Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 117-128
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency

Akash Addiga, Sikha Bagui

Department of Computer Science, University of West Florida, Pensacola, FL, USA
Email: Akash288@hotmail.com, bagui@uwf.edu

How to cite this paper: Addiga, A. and Bagui, S. (2022) Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency. *Journal of Computer and Communications*, **10**, 117-128.
https://doi.org/10.4236/jcc.2022.108008

Received: July 18, 2022
Accepted: August 27, 2022
Published: August 30, 2022

Abstract

This study is an exploratory analysis of applying natural language processing techniques such as Term Frequency-Inverse Document Frequency and Sentiment Analysis on Twitter data. The uniqueness of this work is established by determining the overall sentiment of a politician's tweets based on TF-IDF values of terms used in their published tweets. By calculating the TF-IDF value of terms from the corpus, this work displays the correlation between TF-IDF score and polarity. The results of this work show that calculating the TF-IDF score of the corpus allows for a more accurate representation of the overall polarity since terms are given a weight based on their uniqueness and relevance rather than just the frequency at which they appear in the corpus.

Keywords

Sentiment Analysis, Twitter Data, Term Frequency, Inverse Term Frequency, Term Frequency-Inverse Document Frequency (TF-IDF), Social Media

1. Introduction

Social media has become the primary method of communication. Platforms such as Twitter and Facebook allow users to share ideas and opinions, and engage with their target audience. Being a registered user on Twitter gives users the ability to post tweets, a message which can consist of text with 280 characters, photos, GIFs, and videos. Tweets can then be liked, shared, and replied to by other users. Tweets posted by users result in an abundance of data available for data mining, using the Twitter API. Within the past decade, Twitter has become the foremost intermediary between politicians and the public, therefore it consists of a healthy ecosystem of political discussions. The focus of this study is on applying the natural language processing techniques, specifically Term Frequency-Inverse Term Frequency (TF-IDF) to perform sentiment analysis on

collected tweets published by members of the United States (US) Congress.

For this study, the Twitter API was used to collect posted tweets within a specific time frame of members of the United States Congress. TF-IDF was applied to the collected tweets of each member to identify the most important words and phrases that the collected tweets consisted of. Identification of the most important words and phrases gives context and significance to the topic being discussed within the collected tweets. Importance of the words and phrases used by the member is determined based on the frequency at which they occur in the collected data. Sentiment analysis is then performed on the list of identified words of significance to determine the overall sentiment of the text associated with the member of Congress. The sentiment of the text is then labeled as either positive or negative.

Sentiment analysis, also referred to as opinion mining, can be applied to determine whether a published tweet is objective or subjective. The focus of this study is to identify the overall polarity, positive or negative, of individual members of Congress. The distinctive trait of the approach is applying sentiment analysis to significant words identified through TF-IDF rather than using a bag-of-words approach of sentiment analysis. Applying sentiment analysis to TF-IDF will provide a more accurate classification, as TF-IDF consists of words significant in understanding the context of the data.

The rest of this paper is organized as follows. Section 2 presents the related works on preprocessing Twitter data and sentiment analysis; Section 3 presents the background of this work by discussing how Term Frequency-Inverse Document Frequency was used in this work; Section 4 is the "Material and Methods" section that presents the data as well as the methodology; Section 5 is the "Results and Discussion" section and Section 6 presents the conclusion.

2. Related Works

2.1. Related Works on Preprocessing Twitter Data

Myungsook Klassen [1] applied the methods of normalization, discretization, and transformation for preprocessing Twitter data. Discretization was used to remove noisy data by eliminating errors and small data observation variations while normalization was used to scale data to a standard measure which avoids larger numbers from dominating the data. Data transformation allowed for the mapping of data values into other values using linear and non-linear functions to display the relationships between attributes. Klassen states that using normalization, discretization, and transformation methods of preprocessing improves the classification rates of the data.

Hemalatha *et al.* [2] approach data preprocessing in specific tasks, removing URLs, filtering, questions, special characters, and retweets. Removal of URLs is performed due to them not providing any sentimental context and to reduce data noise. Filtering removes extra letters added to a word as the extra letters are irrelevant. Question words such as what, which, and how are removed as they

are not significant in contributing towards polarity. Special characters are also removed as they cannot be processed. Removal of retweets was done because they are tweets of other users and are not relevant in determining the polarity of the original user.

2.2. Related Works on Sentiment Analysis

Jadon *et al.* [3] apply the approach of sentiment analysis to bigdata using the Hadoop ecosystem, and Naïve Bayes and Support Vector Machine (SVM) algorithms. Sentiment analysis, also known as opinion mining, establishes the overall attitude and viewpoint of users on a specific topic by way of natural language processing and text analysis on a set of data obtained from social media. Primarily used for text classification, the Naïve Bayes classifier is used to filter out the unwanted data by assigning labels to a large dataset and assuming that variables are not correlated to one another. Also used significantly as a classification algorithm for sentiment analysis, Support Vector Machine classifies by finding the hyper-plane which separates classes.

Khan and Malviya [4] propose an approach to sentiment analysis of twitter data using the Hadoop framework along with deep learning. The proposed approach is divided into two steps: the extraction step and the classification step. In the extraction step, the twitter data is feature extracted (the topic being discussed is identified) by the Hadoop cluster and the classification step consists of the features extracted in the previous steps being classified using deep recurrent neural network classification as positive or negative.

Neethu and Rajasree [5] apply the different Machine learning approaches for sentiment analysis to analyze twitter data about certain electronic products. The Machine learning approach is carried out by using a test set and a training set which is used to develop a sentiment classifier. The classification model is produced by using the training set and the test set is used for validation of the model. The approach proposed by the authors consists of a preprocessing step, creation of feature vector step, and the sentiment classification step. The creation of feature vector step is composed of two phases; the first phase involves twitter specific features being extracted to remove unwanted text to normalize the tweet/text and the second phase involves features being extracted from the normalized text for the classification step. In the phase that extracts twitter specific features, emoticons and hashtags are given a weight of "1" if they are positive and "−1" if they are negative. The second phase of the extraction is to account for the tweets which might not consist of any twitter specific features and simple text is used with the unigram approach to determine the classification of the text. The classification step classifies the tweets into positive and negative classes by applying the Naïve Bayes, SVM, Maximum Entropy and Ensemble classifiers. This approach determined the Naïve Bayes classifier to be similar in accuracy with the other classifiers, but the most precise method of classification.

Dhawan *et al.* [6] state the categorization of Artificial Intelligence (AI) me-

thods for grouping data into classes, unsupervised learning, and supervised learning. Unsupervised learning does not consist of classification whereas supervised learning consists of classification. The AI method primarily relevant to sentiment analysis is supervised learning. The AI methods consist of a training set and a test set. The authors propose an algorithm to analyze the sentiments of data retrieved using the Twitter API. The proposition consists of checking the authenticity of users to proceed with sentiment analysis to have accurate results. Like previous articles, the author suggests the first step to be the extraction of twitter specific features to check the sentiment polarity of each tweet based on emoticons. If the sentiment is determined to be neutral, then the polarity is set to "0" and if the sentiment is positive or negative, then the polarity is set to "1" and "−1" respectively.

Shelar and Huang [7] obtain tweets which consist of the keywords "donor," "charity," "donations," and "fundraising," using the Tweepy API, to apply sentiment analysis on the data. Sentiment analysis is done using NLTK 2.0.4 powered text classification process, specifically NLTK's VADER (Valence Aware Dictionary and sentiment Reasoner). NLTK's VADER is stated as a "lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media." VADER categorizes the twitter data into three sentiments: positive, negative, or neutral. The authors focus on the following parameters for further analyzing the data: Date, Location, Keywords, Positive polarity, Negative polarity, and Neutral polarity.

Tiwari *et al.* [8] state the three levels of sentiment analysis: Document Level Analysis, Sentence Level Analysis, and Aspect/Entity Level Analysis. The article focuses on the sentence level analysis aspect of sentiment analysis. Sentence level analysis consists of evaluating the emotion of each individual sentence. The authors provide a flowchart which shows the process from preprocessing the twitter data to using a machine learning algorithm for classification of data. The authors used NLTK's stop word corpus to preprocess the data and remove any unnecessary text from the sentences. The data was also converted to be all lowercase. Elongated words, date expression, and punctuation was extracted during the feature determination process. SVM was used as the classifier.

Bagui *et al.* [9] propose a method of looking at sentiment analysis by using short corpuses taken from Twitter data. In this work, multiple axes were used with respect to a subject, as opposed to using a single positive-negative sentiment axis to classify the text with respect to a subject. This methodology focused on microblogging an entry from Twitter into tokens, identifying the correct axis of the sentiment and then using cosine similarity to generate polarization values for classification of each selection into fine-tuned axis values. Results of this study showed that various axes will have to be combined for better results.

3. Background

3.1. Term Frequency-Inverse Document Frequency

The Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, the

most popular term weighting scheme [10], is used to numerically assess the relevance of a word in a document. The score frequency given to a word with TF-IDF determines the importance of a word for the document(s) based on the frequency of the word. The formulas used to calculate the TF-IDF score step-by-step are:

$$tf\left(w,d\right)=\log\left(1+fw,d\right) \tag{1}$$

$$idf\left(w,D\right)=\log\left(\frac{N}{f\left(w,D\right)}\right) \tag{2}$$

$$tfidf\left(w,d,D\right)=tfw,d*idf\left(w,D\right) \tag{3}$$

The notations used in the formula are: N is the number of documents, d is the given document, D is the total documents used, and w is a word in document d. The first equation is used to calculate the term frequency (TF), where $f(w, d)$ is the number of times word w occurs in document d. The second equation is used to calculate the inverse document frequency (IDF) which is used to increase the weighted score of less frequently occurring terms and lower the weighted score of more frequently occurring terms. The IDF formula is calculated by taking the log of N documents divided by $f(w, d)$, the frequency at which the word w occurs in document d. The final equation consists of multiplying the results of the TF and IDF results to calculate the TF-IDF score.

3.2. Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, finds specific patterns in data to identify the emotion and quantify the data to classify it into the following categories: Negative, Neutral, and Positive. Rule-based sentiment analysis involves running preprocessed text against a sentiment lexicon, which is a set of rules that classifies text as negative, neutral, or positive. This method of sentiment analysis is used to analyze text without training or using ML models.

4. Materials and Methods

4.1. Data

The data was collected using the Twitter API, with the Tweepy Python library, consisting of a combined 250,000 tweets of all members of the United States Congress. The timeframe of the collected data is unique to everyone. A set number of tweets were collected, but the final date of all data is the date on which it was collected: February 14th, 2021. The data was collected in the form of a CSV file with the five columns: timestamp, tweet_text, username, all_hashtags, followers, and location. Although 250,000 tweets were collected, for the purpose of this study, only tweets published by two Senators, one from each political party, were utilized. Each individual tweet data file consisted of 199 tweets. An additional two columns, clean_sentence and clean_words, were added to the data once the data was preprocessed.

4.2. Methodology

Figure 1 graphically presents the overall methodology that was used in this study.

4.2.1. Preprocessing

For preprocessing the data, the following was done:

- Removal of usernames from the mentions

Mentions are tweets which contain another person's username. Usernames were removed from the text as they are not meaningful in analyzing the text of the tweet.

- Removal of "#" symbol

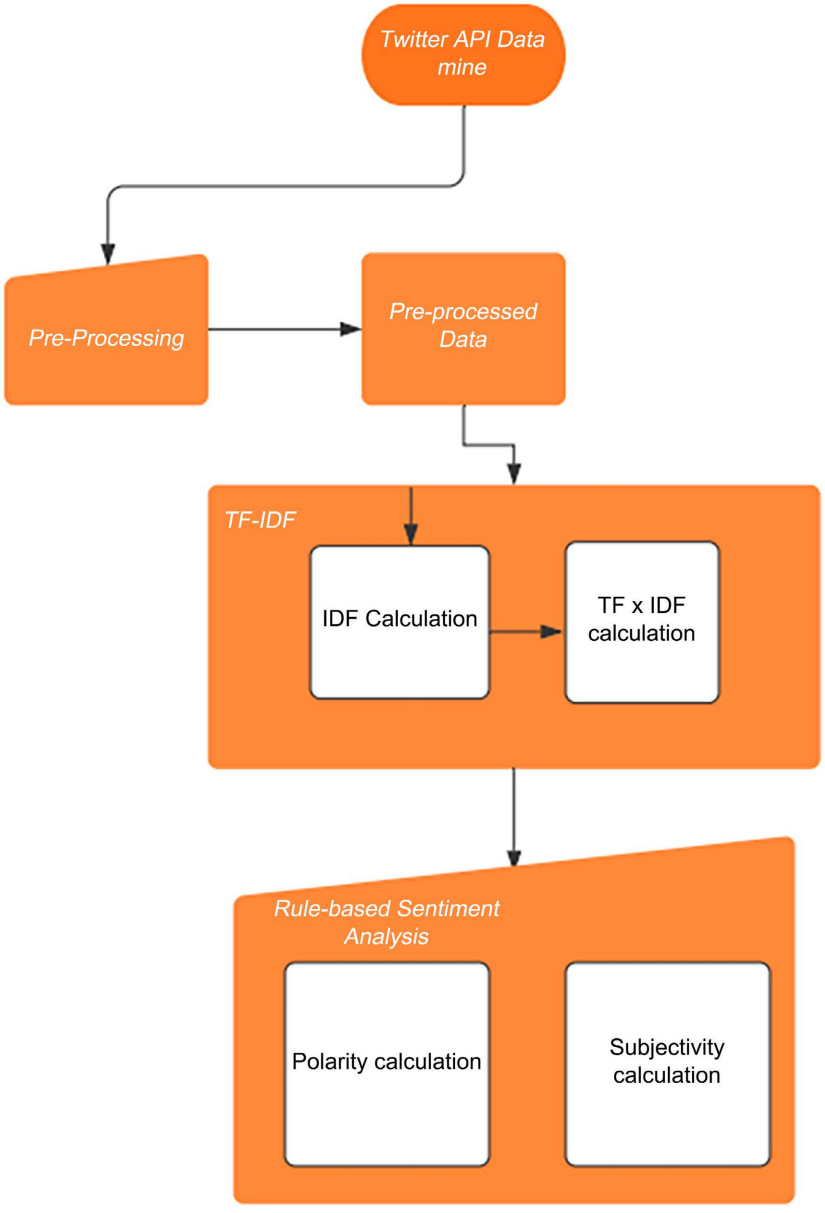

Figure 1. Process flow diagram.

A hashtag is a phrase which consists of "#" followed by the phrase relevant to the topic being discussed. The symbol "#" does not provide any meaningful purpose so it was removed.

- Removal of RT and FAV

Re-tweets consist of "RT" and favorited tweets consist of "FAV." Both were removed since they are not relevant.

- Removal of URLs

URLs were removed from tweets by deleting links that start with "http," "www," "https."

- Removal of punctuation

Punctuation was removed from the tweet as it does not contain any key information.

- Removal of numbers

Numbers were removed as they are not relevant in text processing.

- Removal of stop words

Stop words were commonly used words such as conjunctions. Removal of stop words does not affect the context of the text in the tweet.

- Text normalization

Text normalization transforms text into a single form, lower-case or upper-case, to reduce the randomness.

- Lemmatization

Lemmatization transforms the words in the tweet to its root form. This allows for a simpler and smaller collection of text without taking any context away from it.

4.2.2. Calculating TF-IDF to Identify the Headings

Applying Scikit-learn's CountVectorizer on the values from the "clean_words" column of the preprocessed data transforms the corpus into a vector of terms and count of terms. Having the corpus in count vector form allows for the calculation of the IDF. The IDF is calculated by taking the log of N documents over the frequency of the word occurring in the document. Common words with frequent occurrence consist of a lower IDF value as they are less unique to the document. TF-IDF is computed by multiplying the term frequency by the IDF value. Higher TF-IDF value is given to more unique terms from the document and a lower TF-IDF value is given to more common terms.

4.2.3. Sentiment Analysis

Sentiment analysis was done using NLTK 2.0.4's powered text classification process, specifically using NLTK'S VADER (Valence Aware Dictionary and sentiment Reasoner). The sentiment property of NLTK's VADER returns a polarity range of −1.0 to 1.0 for terms in corpus. This approach allows for analysis of text without training or using Machine Learning models.

5. Results & Discussion

This study is an exploratory analysis of applying natural language processing

techniques such as TF-IDF and Sentiment Analysis on Twitter data. The uniqueness of this work is established by determining the overall sentiment of a politician's tweets based on TF-IDF values of terms used in their published tweets. By calculating the TF-IDF value of terms from the corpus, this work displays the correlation of TF-IDF score and polarity. Calculating TF-IDF score of the corpus allows for a more accurate representation of the overall polarity since terms are given a weight based on their uniqueness and relevance rather than just the frequency at which they appear in the corpus.

Figure 2 presents a sample of the IDF-weights, words, TF-IDF, Subjectivity (relative to the tweet), Polarity and Sentiment (of tweet) for US Senator 1.

Figure 3 presents the correlation of the TF-IDF score, and the polarity of tweets and Figure 4 presents the correlation of IDF score and the polarity of tweets published by US Senator 1. Figure 3 and Figure 4 show that the polarity of the sentiment was highly neutral and more on the positive side.

Figure 5 and Figure 6 present the count and percentage of tweets respectively, by US Senator 1, categorized by sentiment. Figure 5 and Figure 6 show that the neutral tweets were the highest and that there were more positive tweets than negative.

Figure 7 presents a sample of the IDF-weights, words, TF-IDF, Subjectivity (relative to the tweet), Polarity and Sentiment (of tweet) for US Senator 2. Figure 8 presents the correlation of the TF-IDF score, and the polarity of tweets and Figure 9 presents the correlation of IDF score and the polarity of tweets published by US Senator 2. Figure 8 and Figure 9 show that the polarity of the sentiment was highly neutral and more on the positive side.

Figure 8 and Figure 9 present the count and percentage of tweets respectively, by US Senator 2, categorized by sentiment. From Figure 8 and Figure 9 we can see that the neutral tweets were the highest.

Figure 10 and Figure 11 present the count and percentage of tweets respectively, by US Senator 2, categorized by sentiment. Figure 10 and Figure 11 show that the neutral tweets were the highest and that there were slightly more positive tweets than negative period.

	idf_weights	words	tf_idf	Subjectivity	Polarity	Sentiment
_alliance	5.620059	_alliance	0.028303	0.0	0.0	Neutral
mobility	5.620059	mobility	0.028303	0.0	0.0	Neutral
mcleod	5.620059	mcleod	0.028303	0.0	0.0	Neutral
medalxexxcongresss	5.620059	medalxexxcongresss	0.028303	0.0	0.0	Neutral
meeting	5.620059	meeting	0.028303	0.0	0.0	Neutral
...	...	...	...	...	...	...
president	3.094330	president	0.015583	0.0	0.0	Neutral
crisis	3.055109	crisis	0.015386	0.0	0.0	Neutral
wa	3.055109	wa	0.015386	0.0	0.0	Neutral
american	2.786845	american	0.014035	0.0	0.0	Neutral
senate	2.421386	senate	0.012194	0.0	0.0	Neutral

Figure 2. TF-IDF values, in descending order, of tweets of US Senator 1.

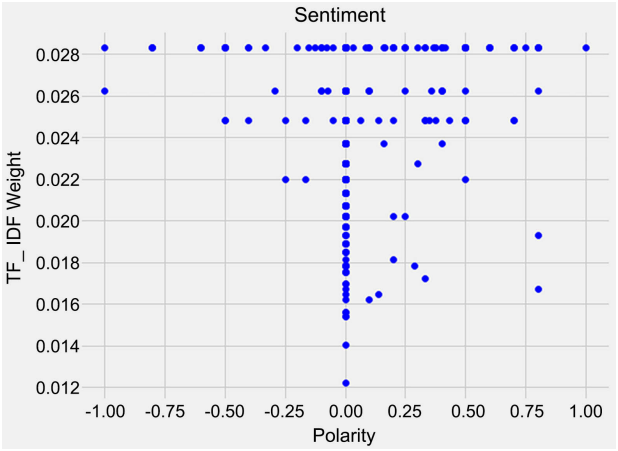

Figure 3. Correlation of TF-IDF score and polarity of tweets published by US Senator 1.

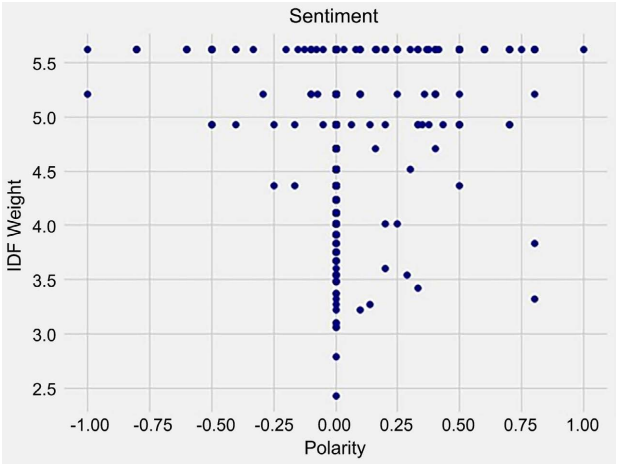

Figure 4. Correlation of IDF score and polarity of tweets published by US Senator 1.

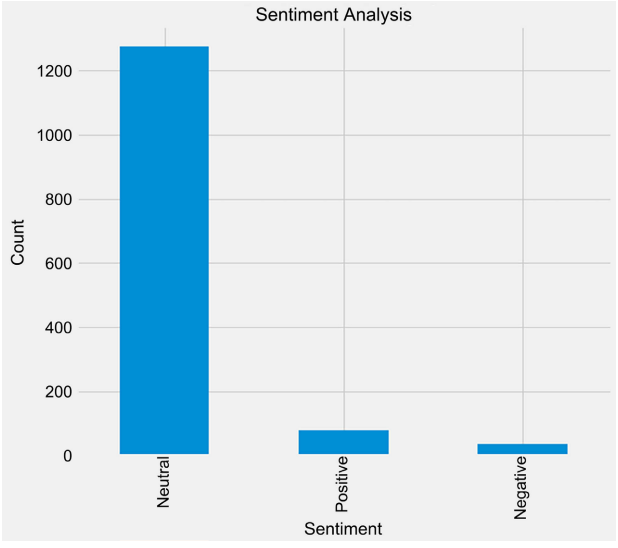

Figure 5. Count of tweets published by US Senator 1 categorized by sentiment.

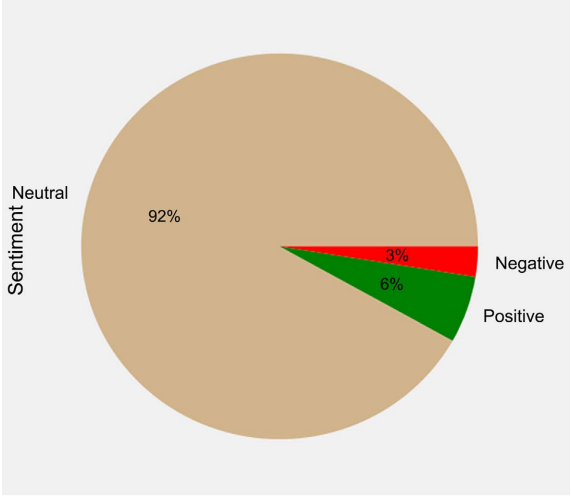

Figure 6. Percentage of tweets published by US Senator 1 categorized by sentiment.

	idf_weights	words	tf_idf	Subjectivity	Polarity	Sentiment
directed	5.610158	directed	0.051744	0.000	0.000	Neutral
sent	5.610158	sent	0.051744	0.000	0.000	Neutral
economic	5.610158	economic	0.051744	0.200	0.200	Positive
saved	5.610158	saved	0.051744	0.000	0.000	Neutral
due	5.610158	due	0.051744	0.375	-0.125	Negative
...	...	...	...	...	...	...
focused	5.610158	focused	0.000000	0.000	0.000	Neutral
focus	5.610158	focus	0.000000	0.000	0.000	Neutral
florida	4.106080	florida	0.000000	0.000	0.000	Neutral
flood	5.610158	flood	0.000000	0.000	0.000	Neutral
youxexxd	5.204693	youxexxd	0.000000	0.000	0.000	Neutral

Figure 7. TF-IDF values, in descending order, of tweets of US Senator 2.

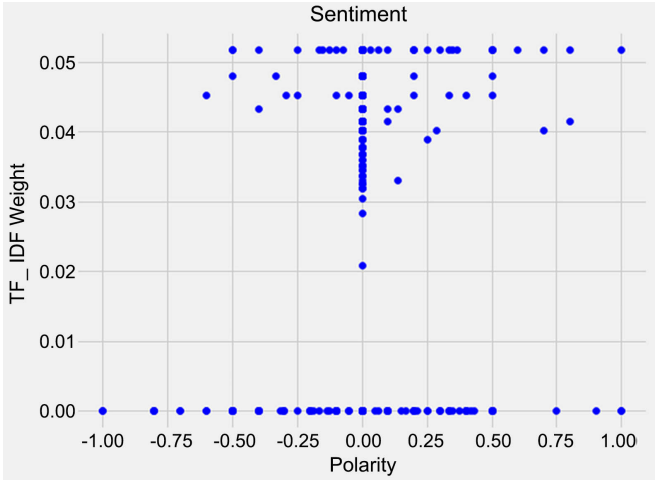

Figure 8. Correlation of TF-IDF score and polarity of tweets published by US Senator 2.

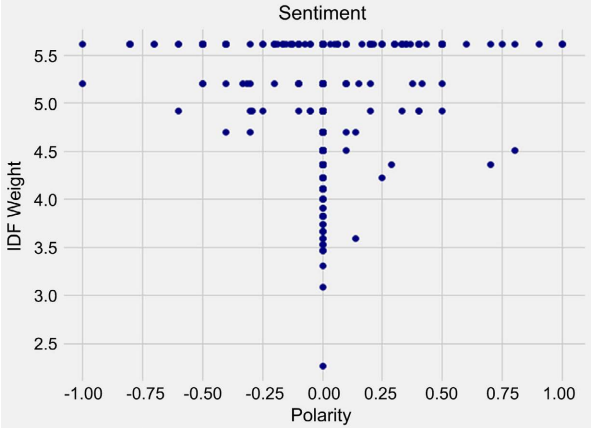

Figure 9. Correlation of IDF weight and polarity of tweets published by US Senator 2.

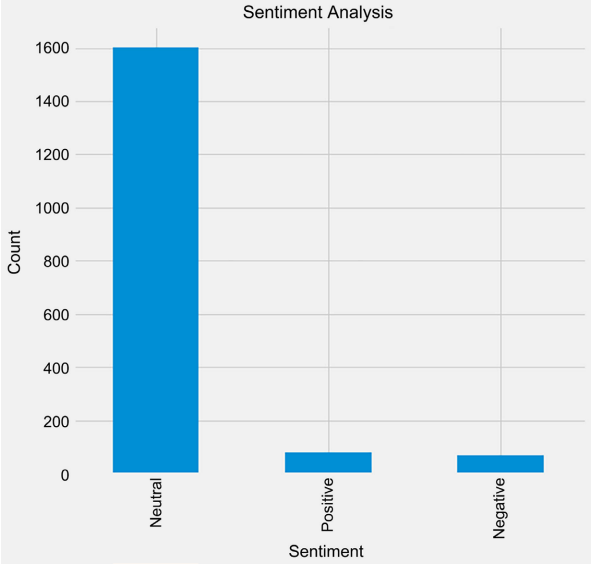

Figure 10. Count of tweets published by US Senator 2 categorized by sentiment.

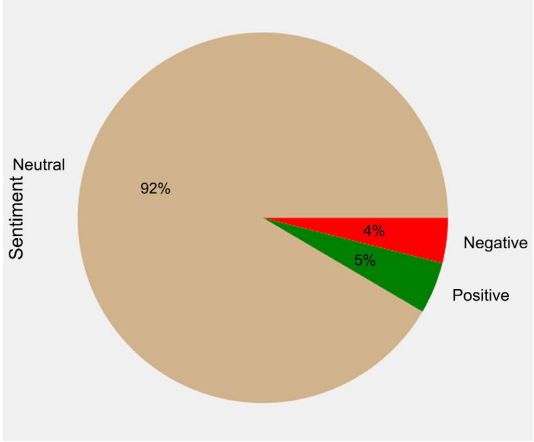

Figure 11. Percentage of tweets published by US Senator 2 categorized by sentiment.

6. Conclusion

In conclusion, calculating the TF-IDF score of the corpus allows for a more accurate representation of the overall polarity since terms are given weights based on their uniqueness and relevance, rather than just the frequency at which they appear in the corpus.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Myngsook, K. (2013) Twitter Data Preprocessing for Spam Detection. 2013 *Future Computing: The Fifth International Conference on Future Computational Technologies and Applications*, May 27-June 1 2013, Thousand Oaks.

[2] Hemalatha, I., Varma, G. and Govardhan, A. (2012) Preprocessing the Informal Text for Efficient Sentiment Analysis. *International Journal of Emerging Trends & Technology in Computer Science*, **1**, 58-61.

[3] Jadon, P., Bhatia, D. and Mishra, D. (2019) A Big Data Approach for Sentiment Analysis of Twitter Data Using Naïve Bayes and SVM Algorithm. 2019 *Sixteenth International Conference on Wireless and Optical Communication Networks* (*WOCN*), 19-21 December 2019, Bhopal. https://doi.org/10.1109/WOCN45266.2019.8995109

[4] Khan, M. and Malviya, A. (2020) Big Data Approach for Sentiment Analysis of Twitter Data Using Hadoop Framework and Deep Learning. 2020 *International Conference on Emerging Trends in Information Technology and Engineering* (*ic-ETITE*), 24-25 February 2020, Vellore. https://doi.org/10.1109/ic-ETITE47903.2020.201

[5] Neethu, M.S. and Rajasree, R. (2013) Sentiment Analysis in Twitter using Machine Learning Techniques. 2013 *Fourth International Conference on Computing, Communications and Networking Technologies* (*ICCCNT*), 4-6 July 2013, Tiruchengode. https://doi.org/10.1109/ICCCNT.2013.6726818

[6] Dhawan, S., Singh, K. and Chauhan, P. (2019) Sentiment Analysis of Twitter Data in Online Social Network. 2019 5*th International Conference on Signal Processing, Computing and Control* (*ISPCC*), 10-12 October 2019, Solan, India. https://doi.org/10.1109/ISPCC48220.2019.8988450

[7] Shelar, A. and Huang, C. (2018) Sentiment Analysis of Twitter Data. 2018 *International Conference on Computational Science and Computational Intelligence* (*CSCI*), 3-4 April 2019, Union. https://doi.org/10.1109/CSCI46756.2018.00252

[8] Tiwari, S., Verma, A., Garg, P. and Bansal, D. (2020) Social Media Sentiment Analysis on Twitter Datasets. 2020 6*th International Conference on Advanced Computing & Communication Systems* (*ICACCS*), 6-7 March 2020, Coimbatore. https://doi.org/10.1109/ICACCS48705.2020.9074208

[9] Bagui, S., Wilber, C. and Ren, K. (2020) Analysis of Political Sentiment From Twitter Data. *Natural Language Processing Research*, **1**, 22-33. https://doi.org/10.2991/nlpr.d.201013.001

[10] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B. and Kochut, K. (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *Proceedings of KDD Bigdas,* August 2017, Halifax.