International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

A CONCEPT OF NATURAL PROCESSING LANGUAGE: SENTIMENT ANALYSIS USING ARTIFICIAL INTELLIGENCE

S. Sindhuja[1], Cindum Sri Pranav[2], R. Mahesh[2], M. Dharani[2], S. Kranthi Kumar Reddyi[2], Akkuloori Akash[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science Engineering
[1,2]Malla Reddy Engineering College and Management Science, Kistapur, Medchal-50140l, Hyderabad, Telangana, India

ABSTRACT

Lexicon algorithm is used to determine the sentiment expressed by a textual content. This sentiment might be negative, neutral, or positive. It is possible to be sarcastic using only positive or neutral sentiment textual contents. Hence, lexicon algorithm can be useful but insufficient for sarcasm detection. It is necessary to extend the lexicon algorithm to come out with systems that would be proven efficient for sarcasm detection on neutral and positive sentiment textual contents. In this paper, two sarcasm analysis systems both obtained from the extension of the lexicon algorithm have been proposed for that sake. The first system consists of the combination of a lexicon algorithm and a pure sarcasm analysis algorithm. The second system consists of the combination of a lexicon algorithm and a sentiment prediction algorithm. Finally, naive bayes is used to predict the sarcasm detection using pretrained features.

Keywords: Lexicon algorithm, sarcasm detection, sentiment prediction algorithm, a pure sarcasm analysis algorithm.

1. INTRODUCTION

The extension of the Lexicon Algorithm for Emoji-Based Sarcasm Detection from Twitter Data is a cutting-edge approach to tackle the nuanced and ever-evolving phenomenon of sarcasm in online social media, with a specific focus on Twitter. Sarcasm in text is notoriously challenging to detect due to its reliance on context [1], tone, and subtlety, making it a perfect testbed for natural language processing and sentiment analysis techniques. The Lexicon Algorithm, a well-established method for sentiment analysis, forms the basis for this extension. In this advanced approach, researchers have incorporated a comprehensive emoji lexicon to enhance the algorithm's performance [2]. Emojis play a vital role in conveying sentiment and emotions in online communication, often being key indicators of sarcasm. By integrating an extensive emoji lexicon into the Lexicon Algorithm, this extension aims to improve the accuracy of sarcasm detection in Twitter data. The algorithm begins by preprocessing the Twitter data, which typically involves tokenization, stemming, and removing stopwords. It then identifies sentences or tweets containing potential sarcasm cues, such as negations, contrastive conjunctions, and sentiment-laden words [3]. The innovation lies in the algorithm's ability to consider emojis alongside these textual cues. Emojis, when used ironically or sarcastically, can completely reverse the sentiment of a sentence, and the algorithm is designed to recognize such patterns.

Furthermore, the extension considers the surrounding context of tweets, considering user-specific patterns and common sarcasm structures in the Twitterverse. Machine learning techniques, such as supervised learning or neural networks [4], may be employed to fine-tune the algorithm's performance and adapt it to the evolving nature of sarcasm on Twitter. The implications of this extended Lexicon Algorithm are far-reaching, with potential applications in sentiment analysis, brand monitoring, and social media trend analysis. Accurately detecting sarcasm in Twitter data can help improve our understanding of public sentiment, social trends, and user engagement, benefiting businesses,

researchers, and social media platforms themselves. However, it's important to note that this field is dynamic, and ongoing research and adaptation are necessary to keep pace with the ever-changing landscape of online communication.

2. LITERATURE SURVEY

Despite sarcasm having long been studied in the social sciences, automated detection of sarcasm in text is a new area of study. Recently, the research community in the domain of natural language processing (NLP) and machine learning (ML) has been interested in automated sentiment classification [7]. An NLP-based technique uses linguistic corpora and language features to understand qualitative data. ML systems utilize unsupervised and supervised classification approaches based on tagged or unlabeled information to interpret sarcastic remarks. An extensive dataset for sarcasm text detection was created by Khodak et al. [8]. Before comparing their findings with methods such as sentence vectorization, bag-of-words, and bag-of-bigrams, the authors performed hand annotation. They discovered that hand sarcasm detection outperformed other methods. Eke et al. [9] evaluated a range of earlier studies on sarcasm recognition. According to this review research, N-gram and part-of-speech tag (POS) methods were the most frequently used feature extraction algorithms. Binary interpretation and word frequencies were used for feature representation, nevertheless. The review also noted that the chi-squared test and information gain (IG) method were frequently employed for feature selection. In addition, the maximum entropy, naive Bayes, random forests (RF), and support vector machine (SVM) classification techniques were used. A review of several "Customized Machine Learning Algorithms" (CMLA) and "Adapted Machine Learning Algorithms" (AMLA) utilized in sarcasm detection research was also published by Sarsam et al. in [10]. Their findings concurred with those of Eke et al. They found that CNN-SVM can perform better when both lexical and personal characteristics are used. The SVM performs better using lexical, frequency, pragmatic, and part-of-speech labeling. Prior to this, models based on machine learning were created; these models primarily acquire language features and train these qualities over classifiers learned by machine learning. Machine learning has been used for content-based features by Keerthi Kumar and Harish [11]. Before submitting the data to the clustering method for various filters, the authors used feature selection approaches such as "Information Gain" (IG), "Mutual Information" (MI), and chi-square. An SVM was used to categorize data at the very end. Pawar and Bhingarkar [12] employed an ML classification model for sarcasm detection on a related subject. They collected data on recurring themes, punctuation, interjections, and emotions. Using the random forest and SVM techniques, these feature sets for classification were learned.

3. PROPOSED METHODOLOGY

Lexicon algorithm is used to determine the sentiment expressed by a textual content. This sentiment might be negative, neutral, or positive. It is possible to be sarcastic using only positive or neutral sentiment textual contents. Hence, lexicon algorithm can be useful but yet insufficient for sarcasm detection. It is necessary to extend the lexicon algorithm to come out with systems that would be proven efficient for sarcasm detection on neutral and positive sentiment textual contents. In this paper, two sarcasm analysis systems both obtained from the extension of the lexicon algorithm have been proposed for that sake. The first system consists of the combination of a lexicon algorithm and a pure sarcasm analysis algorithm. The second system consists of the combination of a lexicon algorithm and a sentiment prediction algorithm. Figure 4.1 shows the proposed system model. The detailed operation illustrated as follows:

Step 1. Twitter Dataset: The research work begins with the collection of a dataset containing tweets from the social media platform Twitter. This dataset serves as the primary source of textual content for

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

analysis. It comprises a mix of tweets with varying sentiments, including positive, neutral, and potentially sarcastic ones.

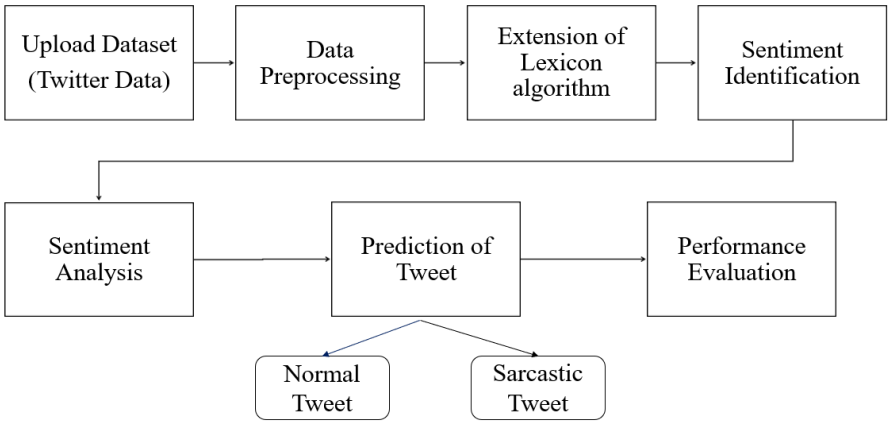

Figure 1. Proposed system model.

Step 2. Data Preprocessing: Before any analysis can take place, the Twitter dataset undergoes a series of preprocessing steps to clean and prepare the text for further processing. Data preprocessing involves tasks such as:

- Tokenization: Breaking down the tweets into individual words or tokens.
- Stemming: Reducing words to their base or root form.
- Stopword Removal: Eliminating common, non-informative words (e.g., "and," "the," "in").
- Special Character Removal: Stripping away punctuation and special characters.

Step 3. Extension of Lexicon Algorithm: In this step, the researchers extend the traditional Lexicon Algorithm, which is primarily used for sentiment analysis. The extension aims to make the Lexicon Algorithm more suitable for detecting sarcasm, even in tweets with positive or neutral sentiments.

Step 4. Sentiment Identification: The extended Lexicon Algorithm is then employed to identify the sentiment expressed in each tweet. This involves determining whether a tweet conveys a positive, negative, or neutral sentiment based on the polarity of the individual terms within the tweet. This sentiment identification provides a baseline understanding of the sentiment in the tweets, which includes both non-sarcastic and potentially sarcastic content.

Step 5. Sentiment Analysis: Building upon the sentiment identification, the researchers perform a more in-depth sentiment analysis. This analysis goes beyond mere polarity determination and seeks to recognize nuanced expressions of sentiment, including sarcasm. The Lexicon Algorithm's extension allows it to consider the possibility of sarcasm within tweets, especially those with positive or neutral sentiments.

Step 6. Prediction of Tweet (Normal and Sarcastic): In this final step, the research work proposes two systems for sarcasm detection within tweets that primarily express positive or neutral sentiments:

> **a. Lexicon Algorithm + Pure Sarcasm Analysis Algorithm:** The first system combines the extended Lexicon Algorithm with a dedicated pure sarcasm analysis algorithm. This combination aims to enhance sarcasm detection in tweets that may not contain obvious negative sentiment cues.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

b. Lexicon Algorithm + Sentiment Prediction Algorithm: The second system combines the extended Lexicon Algorithm with a sentiment prediction algorithm. This combination seeks to predict whether a tweet is normal (non-sarcastic) or sarcastic, even when the sentiment expressed is positive or neutral.

These two systems leverage the extended Lexicon Algorithm's capabilities to identify subtle sarcasm cues within tweets that might otherwise be missed by traditional sentiment analysis techniques.

4. RESULTS AND DISCUSSION

This section gives the detailed analysis of simulation results implemented using "python environment". Further, the performance of proposed method is compared with existing methods using same dataset.

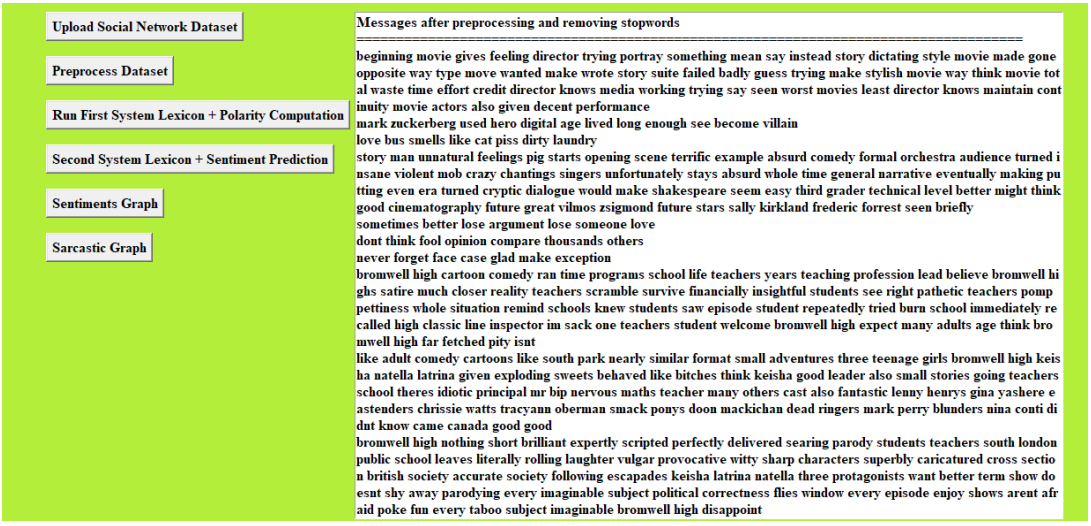

Figure 2. Run First System Lexicon + Polarity Computation

In Figure 2, we can see all messages after removing special symbols and stop words. Now click on 'Run First System Lexicon + Polarity Computation' button to calculate polarity of messages.

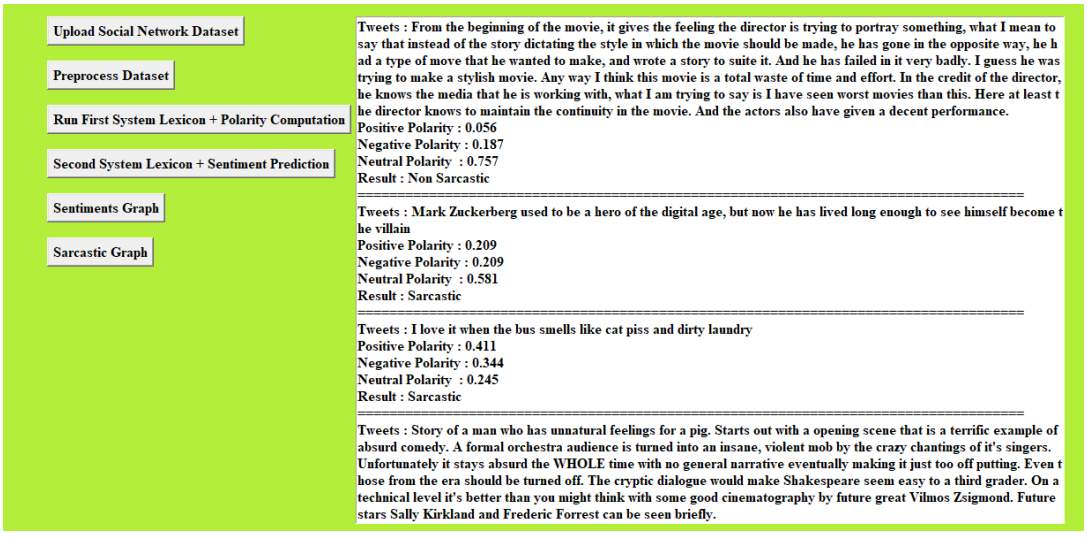

Figure 3. First system predicted outcome.

In Figure 3 for each message, we can see tweet data and positive, negative and neutral polarity score and the message in tweets is sarcastic or non-sarcastic. The tweets will be classified to positive, negative or neutral based on its high score for example in first tweet neutral got high score as 0.757 so tweet will

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

consider as neutral. If that neutral tweet contains some negative words then consider as sarcastic. You can scroll down above screen text area to see all messages details.

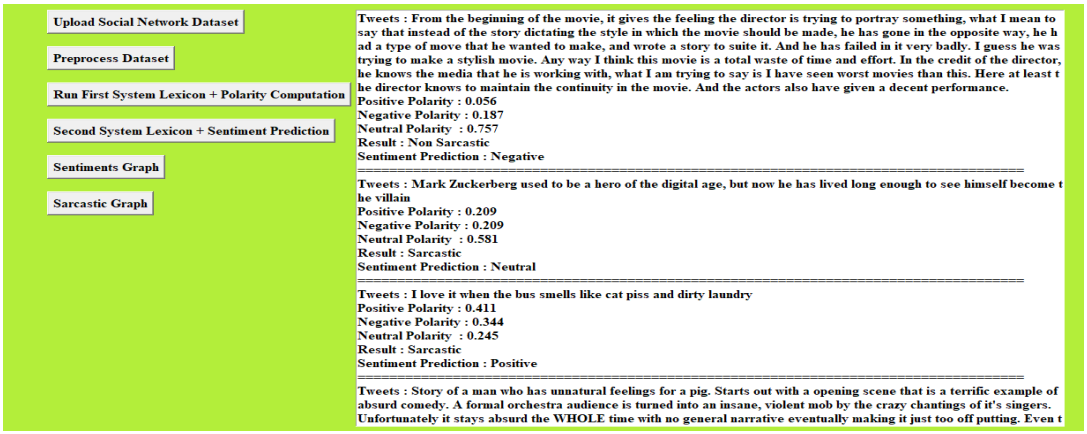

Figure 4. Proposed predicted outcome.

In Figure 4, we can see same results with extra details such as whether tweet/message is positive or negative or neutral. You can scroll down above text area to see all messages. Above Figure shows the results from the sentiment prediction algorithm. For each comment, the result of the sentiment prediction algorithm has been compared with result of the lexicon algorithm. In case these results are similar, the content is considered as non-sarcastic else it is considered as sarcastic. The environment details used in this experiment are assumptions that have been made since the required data were not available and the proposed system was an innovative one at the time of experiment.

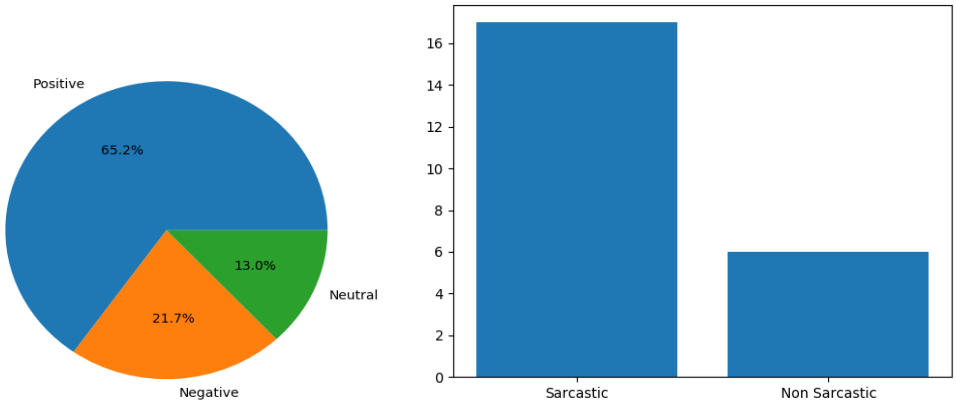

Figure 5. Sentiment graph (left). Count of sarcastic or non-sarcastic tweets (right).

In Figure 5 using pie chart we can see percentage of positive, negative, or neutral tweets, and x-axis represents type of tweets and y-axis represents count of sarcastic or non-sarcastic tweets.

5. CONCLUSION

The aim of this study was to propose ways to extend the lexicon algorithm in order to build systems that would be more efficient for sarcasm detection. This aim had been successfully met as two systems have been developed to address this situation. However, in the first system, it had been noticed that the training set of the sarcasm analysis algorithm must be relevant to the actual data that need to be analyzed to obtain meaningful results and to improve the accuracy of the system. The second system constitutes a vast area of study. Some work needs to be done in order to develop a system that would allow the collection of environment details under which the textual contents would be made on social media

platforms. A consolidated way of computing the sentiment polarity of the environments based on their details should also be developed.

REFERENCES

[1] Edwards, V.V. Sarcasm: What It Is and Why It Hurts Us. 2014. Available online: https://www.scienceofpeople.com/sarcasm-why-it-hurts-us/ (accessed on 5 October 2021).
[2] Rothermich, K.; Ogunlana, A.; Jaworska, N. Change in humor and sarcasm use based on anxiety and depression symptom severity during the COVID-19 pandemic. J. Psychiatr. Res. 2021, 140, 95–100.
[3] Ezaiza, H.; Humayoun, S.R.; Al Tarawneh, R.; Ebert, A. Person-vis: Visualizing personal social networks (ego networks). In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 1222–1228.
[4] Akula, R.; Garibay, I. Viztract: Visualization of complex social networks for easy user perception. Big Data Cogn. Comput. 2019, 3, 17.
[5] Singh, B.; Sharma, D.K. Predicting image credibility in fake news over social media using multi-modal approach. Neural Comput. Appl. 2021, 34, 21503–21517.
[6] Singh, B.; Sharma, D.K. SiteForge: Detecting and localizing forged images on microblogging platforms using deep convolutional neural network. Comput. Ind. Eng. 2021, 162, 107733.
[7] Wallace, B.C. Computational irony: A survey and new perspectives. Artif. Intell. Rev. 2015, 43, 467–483.
[8] Khodak, M.; Saunshi, N.; Vodrahalli, K. A large self-annotated corpus for sarcasm. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 7–12 May 2018.
[9] Eke, C.I.; Norman, A.A.; Shuib, L.; Nweke, H.F. Sarcasm identification in textual data: Systematic review, research challenges and open directions. Artif. Intell. Rev. 2020, 53, 4215–4258.
[10] Sarsam, S.M.; Al-Samarraie, H.; Alzahrani, A.I.; Wright, B. Sarcasm detection using machine learning algorithms in Twitter: A systematic review. Int. J. Mark. Res. 2020, 62, 578–598.
[11] Keerthi Kumar, H.M.; Harish, B.S. Sarcasm classification: A novel approach by using Content Based Feature Selection Method. Procedia Comput. Sci. 2018, 143, 378–386.
[12] Pawar, N.; Bhingarkar, S. Machine Learning based Sarcasm Detection on Twitter Data. In Proceedings of the 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020.