Graphical Abstract

Sentiment analysis with text mining: a study from the Newspaper Contents of Bangladesh

Md. Habibur Rahman, Maisha Mumtaj

Highlights

Sentiment analysis with text mining: a study from the Newspaper Contents of Bangladesh

Md. Habibur Rahman, Maisha Mumtaj

- Research highlight 1
- Research highlight 2

Sentiment analysis with text mining: a study from the Newspaper Contents of Bangladesh

Md. Habibur Rahman[a], Maisha Mumtaj[a]

[a]*Department of Statistics and Data Science, Jahangirnagar University, , Dhaka, 1342, Bangladesh*

Abstract

This study applies sentiment analysis techniques via web scraping approaches to give a thorough examination of sentiment patterns in newspaper data from *The Daily Sun* and *Dhaka Tribune*. The study looked into the common themes and opinions conveyed in the articles using emotion plot visualization, k-means clustering, and classifiers like Random Forest and Naive Bayes. Though there were difficulties during the web scraping procedure, especially with parser selection, this study showed interesting trends in sentiment distribution. Both of the newspaper shows that the articles throughout the year have a neutral sentiment. The accuracy for the Random Forest and Naive Bayes is almost ninety-six and ninety-eight percent respectively for both newspapers. In the future, this effort will involve improving web scraping strategies, investigating sophisticated sentiment analysis tools, and broadening the study's scope to encompass additional newspapers. With the help of news media data, these projects hope to improve the predictive power of sentiment analysis models and deepen this study's understanding of the dynamics of public opinion.

Keywords: Sentiment, Text mining, Web Scraping, Clustering, Classification.

1. Introduction

People are social creatures. People must grow up in organized societies where one is taught the customs, laws, and regulations governing people's

existence and coexistence if people are ever to attain the status of what a person refers to as "human." A human being constantly molds one's behavior and attitudes—often unconsciously—based on these social norms, on the opinions of both the public and private spheres and on events that occur in the world around us. As part of one's daily rituals to learn more, comprehend the world around us, and assimilate into it, someone offers and accepts advice [1]. In addition, social norms that dictate what is generally accepted as appropriate behavior in a given setting also influence our emotional reactions and attitudes toward those circumstances [2, 3]. Newspapers are still essential in the digital age for the dissemination of information, reflecting society's values, and influencing public opinion. Researchers are now looking at the emotional content that is embedded in newspaper articles thanks to the development of sentiment analysis techniques. The study of extracting and interpreting sentiments, opinions, and emotions from textual data is the focus of the natural language processing sub-field of sentiment analysis. Understanding societal attitudes, emotional patterns, and their wider ramifications more deeply can be achieved by applying sentiment analysis to newspaper content. Newspapers frequently report on events that elicit strong feelings from readers, which reflects cultural norms and human emotions. Newspapers act as a reflection of societal problems, affecting public opinion and evoking shared feelings. Text's emotional tone and load are specified by sentiment analysis tools, which classify data as positive, negative, or neutral [4]. In addition to being avid readers, many people also enjoy leaving comments on various reading materials, including blogs, newspapers, magazines, and letters. Their remarks may be viewed as negative or occasionally impartial [5]. Sentiment analysis (SA) is a cognitive procedure that helps users understand and identify their feelings and emotions [6].

The field of study that examines people's opinions, sentiments, assessments, attitudes, and emotions toward entities and their attributes expressed in written text is called opinion mining, or sentiment analysis [7]. The term sentiment analysis falls under the purview of natural language processing and machine learning which is employed to extract, identify, or depict viewpoints from various content structures, such as news, reviews, and articles, and classifies those topics as favorable, impartial, or unfavorable [8]. Various terms and slightly different tasks are also present, such as subjectivity analysis, affect analysis, emotion analysis, review mining, sentiment analysis, opinion mining, opinion extraction, sentiment mining, and so on. But now, sentiment analysis and opinion mining cover the whole thing. Though opinion mining

and sentiment analysis are often used in academia, sentiment analysis is more commonly used in industry [9]. A key element of contemporary data analytics is sentiment analysis, which is essential for gauging a country's mood at momentous occasions like elections. Sentiment analysis can be found at three different levels: feature, document, and sentence levels. The goal is to categorize the opinion into positive and negative sentiments based on the sentence, document, or feature [10]. In business analysis, Sentiment analysis offers numerous benefits [11]. Businesses used the analysis's adaptation to gather public opinion to decide on their next course of action, and political parties used it to influence people through election strategies, newspapers, Twitter, Facebook, and blogs, in addition to applying it to book reviews, movie reviews, and product reviews [12]. A methodical approach to interpreting the complex web of emotions woven throughout journalistic content is provided by sentiment analysis, a rapidly developing field at the nexus of machine learning and natural language processing. Sentiment analysis, which uses computational methods to examine textual data, has the potential to reveal hidden attitudes, convictions, and sentiments that are expressed in newspaper articles. The amount of research on sentiment analysis has increased significantly in recent years, primarily focusing on extremely subjective text types (such as movie or product reviews). These texts differ from news articles primarily in that their purpose is made clear and consistent throughout the text [13]. Subjectivity analysis is the "linguistic expression of somebody's opinions, sentiments, emotions, evaluations, beliefs, and speculations." [14]. The linguist Ann Ban Field provided inspiration for the author's definition. Ban field is defined as subjective "sentences that take a character's point of view" and "that present private states"— that is, states that are closed off to objective observation or verification—of an experience holding an optional attitude toward an object [15]. Movie review classification has been a main test-bed task for sentiment classification. Reviews provide a challenging and fascinating sentiment analysis test case. Plot summaries and other unrelated, possibly misleading text are abundant, and opinions are expressed in a variety of sophisticated ways, including sarcasm and metaphor [16]. Reviews of newspaper articles can therefore address three sub-tasks: accurately identifying the target, separating the positive and negative content from the reviews on the relevant target, and thoroughly assessing the various viewpoints expressed [17]. Sentiment analysis vocation is to classify people's opinions into specific categories to facilitate understanding the behind phenomenon [18]. Numerous approaches to classification exist; some focus solely on positive

versus negative classes, while others address a larger number of more significant classes [19, 20, 21, 22]. SA is divided into three primary classification levels: aspect level, sentence level, and document level. Positive or negative opinions or sentiments are intended to be expressed in an opinion document; this is the goal of document-level SA. The document is viewed as a single-topic basic information unit. Sentiment classification in sentences is the goal of sentence-level SA. Whether a sentence is subjective or objective must be determined first. Sentence-level SA will tell us whether the sentence conveys opinions that are positive or negative if it is subjective [23]. Sentiment analysis has been used extensively in a variety of languages and situations [24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38]. But its application to newspapers allows stakeholders to assess how well journalism shapes public opinion [39]. Large amounts of natural language data can be processed by sentiment analysis models and tools to extract underlying sentiments. These techniques, which include those outlined by Chapelle and Chung [40], combine advanced natural language processing (NLP) and machine learning techniques. This strategic significance is emphasized even more in industries like healthcare [41, 42, 43, 44], where sentiment analysis of social media and other text sources is used to determine consumer and public opinion about medical goods and services. Sentiment analysis, though, faces many obstacles in spite of its potential, such as the need to accurately interpret context, sarcasm, and the nuances of human emotion [45, 46]. It requires that the data be prepossessed carefully, that the right analytical techniques be chosen, and that the outcomes be rigorously assessed. Academic studies on the efficacy of diverse sentiment analysis techniques persist, contrasting the precision of disparate algorithms and models [47]. While some techniques, such as support vector machines, may provide high accuracy, a thorough analysis of the body of research indicates that every algorithm has advantages and uses in the larger context of sentiment analysis [47]. In this study, sentiment analysis is applied to Daily Sun and Dhaka Tribune newspaper data, recognizing the influence of editorial narratives on public sentiment and the feedback loop between public opinion and media coverage. Considering this periodic impact, the research closely monitors methodological rigor, following guidelines for systematic reviews and guaranteeing a thorough analysis of sentiments. The objectives of this study are - to obtain the sentiment through text mining with the help of web scraping, to determine the cluster of each sentiment, to find the most frequent words in each newspaper, and to classify the sentiment scores by using machine learning techniques.

For any Statistical analysis, relevant data is essential to extract new information and findings. Data will be collected from the leading newspapers: The Daily Sun (https://www.daily-sun.com/), and Dhaka Tribune (https://www.dhakatribune.com/) for the period of 1st January 2023 to December 2023.

2. Methods and Methodology

Sentiment analysis has become a vital tool for comprehending public opinion and perception of a wide range of things, from political events and social issues to goods and services. With this methodology, this study offer a thorough process for performing sentiment analysis on articles taken from the Dhaka Tribune and the Daily Sun, two well-known newspapers. Using web scraping techniques, the study collect textual data from the online platforms of these newspapers and use clustering algorithms to analyze and classify the sentiments expressed in the articles.

2.1. Data collection from web scrapping

The first step of our methodology involves web scraping, a process of automatically extracting information from websites. The process of automatically gathering both structured and unstructured data is known as web scraping. Web data extraction and web data scraping are other common names for it. Web scraping has many applications, some of which include price monitoring, price intelligence, news monitoring, lead generation, and market research. In this research work, the data was collected from web scraping using the Python programming language. Spyder from Anaconda Navigator is used for running the program. This study leverages Python libraries such as BeautifulSoup and Scrapy to retrieve articles from the online archives of the Daily Sun and the Dhaka Tribune. By specifying relevant search queries or categories, this study target articles covering diverse topics, including politics, economy, sports, entertainment, and social issues. The extracted textual data undergoes prepossessing to remove noise, such as HTML tags and advertisements, ensuring the quality and integrity of the dataset.

2.2. Text preprocessing

For analyzing the sentiment of the data, this first needs to preprocess and clean the text of the article. The punctuation marks, space, dot lines, symbols, and other non-alphabetic marks are cleaned from the text. After

cleaning and processing the text of the articles, the words from The Daily Sun and Dhaka Tribune have been shown in table 1

3. Data

Table 1: Words collected from Dhaka Tribune and The Daily Sun From 1 January to 31 December 2023.

Dhaka Tribune and The Daily Sun	
Words Count (Dhaka Tribune)	Words Count (The Daily Sun)
8377839	8477550

Before going to sentiment analysis, text preprocessing is a must. For cleaning and preparing textual data for analysis or modeling, text preprocessing is an essential step in natural language processing (NLP) tasks. In this research project, Python libraries, including SpaCy, scikit-learn, and NLTK (Natural Language Toolkit) are used for text preprocessing tools. Several steps are usually involved in the process, such as tokenization, lowercasing, punctuation, stop words, special characters, and stemming or lemmatization to return words to their base forms. While lowercasing guarantees consistency in word representation, tokenization divides the text into individual words or tokens. Eliminating stop words, punctuation, and special characters can help lower data noise. To improve comprehension and analysis, lemmatization and stemming also seek to reduce words to their most basic forms.

3.1. Sentiment analysis technique

The TextBlob library scripted in Python was utilized to compute the sentiment score for each article. This tool provided a compound score indicating the overall sentiment and individual scores for positive, negative, and neutral. The analysis allows us to categorize articles into sentiment scores and word count of the newspaper data. For The Daily Sun and Dhaka Tribune, the sentiment is divided into three categories: Positive, Negative, and Neutral. The words are arranged in these three categories to see the emotional tone of the articles.

3.2. Classification

Three machine learning techniques are used to build the model. Naive Bayes classifier, k-means classifier, and Random forest classifier. The model was constructed using these three classification algorithms, which were also used to compare each model's performance using various metrics like precision, recall, and F_1 score to determine performance and identify the top classification model.

1. k-Means: This partitioning algorithm divides data into 'k' clusters by repeatedly allocating each data point to the closest cluster centroid and then modifies the centroid according to the average of the data points within each cluster.

2. Naive Bayes: A probabilistic classifier predicated on the feature independence assumption given the class and based on Bayes' theorem. gives a set of features, determines the probability of each class, and chooses the class with the highest probability.

3. Random Forest: An ensemble learning technique that constructs several decision trees during training and combines those decisions to produce a prediction that is more reliable and accurate. A random subset of the training data and features is used to train each decision tree, with the mode of all the predictions being used as the final prediction.

3.3. Perfomance matrics

Evaluation metrics are crucial for figuring out how accurate the classification is. A classifier's precision on a test dataset is the percentage of datasets that the classifier correctly classifies. Since the accuracy metric is not useful for assessing the effectiveness of the classifier in text mining, this study searched for other metrics. Precision, Accuracy, and F-measure are the three metrics that are frequently employed. For building the confusion matrix, a few things need to be looked up :

1. *True Positive (TP):* The number of correctly predicted positive instances (i.e., instances that fall into the positive class) is known as the True Positive (TP).

2. *True Negative (TN):*The number of correctly predicted negative instances (i.e., instances that fall into the negative class) is known as the True Negative (TN).

3. *False Positive (FP):* Also referred to as Type I error, it is the number of cases that fall into the negative class but were mistakenly predicted as positive.

4. *False Negative (FN):* Also referred to as Type II error, it is the number of cases that fall into the positive class but were mistakenly predicted as negative.

3.3.1. Precision

Precision in a decision matrix usually refers to the specificity or correctness of the criteria applied while assessing options or alternatives. A systematic method for comparing options according to several criteria or factors is a decision matrix. Every criterion has a weight or relevance ascribed to it, and these criteria are used to compare each alternative. Making sure the criteria in a decision matrix are clear, quantifiable, and pertinent to the choice at hand is necessary to ensure precision. This guarantees the impartiality and consistency of the evaluation procedure. For instance, precision would entail precisely identifying criteria like cost, quality, dependability, and customer service and making sure that these factors can be measured or evaluated impartially in a decision matrix for selecting a new supplier. The Formula for Precision is :

$$Precision = \frac{TP}{TP + FP}$$

3.3.2. Recall

When used in relation to a decision matrix, recall usually denotes how thorough the review process was, particularly in terms of how well the matrix captured all pertinent criteria and options. It makes sure that no important consideration is missed when making decisions. When a decision matrix has a high recall, it contains all relevant criteria and options required to make a thorough and educated choice. Thus, it is important to keep strong recall to guarantee that the decision matrix appropriately depicts the decision-making environment and permits decision-makers to take into account all relevant elements prior to making a decision.

$$Recall = \frac{TP}{TP + FN}$$

3.3.3. Accuracy

A decision matrix's accuracy is crucial for guaranteeing that the assessment process accurately captures the decisions' actual priorities and results. To evaluate alternatives against these criteria, it is necessary to define pertinent criteria in detail, assign weights that appropriately reflect their importance, use trustworthy data, ensure consistency and reliability in evaluations, and, in the end, give decision-makers trustworthy estimates of the outcomes connected to each alternative. Through maintaining precision, a decision matrix helps make well-informed decisions, reducing biases and uncertainties and producing better results. The Formula for accuracy is :

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

3.3.4. F_1-Score

The F_1 score is a metric that integrates recall and precision into one that is frequently used in machine learning and classification tasks. It offers a balance between the two and sheds light on a model or system's overall performance. The precision and recall harmonic means are used to compute the F_1 score. Given that it takes into account both erroneous positives (precision) and false negatives (recall), it is especially helpful in cases when there is an imbalance between the classes being predicted. A high F_1 score means that the model is efficiently recognising the pertinent occurrences while reducing false positives and false negatives. It also means that the model has high precision and good recall.

$$F_1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Using web scraping and clustering techniques, this methodology offers a systematic framework for conducting sentiment analysis on textual data extracted from the Daily Sun and the Dhaka Tribune newspapers. This study hopes to learn more about the dynamics of public opinion and spot noteworthy trends and issues by examining sentiment patterns found in the articles. The results obtained from this analysis can assist a range of stakeholders, such as businesses, journalists, and policymakers, in understanding the pulse of society and making well-informed decisions.

4. Results and Discussion

News distribution has changed dramatically in the rapidly developing field of digital media, radically altering the way opinions and public debates are developed. Newspapers remain influential platforms for the exchange of information throughout this transition, exerting significant influence over political discourse and social narratives. Publications such as the Daily Sun and Dhaka Tribune hold great sway over public opinion and play a major role in shaping national discourse in Bangladesh, a country that boasts a diverse media landscape. Using cutting-edge web scraping techniques, this research project collects a sizable dataset in a predetermined amount of time by conducting a thorough analysis of the opinions stated in news articles taken from these major newspapers. By applying Natural Language Processing (NLP), this research attempts to analyze the dominant emotions expressed in the gathered articles, providing detailed insights into the emotional terrain of news discourse. The study aims to reveal underlying trends, biases, and editorial viewpoints by closely examining the tone and sentiment of news coverage. This will help to clarify the complex relationship between public sentiment and media portrayal. Furthermore, by means of a comparative examination of the Daily Sun and Dhaka Tribune, this study endeavors to clarify the multitude of opinions and editorial positions present in Bangladesh's media landscape, offering significant discernment into the varied voices and viewpoints that mold national stories.

4.1. Overview of collected data

Data Quality: The web scraping process successfully extracted articles from 1 January 2023 to 31 December 2023 from the two leading newspapers of Bangladesh, The Daily Sun and Dhaka Tribune.

Data Diversity: The Dataset encompassed a wide range of topics including politics, economy, culture, business, and sports, providing a comprehensive overview of the year's news landscape.

Data Structure: Articles were stored with key attributes such as headlines, publication date, content, and identified geographical location.

4.2. Sentiment analysis

Sentiment analysis is the most common method for giving a review for a positive, negative, and neutral result. From web-scraping using the Python

toolkit Textblob finds the sentiment score over word count. From scraping The Daily Sun newspaper's website, the result found 8477550 words over the time period 1 January to 31 December 2023. The sentiment score, which varies from -1.00 to 1.00, is represented by the x-axis. Extremely negative sentiment is indicated by a score of -1.00, neutral sentiment is shown by a score of 0, and extremely positive sentiment is indicated by a score of 1.00. The number of articles or content pieces that fit into each sentiment category is shown on the y-axis Which is shown in figure **??**. Most of the articles are neutral, with feelings hovering around 0. This implies that a sizable percentage of the Daily Sun's material has a neutral tone. Positive sentiment is present in a substantial amount of the text, peaking at 0.25. This suggests that a good number of articles have a positive tone. Additionally, there is a tiny peak at roughly -1.0 sentiment, indicating that there are some extremely negative articles that are not as common as neutral or positive ones. With a small slant towards passivity, the Daily Sun newspaper's content seems to have a balanced distribution of sentiment.

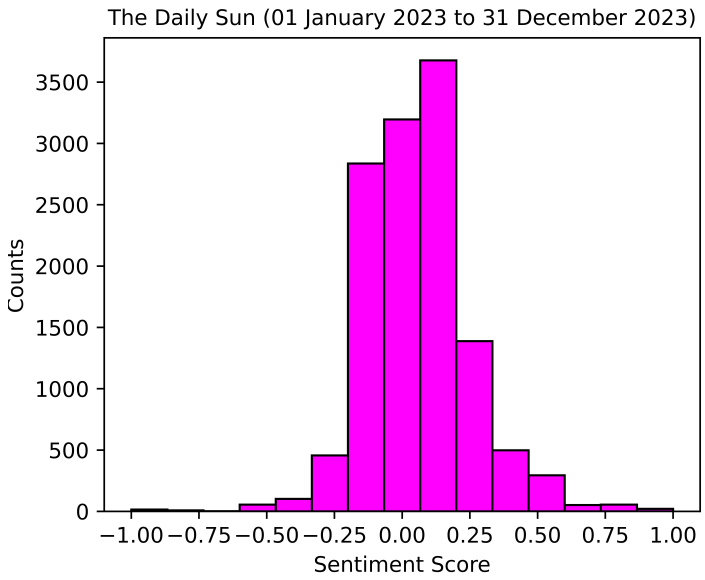

Figure 1: Schematic frequency distribution plot for the sentiment scores are generated from *The Daily Sun* for the period of 01 January 2023 to 31 December 2023 where the X-axis denotes the sentiment score and the Y-axis denotes the count.

The sentiment analysis results for "The Daily Sun" newspaper are dis-

played in a bar chart in **??**. Positive feelings, as indicated by the green bar, are the least common, with fewer than 5,000 instances. On the other hand, the red bar represents negative sentiment, which is slightly greater than positive sentiment but still falls below 10,000. The blue bar on the chart represents neutral sentiments, which are predominant and peak at about 35,000 instances. This suggests that a large percentage of the content in "The Daily Sun" elicited neutral responses, with relatively few instances of negativity and even fewer of positivity. Such a distribution might represent the nature of the news stories covered or imply a fair editorial stance. There is a dearth of positive sentiments, which may indicate fewer uplifting news items, as the sentiment distribution chart illustrates. It also shows a variety of reader reactions to the newspaper's content. On the other hand, the marginally higher number of negative opinions suggests that controversial subjects were covered. The frequency of neutral feelings indicates that a significant amount of the content elicits neither very positive nor negative reactions. The interplay between editorial decisions, reader preferences, and the nature of covered news is reflected in this complexity. The analysis's conclusions highlight the newspaper's propensity for objectivity in reporting and provide important background information for understanding its editorial position and impact on public opinion. Expanding upon the investigation of particular subjects linked to distinct sentiment classifications may enhance comprehension of the dynamics of reader engagement.

The pie chart of The Daily Sun also provides a visual representation of sentiment values which is presented in figure 3, with each sentiment category distinguished by a different color and corresponding percentage. The chart reveals that the majority of the analyzed data, which is color green, precisely 77.3%, is neutral, suggesting neither positive nor negative sentiment. This is followed by positive sentiments, which make up 14.6% of the data, indicating a significant, albeit a smaller, portion of the data is positively inclined with the color blue. Lastly, negative sentiments constitute the smallest portion, at 8.2%, suggesting a relatively minor presence of negative sentiment in the data shown in the color red. This distribution of sentiments could imply that the content or subject under analysis is generally perceived as neutral by its audience.

Again, For Dhaka Tribune, the result found 8377839 words from web scraping the newspaper website from 1 January to 31 December 2023. An informative depiction of the sentiment values connected to the Dhaka Tribune can be seen in the bar graph named "Sentiment Analysis of Dhaka Tribune"

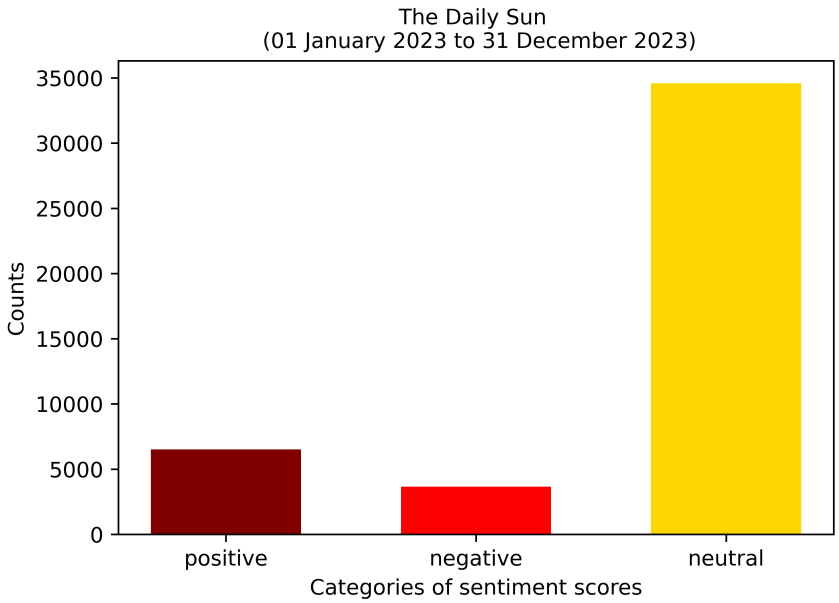

Figure 2: Sentiment into three parts - Positive, Negative, and Neutral of The Daily Sun.

which is shown in 4. Plotted on the x-axis are the sentiment values, which range from -1.00 (very negative) to 1.00 (highly positive). The count of each sentiment value, which ranges from 0 to 4000, is shown on the y-axis. 13 bars, each denoting a distinct sentiment value, make up the graph. almost 3000 words have a negative sentiment which has been shown in the histogram. Most of the bars fall in the range of neutral sentiments (which is nearly zero). There are almost 5000 words that lie in the neutral range of sentiment. Additionally, no extremely positive or negative sentiment. The majority of the articles are unbiased, with sentiments close to zero. This suggests that a significant portion of the content published in Dhaka Tribune has a neutral tone.

The "Sentiment Analysis" bar chart provides information about the sentiment distribution in the "Dhaka Tribune" newspaper in 5. A considerable but not overwhelming presence of positive sentiments in the newspaper's content is indicated by the green bar representing positive sentiment, which extends somewhat above 10,000. The shorter red bar, which represents negative sentiment, on the other hand, points to a lower count—less than 10,000—which suggests that negative sentiment is less common than positive sentiment. The tall blue bar for neutral sentiment is more than 50,000, meaning that

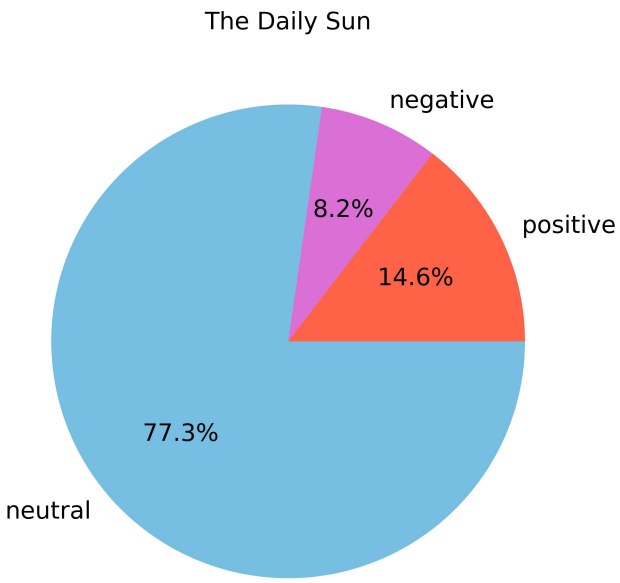

Figure 3: Sentiment pie plot of the Daily Sun Which indicates the positive, negative, and neutral sentiment.

a significant amount of the content is related to neutral sentiments. This distribution suggests that "Dhaka Tribune" evoked mostly neutral feelings, with fewer instances of positive or negative responses.

The pie chart's "Sentiment Analysis Results" graphic illustrates how the diverse sentiments in the analyzed dataset are broken down. It demonstrates that a significant percentage, at 76.9%, expresses neutral opinions in the color green, indicating a general feeling of objectivity in the material being examined. Contrarily, albeit less conspicuously, positivity may be seen in 15.2%, of the dataset which indicates the color blue. On the other hand, in colour red, the percentage of negativity is only 7.9%, which suggests that there aren't many negative feelings. Based on this research, it appears that most of the content takes a neutral position, with sporadic instances of positivity and very little negative information. These sentiment analysis-derived insights provide important viewpoints for additional interpretation and com-

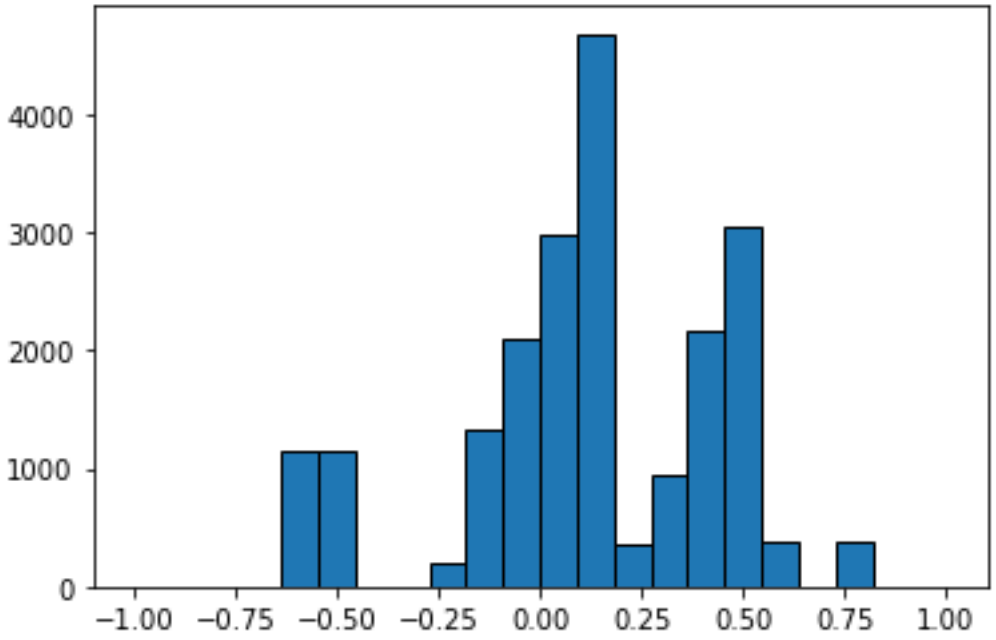

Figure 4: Sentiment plot of Dhaka Tribune from 1 January 2023 to 31 December 2023 where the X-axis denotes the sentiment score and the Y-axis denotes the word count.

prehension by shedding light on the dataset's general tone and sentiment landscape. The neutral tone of the sentiment score says that the news was neither so negative nor so positive throughout the year. The pie-chart of the sentiment scores of Dhaka Tribune is shown in 6.

4.3. Frequency of words

Although, The two leading newspapers has the neutral tone of articles, the top 50 words are different for each paper. After web scraping and utilizing NLTK python script for tokenize and stop words, the top 10 data for each newspaper has been shown in table 2. With an almost neutral sentiment score of 0.1970, the Dhaka Tribune has an average sentiment score. Throughout the newspaper's almost 8377839 words, "to" is used most of the time. On the other hand, the word that appears in The Daily Sun the most frequently is "in," with a frequency of 58665. Additionally demonstrating a neutral feeling, The Daily Sun's average sentiment score is 0.1895. The terms that are used in the newspaper the most, based on word count, seem to be neutral in both. 'In' ranks as the second most often used word in the Dhaka Tribune,

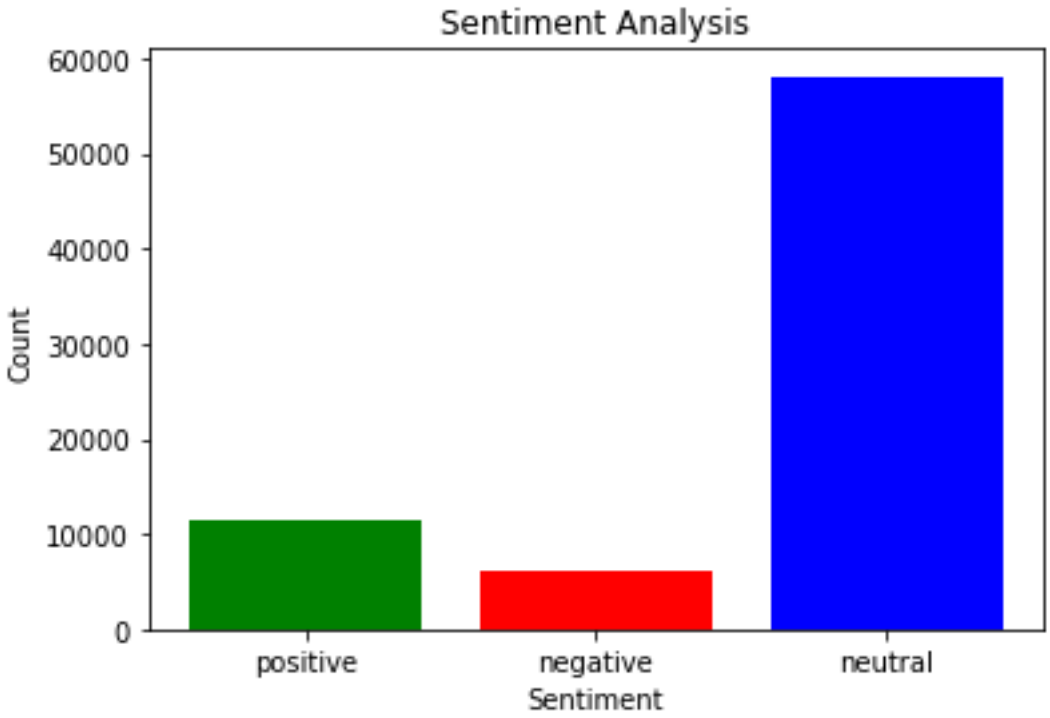

Figure 5: Sentiment into three parts (Positive, Negative, and Neutral of Dhaka Tribune.)

according to a plot of the top 10 words versus their frequency. "of" is the second most used word in The Daily Sun. The plots for the top 10 words of both newspapers are shown in figure 7 and figure 8.

4.4. Clustering

In unsupervised machine learning, clustering is a fundamental technique that groups similar data points together based on their inherent properties or qualities. Without any prior information on group memberships, the objective is to find naturally occurring groupings or clusters within the data. One of the most well-liked and frequently applied clustering methods is k-means clustering. By repeatedly assigning each data point to the closest cluster centroid and then recalculating the centroids based on the mean of the data points allocated to each cluster, it divides the data into a predefined number of clusters (K). This process is carried out until a predetermined number of iterations is reached or convergence occurs, at which point the assignments

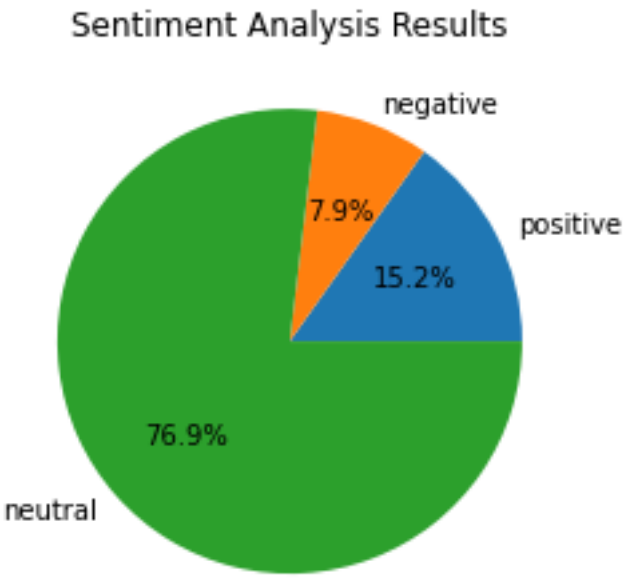

Figure 6: Sentiment pie plot of Dhaka Tribune Which indicates the positive, negative, and neutral sentiment.

no longer appreciably change. The k-means clustering approach is used in this investigation. The results of applying a k-means clustering method on sentiment data extracted from The Daily Sun are shown in the scatter plot in figure 10. The dataset's word index is represented by the word index, which is shown along the X-axis. It falls between 0 and 12000. Plotting sentiment ratings, however, is done along the Y-axis. It is a scale from -1 to 1, where a score of -1 denotes a very negative attitude, a score of 0 denotes neutral sentiment, and a score of 1 denotes a very positive mood. The data can be graphically interpreted into three groups: yellow denotes positive thoughts, purple denotes neutral sentiments, and blue denotes negative opinions. Notably, the top yellow cluster represents words with high emotion ratings, the middle purple cluster represents words with neutral sentiment ratings, and the bottom blue cluster represents phrases with low sentiment ratings. The k-means clustering plot of Dhaka Tribune, the other newspaper, reveals that the purple line from -0.2 to 0.2 explains the neutral sentiment of the articles because those are very close to 0 (Neutral sentiment). The X-axis contains the word index from 0 to 30000, and the Y-axis contains the sentiment plot

Table 2: The table represents the top 50 Words of The Daily Sun and the Dhaka Tribune.

The Daily Sun				Dhaka Tribune			
Word	Count	Word	Count	Word	Count	Word	Count
in	58665	Page	3112	to	64240	exceeds	8395
of	52674	was	3082	in	60590	dead	8395
the	46572	US	3056	of	39420	60	8395
to	44795	Md	3016	minutes	30660	injured	8395
on	33371	accident	2979	the	27375	India	8395
for	22636	held	2953	with	27010	billboard	8395
at	21216	is	2855	on	21900	bakes	8395
a	19563	photo	2758	Lu	20805	parts	8395
and	15104	Bank	2710	paddy	16425	37	8036
with	10439	more	2644	bodies	16060	already	8030
by	9057	Minister	2540	ago	16060	harvested	8030
day	7695	taken	2502	a	14600	Highest	8030
as	7663	killed	2413	Bangladesh	14600	production	8030
from	6954	Sheikh	2401	slash	14235	target	8030
after	5273	against	2339	Donald	13870	decade	8030
over	4783	Dhaka	2217	help	13870	removed	8030
Bangladesh	4634	up	2213	bilateral	13505	from	8030
capital	4582	get	2210	will	13505	wreckage	8030
be	4313	Daily	2207	Dhaka	12410	Four	8030
The	4139	Monday	2190	protected]	9855	more	8030
new	3915	will	2147	[email	9855	are	8030
Sun	3749	Dr	2065	ties	9125	still	8030
Photo	3655	University	1960	by	9125	buried	8030
World	3303	Asia	1954	cultivation	8395	nights	8030
now	3268	an	1935	Rajshahi	8395	become	8030

similar to The Daily Sun. The negative attitudes are explained by a yellow line at -0.8, while the good sentiments are explained by a green line at the top of the plot which is in figure 11. Activities like sentiment analysis, social media monitoring, and reputation management are made easier by this graphic depiction, which offers a comprehensive description of the sentiment distribution in the text data. The k-means clustering approach applied to both newspapers indicates that the feelings that are positive, negative, and

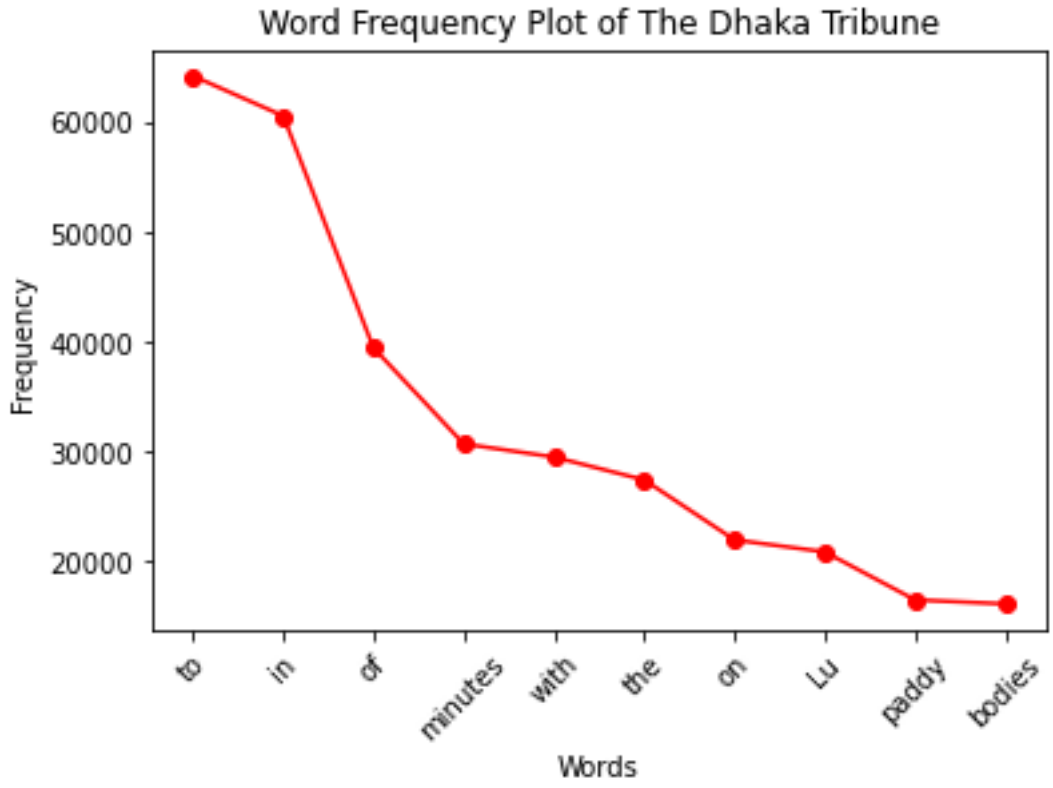

Figure 7: Frequency curve of Dhaka Tribune Where the X-axis contains the top 10 words and the Y-axis contains the Frequency over the year 2023.

neutral are arranged inside three clusters and that both newspapers have neutral sentiments throughout the year.

4.5. Classification

A classifier is a key concept in machine learning that is used to categorize input data into predetermined classes or categories based on their attributes. It is a form of predictive modeling technique employed for classification jobs, with the objective of assigning labels or categories to input data points based on their properties. Classifiers acquire patterns from annotated training data and subsequently employ this information to forecast the labels of unobserved or novel data pieces. These data pieces can undergo training using a range of techniques, including decision tree, random forest, support vector machines (SVM), logistic regression, Naive Bayes, k-nearest neighbors (kNN), and neu-

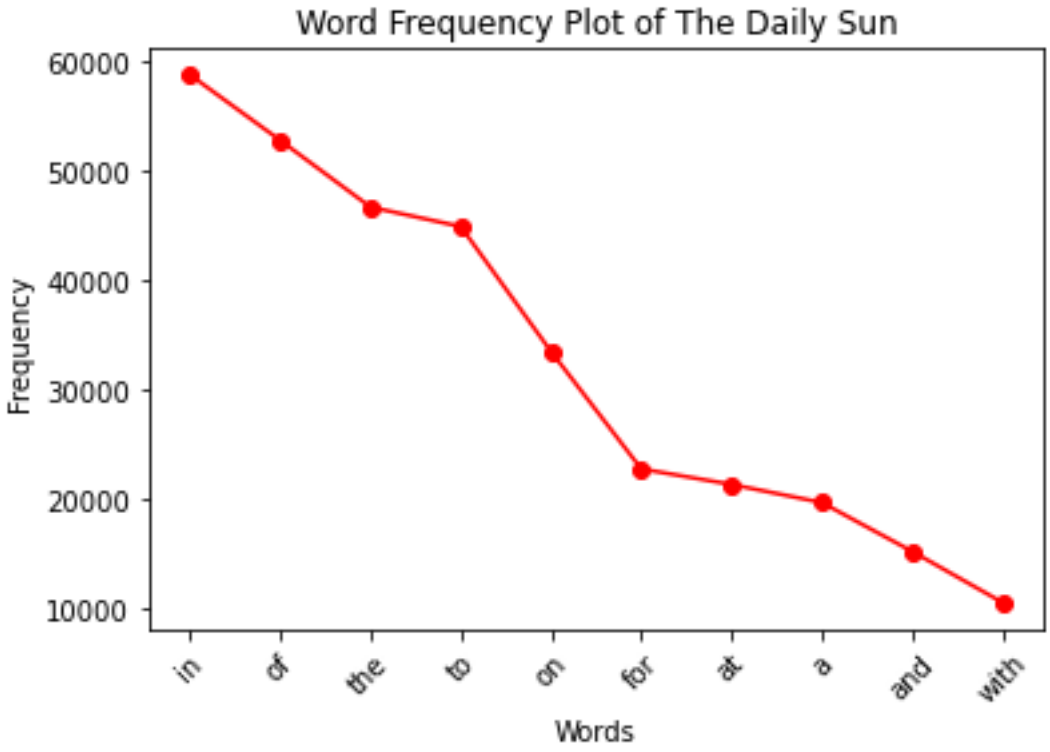

Figure 8: Frequency curve of The Daily Sun Where the X-axis contains the top 10 words and the Y-axis contains the Frequency over the year 2023.

ral networks. Classifier performance is commonly assessed using metrics such as accuracy, precision, recall, F_1 score, and area under the receiver operating characteristic (ROC) curve. These metrics offer valuable information about the classifier's performance in accurately predicting the labels of the input data.

Table 3: Performance metrics for Naive Bayes and Random Forest For The Daily Sun.

Classifier	**Performance Metrics**			
	Precision	**Recall**	**F_1-score**	**Accuracy**
Naive Bayes	1.00	0.88	0.94	0.96
	0.94	1.00	0.97	0.96
Random Forest	1.00	1.00	1.00	0.98
	1.00	1.00	1.00	0.98

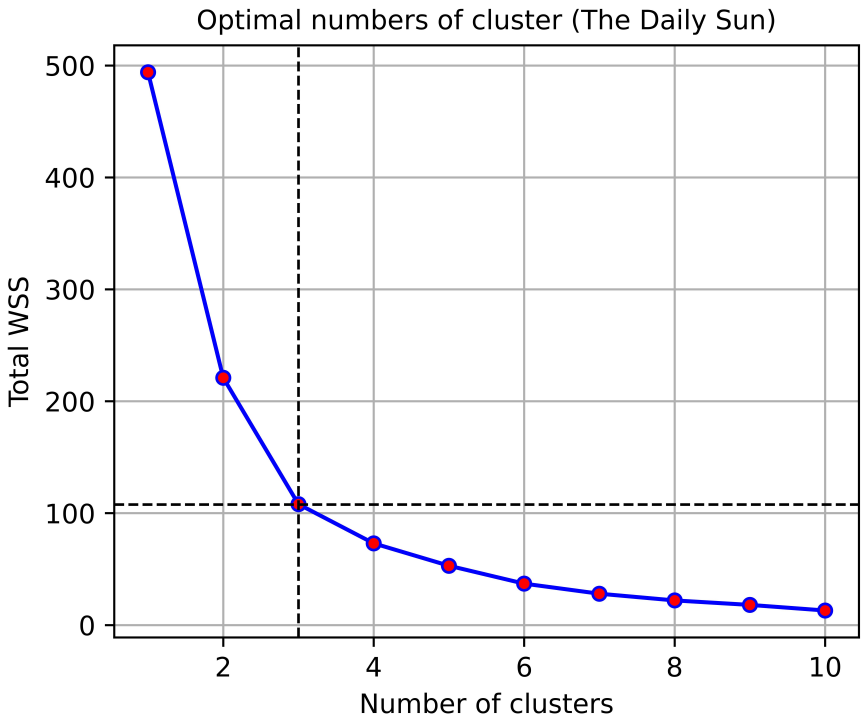

Figure 9: Optimal number of cluster for k-means clustering of sentiment scores generated for The Daily Sun for the period of 01 January 2023 to 31 December 2023.

```
Naive Bayes:
              precision    recall  f1-score   support

    Negative       1.00      0.88      0.94      1816
     Neutral       1.00      1.00      1.00         4
    Positive       0.94      1.00      0.97      3242

    accuracy                           0.96      5062
   macro avg       0.98      0.96      0.97      5062
weighted avg       0.96      0.96      0.96      5062

Accuracy: 0.9563413670485974
cohen_kappa_score(y_test, nb_predictions)
0.9027237814234146
```

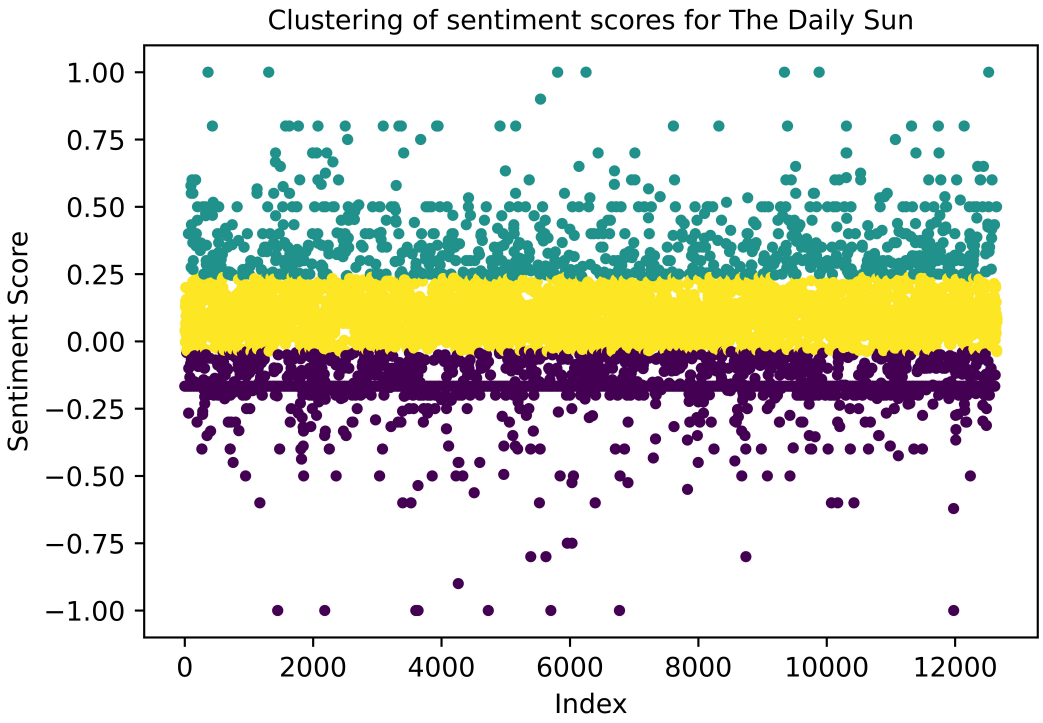

Figure 10: k-means clustering of The Daily Sun where the X-axis denotes the word Index and the Y-axis denotes the sentiment score.

```
print(confusion_matrix(y_test, nb_predictions))
[[1595    0  221]
 [   0    4    0]
 [   0    0 3242]]
```

SVM:

	precision	recall	f1-score	support
Negative	1.00	1.00	1.00	1816
Neutral	1.00	1.00	1.00	4
Positive	1.00	1.00	1.00	3242

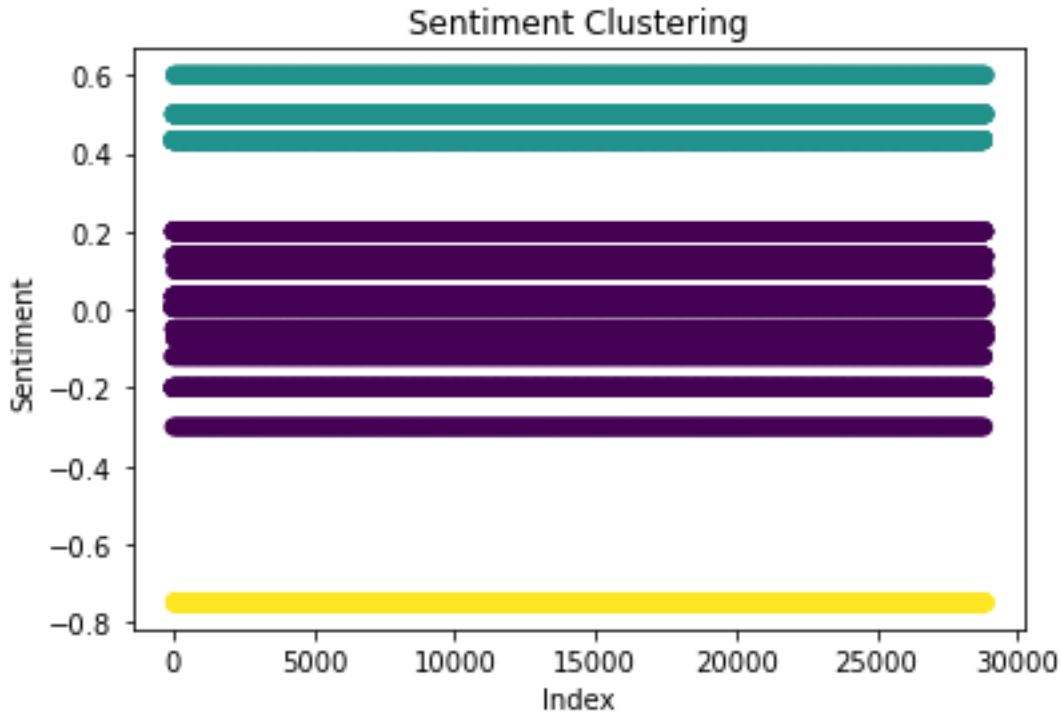

Figure 11: k-means clustering of Dhaka Tribune where the X-axis denotes the word Index and the Y-axis denotes the sentiment score.

```
    accuracy                           1.00      5062
   macro avg       1.00      1.00      1.00      5062
weighted avg       1.00      1.00      1.00      5062

Accuracy: 0.9990122481232714
Cohen Kappa: 0.9978565884034912
[[1811    0    5]
 [   0    4    0]
 [   0    0 3242]]

Random Forest:
              precision    recall  f1-score   support

    Negative       1.00      1.00      1.00      1816
     Neutral       1.00      1.00      1.00         4
```

Table 4: Results from Naive Bayes

Confusion matrix			
Category	Negative	Neutral	Positive
Negative	1595	0	221
Neutral	0	4	0
Positive	0	0	3242
Classification report			
Accuracy measures	Precision	Recall	F_1-score
Negative	1.00	0.88	0.94
Neutral	1.00	1.00	1.00
Positive	0.94	1.00	0.97
Macro average	0.98	0.96	0.97
Weighted average	0.96	0.96	0.96
Accuracy			0.96
Cohen Kappa			0.90

```
    Positive       1.00      1.00      1.00      3242

    accuracy                           1.00      5062
   macro avg       1.00      1.00      1.00      5062
weighted avg       1.00      1.00      1.00      5062

Confusion Matrix:
 [[1816    0    0]
 [   0    4    0]
 [   2    0 3240]]
Accuracy: 0.9996048992493086
Cohen Kappa: 0.9991433595131601

Logistic Regression:
              precision    recall  f1-score   support

    Negative       1.00      0.94      0.97      1816
     Neutral       0.00      0.00      0.00         4
    Positive       0.97      1.00      0.98      3242

    accuracy                           0.98      5062
```

```
   macro avg       0.66      0.65      0.65      5062
weighted avg       0.98      0.98      0.98      5062

Accuracy: 0.9794547609640458
Cohen Kappa: 0.9548716049636709

Confusion Matrix:
 [[1716    0  100]
 [   4    0    0]
 [   0    0 3242]]
```

In this study, random forest and Naive Bayes is used to classify the sentiment data. The Daily Sun newspaper shows high accuracy which is around 96% The performance of the Naive Bayes and Random Forest classifiers is indicated by the results obtained for key metrics. Naive Bayes demonstrates a high level of precision and recall when classifying examples as belonging to class 1, indicating accurate identification. However, there is a minor decrease in precision and recall when classifying instances as belonging to class 0. Nevertheless, its impressive overall accuracy of 96% showcases its robust predictive capacity. On the other hand, Random Forest demonstrates exceptional precision, recall, and F_1-score for both classes, coupled with a flawless accuracy of 100%. While Naive Bayes exhibits decent performance, Random Forest's greater accuracy and balanced performance across all criteria make it the favored choice based purely on these data. Nonetheless, more considerations, such as model complexity and interpretability, are important for a full examination before making a final decision. The result is shown in 3.

Table 5: Performance metrics for Naive Bayes and Random Forest For Dhaka Tribune.

Classifier	**Performance Metrics**			
	Precision	**Recall**	**F_1-score**	**Accuracy**
Naive Bayes	1.00	0.70	0.82	0.96
	0.96	1.00	0.98	0.96
Random Forest	1.00	1.00	1.00	0.98
	1.00	1.00	1.00	0.98

An understanding of the predictive power of the Naive Bayes and Random Forest classifiers is provided by the performance metrics table on Dhaka

Tribune data. With a precision of 1.00, Naive Bayes is always correct when predicting a data point to belong to a particular class, but with a recall of 0.70, it is only accurate 70% of the time. Its performance is matched by its F_1-score of 0.82, which analyzes recall and precision into a single statistic. With an overall accuracy of 0.96, Naive Bayes guesses the class label accurately 96% of the time. Corresponding to this, Random Forest exhibits perfect recall, F_1-score, and precision for both classes, demonstrating excellent performance in predicting positive and negative cases. The classifier table shown is 5.

The combination of web scraping and sentiment analysis, utilizing advanced techniques such as k-means clustering, Naive Bayes, and Random Forest classifiers, has provided valuable insights into the sentiment patterns of the analyzed data. By utilizing web scraping techniques, a substantial volume of textual data was gathered, facilitating a thorough comprehension of sentiment patterns. The application of k-means clustering effectively grouped the data into distinct clusters, based on similarities in sentiment. This facilitated the identification of overarching sentiment trends within the dataset. The Naive Bayes and Random Forest classifiers improved the analysis by effectively categorizing sentiments and forecasting sentiment labels for new data. The results indicate the efficacy of these methodologies in extracting sentiment-related information from textual data obtained through web scraping. Utilizing sophisticated methods like k-means clustering, Naive Bayes, and Random Forest classifiers, the combination of web scraping and sentiment analysis has yielded insightful information about the sentiment patterns in the studied data. A significant amount of textual data was collected through the use of web scraping techniques, which made it possible to fully understand sentiment patterns. By using k-means clustering, the data was successfully divided into different groups according to sentiment similarities. This made it easier to find the dataset's overarching sentiment trends. The analysis was enhanced by the Naive Bayes and Random Forest classifiers, which successfully categorized sentiments and predicted sentiment labels for fresh data. The outcomes show how effective these techniques are at removing sentiment-related information from textual data that is collected via web scraping.

5. Conclusion

In this work, this study used web scraping techniques to obtain newspaper data from The Daily Sun and Dhaka Tribune and then conducted sentiment

analysis on the data. To analyze the sentiment trends found in the articles, this study used two classifiers, Random Forest and Naive Bayes, along with sentiment plot visualization and k-means clustering. Our analysis of the news items' sentiment distribution patterns identified interesting trends that provide light on the dominant tenors and viewpoints conveyed in the pieces. However, there were obstacles this study had to overcome when online scraping, especially when choosing the right parser to precisely pull the pertinent data from the websites. There are other directions to pursue in this area in the future. First off, streamlining the web scraping procedure by honing the parser selection technique or looking into other options could improve data collecting accuracy and efficiency. The robustness and forecast accuracy of the sentiment analysis models may also be improved by combining more sophisticated sentiment analysis techniques and investigating other classifiers or ensemble approaches. Furthermore, broadening the study's scope to incorporate more newspapers or other media outlets would yield a more thorough knowledge of sentiment trends in news reporting. Finally, additional insights into the significance of the findings could be obtained by performing qualitative analysis or applying domain-specific expertise to understand the sentiment analysis results in relation to socio-political events or public sentiment. Overall, this work establishes the foundation for future investigations that will use sentiment analysis methods to mine newspaper data for insightful information and deepen our comprehension of the dynamics of public sentiment.

References

[1] A. Balahur, R. Steinberger, et al., Rethinking sentiment analysis in the news: from theory to practice and back, Proceeding of WOMSA 9 (2009) 1–12.

[2] C. Ratner, A cultural-psychological analysis of emotions, Culture & Psychology 6 (1) (2000) 5–39.

[3] D. Goleman, Emotional intelligence. bantam books: N. y (1995).

[4] N. Bondarchuk, N. Hrytsiv, I. Bekhta, O. Melnychuk, Sentiment analysis of weather news in british online newspapers, Amazonia Investiga 12 (63) (2023) 99–108.

[5] M. Puteh, N. Isa, S. Puteh, N. A. Redzuan, Sentiment mining of malay newspaper (samnews) using artificial immune system, in: Proceedings of the World Congress on Engineering, Vol. 3, 2013, pp. 1498–1503.

[6] M. D. Devika, C. Sunitha, A. Ganesh, Sentiment analysis: a comparative study on different approaches, Procedia Computer Science 87 (2016) 44–49.

[7] Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions, Journal of King Saud University - Computer and Information Sciences 35 (9) (2023) 101776. doi:https://doi.org/10.1016/j.jksuci.2023.101776.

[8] P. Sharma, T.-S. Moh, Prediction of indian election using sentiment analysis on hindi twitter, in: 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1966–1971. doi:10.1109/BigData.2016.7840818.

[9] B. Liu, Sentiment analysis and opinion mining, Springer Nature, 2022.

[10] Z. Drus, H. Khalid, Sentiment analysis in social media and its application: Systematic literature review, Procedia Computer Science 161 (2019) 707–714.

[11] Sentiment analysis methods, applications, and challenges: A systematic literature review, Journal of King Saud University - Computer and Information Sciences 36 (4) (2024) 102048. doi:https://doi.org/10.1016/j.jksuci.2024.102048.

[12] A. Hossain, M. Karimuzzaman, M. M. Hossain, A. Rahman, Text mining and sentiment analysis of newspaper headlines, Information 12 (10) (2021) 414.

[13] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, J. Belyaeva, Sentiment analysis in the news (2013). arXiv:1309.6202.

[14] J. Wiebe, Tracking point of view in narrative, Computational Linguistics 20 (07 2002).

[15] B. McHale, Unspeakable sentences, unnatural acts: Linguistics and poetics revisited, Poetics Today 4 (1) (1983) 17–45.
URL http://www.jstor.org/stable/1772149

[16] C. Whitelaw, N. Garg, S. Argamon, Using appraisal groups for sentiment analysis, in: Proceedings of the 14th ACM international conference on Information and knowledge management, 2005, pp. 625–631.

[17] A. Pandiaraj, C. Sundar, S. Pavalarajan, Sentiment analysis on newspaper article reviews: contribution towards improved rider optimization-based hybrid classifier, Kybernetes 51 (1) (2021) 348–382.

[18] H. Rahab, A. Zitouni, M. Djoudi, Sana: Sentiment analysis on newspapers comments in algeria, Journal of King Saud University-Computer and Information Sciences 33 (7) (2021) 899–907.

[19] S. Atia, K. Shaalan, Increasing the accuracy of opinion mining in arabic, in: 2015 first international conference on arabic computational linguistics (ACLing), IEEE, 2015, pp. 106–113.

[20] H. Rahab, A. Zitouni, M. Djoudi, Siaac: Sentiment polarity identification on arabic algerian newspaper comments, in: Applied Computational Intelligence and Mathematical Methods: Computational Methods in Systems and Software 2017, vol. 2, Springer, 2018, pp. 139–149.

[21] W. Cherif, A. Madani, M. Kissi, Towards an efficient opinion measurement in arabic comments, Procedia Computer Science 73 (2015) 122–129.

[22] A. Ziani, Y. Tlili Guaissa, A. Nabiha, Détection de polarité d'opinion dans les forums en langues arabe par fusion de plusieurs svms, vol 7 (2013) 17–21.

[23] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams engineering journal 5 (4) (2014) 1093–1113.

[24] Z. U. Rehman, I. S. Bajwa, Lexicon-based sentiment analysis for urdu language, in: 2016 Sixth International Conference on Innovative Computing Technology (INTECH), 2016, pp. 497–501. doi:10.1109/INTECH.2016.7845095.

[25] S. L. Lo, E. Cambria, R. Chiong, D. Cornforth, Multilingual sentiment analysis: from formal to informal and scarce resource languages, Artificial Intelligence Review 48 (2017) 499–527.

[26] H. Peng, E. Cambria, A. Hussain, A review of sentiment analysis research in chinese language, Cognitive Computation 9 (2017) 423–435.

[27] S. Rani, P. Kumar, A journey of indian languages over sentiment analysis: a systematic review, Artificial Intelligence Review 52 (2019) 1415–1462.

[28] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, S. Ahmad, Creating sentiment lexicon for sentiment analysis in urdu: The case of a resource-poor language, Expert Systems 36 (3) (2019) e12397.

[29] A. Amram, A. B. David, R. Tsarfaty, Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from modern hebrew, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2242–2252.

[30] A. Hassan, M. R. Amin, A. K. Al Azad, N. Mohammed, Sentiment analysis on bangla and romanized bangla text using deep recurrent models, in: 2016 International Workshop on Computational Intelligence (IWCI), IEEE, 2016, pp. 51–56.

[31] P. Sharma, T.-S. Moh, Prediction of indian election using sentiment analysis on hindi twitter, in: 2016 IEEE international conference on big data (big data), IEEE, 2016, pp. 1966–1971.

[32] M. Saraee, A. Bagheri, Feature selection methods in persian sentiment analysis, in: Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings 18, Springer, 2013, pp. 303–308.

[33] M. Abdul-Mageed, M. Diab, S. Kübler, Samar: Subjectivity and sentiment analysis for arabic social media, Computer Speech & Language 28 (1) (2014) 20–37.

[34] M. A. Paredes-Valverde, R. Colomo-Palacios, M. d. P. Salas-Zárate, R. Valencia-García, et al., Sentiment analysis in spanish for improvement of products and services: A deep learning approach, Scientific Programming 2017 (2017).

[35] H. Ghorbel, D. Jacot, Sentiment analysis of french movie reviews, in: Advances in Distributed Agent-Based Retrieval Tools, Springer, 2011, pp. 97–108.

[36] M. Ptaszynski, R. Rzepka, K. Araki, Y. Momouchi, Automatically annotating a five-billion-word corpus of japanese blogs for sentiment and affect analysis, Computer Speech & Language 28 (1) (2014) 38–55.

[37] M. Song, H. Park, K.-s. Shin, Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in korean, Information Processing & Management 56 (3) (2019) 637–653.

[38] A systematic literature review of arabic dialect sentiment analysis, Journal of King Saud University - Computer and Information Sciences 35 (6) (2023) 101570. doi:https://doi.org/10.1016/j.jksuci.2023.101570.

[39] D. C. Mutz, J. Soss, Reading public opinion: The influence of news coverage on perceptions of public sentiment, Public Opinion Quarterly (1997) 431–451.

[40] C. A. Chapelle, Y.-R. Chung, The promise of nlp and speech processing technologies in language assessment, Language Testing 27 (3) (2010) 301–315.

[41] L. Abualigah, H. E. Alfar, M. Shehab, A. M. A. Hussein, Sentiment analysis in healthcare: a brief review, Recent advances in NLP: the case of Arabic language (2020) 129–141.

[42] I. Aattouchi, S. Elmendili, F. Elmendili, Sentiment analysis of health care, in: E3S Web of Conferences, Vol. 319, EDP Sciences, 2021, p. 01064.

[43] A. Zunic, P. Corcoran, I. Spasic, Sentiment analysis in health and well-being: systematic review, JMIR medical informatics 8 (1) (2020) e16023.

[44] K. Denecke, Y. Deng, Sentiment analysis in medical settings: New opportunities and challenges, Artificial intelligence in medicine 64 (1) (2015) 17–27.

[45] H. Pandey, A. K. Mishra, D. N. Kumar, Various aspects of sentiment analysis: a review, in: Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE), 2019.

[46] R. Varghese, M. Jayasree, A survey on sentiment analysis and opinion mining, International journal of Research in engineering and technology 2 (11) (2013) 312–317.

[47] T. Shivaprasad, J. Shetty, Sentiment analysis of product reviews: A review, in: 2017 International conference on inventive communication and computational technologies (ICICCT), IEEE, 2017, pp. 298–301.