information

Article

Text Mining and Sentiment Analysis of Newspaper Headlines

Arafat Hossain [1], Md. Karimuzzaman [1], Md. Moyazzem Hossain [1,*] and Azizur Rahman [2,*]

[1] Department of Statistics, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh; arafathossen239@gmail.com (A.H.); karimuzzaman.statju@gmail.com (M.K.)
[2] School of Computing, Mathematics and Engineering, Charles Sturt University, Wagga Wagga, NSW 2678, Australia
* Correspondence: hossainmm@juniv.edu (M.M.H.); azrahman@csu.edu.au (A.R.)

Abstract: Text analytics are well-known in the modern era for extracting information and patterns from text. However, no study has attempted to illustrate the pattern and priorities of newspaper headlines in Bangladesh using a combination of text analytics techniques. The purpose of this paper is to examine the pattern of words that appeared on the front page of a well-known daily English newspaper in Bangladesh, *The Daily Star*, in 2018 and 2019. The elucidation of that era's possible social and political context was also attempted using word patterns. The study employs three widely used and contemporary text mining techniques: word clouds, sentiment analysis, and cluster analysis. The word cloud reveals that election, kill, cricket, and Rohingya-related terms appeared more than 60 times in 2018, whereas BNP, poll, kill, AL, and Khaleda appeared more than 80 times in 2019. These indicated the country's passion for cricket, political turmoil, and Rohingya-related issues. Furthermore, sentiment analysis reveals that words of fear and negative emotions appeared more than 600 times, whereas anger, anticipation, sadness, trust, and positive-type emotions came up more than 400 times in both years. Finally, the clustering method demonstrates that election, politics, deaths, digital security act, Rohingya, and cricket-related words exhibit similarity and belong to a similar group in 2019, whereas rape, deaths, road, and fire-related words clustered in 2018 alongside a similar-appearing group. In general, this analysis demonstrates how vividly the text mining approach depicts Bangladesh's social, political, and law-and-order situation, particularly during election season and the country's cricket craze, and also validates the significance of the text mining approach to understanding the overall view of a country during a particular time in an efficient manner.

Keywords: newspaper; headlines pattern and context; word cloud; cluster analysis; sentiment analysis; Bangladesh

Citation: Hossain, A.; Karimuzzaman, M.; Hossain, M.M.; Rahman, A. Text Mining and Sentiment Analysis of Newspaper Headlines. *Information* **2021**, *12*, 414. https://doi.org/10.3390/info12100414

Academic Editor: Byung-Won On

Received: 1 April 2021
Accepted: 12 August 2021
Published: 9 October 2021

1. Introduction

Text mining is a technique for extracting information from text by recognizing patterns and trends. The term text mining, text analytics, or text analysis refers to the process of retrieving information through lexical resources, tagging or annotation, and techniques such as association, visualization, and prediction. After successfully developing basic natural language processing (NLP) in the 1960s, different adoptions of techniques such as dimension reduction, latent factor identification, and database text processing have contributed to the flourishing of the new era of information retrieval. Moreover, the topic model or latent semantic analysis and machine learning algorithms seemingly gave a more substantial base after the 1990s—sentiment analysis and opinion mining methods have emerged from analysing the sentiment of humans from text, which enthrals intellectual fields including computer science, statistics, linguistics, and social science. Additionally, a successful implication for the analysis of journals, social network services, and online customer reviews, along with email filtering, product suggestions, fraud detection, search engines, and bankruptcy predictions, has increased its significance in all aspects [1–7].

Interestingly, a distinct feature of elucidating the insights of text makes the text mining procedure more convenient. The term word cloud is derived from the word tag cloud, firstly used by the community-oriented website Flickr, which uses several techniques, including text shortening (a most crucial feature of tag clouds) and visualization [8]. Indeed, the word cloud cannot give an accurate statistical summary and/or the context, emotion, and linguistic knowledge, but it elicits the overall concept of the text [4]. Several advanced methodologies of word cloud data handling have emerged in the last two decades, such as Kaser and Lemire's rectangular form to reduce and balance the white space [9], Seifert et al.'s (2014) polygon type layout for space-filling word clouds, along with Wordle, Tagul, and Taxedo [10]. For instance, ConcentriCloud uses a concentric design to implement a simple layout over the data; nevertheless, several layouts need to be allotted the weighted words and assigned the word-by-word relationships [11].

Another imperative method of analysing the emotions of humans from text is sentiment analysis—a procedure of detecting or defining eight primary emotions of humans and animals, including anger, fear, sadness, disgust, surprise, anticipation, trust, and joy emotions (defined by the eminent psychologist Robert Plutchik) through words. In addition to emotional states, another straightforward approach is the polarity of a document, which develops the method of detecting polarity, particularly positive and negative explication, through distinct algorithms [12–14]. Moreover, the adaptation of this analysis led businesses to accumulate people's opinions to determine their following business policy and political parties to manipulate people through election strategies, newspapers, Twitter, Facebook, and blogs, along with the application to novels, movie reviews, and product reviews [15–17]. Similar to the word cloud, sentiment analysis concedes constraints, such as the number of data sets, measurements, distinct meanings for identical text, different languages, shortcut words written by the author, typing mistakes, and the rhetoric of writing [1,12,18]. However, besides these problems, sentiment analyses are still recognized as one of the most precise techniques to analyse texts in terms of emotion, where lexicon-based analysis techniques are primarily used [13]. Correspondingly, cluster analysis is an unsupervised machine learning technique of classifying or grouping the text or documents through distance or similarity-based algorithms. The reduction of dimensionality of a massive document may include millions of words through clustering by having a meaningful and interpretable number of similar groups or clusters that may occur naturally in the data. However, a significant improvement of text mining models has been observed in the recent past with the intense use of modern supervised and unsupervised machine learning algorithms, especially for the high dimensional setting of data [19]. For instance, NLP frameworks for online opinion mining based on GA and ontology [20], a prediction model for creativity education using clustering methods based on discussion and records [21], and forums hotspot detection and forecast through a newly developed mixed unsupervised machine learning-based text mining and sentiment technique [22] have been observed in the literature.

Moreover, in many cases, social scientists rely on conventional qualitative research techniques, e.g., FGD, interviews, and others. Most of these cases also depend on few respondents because of constraints, including money, time, and others. However, several social researchers adopt this cutting-edge technology to find the social pattern, sometimes as an alternative or mixed with conventional qualitative methods [23–27].

2. Related Work

Electronic and press media have increasingly usurped newspapers as the primary source of news for the general public. Additionally, they are occasionally referred to as a society's mirror. One can obtain an overview of the entire country at any particular era by simply reading a few pages or headlines of a newspaper. Summarizing various reports on several subject matter occurrences helps news readers to quickly browse through news topics. Since readers tend to follow events, keywords, and subjects, presentation and identification are significantly implied techniques for all news. By synthesizing large

bodies of newspaper text information into summaries, a proliferation of studies has been seen that have employed both long-form and unstructured newspaper data with social media, such as Twitter and Tumblr, to retrieve the opinion of the general public [16,22,28].

In addition, new technologies of automatic identification, especially in the fields of online finance and health, economic news article mining for understanding the economic consequences of Turkey [29], empirical financial modelling and decision making [28–30] in order to predict and recommend stock market trades [30–33], and risk management [34] have also been seen in the literature. In addition, consumer general and brand sentiment [35–37] and the economic impact of product reviews [38] have also been studied to understand, predict, and make decisions for furthering business strategy through a text mining approach [39]. Moreover, the study of media coverage in times of political crisis, along with measuring online political dialogue [33,40] and ethnographic assessment [26], is also found in the literature.

In Bangladesh, as in many other countries, most of the text mining approaches are primarily built for the language English. However, researchers are currently trying to build their own corpus and sentiment using modern machine learning and deep learning techniques; for instance, a group of scholars recently proposed an algorithm for Multiclass classification of Bangla Newspaper tag using level data augmentation [41]. Similarly, several proposed algorithms for Bangla newspaper sentiment analysis have been seen in the literature, including Supervised Machine Learning with Extended Lexicon Dictionary [42], Support Vector Machine (SVM) and Logistic Regression [41,43], Long Short-Term Memory (LSTM) Recurrent Neural Network [44]. Moreover, analyzing cricket supporter sentiment, online shopping reviews (Ismail Siddiqi) [45], microblog posts [46], movie reviews [12], E-commerce business [47] sentiment through text mining approaches have been seen in the literature.

Nevertheless, a recent review revealed that most sentiment analysis-based research analyzes social media and microblogging sites and are based on the use of lexicons [13]. Conventional qualitative analysis, such as thematic, discourse and content analysis of agriculture news [48], crime monitoring [49], and extrajudicial execution [50], have also been observed in some academic studies. The text mining of English newspaper editorials, in particular, the daily English newspaper, *The Daily Star*, over the period 1 January 2018 to 30 June 2018 [51], use only the word cloud and frequency distribution by expressing the graphs—not others manuals of text mining including sentiment and cluster analysis.

However, most Bangladeshi text mining-related studies propose their own algorithm or analyze it through a particular frequency, sentiment, or classification method. Furthermore, to the best of our knowledge, no study involved word cloud, sentiment, and cluster analysis combined to find the patterns and overview of the news and link it with the social and political context. Most of them focus only on the Bangla newspaper and short-period data. At the same time, a study finds that English newspapers' enormous impact and readability with a short circulation [52] set the significance of analyzing English newspapers. Moreover, social scientists can also adopt this study as an example of modern technological advancement as no study has been found in Bangladesh where these methods are used in their analysis, to the best of our knowledge. Therefore, this paper attempts to explore the newspaper headlines with the help of text analytics techniques to find the pattern of the word and link them with the social context of that time. The word cloud, alongside sentiment and cluster analysis have been implemented to find the most frequent words used, emotion base (positive or negative), and similar words in the headline and then visualizes this data to create a supposition about the overall situation of Bangladesh over the last two years.

3. Materials and Methods

3.1. Data Preparation

This paper focuses on the writing of newspaper front-page headlines for Bangladesh's most popular daily English newspaper, *The Daily Star*, as front-page headlines of the

newspaper contain vital issues. The texts were collected manually for two years, from 1 January 2018 to 31 December 2019, from the newspaper's official website (www.thedailystar.net, accessed on 11 July 2020) [53]. Then, the data were pre-processed and analyzed using several R-packages for text mining, as detailed in the following section.

3.2. Methods

Text mining is a process that has been making our daily life smoother by using various methods of it through deciding on the text of related objective. The most prominent methods of text mining are word cloud and sentiment analysis. The following flowchart exhibit the popular text mining process (Figure 1).

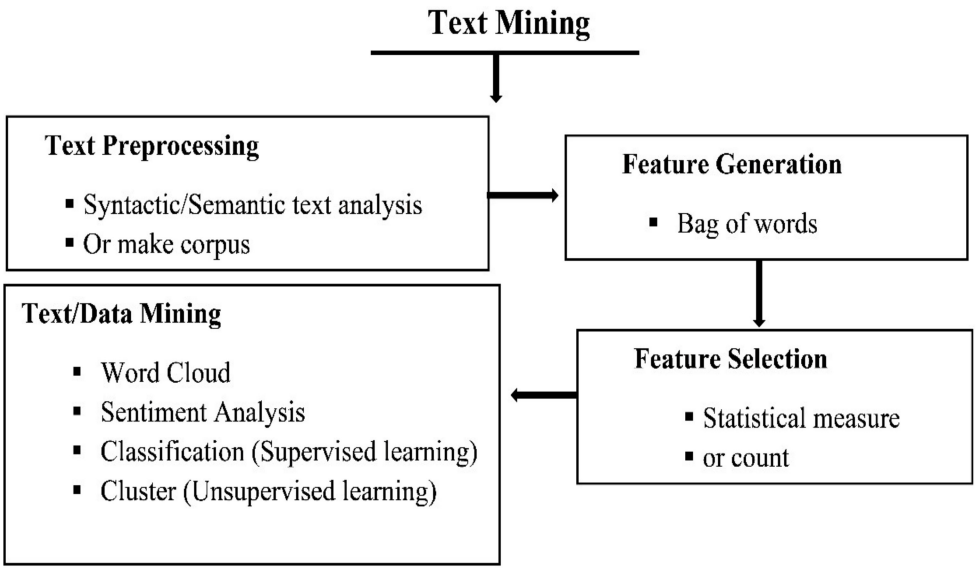

Figure 1. Overview of Text Mining, Source [50].

Essentially, text pre-processing operates on the raw text and removes superfluous information. Tokenization, standardization, and cleansing and stop word removal and stemming or lemmatization constitute the pre-processing stage. The procedure is critical in gathering information and simplifying it. Typically, the preposition, URL, numbers, conjunction, and other superfluous elements that have no bearing on the sentence are omitted. The following diagram depicts the pre-processing of text data, which is the most critical step before beginning any text analytics procedure (Figure 2).

3.2.1. Word Cloud

Word clouds or tag clouds are a type of text data visualization technique that conveys a concept about a subject through the visual representation of words and being an effective technique for analysis, survey, and the collection of written opinions, among other things. Among the other methods for visualizing text data in a graphical format, the word cloud is the most beneficial and straightforward process; not only does a word cloud represent an image of some words, but it also represents something other than the word [1,54,55]. Generally, the term "cloud" has been used to refer to the initial stage of a text document's analysis. Therefore, the frequency of words is the primary consideration when categorizing or visualizing a word cloud.

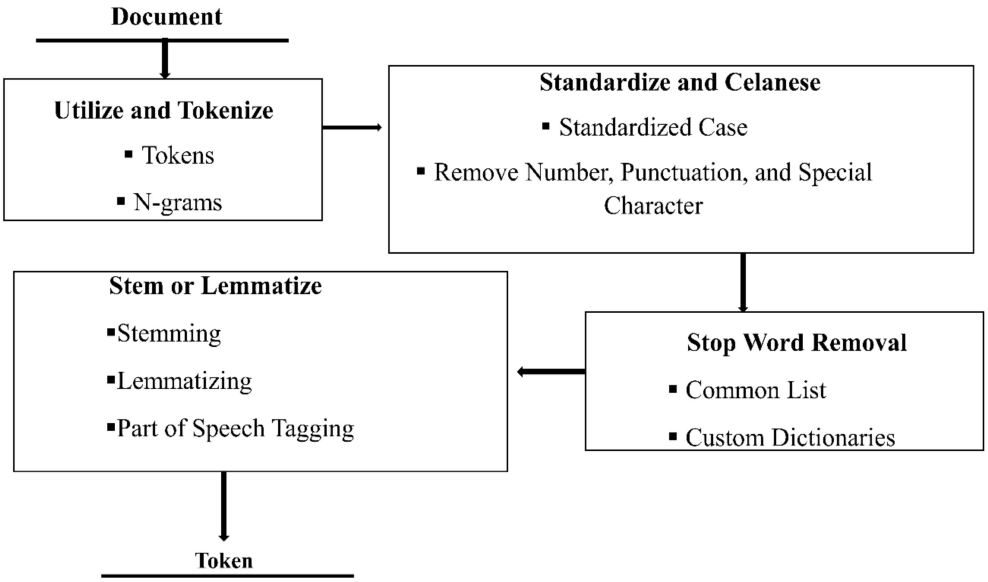

Figure 2. The Text Data Preprocessing, Source [1].

This study used the R package 'quanteda' to pre-process the data after importing the texts, particularly for word cloud and frequency visualization [56]. After removing numbers, punctuation, symbols and hyphens, a lower-case conversion of tokens has been performed with the application of the stemmer. Quanteda, on the other contrary, can automatically use a wrapper for a specified stemming algorithm, such as Martin Porter's. The matrix conversion of the pre-processed data, on the other hand, has been used to apply the subsequent methods. Furthermore, the same package has visualized the word that appeared more than 40 times using a bar chart. The R package 'wordcloud2' has been used in our analysis [57]. We use a circular shape cloud with the size of 0.5.

3.2.2. Sentiment Analysis

When a human peruses a text, sentiment analysis assists in analyzing the human emotion in order to determine its polarity. However, the author of this article follows the tidy text procedure by utilizing a sentiment lexicon and visualization as described in Julie Siege's famous book on the tidy text procedure [18]. The 'tidytext' [58] package in R includes several methods and dictionaries for detecting the presence of defined opinion or emotion in the text by utilizing a sentiment lexicon. In contrast, unigram-based lexicons split lexicons into three categories: AFINN, BING, and NRC [18]. Each of these lexicons has a unique feature for analyzing the output. AFINN is the smallest lexicon developed by Finn Arup Nielsen; It contains 2476 English words with valance scores ranging from minus five to five [59]. On the other hand, Bing is the world's most extensive lexicon, developed by Bind Liu and collaborators. It contains 6788 English words and words with a score of -1 for negative words and +1 for each word. In contrast, NRC is a critical lexicon for various aspects of human sentiment and can significantly impact analysis. Moreover, Saif Mohammad and Peter Tumey classified the NRC lexicon through 6468 words as positive or negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust [36,60].

The analysis made use of several R packages (dplyr, ggplot2, tidyr, and textdata) [61–64] as well as the core package 'tidytext.' Initially, the raw data were converted to tidy text, and then the NRC lexicon was applied to the data, resulting in a bar chart of about ten specified emotions for each different year. Additionally, the Bing lexicon was used in this study because it contains the most significant number of containable English words. After applying the BING algorithm to the tidy approach, the most frequently occurring positive and negative words were visualized.

3.2.3. Cluster Analysis

In-text analytics and clustering can be accomplished using two popular terms: term-document matrix (TDM) or document-term matrix (DTM). The distance and similarity between terms or documents are typically calculated. However, the procedure is primarily implemented as a clustering concept similar to the multivariate statistical tool or data mining concept. Since text analytics utilizes conventional hierarchical clustering techniques, the technical procedures have not been discussed in detail. Readers are recommended to consult the referred book for a more comprehensive and concise review of the clustering methodology [1,18].

Similar to the sentiment analysis, the cluster analysis was implemented using several supporting R packages, including cluster, fpc, tm, clue, dbscan, skmeans, proxy, and colorspace [65–71]. Pre-processing the data entails creating a corpus of the data by removing punctuation, numbers, and other unnecessary elements. Then, following creating of a vector space, a matrix of the final data is created following creating a term-document matrix and removing sparse terms. However, the clusters were visualized after calculating the normalized distance using the hierarchical clustering method 'ward.D2'. These specific methods were chosen due to their efficacy in visualizing graphs following a trial of all other methods.

The entire procedure of the analysis and methods have illustrated in Figure 3.

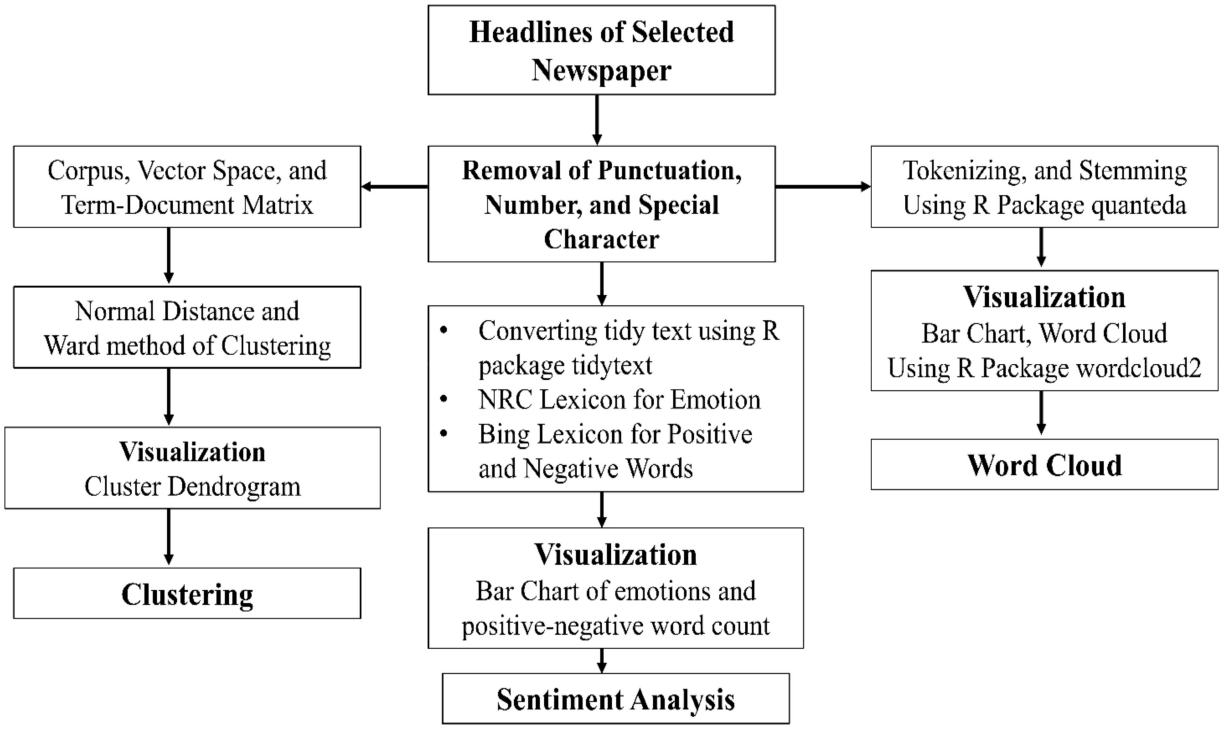

Figure 3. Flowchart of the working procedure.

4. Results

After pre-processing and generating a bag of words, the word cloud was applied because its appearance aids in the precise analysis and word size, position, and boldness. Apart from the visual representation of the word cloud, the top 15 words that appeared more than 40 times were represented by a bar diagram, even though both the bar and word cloud serve the same purpose. For example, in 2018, the "BNP, Awami League (AL), Kill, Rohingya, Khaleda, Tigers, case, bank, government, and attack" were the most frequently mentioned terms (Figure 4a,c). In comparison, the most frequently published front-page title words in 2019 were "Kill, Rohingya, Murder, Tiger, PM, Road, Dhaka, Dengue, Rape, and other" related terms (Figure 4b,d).

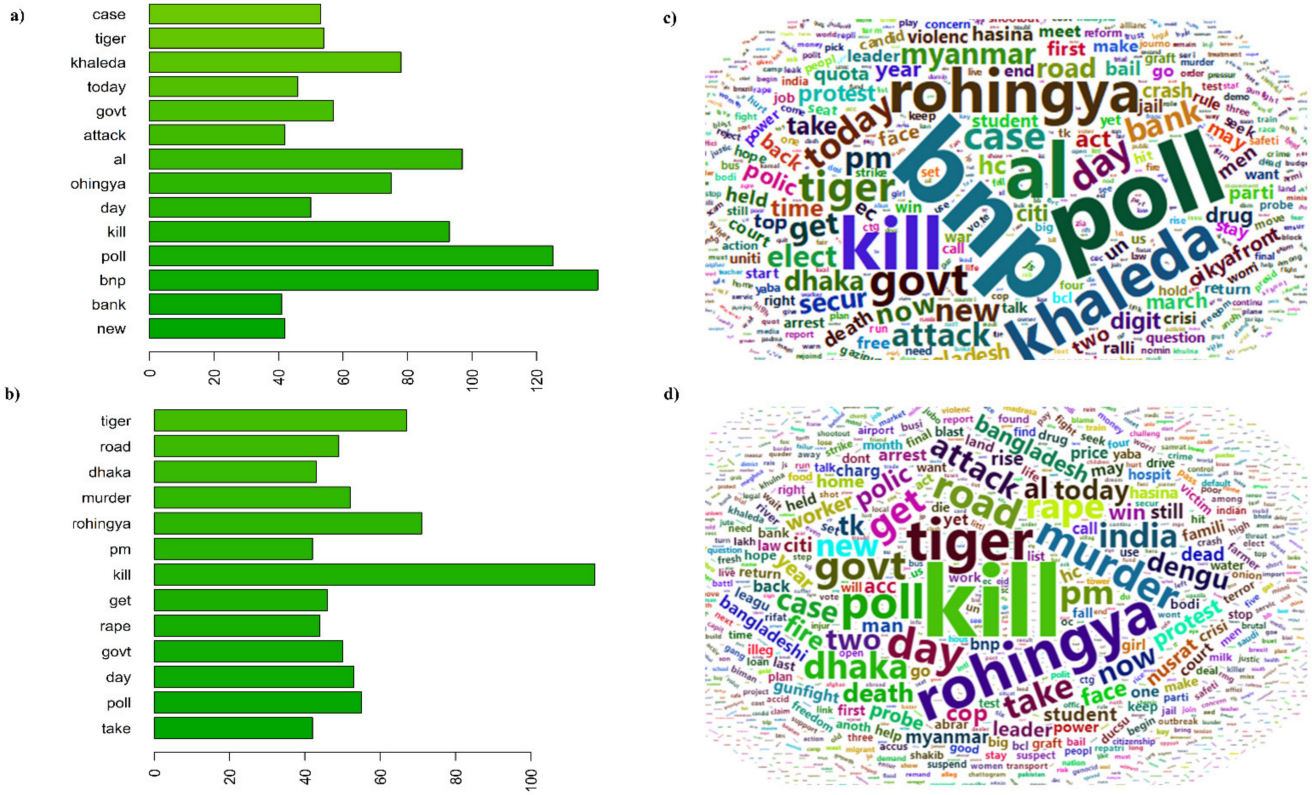

Figure 4. Bar diagram of most frequent words (2019 (**a**) and 2018 (**b**) bar diagram and 2019 (**c**) and 2018 (**d**) word cloud).

The findings presented in word cloud signifying the country's cricket craziness, political turmoil, and Rohingya-related issues. In general, this analysis demonstrates how vividly the text mining approach depicts Bangladesh's social, political, and law-and-order situation, particularly during election season and the country's cricket craze, and also validates the significance of the text mining approach to help understand the overall view of a country during a particular time, in efficient manner.

Additionally, sentiment analysis utilizes the NRC lexicon to express emotions, such as anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive, and negative. Negative, positive, and fearful sentiments ranked highest in both years, while Anger, Sadness, Trust, Anticipation, and Disguised sentiments ranked second in 2018. In comparison, the type words Trust, Anger, Sadness, Anticipation, and Joy appeared in the year 2019 (Figure 5a,b). The BING lexicon was used to represent two different group sentiments for words, e.g., positive and negative. In general, positive words outnumbered negative words in 2018. However, among the positive words, "win, freedom, top, fresh, support, and safe" appear to have been the most frequently used in 2019. On the other hand, the terms "top, free, unity, win, reform, freedom, trust, fair, support, and safe" appear to be the most frequently used positive terms. In comparison, the most frequent negative words in 2018 were "killed, death, crisis, crash, attack, and concern", whereas the most frequent negative words in 2019 were "killed or killing or kills, murder, rape and raped, death, attack or attacks, dead, and graft" (Figure 5c,d).

However, by grouping the words and calculating their similarity using a clustering algorithm, one can gain additional insight into the appeared words. The Wards method of hierarchical clustering was used and produced some exciting results for both years in order to determine the similarities and roots of the words. In 2018, election-related terms such as "oikyafront, election, Khaleda, BNP, polls, killed, bail, case, and police" were also the most similar (Figure 6 (top)). In comparison, terms such as "assault, cops or police, city, and rape; fire and death; dengue and new; and road, and murder; government, poll, killed, tigers" exhibit the most significant similarity in 2019 (Figure 6 (bottom)).

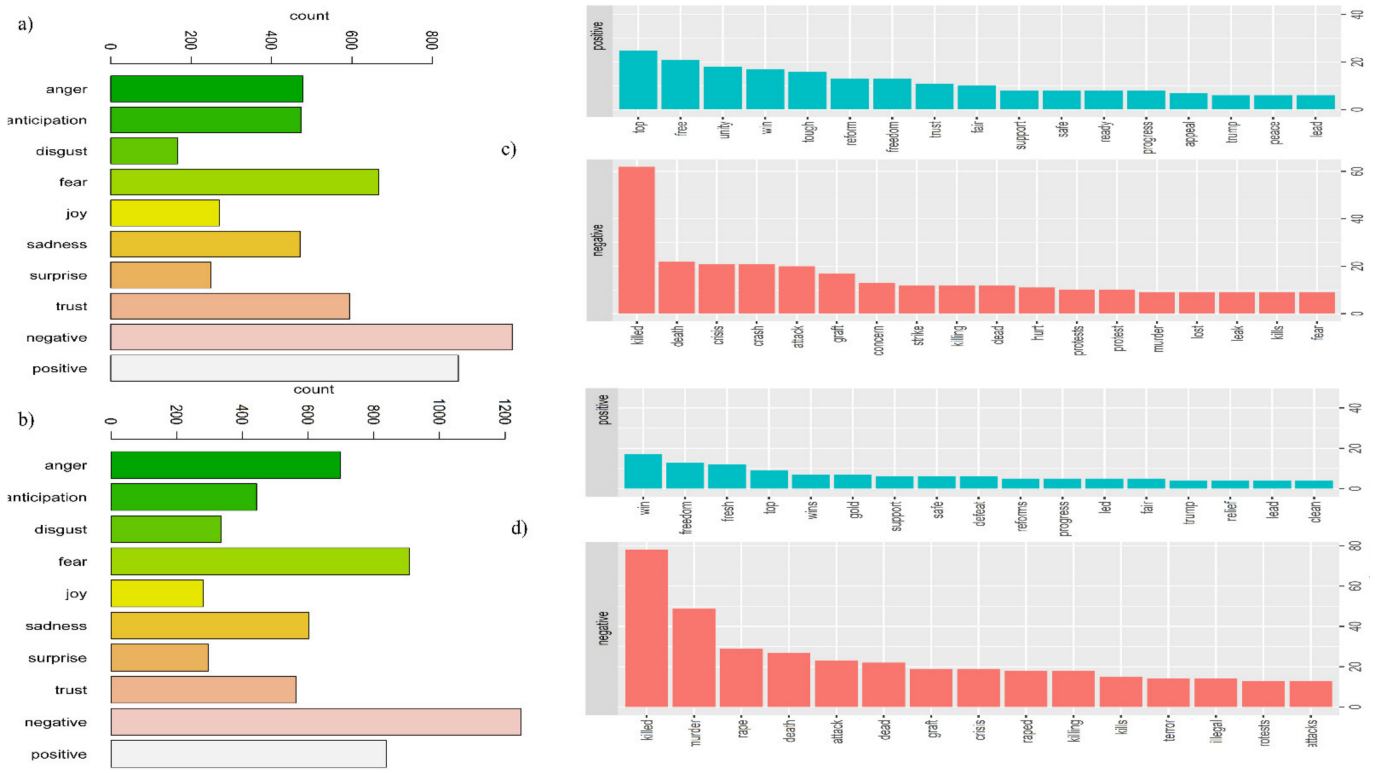

Figure 5. Sentiment Analysis in emotions (left panel, 2019 (**a**) and 2018 (**b**)) and words of sentiment (right panel, 2019 (**c**) and 2018 (**d**)).

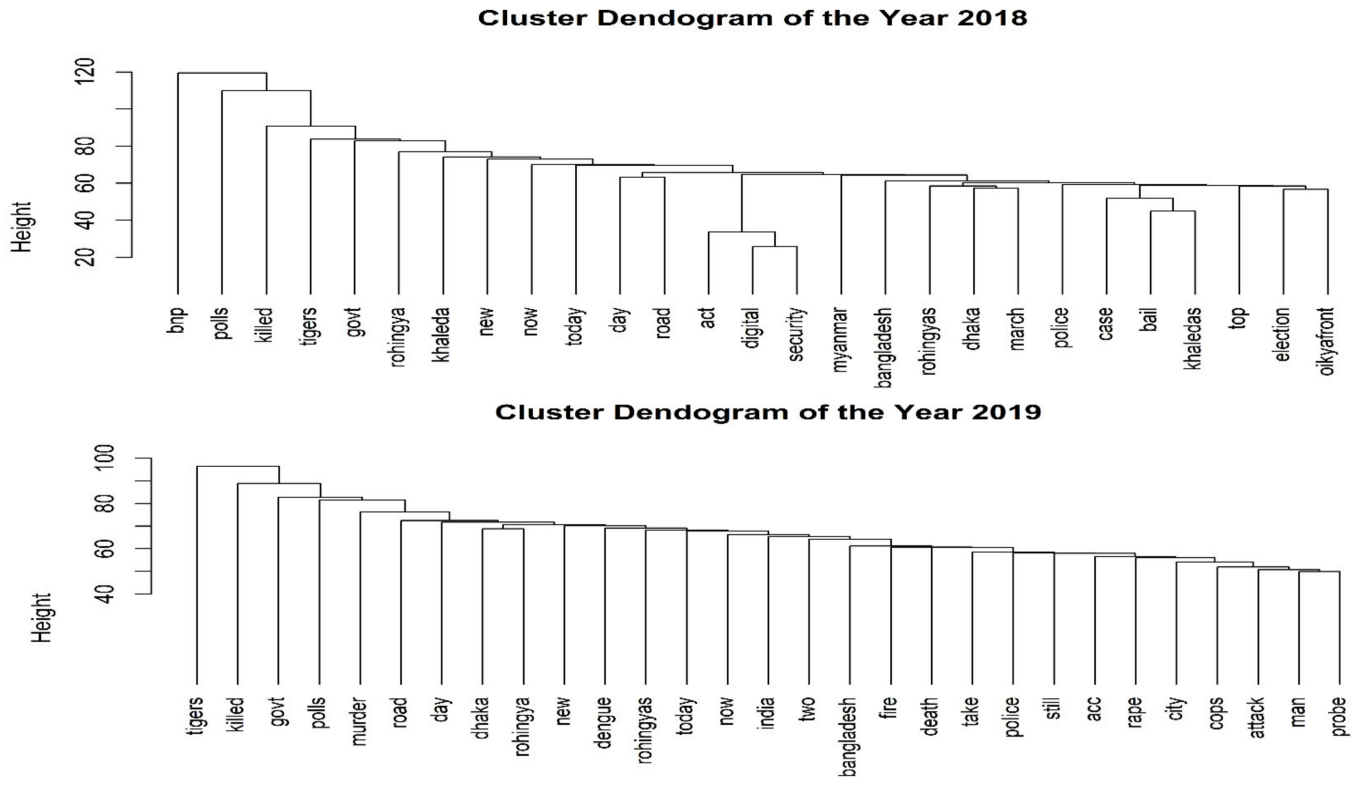

Figure 6. Clustering of the Headlines of Two Years.

5. Discussion

After extracting text from Bangladesh's daily newspaper headlines, *The Daily Star*, the data was pre-processed to create a bag of words. Then, the discussed word cloud, sentiment analysis, and clustering methods were used to determine the most frequently occurring words, the sentiments associated with the appeared words, and clusters or similarities between the words, respectively. The bar diagram and word cloud indicate that the most frequently used terms in 2018 were Awami League (AL), Rohingya, Kill, BNP, Poll, Khaleda, Tigers, case, and bank. Furthermore, the most frequently used words in 2019 were Tiger, PM, Road, Dhaka, Dengue, Kill, Rohingya, Murder, and Rape. As 2018 was Bangladesh's national election year, the political terms BNP (Bangladesh Nationalist Party), Awami League (AL), Khaleda (BNP Chairperson), and Poll have appeared most frequently in newspapers. The terms Rohingya and Kill may appear in reference to the Myanmar ethnic cleaning issue and migration of Rohingyas. Additionally, the term "Tiger" may appear due to the burgeoning popularity and excitement surrounding cricket in the country. In comparison, Kill, Murder, Rohingya, and Tiger may reappear for the same reason they did in 2018.

Newspapers and social media have always been significant sources of information about political and social issues, and their impact is magnified during times of political crisis; for example, the Cambridge Analytica controversy is widely known throughout the world [72,73]. Studies are being conducted to ascertain the impact of newspaper political crises and online political debates [26,33,40]. However, only one study, specifically for Bangladesh, attempts to comprehend the editorial of the same English newspaper, which did not elaborate on the reason for the word's frequency, to the best of our knowledge [51]. However, the appearance of the word Dengue may indicate that dengue fever is flourishing during the mild months of the year, and Rape may come forward due to some immoral rape incidents throughout the year. Finally, the word Road may appear due to an internal road accident or construction issues, as Bangladesh is one of the most dangerous countries in South Asia [74–76], and the word Dhaka may indicate the capital's importance in various issues.

On the other hand, using the BING lexicon, the tidy approach of sentiment analysis defined the topmost positive and negative words, whereas NRC defined the emotion. The visualization used the top fifteen words that contribute to newspaper front-page headlines' positive and negative sentiment. Similarly, in both years, the most prevalent sentiments were Negative, Positive, and Fear. As mentioned previously, the political crisis dominated both years. This may explain why negative and fear-based emotions rank first while positive words rank similarly. Additionally, sentiment analysis using a tidy process reveals that writers frequently use the most hurtful words in headlines, including Killed, Murder, Death, Attack, Crisis, Graft, Dead, Raped, Terror, Illegal, and Attack. This term may elucidate the period's crisis, as elections have historically been a bad time for Bangladesh, dating all the way back to the country's independence. Furthermore, these words reflect the state of the country's law and order at the time.

In addition, positive words such as Top, Fresh, Support, Safe, Win, Freedom, and others appear to be more prevalent in 2019, whereas negative words appear to be much less prevalent in 2018 than in 2019. Similar terms such as Top, Free, Unity, Reform, Liberty, Trust, Fair, Support, Safe, Additionally, Prepared, appear to be prevalent in both years. However, negative terms such as killed, death, crisis, crash, attack, and concern appear to have appeared more frequently in 2018. Newspaper and social media sentiment analysis is relatively common in studies. For example, Zulfadzli Drus's study elucidates 24 related research findings and applications toward text analysis and pattern discovery in a variety of fields, including community development, airport service, threat and fear, business performance, depression level, security, employment, food habits, stock price, and government election, among others [13]. Other studies are also conducted to ascertain a variety of social, political, and business issues. However, while the majority of Bangladeshi newspaper-related studies attempt to provide an algorithm or method for extracting

Bengali newspapers, only two studies have been identified where one attempts to identify patterns of crime according to time and location, and the other proposes a new method for precisely understanding positive and negative news. Thus, our study is unique because it elucidates emotion using well-validated lexicons and attempts to develop a rationale for the analysis results.

Clustering was completely absent from the literature for Bangladeshi articles, even though classification and machine learning models are widely used in international competitions. Additionally, one of our objectives was to determine the pattern of words and their sentiment and frequency. However, the cluster dendrogram illustrates the same thing as the word cloud and sentiment analysis discussed previously. Notably, in 2018, the dendrogram demonstrated the highest degree of similarity between the terms Oikyafront (Political parties' collaboration) and Election. The following similar terms also appeared: Khaleda, Bail, Case, and Police.

Furthermore, BNP, Polls, and, Murder appeared to be most similar, as 2018 was Bangladesh's election year. These facts corroborate the election-related issues, and political crises described previously. However, the Rohingya issue and the digital security act appear to be synonymous, as Bangladesh is currently facing a massive refugee crisis following ethnic cleansing in Myanmar. The year 2019 was fraught with controversy, as the dendrogram words Attack, Cops or Police, City, and Rape demonstrate. Alternatively, fire and death occurred in the same group, indicating a high rate of fire accidents. However, dengue fever was the hot topic in 2019, with new cases reported daily in the newspaper, as reflected in the dendrogram. Rohingya issues were given similar prominence to those in 2018, as they appear to be in a similar cluster. In addition, road accidents and poll-related deaths appeared to be occurring concurrently, as shown in the dendrogram for 2018. Finally, this analysis initially tries to bridge three distinct text analytics techniques to find the pattern and in-depth view of the words, which also linked and discussed with the social and political context of that time. This research can also be an example of capturing an overall summary of a particular time period of a country. Social scientists can also use these techniques to identify the social norms and adopt their research along with conventional research techniques, including interviews, FGD, and others.

6. Conclusions

Text analytics is a growing field that aims to improve people's perceptions of various issues by revealing important hidden information and patterns in text. The rapid growth of natural language processing models and implementations enables the prediction, classification, and identification of real-time information [77]. This research uncovered the pattern and similarity of words on the front page of Bangladesh's *The Daily Star* newspaper over the last two years. The texts were pre-processed using one of the numerous pre-processing and text mining packages available in R. However, among the methods used, the word cloud technique was used to determine the most informative word, indicating that Election, Politics, Cricket, and Rohingya-related terms appeared most frequently in 2018. In comparison, PM, Rohingya, Deaths, Road, Rape, Tiger, Dengue, Poll, and Kill were the most frequently used terms in 2019. The mutual or frequent occurrence of the terms Tiger, Poll, and Rohingya may indicate that in addition to cricket craziness, political and Rohingya-related issues dominated the front page.

On the other hand, using the tidy approach, sentiment analysis was used to determine the most positive and negative words. It reveals that, among other emotions, words associated with Negative, Positive, and Fear emotions were most frequently used in both years. Apart from the preceding, the words "Anger" and "Sadness" ranked second and third, respectively, in 2019. Additionally, trust and anger-related terms ranked second and third, respectively, in 2018. Finally, the clustering methods illustrate similar groups based on the word's distance, indicating that Election, Politics, Deaths, the digital security act, Rohingya, and cricket-related terms shared similarities in 2018. The similarity of the rape, death, road and fire-related word groups added in 2019 complements the 2018 group.

However, this study vividly depicts Bangladesh's social, political, and law-and-order situation, particularly during election time. Interestingly, the refugee crisis and other issues are brought up, demonstrating the similarity. Moreover, social science researchers can also use these methods of analysis as an alternative and adopt them into their qualitative research as it portrays a lucid view of society at a particular time. However, there is scope for applying classifications, models, and a newly proposed algorithm to newspaper text analysis. Additionally, researchers can also use the Bengali newspaper sentiment and text processes. The comments on social media and the online version of the news can be used to analyze and cross-validate the conspicuous pattern of people's views and newspaper sentiment in the future. Moreover, more years and texts can also be considered in further study.

Author Contributions: Conceptualization: A.H., M.K., M.M.H., A.R.; Data curation: A.H.; Formal analysis and Methodology: A.H., M.K. with the supervision of A.R. and M.M.H.; Writing: A.H., M.K., M.M.H.; Revision: M.K., M.M.H., A.R.; Critical Review: M.M.H., A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and materials will be made available upon the request to the authors.

Acknowledgments: We are grateful to the well-wishers and their peers to motivate us for doing this research. Last but not least, the authors would like to sincerely thank the three reviewers, the Editor, and Academic Editor, for their valuable comments and suggestions, which have been used to improve the quality and readability of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

AL	Awami League
AFINN	Lexicon assigns words with a score that runs between −5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment
BING	Lexicon categorizes words in a binary fashion into positive and negative categories
BNP	Bangladesh Nationalist Party
DTM	Document-Term Matrix
NRC	Lexicon categorizes words in a binary fashion ("yes"/"no") into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust
PM	Prime Minister
R	Is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing
URL	Uniform Resource Locator

References

1. Anandarajan, M.; Hill, C.; Nolan, T. Text Preprocessing. In *Practical Text Analytics: Maximizing the Value of Text Data*; Anandarajan, M., Hill, C., Nolan, T., Eds.; Advances in Analytics and Data Science; Springer International Publishing: Cham, Switzerland, 2019; pp. 45–59. ISBN 978-3-319-95663-3.
2. Chen, R.; Xu, W. The Determinants of Online Customer Ratings: A Combined Domain Ontology and Topic Text Analytics Approach. *Electron. Commer. Res.* **2017**, *17*, 31–50. [CrossRef]
3. Cho, Y.-J.; Fu, P.-W.; Wu, C.-C. Popular Research Topics in Marketing Journals, 1995–2014. *J. Interact. Mark.* **2017**, *40*, 52–72. [CrossRef]
4. Heimerl, F.; Lohmann, S.; Lange, S.; Ertl, T. Word Cloud Explorer: Text Analytics Based on Word Clouds. In Proceedings of the 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 6–9 January 2014; pp. 1833–1842.

5. Michelson, M.; Macskassy, S.A. Discovering Users' Topics of Interest on Twitter: A First Look. In Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data, Toronto, ON, Canada, 26–30 October 2010; Association for Computing Machinery: New York, NY, USA; pp. 73–80.
6. Qiao, Z.; Zhang, X.; Zhou, M.; Wang, A.; Fan, W. A Domain Oriented LDA Model for Mining Product Defects from Online Customer Reviews. In Proceedings of the Annual Hawaii International Conference on System Sciences 2017, Waikoloa, HI, USA, 4–7 January 2017; pp. 1821–1830. [CrossRef]
7. Scanfeld, D.; Scanfeld, V.; Larson, E.L. Dissemination of Health Information through Social Networks: Twitter and Antibiotics. *Am. J. Infect. Control* **2010**, *38*, 182–188. [CrossRef]
8. Text Mining. 2020. Available online: https://en.wikipedia.org/wiki/Text_mining (accessed on 10 July 2020).
9. Kaser, O.; Lemire, D. Tag-Cloud Drawing: Algorithms for Cloud Visualization. *arXiv* **2007**, arXiv:cs/0703109.
10. Seifert, C.; Jurgovsky, J.; Granitzer, M. FacetScape: A Visualization for Exploring the Search Space. In Proceedings of the 2014 18th International Conference on Information Visualisation, Paris, France, 16–18 July 2014; pp. 94–101.
11. Lohmann, S.; Heimerl, F.; Bopp, F.; Burch, M.; Ertl, T. Concentri Cloud: Word Cloud Visualization for Multiple Text Documents. In Proceedings of the 2015 19th International Conference on Information Visualisation, Barcelona, Spain, 22–24 July 2015; pp. 114–120.
12. Chowdhury, R.R.; Shahadat Hossain, M.; Hossain, S.; Andersson, K. Analyzing Sentiment of Movie Reviews in Bangla by Applying Machine Learning Techniques. In Proceedings of the 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, Bangladesh, 27–28 September 2019; pp. 1–6.
13. Drus, Z.; Khalid, H. Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Comput. Sci.* **2019**, *161*, 707–714. [CrossRef]
14. Medhat, W.; Hassan, A.; Korashy, H. Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
15. Emam, A.; Alzahrani, M. Opinion Mining Techniques and Tools: A Case Study on an Arab Newspaper. In Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 14–16 December 2017; pp. 292–296.
16. Li, J. From Tweets and Newspapers to Polls A Sentiment Study on 2017 United Kingdom General Election. Available online: http://localhost/handle/1874/373203 (accessed on 11 July 2021).
17. Patodkar, V.N.; Sheikh, I.R. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Int. J. Adv. Res. Comput. Commun. Eng.* **2016**, *5*, 320–322. [CrossRef]
18. Silge, J.; Robinson, D. *Text Mining with R: A Tidy Approach*; O'Reilly Media, Inc.: California, CA, USA, 2017; ISBN 978-1-4919-8162-7.
19. Hu, Z.; Wei, Z.; Sun, H.; Yang, J.; Wei, L. Optimization of Metal Rolling Control Using Soft Computing Approaches: A Review. *Arch. Comput. Methods Eng.* **2021**, *28*, 405–421. [CrossRef]
20. Manik, S.; Gurvinder, S.; Rajinder, S. Design of GA and Ontology Based NLP Frameworks for Online Opinion Mining. *Recent Pat. Eng.* **2019**, *13*, 159–165.
21. Chien, Y.-C.; Liu, M.-C.; Wu, T.-T. Discussion-Record-Based Prediction Model for Creativity Education Using Clustering Methods. *Think. Ski. Creat.* **2020**, *36*, 100650. [CrossRef]
22. Li, N.; Wu, D.D. Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast. *Decis. Support Syst.* **2010**, *48*, 354–368. [CrossRef]
23. Introduction to Text Mining for Social Scientists. Available online: https://campus.sagepub.com/blog/introduction-to-text-mining-for-social-scientists (accessed on 19 July 2021).
24. Karlgren, J.; Li, R.; Milgrom, E.M.M. Text Mining for Processing Interview Data in Computational Social Science. *arXiv* **2020**, arXiv:2011.14037.
25. Nguyen, D.; Liakata, M.; DeDeo, S.; Eisenstein, J.; Mimno, D.; Tromble, R.; Winters, J. How We Do Things with Words: Analyzing Text as Social and Cultural Data. *Front. Artif. Intell.* **2020**, *3*, 62. [CrossRef]
26. Carley, K.M.; Bigrigg, M.W.; Diallo, B. Data-to-Model: A Mixed Initiative Approach for Rapid Ethnographic Assessment. *Comput. Math. Organ. Theory* **2012**, *18*, 300–327. [CrossRef]
27. Lee, C.; Cheng, C.-I.; Zeleke, A. Can Text Mining Technique Be Used as an Alternative Tool for Qualitative Research in Education? In Proceedings of the 15th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Las Vegas, NV, USA, 30 June–2 July 2014; pp. 1–6.
28. Kumar, A.; Jaiswal, A. Empirical Study of Twitter and Tumblr for Sentiment Analysis Using Soft Computing Techniques. In Proceedings of the World Congress on Engineering and Computer Science 2017 Vol I WCECS 2017, San Francisco, CA, USA, 25–27 October 2017. Available online: http://www.iaeng.org/publication/WCECS2017/WCECS2017_pp472-476.pdf (accessed on 2 May 2021).
29. Özyirmidokuz, E.K. Mining Unstructured Turkish Economy News Articles. *Procedia Econ. Financ.* **2014**, *16*, 320–328. [CrossRef]
30. Hagenau, M.; Liebmann, M.; Neumann, D. Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Capturing Features. *Decis. Support Syst.* **2013**, *55*, 685–697. [CrossRef]
31. Ammann, M.; Frey, R.; Verhofen, M. Do Newspaper Articles Predict Aggregate Stock Returns? *J. Behav. Financ.* **2014**, *15*, 195–213. [CrossRef]

32. Geva, T.; Zahavi, J. Empirical Evaluation of an Automated Intraday Stock Recommendation System Incorporating Both Market Data and Textual News. *Decis. Support Syst.* **2014**, *57*, 212–223. [CrossRef]
33. De Fortuny, E.J.; De Smedt, T.; Martens, D.; Daelemans, W. Media Coverage in Times of Political Crisis: A Text Mining Approach. *Expert Syst. Appl.* **2012**, *39*, 11616–11622. [CrossRef]
34. Groth, S.S.; Muntermann, J. An Intraday Market Risk Management Approach Based on Textual Analysis. *Decis. Support Syst.* **2011**, *50*, 680–691. [CrossRef]
35. Bai, X. Predicting Consumer Sentiments from Online Text. *Decis. Support Syst.* **2011**, *50*, 732–742. [CrossRef]
36. Mohammad, S.M.; Turney, P.D. Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [CrossRef]
37. Mostafa, M.M. More than Words: Social Networks' Text Mining for Consumer Brand Sentiments. *Expert Syst. Appl.* **2013**, *40*, 4241–4251. [CrossRef]
38. Ghose, A.; Ipeirotis, P.G. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1498–1512. [CrossRef]
39. Al-Hasan, A.; Yim, D.; Lucas, H. A Tale of Two Movements: Egypt during the Arab Spring and Occupy Wall Street. *IEEE Trans. Eng. Manag.* **2018**, *66*, 84–97. [CrossRef]
40. Serrano-Contreras, I.-J.; García-Marín, J.; Luengo, Ó.G. Measuring Online Political Dialogue: Does Polarization Trigger More Deliberation? *Media Commun.* **2020**, *8*, 63–72. [CrossRef]
41. Hossain, M.S.; Jui, I.J.; Suzana, A.Z. Sentiment Analysis for Bengali Newspaper Headlines. BSc Thesis, BRAC University, Dhaka, Bangladesh, 2017.
42. Bhowmik, N.R.; Arifuzzaman, M.; Mondal, M.R.H.; Islam, M.S. Bangla Text Sentiment Analysis Using Supervised Machine Learning with Extended Lexicon Dictionary. *Nat. Lang. Process. Res.* **2021**, *1*, 34–45. [CrossRef]
43. Arafin Mahtab, S.; Islam, N.; Mahfuzur Rahaman, M. Sentiment Analysis on Bangladesh Cricket with Support Vector Machine. In Proceedings of the 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, Bangladesh, 21–22 September 2018; pp. 1–4.
44. Ahmed, A.; Yousuf, M.A. Sentiment Analysis on Bangla Text Using Long Short-Term Memory (LSTM) Recurrent Neural Network. In Proceedings of the International Conference on Trends in Computational and Cognitive Engineering, Dhaka, Bangladesh, 17–18 December 2020; Kaiser, M.S., Bandyopadhyay, A., Mahmud, M., Ray, K., Eds.; Springer: Singapore, 2021; pp. 181–192.
45. Emon, I.S.; Ahmed, S.S.; Milu, S.A.; Mahtab, S.S. Sentiment Analysis of Bengali Online Reviews Written with English Letter Using Machine Learning Approaches. In Proceedings of the 6th International Conference on Networking, Systems and Security, Dhaka, Bangladesh, 17–19 December 2019; Association for Computing Machinery: New York, NY, USA; pp. 109–115.
46. Chowdhury, S.; Chowdhury, W. Performing Sentiment Analysis in Bangla Microblog Posts. In Proceedings of the 2014 International Conference on Informatics, Electronics Vision (ICIEV), Dhaka, Bangladesh, 23–24 May 2014; pp. 1–6.
47. Mahmud, K.A.; Ahmed, G.T. *Sentiment Analysis on E-Commerce Business in Bangladesh Perspective*; Daffodil International University: Dhaka, Bangladesh, 2019; Report for Bachelor of Science in Computer Science and Engineering.
48. Content Analysis of Agricultural News in the Mainstream Newspapers of Bangladesh. Available online: http://www.ijbssr.com/journal/details/content-analysis-of-agricultural-news-in-the-mainstream-newspapers-of-bangladesh-140132914 (accessed on 2 May 2021).
49. Chowdhury, S.M.M.H.; Tumpa, Z.N.; Khatun, F.; Rabby, S.K.F. Crime Monitoring from Newspaper Data Based on Sentiment Analysis. In Proceedings of the 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 22–23 November 2019; pp. 299–304.
50. Rasmussen, J.; Farhad, A.T.M. Media and Communication Studies. MS Thesis, School of Humanities, Education and Social Sciences, Örebro University, Örebro, Sweden, 2016.
51. Manir, T.I.; Hossain, M.M. Application of Text Mining on the Editorial of a Newspaper of Bangladesh. *Int. J. Comput. Appl.* **2019**, *178*, 23–29.
52. Genilo, J.W.; Asiuzzaman, M.; Osmani, M.M.H. Small Circulation, Big Impact: English Language Newspaper Readability in Bangladesh. *Adv. J. Commun.* **2016**, *4*, 127–148. [CrossRef]
53. The Daily Star. Available online: https://www.thedailystar.net/ (accessed on 11 July 2021).
54. Segall, R. Web-Based Text Mining of Hotel Customer Comments Using SAS ® Text Miner and Megaputer Polyanalyst ®. Available online: https://www.semanticscholar.org/paper/Web-Based-Text-Mining-of-Hotel-Customer-Comments-%C2%AE-Segall/989d52db9226bdba077733f43f0f77d024e78d52 (accessed on 11 July 2021).
55. Chowdhury, S.M.M.H.; Ghosh, P.; Abujar, S.; Afrin, M.; Hossain, S. Sentiment Analysis of Tweet Data: The Study of Sentimental State of Human from Tweet Text. In *Emerging Technologies in Data Mining and Information Security*; Abraham, A., Dutta, P., Mandal, J., Bhattacharya, A., Dutta, S., Eds.; Springer: Singapore, 2018; Volume 813, Advances in Intelligent Systems and Computing.
56. Benoit, K.; Watanabe, K.; Wang, H.; Nulty, P.; Obeng, A.; Müller, S.; Matsuo, A. Quanteda: An R Package for the Quantitative Analysis of Textual Data. *JOSS* **2018**, *3*, 774. [CrossRef]
57. Holtz, Y. The Wordcloud2 Library. Available online: https://www.r-graph-gallery.com/196-the-wordcloud2-library.html (accessed on 19 July 2021).
58. Tidytext: Tidytext: Text Mining Using "Dplyr", "Ggplot2", and Other. in Tidytext: Text Mining Using "Dplyr", "Ggplot2", and Other Tidy Tools. Available online: https://rdrr.io/cran/tidytext/man/tidytext.html (accessed on 11 July 2021).

59. Nielsen, F. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, 30 May 2011; pp. 93–98.
60. Zhang, L.; Wang, S.; Liu, B. Deep Learning for Sentiment Analysis: A Survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1253. [CrossRef]
61. Hvitfeldt, E.; Silge, J. Textdata: Download and Load Various Text Datasets, R Package Version 0.4.1. 2020. Available online: https://cran.r-project.org/web/packages/textdata/index.html (accessed on 11 July 2021).
62. Wickham, H. RStudio Tidyr: Tidy Messy Data. 2021. Available online: https://tidyr.tidyverse.org/reference/tidyr-package.html (accessed on 11 July 2021).
63. Wickham, H.; François, R.; Henry, L.; Müller, K. RStudio Dplyr: A Grammar of Data Manipulation. 2021. Available online: https://dplyr.tidyverse.org/reference/dplyr-package.html (accessed on 11 July 2021).
64. Wickham, H.; Chang, W.; Henry, L.; Pedersen, T.L.; Takahashi, K.; Wilke, C.; Woo, K.; Yutani, H.; Dunnington, D. RStudio Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics; 2021. Available online: https://cran.r-project.org/web/packages/ggplot2/index.html (accessed on 11 July 2021).
65. Hahsler, M.; Piekenbrock, M.; Arya, S.; Mount, D. Dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms, R Package Version 1.1-8. 2021. Available online: https://cran.r-project.org/web/packages/dbscan/index.html (accessed on 11 July 2021).
66. Hennig, C. Fpc: Flexible Procedures for Clustering, R Package Version 2.2-9. 2020. Available online: https://cran.r-project.org/web/packages/fpc/index.html (accessed on 11 July 2021).
67. Hornik, K.; Böhm, W. *Clue: Cluster Ensembles, R Package Version 0.3-59*. 2021. Available online: https://cran.r-project.org/web/packages/clue/index.html (accessed on 11 July 2021).
68. Ihaka, R.; Murrell, P.; Hornik, K.; Fisher, J.C.; Stauffer, R.; Wilke, C.O.; McWhite, C.D.; Zeileis, A. Colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes, R Package Version 2.0-2. 2021. Available online: https://cran.r-project.org/web/packages/colorspace/index.html (accessed on 11 July 2021).
69. Maitra, R.; Ramler, I.P. A k -Mean-Directions Algorithm for Fast Clustering of Data on the Sphere. *J. Comput. Graph. Stat.* **2010**, *19*, 377–396. [CrossRef]
70. Meyer, D.; Buchta, C. Proxy: Distance and Similarity Measures, R Package Version 0.4-26. 2021. Available online: https://cran.r-project.org/web/packages/proxy/index.html (accessed on 11 July 2021).
71. Tm Package—RDocumentation. Available online: https://www.rdocumentation.org/packages/tm/versions/0.7-8 (accessed on 11 July 2021).
72. Facebook's New Controversy Shows How Easily Online Political Ads Can Manipulate You. Available online: https://time.com/5197255/facebook-cambridge-analytica-donald-trump-ads-data/ (accessed on 10 July 2021).

73. Radio, C.B.C. Data Mining Firm behind Trump Election Built Psychological Profiles of Nearly Every American Voter | CBC Radio. Available online: https://www.cbc.ca/radio/day6/episode-359-harvey-weinstein-a-stock-market-for-sneakers-trump-s-data-mining-the-curious-incident-more-1.4348278/data-mining-firm-behind-trump-election-built-psychological-profiles-of-nearly-every-american-voter-1.4348283 (accessed on 10 July 2021).
74. Road Safety in South Asia. Available online: https://www.worldbank.org/en/region/sar/publication/road-safety-in-south-asia (accessed on 10 July 2021).
75. In South Asia, the Case for Road Safety Investment is Stronger than Ever. Available online: https://blogs.worldbank.org/transport/south-asia-case-road-safety-investment-stronger-ever (accessed on 10 July 2021).
76. Road Safety. Available online: https://www.who.int/bangladesh/news/detail/12-05-2019-road-safety (accessed on 10 July 2021).
77. Rahman, A. Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. *Int. J. Artif. Intell.* **2019**, *17*, 44–65.