is having higher negative value than positive , and the rest of the features have Higer positive values which shows that the users are happy with these features.

VIIICONCLUSIONANDFUTURESCOPE

Sentimental Analysis is an easier and cost-effective way to understand how the people are feeling about a particular subject of matter [5].. The Naive Bayes classifier used in the algorithm and polarity algorithm results depicted clearly the sentiment of the buyers and thus interpret the data.

Ithasbeenobservedthatthepre-processingofthedatagreatly affects quality of detected sentiments. We find the sentiments for each category of features separately. The polarity algorithm finds score of each words. Then sentiments are classified as positive, negative. The analysis for each attribute of the product results in finding what is people's opinion about the various product features which can be represented graphically in experimental results section. These results can guide the owners of the product to detect customer attitude and improve on the aspects that seem negative or is disliked by the targeted audience and can improve their online reputation.

ACKNOWLEDGMENT

I express my sincere gratitude to my guide Dr. Smita Suresh Daniel for her valuable guidance, insightful inputs, and whole-hearted cooperation throughout the development of this work.

REFERENCES

[1] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes and k-NN Classifiers," *International Journal of Information Engineering and Electronic Business*, vol. 4, pp. 54–62, 2016.

[2] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, 1997, pp. 412–420.

[3] **Khan, A., Zhang, H., Boudjellal, N., Ahmad, A., & Khan, M. (2023).** *Improving Sentiment Analysis in Election-Based Conversations on Twitter with Elec BERT Language Model.* **Computers, Materials & Continua, 76(3), 3345–3361.** DOI: 10.32604/cmc.2023.041520.

[4] **Fathir, F., Rizk, A., Yuliyanti, Y., & Mutmainah, S. (2023).** ***Comparative Sentiment Analysis of Election News Articles with SMOTE using Classification Algorithm.*** **Journal Kridatama Sains dan Teknologi, 6(2), 1253.**

[5] **Pradipta, N. Y., & Soetanto, H. (2024).** ***Sentiment Classification of General Election 2024 News Titles on Detik.com Online Media Website Using Multinomial Naive Bayes Method.*** **Journal of Applied Science, Engineering, Technology, and Education, 6(1), 43–55**

Authors Profile

Ms. Divya Jaiswal is currently pursuing a Master of Science in Information Technology (M.Sc. IT) at St. Thomas College, Bhilai, affiliated with Durg University, Chhattisgarh, India. She has a strong interest in Natural Language Processing (NLP), Sentiment Analysis, and Machine Learning. This research work on text classification and sentiment analysis is part of her final semester major project, focusing on practical applications of supervised learning models in the domain of textual data analytics.

Dr. Smita Suresh is working as Assistant Professor in Department of Computer Science at St. Thomas College, Bhilai, Hemchand Yaday University, Chhattisgarh, she is actively engaged in academic and research activities in the field of computer science.