IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104239 |

Sentiment Analysis for Comments in social Media Using Machine Learning Techniques

Kanishq Mehta, Mithil N, Nidhi V Jain, Sanjana B S, Prof.Naresh Patel K M, Prof.Anusha N

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India

B.E Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India

Assistant Professor, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India

Assistant Professor, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India

ABSTRACT: Sentiment analysis, known as opinion mining, is a technique to determine the emotional undertone of a document. This is a common method used by organizations to identify and group ideas regarding a certain good, service, or concept. In recent years, the use of social media platforms has increased significantly people tend to use social media to express their sentiments and emotions in the form of text and emojis in a more clear and efficient way. In the current scenario, there is limited research done on an understanding of the text with emojis which are used to show sentiments or emotions in an accurate manner. To overcome this problem,we will use Machine Learning algorithms to generate more accurate results which dramatically improves the performance of sentiment polarity prediction.

KEYWORDS: Aspect-based opinion mining, Bidirectional long short-term memory,Machine learning Sentiment analysis, Sentiment intensity lexicon.

I. INTRODUCTION

Billions of social media posts are created every day. With the increase in social media, there has been an increase in the use of text and emojis. Emoji is a pictorial representation and Emojis are widely used in social media to express emotions, moods, and ideas, especially when people struggle to express their emotions through pure text. These emojis have great value as they have a big impact on expressing the sentiment of a sentence. Sentiment analysis is an NLP technique for analysing a text and determining the attitude that lies beneath it. Emojis are not only being used to improve sentiment analysis but also for other composite tasks like sarcasm detection. The polarity of like/dislike feelings is generally the focus of current sentiment analysis approaches. It is relatively easy to define simply positive or negative emotions, but it is far more challenging to define a complete and distinct collection of emotions. While text-based communication may not provide auditory or visual signals, alternative ways can convey emotion. Emojis and emoticons are extensively used in social media to communicate emotions, moods, and thoughts.These emojis and emoticons are extremely valuable since they significantly influence communicating a sentence's sentiment polarity. Emojis and emoticons are thus valuable tools for sentiment analysis. The widespread usage of emojis has piqued the interest of academics because they provide essential semantical and sentimental information to visually complement textual data, which is crucial for deciphering the inherent emotional signals in texts. Emoji embeddings, for example, have been proposed to better comprehend the semantics of emojis and embedding vectors may be used to display and

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104239 |

forecast emoji usage based on their contexts [1].Emoji research currently has two significant drawbacks. As a result, prior emoji embedding approaches fail to manage situations where the learned emoji embedding semantics or emotions contradict information from the associated contexts or when emojis express various meanings and sentiments. They treat emoji and plain text emotions separately, failing to properly investigate the influence of emojis on text sentiment polarity.Emojis have a significant impact on the polarity of plain text sentiments. This study aims to see how emojis affect text sentiment polarity to predict the sentiment polarity of a microblog post as a whole [13]. For opinion mining, an Emoji-Based Opinion Mining (EBOM) model is presented to model the influence of emojis on text sentiment polarity. Emojis and words in microblog posts are combined to create emoji representations that include contextual information. According to the results, this method dramatically improves the performance of sentiment polarity prediction.Emojis are not only being used to improve sentiment analysis but also for other composite tasks like sarcasm detection. The idea that one single emoji could have such an impact on the sentiment of a statement was the motivation forthis paper. This paper aims to find how emojis are used insocial media, how emojis affect the sentiment detection of asentence, and how to improve the sentiment analysis accuracy by using emojis.

II. LITERATURE SURVEY

With the usage of various tokenization methods, machine learning algorithms, and word embedding modelsByungkyuYoo and Julia Rayz [7], in an attempt to gain a better understanding of emojis and increase the sentiment analysis score. There were several findings that we saw from this research paper. Emojis generally replace emoticons; they are not used together. Emojis show stronger sentiments compared to words. Emojis that are next to each other show a stronger sentiment ratio and, using this knowledge, sentiment analysis accuracy can be slightly increased. Emojis are generally close to other emojis or modern phrases like “lol” or “haha,” but these words should not be used to replace emojis while doing sentiment analysis. Replacing emojis with one or a few words leads to the highest accuracy as sentiments of emoji help train other English words.

Jagadishwari V, Indulekha A, Kiran Raghu and, Harshini P [8] highlights the Twitter data set which are used to train and test the modelsand also machine learning models that were implemented used were Bernoulli Bayes, multinomial Bayes, regression, and SVM. The models were trained only with the text message removing the emoticons and emojis and then tested and their performance was evaluated. The experimental results proved that the Bayes family of classifiers perform well in sentiment analysis giving a very high accuracy and also it was evident that emoticons have a negligible effect on the accuracy of sentiment analysis.

Kiran Raghu andHarshini P[11], aims at detecting the sentiments expressed in social media posts. The machine learning models namely BernoulliBayes, multinomial Bayes, regression, and SVM were implemented. All these models are trained and tested with Twitter data sets. The effect of emoticons present in the tweet is also analysed. The models are first trained only with the text and then they are trained with text and emoticons in the tweet.

Emojis and emoticons can help users express their sentiments on certain topics. According to Amalia Anjani Arifiyanti& Eka Dyar Wahyuni [10], although some emojis and emoticons cannot be grouped into certain sentiment groups, the algorithm can still create a classification model with a reasonably good performance. Svm has the best and most stable performance model of the three learning algorithms used in this study compared to bnb and mnb. This was followed by bnb and mnb in the last position. Converting emojis and emoticons into sentiment categories reduces the large variety of emoji and emoticon features, which impacts more stable model performance

Muhammed Sinan Başarslan&FatihKayaalp[12],chooses to use two datasets.The first one is the user reviews about movies from the IMDb, which has been labeled by Kotzias, and the second one is the Twitter tweets, including the tweets of users about health topics in English in 2019, collected using the TwitterAPI. The Python programming language was used in the study both for implementing the classification models using the Naïve Bayes (NB), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) algorithms and for categorizing the sentiments as positive, negative, and neutral. The feature extraction from the dataset was performed using term frequency-inverse document frequency (tf-idf) and word2vec (w2v) modeling techniques.

The key idea of Yequan Wang, Minlie Huang, Xiaoyan Zhu, Li Zhao [1], are to learn aspectembeddings and let aspects participate in computingattention weights.Our proposed models can concentrate on different parts of a sentence when different aspects are given so that they aremore competitive for aspect-level classification. Experimentsshow that

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104239 |

our proposed models, AE-LSTM andATAE-LSTM, obtain superior performance over thebaseline models.Though the proposals have shown potentials foraspect-level sentiment analysis, different aspects areinput separately.

Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., & González-Castaño's [9] paper describes an unsupervised strategy based on semantic dependencies, called pad, enhanced with descriptions by emoji creators from Emojipedia, with the objective of creating a fully unsupervised emoji sentiment lexicon. This lexicon is then improved in different variants that take advantage of the sentiment distribution of informal texts including emojis. In all cases, suspension guarantees that neither labeling nor training is necessary. Our approach analyzes dependencies between lemmatized tagged words using a sentiment propagation algorithm that considers key linguistic phenomena, namely intensification, modification, negation, and adversative and concessive relations.

It is desirable to integrate the connections between target word and context words when building a learning system. In this paper, we develop two target dependent long short-term memory (LSTM) models, where target information is automatically taken into account.Duyu Tang, Bing Qin, Ting Liu[2], have evaluated methods on a benchmark dataset from Twitter. Empirical results show that modeling sentence representation with standard LSTM does not perform well. Incorporating target information into LSTM can significantly boost the classification accuracy. The target-dependent LSTM models achieve state-of-the-art performances without using syntactic parser or external sentiment lexicons.

Deep learning techniques have achieved success in aspect-based sentiment analysis in recent years.Shiliang Zheng, Rui Xia [3], proposes an approach, called left-center-right separated neural network with rotatory attention (LCR-Rot), to better address the two problems.This approach has two characteristics: 1) it has three separated LSTMs, i.e., left, center and right LSTMs, corresponding to three parts of a review (left context, target phrase and right context); 2) it has a rotatory attention mechanism which models the relation between target and left/right contexts. The target2context attention is used to capture the most indicative sentiment words in left/right contexts. Subsequently, the context2target attention is used to capture the most important word in the target. This leads to a two-side representation of the target: left-aware target and right-aware target.

II. PROPOSED SYSTEM

This model aims to develop a system for the evaluation of the sentiment analysis score based on social media posts or tweets. Different Machine Learning techniques are used (SVN, RF, NBs) to evaluate the sentiment from the posts,comments, or tweets in social media.First data will be preprocessed by procedures of preprocessing which provides the output of text with emoji set.Using text and emoji lexicons, feature extraction (uses tf-idf,n-gram, and CBOW) selection will be done.Later by applying ML algorithms results will be provided which will be classified into positive, negative, and neutral comments or posts.

III. METHODOLOGY

Dataset collection
The sample data is collected from the social media comment section which consists of all types of text and emojis. Itis done through different appson social media platforms and a standard emoji set dataset is also considered.

Data pre-processing
The preprocessing stage is then used to clean up the data before converting the retrieved information into a structured format so that the patterns (both apparent and hidden) can be analyzed.The natural language toolkit was used to prepare and clean the corpus.
Spelling correction: The correction of words in texts is known as spelling correction. This functionality is provided by the Python package spellchecker, which finds words that may have been misspelled and suggests plausible words.

Tokenization: The text is divided into smaller pieces in this stage and either sentence or word tokenization is utilized.
Stopword removal: Stopwords are words that appear repeatedly in a text yet provide no meaningful information. Stopwords include words like they, there, this, where, and others. With around 180 stopwords eliminated, the nltk library is a commonly used library for eliminating stopwords.

Remove punctuations & numbers: The first decision to make when deciding how to preprocess a corpus is which character and markup classes are to be considered as real text. The most comprehensive way is to preprocess all text, including

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104239 |

numbers, any markup (HTML) or tags, punctuation, special characters ($, %, &, etc), and additional white-space characters. These non-letter characters and markup may be beneficial in some analyses (such as hashtags in Twitter data), but they are considered uninformative in many others. As a result, removing them is a normal procedure. Punctuation is the most commonly removed of these character types. The first preprocessing decision is whether or not to include or delete punctuation.

Stemming / Lemmatization: stemming is the process of reducing a word to its essential stem. For example, run, running, runs, and are all formed from the same word. In simple terms, stemming is the process of eliminating the prefix or suffix from words like ing, s, and es. The words are stemmed with the help of the nltkpackage.Lemmatization is related to stemming in that it is used to break down words into their source words, but it works differently. Actually, lemmatizationis a process that compares words to a lexicon in order to reduce them to their lemma. It stems from the word while keeping its meaning.

Feature extraction and selection

After the preprocessing steps are completed, a dictionary of words is created to map each word to its unique id. This dictionary is transformed to vector representation with the help of the cbow model. The cbow model representation is mapped to another vector space using the tf-idf model. With the help of tf-idf, based on the occurrence of words, score will be given to a word in a collection of documents. This value of score tells how important the word is in a given document. Tf is the ratio of word count in documents to the total words count in all documents. Idf is the logarithmic ratio of total document count in the corpus to the number of documents in which the word appears. If a word occurs in less number of documents then it gives high idf value and it gives less idf value if a word occurs in more number of documents.

Emoji lexicon: emoji sentiment is calculated using the sentiment of tweets. Sentiment labels have three possible values: negative, neutral, and positive. A sentiment label, c, is a discrete, three-valued variable with the values $c \in \{-1, 0, +1\}$. An emotion is attributed to an emoji based on all of the tweets in which it appears. First, a discrete probability distribution (p-,p0,p+) is created for each emoji.

Performance evaluation

Training and testing:for better model validation, the dataset in the present study was split into training and testing with the help of the scikit library. It contains a class called imputer which will help us take care of the missing data. We will train our machine learning models on our training set, I.E., our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% of the dataset to the training set and therefore the remaining 20% to the test set.

Performance evaluation:the performance measurement of the model was evaluated with the help of various metrics like accuracy, sensitivity, f1-score and precision.The best algorithm based on the performance parameters was selected to predict the sentiment analysis from social media comments. Based on the details collected by the dataset which contains text and emojis could be predicted and the result would be displayed along with the suggestions for further improvement.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104239 |

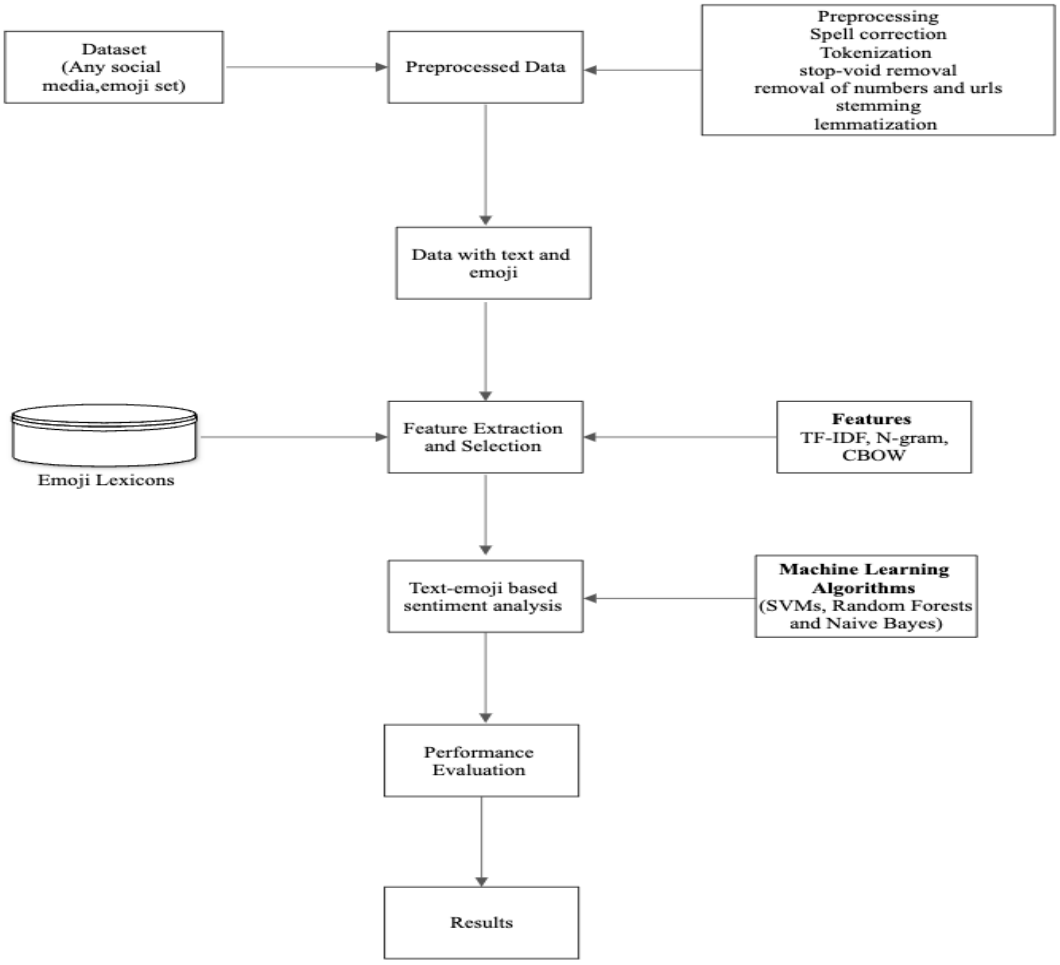

Fig 4.1: Methodology diagram

IV. CONCLUSION

This paper is employed to search to gain knowledge about various tokenization methods, and machine learning algorithms in an attempt to gain a better understanding of text and emojis and increase the sentiment analysis score. There were several findings that we saw that emojis generally replace emoticons; they are not used together. Emojis show stronger sentiments compared to words. Emojis next to each other show a stronger sentiment ratio and, using this knowledge, sentiment analysis accuracy can be slightly increased. Emojis are generally close to other emojis or modern phrases like “lol” or “haha,” but these words should not be used to replace emojis while doing sentiment analysis. Replacing emojis with one or a few words leads to the highest accuracy as sentiments of emoji help train other English words.

REFERENCES

1. Yequan Wang, Minlie Huang, Xiaoyan Zhu, Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In: EMNLP. pp. 606–615.
2. Duyu Tang, Bing Qin, Ting Liu. 2016a. Effective LSTMs for Target-Dependent Sentiment Classification. In: Proceedings of the 26th international conference on computational linguistics (COLING 2016). pp. 3298–3307.
3. Shiliang Zheng, Rui Xia. 2018. Left-Center-Right Separated Neural Network for Aspect-based Sentiment Analysis with Rotatory Attention. arxiv preprint arXiv:1802.00892.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104239 |

4. Huang, B., Yanglan, O.u., Carley, K.M., 2018. Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks. In: In: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pp. 197–206.
5. Maria Pontiki, DimitriosGalanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis.
6. Nasukawa, T., Yi, J.2003. Sentiment analysis: Capturing favorability using natural language processing. In: Proceedings of the 2nd international conference on Knowledge capture. pp. 70–77.
7. ByungkyuYoo, Julia Rayz."Understanding Emojis for Sentiment Analysis". Purdue University West Lafayette, IN ,2021.
8. Jagadishwari V, Indulekha A, Kiran Raghu, Harshini P. "Sentiment analysis of Social Media Text-Emoticon Post with Machine learning Models Contribution Title".ICAPSM 2021.
9. Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., & González-Castaño, F. (2018)."Creating emoji lexica from unsupervised sentiment analysis of their descriptions"
10. Amalia Anjani Arifiyanti& Eka Dyar Wahyuni. "Emoji and Emoticon in Tweet Sentiment Classification" October 14-16, 2020.
11. Kiran Raghu, Harshini P -CMR Institute of Technology "Sentiment analysis of social media text-emoticon post with machine learning models contribution title" 2021.
12. Muhammed Sinan Başarslan, FatihKayaalp "Sentiment analysis with machine learning methods on social media" 2020.
13. Nirmal Varghese Babul·E. Grace Mary Kanaga1 " Sentiment analysis in social media data."Dec 2020.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details