International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8s
DOI: https://doi.org/10.17762/ijritcc.v11i8s.7174
***Article Received**: 24 April 2023 **Revised:** 07 June 2023 **Accepted:** 26 June 2023*

Sentiment Analysis Framework and Its Application in Geopolitical Scenarios

Dr. Niharika Prasanna Kumar[1], Dr. R. Rajkumar[2]
[1]Associate Professor, Department of ISE, R. V. Institute of Technology and Management, Bangalore, India
niharikapk.rvitm@rvei.edu.in
[2]Associate Professor, Department of ISE, RNS Institute of Technology, Bangalore, India
rajkumar.r@rnsit.ac.in

Abstract—Sentiment analysis or opinion mining involves systamtically extracting, identifying and quantifying subjective information using text analysis and natural language processing algorithms. The paper explores the recent research work in the field of sentiment analysis and categorizes the work into five unique domains. This paper proposes a three stage sentiment analysis framework involving the data gathering, data preparation and the sentiment analysis phases. The proposed framework is applied to the tweets on the Russia-Ukraine conflict in order to understand the current sentiment of the twitterati towards the conflict. The analysis was performed using lexicon based approach and machine learning based approach. The results of the analysis indicate that the machine learning approach provides better performance compared to lexicon based approach. Sentiment analysis also shows that there is still an overall negative sentiment towards the war.

Keywords-sentiment analysis, twitter, machine learning, russia-ukraine conflict, lexicon

I. INTRODUCTION

A sentiment can be defined as a feeling that a person develops based on sensory inputs. Hence, sentiment analysis deals with analyzing the thoughts and judgments of people that manifest themselves as a set of finite sentiments in response to audio, visual or textual inputs. This paper explores sentiment analysis of textual information obtained from varying sources including digital mediums like blogs, social media sites etc.

The paper has been divided into following sections. Section II explores the current literature on sentiment analysis and segregates the research into different domains. The paper proposes a sentiment analysis framework in Section III. Section IV applies this sentiment analysis framework to the Russia-Ukraine war and provides the results of the sentiment analysis. Section V concludes the paper with the proposed future work on this topic.

II. SENTIMENT ANALYSIS DOMAINS

Sentiment analysis has found application in varying domains. This section explores these different domains that are currently under active research by researchers.

A. Health

Health and healthcare are areas that impact people. There are couple of ways to apply sentiment analysis in the field of Health. One way is to analyze the health record of patients, doctor notes on patient's medical conditions as well as the medication details and then infer the sentiment expressed by the health practitioners about the health of patients. However, healthcare data is confidential information hence accessing such information in order to perform sentiment analysis is a challenging endeavor.

However public health and infectious diseases is an area that can be explored for sentiment analysis. Social media content like the twitter messages can be used for sentiment analysis. In [1] authors embark on an explorative analysis of the tweets on Covid-19 as the pandemic spread across different regions of the world. The paper describes the change in discourse over time. Using topic modelling authors provide results on how the topics associated with the pandemic changed. Using sentiment polarity analysis, the paper further provides details on the shift in sentiment from positive valence during the first lockdown to a negative valence post the reopening. The authors conduct a subjectivity analysis to prove that, over time, the tweets started gyrating more towards subjective opinion on Covid19 rather than looking objectively at the pandemic. Sentiment analysis was conducted using TextBlob library. The tweets were divided into blocks of 100 days and the average polarity of the tweets were analyzed over these 100-day blocks. Authors were able to fit a polynomial curve of second degree to the result data. The resulting curve is an inverted U curve indicating a positive sentiment at the beginning of the 100-day period which turns to a neutral sentiment which then transforms into a negative sentiment. The negativity mostly stemmed from the severe lockdown and its impact on the livelihood and cashflow of people. Authors list sarcastic tweets as a major limitation of their research work. Sarcastic tweets which seem to be positive were in fact negative and such tweets could color their results.

In [13], authors explore natural language processing (NLP) techniques for sentiment analysis and topic discovery on online discussions related to Covid-19. The authors used a dataset of tweets related to Covid-19 and applied pre-processing techniques such as tokenization, stop-word removal, and stemming. They then trained a Long Short-Term Memory (LSTM) recurrent neural network (RNN) model to perform sentiment analysis and topic discovery on the pre-processed dataset. The sentiment analysis task involved classifying each tweet as either positive, negative, or neutral, based on the sentiment expressed in the text. The topic discovery task involved identifying the main topics discussed in the tweets. The results of the study showed that the LSTM RNN model achieved high accuracy in both sentiment classification and topic discovery tasks. The model was able to accurately classify the sentiment of tweets and identify the main topics discussed in the dataset.

In [14] authors explore the usage of sentiment analysis in healthcare to analyze patient feedback, social media data, and online reviews to gain insights into patient satisfaction, opinion, and sentiment towards healthcare services and providers. The paper argues that sentiment analysis can help the healthcare organizations in improving patient care and enhance patient experience by identifying areas that require improvement. The authors also discuss the challenges of applying sentiment analysis in healthcare, including the need for high-quality data, the complexity of healthcare language, and the need for domain expertise. Authors suggest that machine learning techniques, such as supervised and unsupervised learning, can help address these challenges and improve the accuracy and effectiveness of sentiment analysis in healthcare. The paper also highlights several opportunities for sentiment analysis in healthcare, including predicting patient satisfaction, identifying trends and patterns in patient feedback, and improving healthcare quality and outcomes. The authors conclude that sentiment analysis can help healthcare organizations make data-driven decisions and improve patient outcomes by providing valuable insights into patient sentiment and opinion

B. Countries and Politics

One of the most widely followed event of the year 2022 was the war between Russia and Ukraine. The war was prominently discussed on twitter with varying viewpoints and sentiments expressed by the participants. In [3] authors keep collating chronologically all the tweets related to the Russia-Ukraine war in the form of twitter IDs (https://github.com/echen102/ukraine-russia). Tools like Hydrator can then be used to download the tweet matching these twitter ID. Ref. [4] provides another source for tweets on the Russia-Ukraine war.

The transition of political power in Afghanistan and its impact on the lives and livelihood of people under Taliban rule is another topic that has seen a lot of discussion in social media. In [8], the authors perform sentiment analysis of twitter message on various facets of Afghanistan and the Taliban. Authors propose a five-phase process comprising of Data Extraction, Data Annotation, Feature Engineering, ML Model Generation and Prediction phase. In the Data Extraction phase, relevant tweets are cleaned. As a part of Data Annotation phase the tweets are annotated with positive, negative, and neutral sentiment using TextBlob library. This annotated text forms the dataset that shall be used for ML Model Generation. Before the ML model can be generated, the data is subjected to feature engineering. Machine Learning models work with numbers and the annotated data is in textual form. Hence authors propose four methods of feature engineering, namely, Bag of Words (BOW), Term Frequency - Inverse Document Frequency (TF-IDF), Word2Vec and Hybrid Feature Engineering (using Chi2). Experimental results prove that Hybrid feature engineering provides better result. This “feature-engineered” dataset is then used to train five machine learning models, namely, Logistic Function based linear model (LR), Tree based ensemble model (ETC), Support Vector Machine (SVM), Gaussian normal distribution based Naïve Bayes model and K-Nearest Neighbor model (KNN). Among these five models, SVM had the best accuracy, precision, recall and F1-score. In addition to the machine learning models, the paper explores deep learning models like CNN-LSTM, LSTM and GRU for sentiment analysis. Among these three deep learning approaches, CNN-LSTM had better performance values for accuracy, precision, recall and F1-score.

In [15], the authors explore the possibility of applying sentiment analysis to the political security threats by analyzing the sentiment expressed in news articles, social media posts, and other online sources. They propose a framework that combines a lexicon-based approach with machine learning techniques to analyze text data and predict political security threats. The lexicon-based approach involves using a pre-defined set of words and phrases to identify the sentiment expressed in the text. The authors developed a custom lexicon for political security threats and used it to analyze text data. Several machine learning models like Decision Tree, Random Forest, and Support Vector Machine are then trained to predict political security threats based on the sentiment analysis results. The authors evaluated the performance of the framework using a dataset of news articles related to political security threats. Metrics like the accuracy, precision, recall, and F1-score for each of these models were analyzed and it was found that the Random Forest model performed better than the other models. The results of the study show that the framework can accurately predict political security threats using a combination of lexicon-based approach and machine learning techniques.

The paper [19] explores how Natural Language Processing (NLP) techniques can be used to analyze the perception of world leaders on the popular social media platform, Reddit. The paper uses state-of-the-art NLP algorithms such as Flair, DistilBERT, and Text Blob Analysis to classify user comments collected from Reddit's API into positive, negative, or neutral sentiments. The end goal is to rank the chosen world leaders based on their likability. The paper highlights the potential of this approach in predicting public opinion, as well as the outcomes of elections or polls. As more people spend increasing hours on social media platforms, such as Reddit, analyzing user-generated content anonymously can be a powerful tool to understand the public's perception of public personalities.

C. eCommerce

Reviews on ecommerce sites like amazon can help sway the purchasing decisions of users. Hence it is important to filter the reviews and weed out the reviews that appear to be fake. In [10] authors use a multistep approach to detect fake review including detecting fake positive and fake negative sentiment. The initial screening process discards product reviews that (a) comprise same review comment repeated by multiple users (b) have usernames made up of only numbers (c) have just a rating and no review content (d) give a very high rating but have a negative sentiment or vice versa (e) have very short content. Creating a labelled dataset is laborious and time consuming, hence authors propose an active learning process wherein, unlabeled data is fed to a semi-supervised learning process wherein the learning algorithm labels the data that is straightforward. However, it seeks feedback from the users when it is not able to classify the text as fake or genuine. This process leverages the sentiment of the comment (positive or negative) to determine the veracity of the review. Once the data is labelled via semi-supervised learning, the labelled data is then used to create a machine learning model using Decision Tree and Random Forest. The learnt model is then used on unlabeled reviews to detect fake reviews.

In [12] authors explore sentiment analysis of the products reviews on ecommerce site Amazon. Authors rely on the filtering efficiency of Amazon’s screening as well as the ratings provided in the reviews as a precursor for the algorithm. Authors propose steps to clean the data and then compute the sentiment score. Authors collect 5.1 million product reviews spanning across beauty, books, electronics, and home segments. A word or a phrase that conveys a sentiment is called sentiment token. For each of the reviews, authors compute the average sentiment score of the sentiment tokens and correlate it against the review ratings. A positive sentiment should have a rating of 3 stars or more. Sentiment polarity categorization involves categorizing the text as positive or negative. Authors propose two step sentiment polarity categorization (1) Sentence level categorization and (2) review level categorization. For sentence level categorization, bag-of-words approach is used i.e., if a sentence has more positive words than negative words, then the sentence is considered positive. Else it is marked as negative. The resulting dataset is then converted to feature vectors and used to train machine learning model. Three machine learning models based on Random Forest, SVM and Naïve Bayes were trained for sentiment analysis. Experimental results show that, for sentence level categorization, as the number of feature vectors increase, Random Forest performs better than Naïve Bayes and SVM with an F1 score of 0.95. But at the review level, SVM and Naïve Bayes perform better than random forest with an F1 score of 0.81 and 0.78 respectively.

Reviews are not only important for consumer durables on these ecommerce sites, they play an equally crucial role during restaurant selection for a meal. In [11] authors attempt to detect fake positive and negative reviews of restaurants on Yelp website. A total of 5853 reviews of 201 hotel by 37083 reviewers were used in the modelling the system. Yelp’s own fake review classification results were used to label the data. 4,709 reviews were labelled as real and the rest 1,144 reviews were labelled as fake. The reviews were split into 70-30 where 70% reviews were used to train the models and the rest 30% were used to test the model. Five machine learning models, namely, SVM, Random Forest, Logistic Regression, K-Nearest Neighbor and Naïve Bayes were trained. Performance comparison of the different machine learning algorithms shows that KNN (with k =7) generates the best results with an F1-score of 82.4%. Authors generate a confusion matrix for the fake detection logic to tabulate the different values of the confusion matrix for all the machine learning algorithms that were used for the fake review detection process.

The authors in [16] explore the application of sentiment analysis of opinion of users on various products or services. The proposed model uses a combination of CNN and BiGRU to capture both local and global features of the text data and an attention mechanism to focus on the most important parts of the text. The authors collected a dataset of comment texts related to online products and services and pre-processed the data by tokenizing and cleaning the text. They then trained the CNN-BiGRU-Attention model on the dataset to classify the sentiment of the comments as positive, negative, or neutral. The authors evaluated the performance of the model using several metrics, including accuracy, precision, recall, and F1-score. The performance of these models are compared against several baseline models, including Support Vector Machine and Naive Bayes. The results of the study show that the CNN-BiGRU-Attention model outperformed the baseline models and achieved 3 to 5% higher accuracy, demonstrating the effectiveness of the proposed approach for sentiment analysis of comment texts.

The paper [20] describes a new sentiment analysis model that is specifically designed to handle e-commerce product review data. The traditional methods of sentiment analysis have been found to be inadequate in mining this type of data, hence the need for a new approach. The proposed model is based on deep learning and utilizes a convolutional neural network (CNN) to mine the deep association between the feature set and emotion tag, which is then used to train an emotion classifier. The text is first divided into words, and the word vector is combined with the word frequency to be input into the neural network for training. The comments are then categorized into positive and negative categories. The results of the experiments demonstrate that this model is highly effective in product feature extraction and sentiment classification, making it a useful tool for online review analysis.

D. Humanities

Humanities [14] includes the study of literature, arts, history and philosophy. Hence humanities is an interesting area for sentiment analysis.

Translation of text from one language to another is an important aspect of literature. In order to reach a wider audience, literary work is usually translated in to different languages. In [2] the authors compare the translation of Bhagavad Gita (ancient literary work from India) by three different translators and try to explore the sentiment expressed by these three translators. A deep learning-based language model known as Bidirectional Encoder Representations from Transformers (BERT) is used to conduct the sentiment analysis. Authors conclude that even though the vocabulary used by the three translators is different, the sentiment expressed in the text is similar. Hence the emotion conveyed in all the three translations is similar and comparable. Senwave dataset comprising of 10,000 tweets spanning 10 sentiments like “optimistic”, “pessimistic”, “anxious”, “thankful” etc are used to train the BERT model. The sentiment analysis on the translated text shows that the bigram [pleasure, pain] is the commonly observed positive sentiment across the three translations. The sentiments "annoyed", "optimistic" and "surprised" were the most expressed sentiments in the text. Sentiments like "denial", "anxious", "thankful" and “empathetic” were the least expressed sentiment.

E. Social Networking

Social media and social content have become all pervasive and touch many aspects of people’s life. In [5] authors propose a machine learning based approach for sentiment analysis of news related to Ukraine and Russia. This paper predates the Russia-Ukraine war. Authors experiment with four models based on Naïve Bayes, DMNBtext, NB Multinomial, SVM Machine Learning respectively. In order to train the models the Wordnet-Affect dataset was enhanced to incorporate words from Ukrainian language. Wordnet-Affect is a collection of emotion related words that have been classified as positive, negative and neutral. These words include nouns, verbs, adjectives and adverbs. University of Moldova has translated these words into Russian Language. Authors used this translated word bag and enhanced it by adding Ukrainian words. Authors used the text from Ukrainian site (https://tsn.ua/) and Russian site (http://censor.net.ua/) as the training data. Annotators were provided this data and were asked to annotate the news items as positive, negative and neutral. A subset of the annotated data, with annotations from at least three annotators, were chosen to train the models using machine learning algorithms provided by WEKA (Waikato Environment for Knowledge Analysis) toolkit. Among the four algorithms, Naïve Bayes performed the best on both Ukrainian and Russian news items with an F1 score of 0.82.

Languages have their own nuances and hence language specific sentiment analysis involves understanding these nuances and interpreting the sentiment based on the subtly of the language. In [6] authors perform sentiment analysis on text from Chinese microblogging sites. Authors propose multiple dictionaries for this purpose which includes an original sentiment dictionary, an adverb dictionary, a conjunction dictionary, an emoji dictionary, a negative and double negative dictionary. Sometimes new and unseen words find a fan following on such sites. Hence an additional dictionary called the new word sentiment dictionary is proposed as well. These dictionaries are applied in conjunction with the semantic rules onto the microblogging text to generate sentiment values. Based on the sentiment value the text is categorized as positive, negative or neutral. In Chinese language, text can be broken down into multiple complex sentences. Each complex sentence can further be split into multiple single sentences. Each of these single sentences is called a clause. Hence sentiment analysis of Chinese text involves applying semantic rules comprising of sentence pattern rules and inter sentence rules to the text and generating the sentiment values. The proposed technique was able to achieve a precision score of 84.9%, 81.8% and 79.7% respectively for the positive, negative and neutral text. The recall scores were 84.3%, 85.3% and 78.5% for positive, negative and neutral text. The F-value was observed to be 84.6%, 80.1% and 82.4% for positive, negative and neutral text respectively.

One of the important challenges of sentiment analysis is to infer the implicit sentiment. There are many approaches for explicit sentiment analysis. However, in order to extract the implicit sentiment in a given sentence, one must look at the context of the text. In [7] authors note that implicit sentiment constitutes about 15-20% of the overall sentiment and are generally expressed via facts and metaphors. Pretrained models like BERT, RoBERTa, XLNet perform well when subjected to explicit sentiment but they fare poorly when tried over text containing implicit sentiment. The paper proposes a multi-layer model called contextBERT that comprises of an Embedding Layer, Encoder Layer and Implicit Sentiment Query Attention Layer. Authors use two datasets namely SMP2019-ECISA that was created by crawling through Chinese sites like Weibo, Mafengwo (tourism), Ctrip (tourism), Autohome (automobile) and author modified dataset called EmoContext-Implicit (which is based on EmoContext dataset). Authors train models based on three well known machine learning models, namely BERT, ROBERTa and ROBETA-Large. In case of BERT Authors use batch-size of 64 and AdamW Optimizer. The learning rate is set to 2e-5. The results are compared against other Well-known models like LTSM, BiLTSM, CsHGCN, NELEC. Performance analysis shows that fine-tuned BERT, ROBERTa and ROBETA-Large were able to achieve an F1 score for neutral, positive and negative score between 0.893 to 0.902, 0.761 to 0.773, and 0.828 to 0.843 respectively.

Danmaku is a video commenting system wherein users can enter comments on a video frame or a scene. Danmaku is very popular on many streaming websites like iQIYI, Tencent Video, and Mango. In [9] authors perform sentiment analysis of Danmaku comments by introducing a new sentiment dictionary that is used for feature extraction and subsequently apply Chinese emotional vocabulary ontology database to the Naïve Bayes algorithm to classify the comments as positive, negative or neutral. The sentiment dictionary was created using the Chinese emotional vocabulary ontology database from the Dalian University of Technology (DUT) that identifies seven sentiment dimensions, namely, happiness, surprise, fear, anger, sadness and disgust. The polarity of happiness and sadness was further divided into seven levels of positive and negative values.

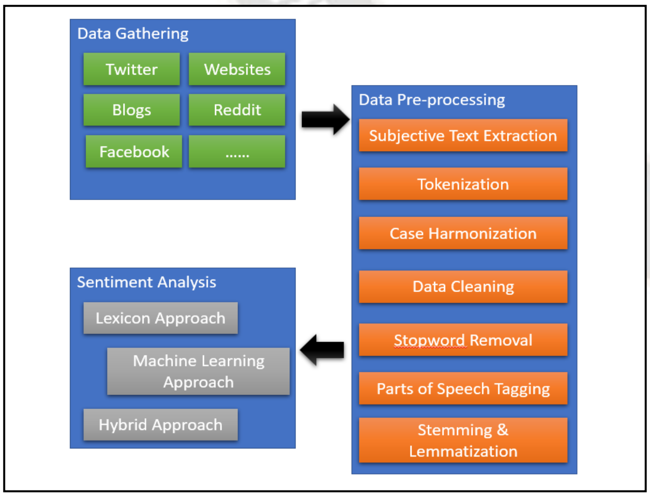


Figure 1. Sentiment Analysis Framework

Authors also experimented with three more algorithms, namely N-gram based Naïve Bayer, N-gram based Support Vector machine model and Convolutional Neutral Network based model. Authors observed that the Sentiment Dictionary based Naïve Bayes model showed the best performance with precision, recall and F1 score of 0.867, 0.783 and 0.823 for positive sentiment and precision recall and F1 score of 0.889, 0.936 and 0.912 for negative sentiment. The overall accuracy for this model was observed to be 0.882.

The study in [17] proposes a novel approach for identifying influential agents in opinion formation by analyzing the sequence of publicly exchanged beliefs in an adaptive social learning protocol. The authors present an algorithm for learning the agents' informativeness and identifying a combination graph, which determines the probability of error of the true hypothesis estimator. The approach also introduces a notion of global agent influence, which quantifies individuals' contribution to learning. The study demonstrates that the suggested approach enables the identification of the most influential agents in the opinion formation process. The authors also describe how to apply the algorithm to Twitter data and illustrate its accuracy in finding global influencers and learning the underlying graph through experiments on both synthetic data and Twitter data.

The authors in [18] investigates the persuasion process and its effects using a novel opinion dynamics model based on social judgment theory (SJT). The authors divided the individual's opinion region into three regions and considered the assimilation phenomenon in the communication process. Numerical experiments were conducted to explore the model's performance in different network topologies, sizes, and initial opinion distributions. The proposed model outperformed other recently proposed models, with a shorter steady-state time, and theoretical proofs showed that it can converge to a stable state in a finite time. The effectiveness of the model was verified using real social platform data and global vaccination data in the decision-making process.

III. SENTIMENT ANALYSIS FRAMEWORK

Sentiment analysis is a multi-stage process. This paper proposes a sentiment analysis framework that shall comprise three stages as shown in Fig. 1

- Data Gathering,
- Data Pre-processing
- Sentiment Analysis

The data gathering and data pre-processing stages can be collectively called as data preparation phase. The next three subsections shall dive deeper into these three stages.

A. *Data Gathering*

The first step in sentiment analysis is to prepare the data and the precursor for data preparation is data gathering. The data for sentiment analysis can come from various sources. Social networks and blogs can provide a trove of data. Hence the data can be collected from various sources like Twitter [1][2][3][4][8], Digital text [2], Websites [5][7][10][11][12], Blogs [6], Redditt, Embedded text in streaming data [9] etc.

B. Data Pre-processing

The data from various data sources cannot be used as-is. The data needs to be pre-processed before it can be used for sentiment analysis. The below subsections describe the various data pre-processing steps. Not all the steps are relevant for all types of problems. Based on the nature of problem being solved, a subset of these steps can be incorporated as a part of data pre-processing phase.

1) Subjective Text Extraction

Subjective text extraction involves selecting sentences that have subjectivity in them. A sentence like “Sun rises in the east” is objective and hence cannot have positive or negative sentiment. Hence, subjective sentences auger well for sentiment analysis.

2) Tokenization

Tokenization is the process of splitting sentences into their respective words. Tokenization is one of the basic steps of pre-processing.

3) Case Harmonization

Text in certain languages can have lower-case and upper-case alphabets. Case conversion is the process of converting all the alphabets and words to a specific case.

4) Data Cleaning

Data can contain Unicode characters, hyperlinks, hashtags, “@” references, punctuation marks and numbers. Such data does not contribute towards subjectivity of the sentence and hence do not aid in sentiment analysis. The data cleaning step involves removing such text from the data.

5) Stopword removal

Words like “a”, “an”, “the” etc. are called stop words. Since these words do not contribute in sentiment analysis, such words are removed from the data. However, in [2], authors deliberately retain stop words as they help in understanding the context of the sentiment

6) Parts of Speech Tagging

English language contains seven parts of speech, namely, noun, pronoun, verb, adverb, adjective, conjunction and interjection. Among the seven parts of speech, Verbs, Adverbs and Adjectives contribute to the sentiment analysis hence these can be retained for sentiment analysis and the rest can be dropped from the data.

7) Stemming and Lemmatization

Stemming and Lemmatization involves converting the words to their basic forms, for example, car, cars, car’s and cars’ are all converted to car.

8) Feature Extraction

If a machine learning based sentiment analysis is being attempted then all the words within a sentence cannot be used as input to the machine learning algorithm. Some of them may need to be omitted. The process of removing certain features (words) that are not important and retaining certain features (words) that are important is called feature extraction. There are many ways of performing feature extraction, (a) using sentiment count values or (b) using cosine similarity between the words in a text (c) using TF-IDF (i.e. True False - Inverse Document). Every word is associated with a TF Score and IDF score. The product of these two values generates a combined TF-IDF score which helps in deciding whether a particular word needs to be part of the feature set or not.

At the end of data pre-processing phase, the data is ready for sentiment analysis.

C. Sentiment Analysis Methods

There are three approaches to sentiment analysis that are described below:

1) Dictionary or Lexicon or Rule based Approach

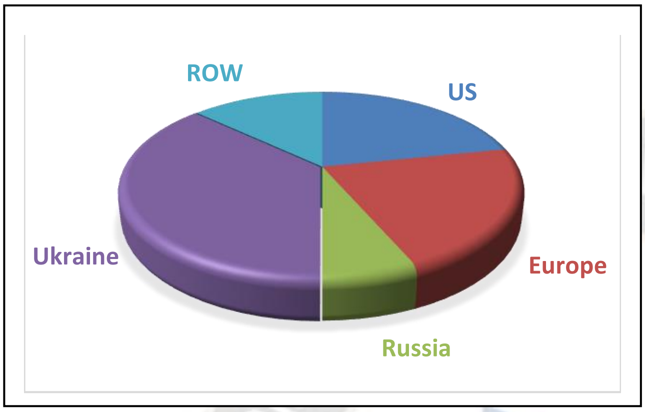
In a dictionary-based approach, a dictionary of words representing positive and negative sentiment is used to analyze the input data. The processed data is passed through a lexicon-based sentiment analyzer. The analyzer parses the text and based on the specified rules, the text is classified as positive, negative or neutral [1]. This is also called bag of words approach. There are different rules to determine the positiveness or the negativity of a sentence. A simple rule is to count the number of positive and negative sentiment words within the text and the majority count determines the sentiment for the text. Some of the widely used lexicon-based sentiment analysis libraries are: TextBlob, VADER

2) Machine Learning based approach

In machine learning-based approach, a data model based on either traditional machine learning algorithms (like Support Vector Machine, Random Forest etc) or deep learning models (like CNN) is trained to detect the sentiment within a specific language. This model is then deployed to perform sentiment analysis on test data as well as live data.

Some of the machine learning algorithm that are widely used in sentiment analysis are: Support Vector Machine (SVM) [5][8][12], K-Nearest Neighbour [8], Naïve Bayes [5][8][9][11][12], Decision Tree [10], Ransom Forest [10][12].

Some of the deep learning algorithms that have found wider application in sentiment analysis are: CNN [8][9], LTSM [7][8], GRU [8], BERT [7], RoBERTa [7], XLNET [7].



3) Hybrid approach

Hybrid approach involves a combination of Dictionary based approach and Machine learning based approach. In a hybrid approach, the dictionary-based sentiment analysis is used to label the data with appropriate sentiments. This labelled data acts as the dataset which is then split into training and test set in order to train the machine learning model [8]. The models thus generated are used on live data to perform sentiment analysis.

IV. SENTIMENT ANALYSIS OF RUSSIA-UKRAINE CONFLICT

The war between Russia and Ukraine has occupied the mindspace of a large cross-section of the society. The impact has been felt in the lives of many people in Eurasia region. Political discourse and the print space of the media has been abuzz with the happenings of the war. Over the past year, the social media has been discussing the topic at length. This paper looks at the current sentiment on the war as the war reaches its first anniversary.

A. Data Source

Twitter messages were used to gauge the sentiment of the larger public about the war. In order to sample the worldview about the war, the messages were chosen such that they span across the globe. To obtain diverse set of opinions, user base was chosen such that it captures both verified and unverified users of twitter.

B. Data Pre-processing

Before initiating the process of sentiment analysis, the data needs to be cleaned. As a part of the data cleaning process, the following steps were undertaken

- Removal of Unicode characters, hyperlinks, hashtags, @username
- Removal of emojis and punctuation marks
- Stopword removal
- Stemming and lemmatization

C. Sentiment Analysis

Sentiment analysis was performed using a dictionary (or lexicon) based approach and machine learning based approach. TextBlob library was used for sentiment analysis as it supports both lexicon-based as well as machine learning based approach.

D. Results and Discussion

This section describes the results of the data analysis as well as sentiment analysis.

1) Geographical Distribution of Tweets

Figure 2. Geographical distribution of tweets.

Fig. 2 shows the geographical distribution of the tweets related to the war. Many tweets originate from Ukraine. One reason for this could be the fact that many news reporters and media units would be stationed in Ukraine leading to a higher percentage of tweets from Ukraine

2) Verified Users v/s Unverified Users

Fig. 3 shows the breakup of tweets with respect to verified users (either blue or golden tick) versus the unverified users. Tweets from verified users, to a certain extent, add veracity to news content. Tweets from unverified users can at times be opinions that can be exaggerated viewpoints. The results clearly showed that a lot more unverified users express their opinion about the conflict.

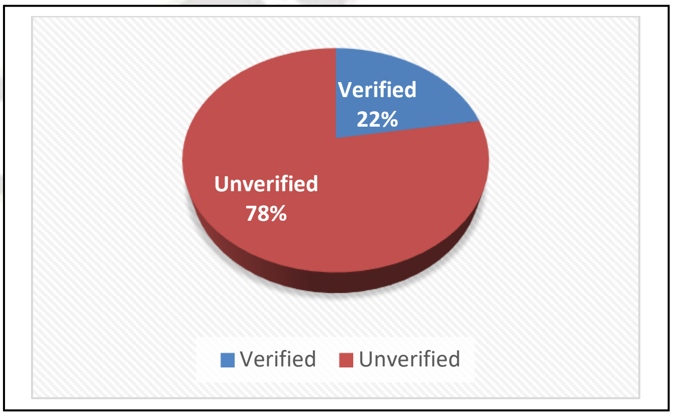


Figure 3. Verified v/s Unverified users

3) Sentiment Analysis

Sentiment analysis was performed on the processed data. TextBlob library was used to perform the sentiment analysis. In order to compare the performance of lexicon based approach and the machine learning based approach, the pattern analyzer and the naïve bayes analyzer was used during the sentiment analysis.

a) Negative Sentiment

In case of the lexicon-based approach, the pattern analyzer was able to detect negative sentiment with an average polarity score of -0.15. A value of -1 would mean the sentiment was completely negative. Hence, in case of tweets with negative sentiment, the pattern analyzer was moderately successfully in detecting the polarity.

However, the Naïve Bayes based analyzer showed a score of 0.56. A value of 1 would mean the sentiment was completely negative. This shows that Naïve Bayes algorithm is able to better predict negative sentiment.

b) Positive Sentiment

In case of the lexicon-based approach, the pattern analyzer was able to detect positive tweets with an average polarity score of 0.16. This seems to be on the lower side since an overwhelmingly positive sentiment would have garnered a polarity of +1.

On the other hand, the machine learning based approach involving the Naïve Bayes algorithm was able to obtain a score of 0.63. Hence the Naïve Bayes algorithm can identify positive sentiment an order of magnitude better than the pattern analyzer.

Hence machine learning based model performed better when compared to the lexicon-based approach. Fig. 4 captures the overall sentiment about the conflict. The overall sentiment is negative which was expected. There could be multiple reasons for this sentiment, for example, loss of lives and livelihood, emotional pain due to separation from near and dear ones etc. On analyzing the tweets to understand the reason for the positivity, it was found that most of the positive tweets were about rebuilding efforts post the ravages of the conflict and regarding the emotional relief and happiness felt among the family members when soldiers return safely back home etc.

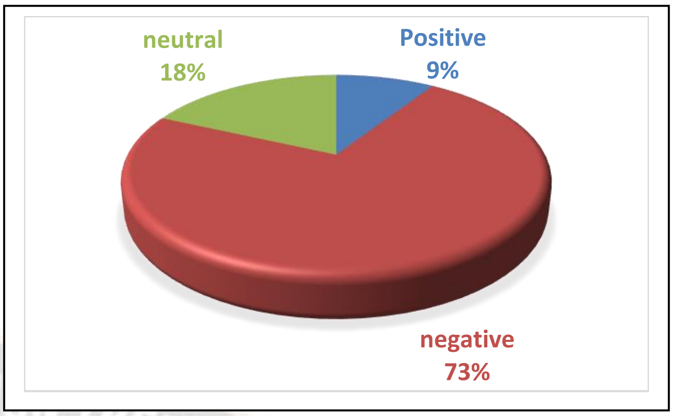


Figure 4. Positive v/s negative tweet

E. Challenges

In case of the Russia-Ukraine related tweets, the sentiment analysis does not perform well when there was sarcasm in the tweets. At times, negative sarcastic tweets were classified as positive and vice versa. This was observed in case of both pattern analyzer as well as the naïve bayes analyzer.

In addition to the above issue, the Pattern analyzer-based sentiment analyzer, at times could not clearly identify subtle positivity in the tweets. For example, the tweet “Ukrainians returned home from Russian captivity” was classified as a neutral tweet whereas there is an inherent positive sentiment within the tweet. The Naïve bayes analyzer was able to classify this tweet as a positive tweet, however the sentiment score was not heavily tilted towards positiveness though.

V. CONCLUSION

This paper explores the recent literature on sentiment analysis. The literature survey reveals that the current research is focused on few domains like Health, Nations and Politics, eCommerce, Humanities and Social Networking. Based on current literature a sentiment analysis framework is proposed that comprises multiple phases like, Data gathering, Data Preparation and Sentiment analysis. Data preparation process involves multiple steps of data processing. Based on problem statement, researchers can choose a subset of these steps. Sentiment analysis can be performed using a lexicon based or a machine learning based approach. This paper applies the proposed Sentiment analysis framework on twitter message dataset on the Russia-Ukraine war. The results of the sentiment analysis reveal that the overall sentiment is still negative on the ongoing war.

A. Future Work

The research work shall be extended to analyze the sentiment of Russia-Ukraine war using deep learning models like CNN, LTSM etc.

REFERENCES

[1] P. Wicke, and M. M. Bolognesi, "Covid-19 Discourse on Twitter: How the topics, sentiments, subjectivity, and figurative frames changed over time," Front. Commun., vol. 6, pp. 1–20, March 2021, doi: 10.3389/fcomm.2021.651997

[2] R. Chandra, and V. Kulkarni, "Semantic and sentiment analysis of selected Bhagavad Gita translations using BERT-based language framework," IEEE Access, vol. 10, pp. 21291 - 21315, February 2022, doi: 10.1109/ACCESS.2022.3152266

[3] E. Chen, and E. Ferrara, "Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia," arXiv, March 2022, doi: 10.48550/arXiv.2203.07488

[4] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, P. Pratikakis, and S. Ioannidis, "Twitter dataset on the Russo-Ukrainian war," arXiv, April 2022, doi:10.48550/arXiv.2204.08530

[5] V. Bobichev, O. Kanishcheva, and O. Cherednichenko, "Sentiment Analysis in the Ukrainian and Russian news," Proc. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 1050-1055, May 2017, doi: 10.1109/UKRCON.2017.8100410

[6] J. Wu, K. Lu, S. Su, and S. Wang, "Chinese micro-blog sentiment analysis based on multiple sentiment dictionaries and semantic rule sets," IEEE Access, vol. 7, pp. 183924-183939, December 2019, doi: 10.1109/ACCESS.2019.2960655

[7] W. Yin, and L. Shang, "ContextBert: Enhanced implicit sentiment analysis using implicit-sentiment-query attention," Proc. 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, July 2022, doi: 10.1109/IJCNN55064.2022.9892878

[8] E. Lee, F. Rustam, I. Ashraf, P. B. Washington, M. Narra, and R. Shafique, "Inquest of current situation in Afghanistan under Taliban rule using sentiment analysis and volume analysis," IEEE Access, vol. 10, pp. 10333-10348, January 2022, doi: 10.1109/ACCESS.2022.3144659

[9] Z. Li, R. Li, and G. Jin, "Sentiment analysis of Danmaku videos based on Naïve Bayes and sentiment dictionary," IEEE Access, vol. 8, pp. 75073-75084, April 2020, doi: 10.1109/ACCESS.2020.2986582

[10] Saini, D. J. B. ., & Qureshi, D. I. . (2021). Feature Extraction and Classification-Based Face Recognition Using Deep Learning Architectures. Research Journal of Computer Systems and Engineering, 2(1), 52:57. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/23

[11] R. P. Kashti, and P. S. Prasad, "Enhancing NLP Techniques for Fake Review Detection," International Research Journal of Engineering and Technology (IRJET), Vol. 06(02), pp. 241-245, February 2019

[12] A. M. Elmogy, U. Tariq , A. Ibrahim, and A. Mohammed, "Fake Reviews Detection using Supervised Machine Learning," International Journal of Advanced Computer Science and Applications, Vol. 12(1), pp. 601-606, January 2021, doi: 10.14569/IJACSA.2021.0120169

[13] X. Fang, and J. Zhan, "Sentiment analysis using product review data," Journal of Big Data, Vol 2(5), June 2015, doi: 10.1186/s40537-015-0015-2

[14] H. Jelodar , Y. Wang , R. Orji, and S.Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach", IEEE Journal of Biomedical and Health Informatics, Vol. 24(10), pp. 2733-2742, October 2020

[15] S. T. Lai and R. Mafas, "Sentiment Analysis in Healthcare: Motives, Challenges & Opportunities pertaining to Machine Learning," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, pp. 1-4, 2022, doi: 10.1109/ICDCECE53908.2022.9792766

[16] Raj, G. ., Verma, A. ., Dalal, P. ., Shukla, A. K. ., & Garg, P. . (2023). Performance Comparison of Several LPWAN Technologies for Energy Constrained IOT Network. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 150–158. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2487

[17] N. A. M. Razali et al., "Political Security Threat Prediction Framework Using Hybrid Lexicon-Based Approach and Machine Learning Technique," in IEEE Access, vol. 11, pp. 17151-17164, 2023, doi: 10.1109/ACCESS.2023.3246162

[18] Y. Dai, Z. Wu and H. Zhang, "Sentiment Analysis of Comment Texts Based on CNN-BiGRU-Attention," 2021 China Automation Congress (CAC), Beijing, China, 2021, pp. 2749-2754, doi: 10.1109/CAC53003.2021.9728140

[19] V. Shumovskaia, M. Kayaalp, M. Cemri and A. H. Sayed, "Discovering Influencers in Opinion Formation Over Social Graphs," in IEEE Open Journal of Signal Processing, vol. 4, pp. 188-207, 2023, doi: 10.1109/OJSP.2023.3261132

[20] Pekka Koskinen, Pieter van der Meer, Michael Steiner, Thomas Keller, Marco Bianchi. Automated Feedback Systems for Programming Assignments using Machine Learning. Kuwait Journal of Machine Learning, 2(2). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/190

[21] H. Xu, X. Xiao, M. Xu and B. Wang, "How Does Persuasion Happen? A Novel Bounded Confidence Opinion Dynamics Model Based on Social Judgment Theory," in IEEE Systems Journal, vol. 17, no. 1, pp. 708-719, March 2023, doi: 10.1109/JSYST.2022.3205724

[22] Muhammad Rahman, Automated Machine Learning for Model Selection and Hyperparameter Optimization , Machine Learning Applications Conference Proceedings, Vol 2 2022.

[23] V. R. Sekar, T. K. R. Kannan, S. N and P. Vijay, "Hybrid Perception Analysis of World Leaders in Reddit using Sentiment Analysis," 2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS), Kochi, India, pp. 1-5, 2023, doi: 10.1109/AICAPS57044.2023.10074005

[24] L. Rong, Z. Weibai and H. Debo, "Sentiment Analysis of Ecommerce Product Review Data Based on Deep Learning," 2021 IEEE 4th Advanced Information Management,

Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2021, pp. 65-68, doi: 10.1109/IMCEC51613.2021.9482223.

[25] Humanities Britannica, https://www.britannica.com/topic/humanities