IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
Research Paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, 2021

EXTENSION OF LEXICON ALGORITHM FOR EMOJI-BASED SARCASM DETECTION FROM TWITTER DATA

[1]MD Reshma, [2]Dr A Nagamalleshwar Rao, [3]B Sandhya

[1,3]Assistance Professor, Department of CSE, Sree Dattha Institute of Engineering and Science

[2]Associate Professor, Department of CSE, Sree Dattha Institute of Engineering and Science

ABSTRACT

Sarcasm detection can help refine traditional sentiment analysis tools by identifying instances where positive or neutral language conveys negative sentiment. This can lead to more accurate insights into public opinion and brand perception on social media. Practical applications of our project include the development of customer service automation, market research, sentiment analysis, and content moderation. The existing system of detecting sarcasm tweets is done by Humans inspection and it has some limit to do the sarcasm tweets and do not reach the milestone to detect the accuracy. Traditional methods lack the ability to discern sarcastic intent, resulting in misclassifications and misinterpretations. To predict the accurate detection and interpretation of sarcasm. It often relies on fine linguistic cues and context, making it difficult for conventional lexical analysis techniques to identify. As a result, there is a need for a more nuanced approach to lexical analysis that can effectively detect and interpret sarcastic expressions. the proposed method, we utilize natural language processing (NLP) techniques for text preprocessing we employ TF-IDF for feature extraction. We utilize a lexical algorithm to detect sarcasm in tweets. Finally, we evaluate the accuracy of sarcasm detection to assess the system's performance. our project aims to introduce the lexical algorithm that addresses the limitations of existing methods and offers improved accuracy and nuanced interpretation of text.

Keywords: Sarcasm Detection, Sentiment Analysis, Lexical Analysis, NLP, TF-IDF.

1. INTRODUCTION

1.1 Overview

The project introduces an extension to the lexicon algorithm for sarcasm detection, specifically tailored to analyse Twitter data. With the widespread use of emojis on social media platforms like Twitter, the study aims to leverage these visual cues alongside textual content to improve sarcasm detection accuracy. Acknowledging the pervasive nature of sarcasm and the intricate nuances involved in its portrayal via emojis on social media platforms such as Twitter, the proposed extension seeks to seamlessly integrate emoji analysis into the established lexicon-based approach. Through meticulous creation of an emoji lexicon and the subsequent assignment of sentiment scores, grounded in the nuanced contextual usage within sarcastic tweets, the augmented algorithm aims to significantly enhance the precision and efficacy of sarcasm detection. The implementation process will necessitate a comprehensive suite of procedures, encompassing meticulous data preprocessing, lexicon construction, sentiment scoring, and sarcasm identification, all facilitated by cutting-edge computational tools and sophisticated natural language processing methodologies. The efficacy and robustness of the proposed approach will be rigorously evaluated using a diverse range of metrics and meticulously curated datasets, with the overarching objective of not only refining sarcasm detection capabilities but also fostering deeper insights into the complexities of online communication dynamics. Ultimately, this endeavour seeks to propel advancements in natural language understanding and computational linguistics research, thereby offering invaluable contributions to the broader fields of sentiment analysis and social media analytics.

1.2 Problem Statement

The problem lies in the inadequacy of existing lexicon-based algorithms to effectively detect sarcasm in Twitter data, especially when emojis are involved. This limitation undermines sentiment analysis efforts and hampers the reliability of automated systems in various domains. By failing to capture the nuanced interplay between emojis and text, current methods fall short in accurately interpreting sarcasm, thereby impacting decision-making processes reliant on sentiment analysis. Hence, there is a pressing need to develop an extension to lexicon algorithms specifically tailored for emoji-based sarcasm detection in Twitter data. This extension would bridge the gap between text and emojis, enhancing the accuracy of sarcasm detection and empowering stakeholders with deeper insights into online communication dynamics.

Despite the growing prevalence of sarcasm in social media discourse, particularly on platforms like Twitter, current approaches to sarcasm detection often struggle to accurately interpret nuances, especially when sarcasm is conveyed through emojis. The existing lexicon-based algorithms, while effective in certain contexts, face significant limitations in capturing the subtleties of sarcasm in text containing emojis. The absence of a robust methodology for integrating emoji analysis into lexicon-based approaches hampers the accuracy and reliability of sarcasm detection in Twitter data. Consequently, there is a pressing need to develop an extension to lexicon algorithms specifically tailored to incorporate emoji analysis, enabling more precise detection of sarcasm in Twitter data. This extension should address challenges such as the diverse interpretations of emojis, contextual nuances, and the dynamic nature of language in social media, ultimately enhancing sentiment analysis and natural language understanding in online communication platforms. Therefore, the problem statement revolves around designing and implementing an effective extension to lexicon algorithms for emoji-based sarcasm detection in Twitter data, with the overarching goal of improving the accuracy and reliability of sarcasm detection in social media discourse.

1.3 Research Motivation

The motivation for extending lexicon algorithms to incorporate emoji-based sarcasm detection in Twitter data stems from the increasing importance of accurately understanding and analysing online communication. Sarcasm is a prevalent form of expression on social media platforms like Twitter, and its detection is crucial for various applications, including sentiment analysis, brand monitoring, and customer feedback analysis. However, traditional lexicon-based approaches often struggle to accurately detect sarcasm, particularly when it is

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, 2021

conveyed through emojis. Emojis play a significant role in conveying emotions and nuances in text, and their usage has become integral to online communication. Hence, there is a clear need to develop more sophisticated algorithms capable of interpreting the complex interplay between text and emojis to accurately detect sarcasm in Twitter data. By addressing this gap, the proposed research aims to enhance the effectiveness of sentiment analysis algorithms, leading to more accurate insights into public opinion, sentiment trends, and user behaviour on social media platforms. Additionally, advancements in emoji-based sarcasm detection have broader implications for natural language processing research, contributing to the development of more context-aware and socially intelligent systems capable of understanding the nuances of human communication in online environments.

2. LITERATURE SURVEY

Kumar, et al. [1] model was validated on a Hindi tweets dataset, Sarc-H, manually annotated with sarcastic and non-sarcastic labels. The preliminary results clearly depicted the importance of using emojis for sarcasm detection, with Kumar model attaining an accuracy of 97.35% with an F-score of 0.9708. The research validated that automated feature engineering facilitates efficient and repeatable predictive model for detecting sarcasm in indigenous, low-resource languages.

Chauhan, et al. [2] proposed an emoji-aware-multimodal multitask deep learning framework for sarcasm detection (i.e., primary task) and sentiment and emotion detection (i.e., secondary task) in a multimodal conversational scenario. Experimental results on the SEEmoji MUStARD show the efficacy of Chauhan proposed emoji-aware-multimodal approach for sarcasm detection over the existing models.

Subramanian, et al. [3] studied the novel problem of exploiting emojis for sarcasm detection on social media. They proposed a new framework ESD, which simultaneously captures various signals from text and emojis for sarcasm detection. Experimental results on real-world datasets demonstrate the effectiveness of the proposed framework.

Prasad, et al. [4] chosen the best classifier and paired with various pre-processing and filtering techniques using emoji and slang dictionary mapping to provide the best possible accuracy. The emoji and slang dictionary being the novel idea introduced in this paper. The obtained results can be used as input to other research and applications.

Lemmens, et al. [5] proposed models consisted of an LSTM with hashtag and emoji representations; a CNN-LSTM with casing, stop word, punctuation, and sentiment representations; an MLP based on Infer sent embeddings; and an SVM trained on stylometric and emotion-based features. All proposed models used the two conversational turns preceding the response as context, except for the SVM, which only used features extracted from the response. The ensemble itself consisted of an adaboost classifier with the decision tree algorithm as a base estimator and yielded F1-scores of 67% and 74% on the Reddit and Twitter test data, respectively.

Gregory, et al. [6] proposed a model that was an ensemble of transformer models including BERT, RoBERTa, XLNet, RoBERTa-large, and ALBERT. This research was performed in conjunction with the sarcasm detection shared task section in the Second Workshop on Figurative Language Processing, co-located with ACL 2020.

Agrawal, et al. [7] proposed a novel model for automated sarcasm detection in text, called Affective Word Embeddings for Sarcasm (AWES), which incorporates affective information into word representations. Extensive evaluation on sarcasm detection on six datasets across three domains of text (tweets, reviews and forum posts) demonstrates the effectiveness of the proposed model. The experimental results indicate that while sentiment affective representations yield best results on datasets comprising of short length text such as tweets, richer representations derived from fine-grained emotions are more suitable for detecting sarcasm from longer length documents such as product reviews and discussion forum posts.

Li, et al. [8] proposed method employs an intermodal emotional inconsistency detection mechanism, a contextual scenario inconsistency detection mechanism, and a cross-modal and segmented attention_mechanism. These innovations enable a richer and more nuanced feature representation, capturing the essence of sarcasm more effectively. Experimental results on the dataset MUStARD Extended confirm the superiority of our approach, establishing it as the new state-of-the-art in sarcasm detection compared to existing models.

Qiang, et al. [9] introduced a combination incongruity fusion layer and cross-modal contrastive loss to fuse fact incongruity and sentiment incongruity for further enhancing the incongruity representations. Extensive experiments and further analyses on publicly available datasets demonstrated the superiority of thier proposed model.

Prasanth, et al. [10] proposed the modification of the logistic loss function to detect tweet-level stress by availing the information of sarcasm that existed in the tweet-content. The experimental results showed that the proposed STSD model, when applied along with kernel PCA, recorded a significant improvement in accuracy by a minimum of 5.25%maximum of 9.19% over baseline models. Also, there was an increment in F1-score by at least 0.085 points and a maximum of 0.164 points when compared to the baseline models.

Rustagi, et al. [11] performance of the proposed system was evaluated using annotator-agreement methods with metrics such as F1 score, Precision, Recall, and Accuracy. The performance showed that integrating more features enhanced the accuracy by a considerable margin compared to previously defined methodologies.

Cui, et, al. [12] proposed system suggested that stereotypical information about the sender could drive the interpretation of potentially sarcastic statements, and the contextual information modulated the effect of sender occupation on sarcasm interpretation.

Sukhavasi, et, al. [13] proposed model, the attention-based transformer model, was developed, which showed effective performance in analysing both the emoji and text data. the method was compared to many current methodologies regarding various performances. The English Twitter dataset attained 99.1% accuracy, 99.2% precision, 99.1% recall, 99.1% F-measure, an execution time of 56.66 s, and an average threshold of 12364.365 s. The accuracy, recall, precision, and F-measure of the Hindi Twitter dataset were 98.1%, 98.41%, 98.2%, and 69.6%, respectively.

Garcia ,et al. [14] proposed method used an online rating task to investigate the influence of the winking face emoji on both the interpretation and perception of message intent for sarcastic or literal criticism or praise. Results revealed that older adults, in comparison to their younger counterparts, demonstrated deficient ability in interpreting and perceiving sarcastic intent.

Farnham, et al. [15] proposed sarcasm detection pipeline successfully facilitated the task using a GRU neural network with sentiment-aware attention, achieving an accuracy of 73% and showing promising indications regarding model robustness as part of a framework that was easily scalable for the inclusion of any future emojis released. Both enhanced sentiment information to supplement context in addition to consideration of the emoji were found to improve outcomes for the task.

3. PROPOSED METHOD

3.1 Overview

Twitter have become the largest platform where people express their feelings, opinions, view and real time events such as live tweets etc. With respect to previous years, the data of twitter has increased much and thus forming big data. Twitter has 315 million monthly active users, eighty two percent of active users on mobile and millions of tweets are being circulated through twitter every day. Various organizations as well as companies are interested in twitter data for finding the views of various people towards their products or events. Twitter is also used to find out various views of people towards political events, movies etc.

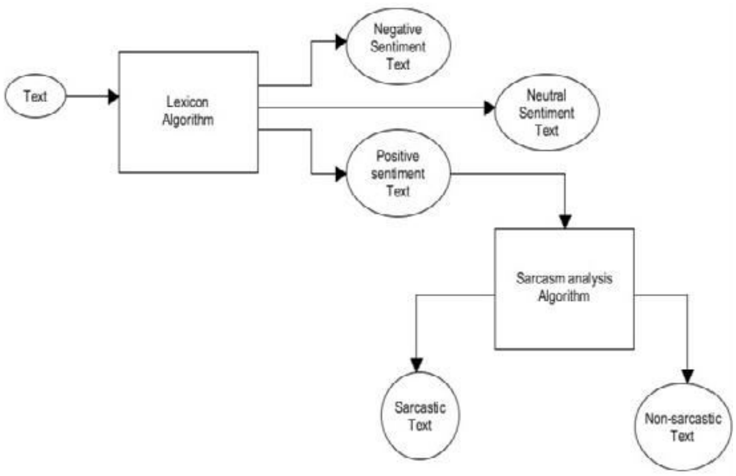

Figure 3.1: Proposed Block diagram of Lexical Algorithm

The Project aims to develop a tool that can automatically detect sarcasm and analyze sentiment in Twitter data. Sarcasm and sentiment analysis are crucial in understanding the context and emotions expressed in social media text, which has numerous applications in market research, social trend analysis, and sentiment monitoring.

— **Graphical User Interface (GUI):** The GUI is developed using Tkinter, a standard Python library for creating graphical interfaces. It provides users with an intuitive platform to interact with the application. The interface includes buttons for uploading datasets, preprocessing data, running algorithms, and visualizing results. Text areas are used to display the dataset, preprocessing outputs, and analysis results, ensuring clear communication with the user.

— **Text Preprocessing:** Before analyzing the data, text preprocessing is performed to clean and prepare the text for analysis. This involves converting text to lowercase, removing punctuation, stopwords, and non-alphabetic characters. NLTK's stopwords corpus and string library are utilized for this purpose. The cleaned text is then ready for sarcasm detection and sentiment analysis.

— **Sarcasm Detection and Sentiment Analysis:** Two algorithms are implemented for sarcasm detection and sentiment analysis. The first algorithm combines lexicon-based analysis with polarity computation using the VADER sentiment analysis tool. It determines sarcasm based on polarity scores and the presence of negative words. The second algorithm also uses lexicon-based analysis with sentiment prediction, categorizing tweets as positive, negative, or neutral based on compound polarity scores.

— **Data Visualization:** Matplotlib is utilized for data visualization, enabling users to gain insights from the analysis results. Bar graphs and pie charts are generated to visualize the distribution of sarcastic and non-sarcastic tweets and sentiment analysis outcomes. These visualizations facilitate the interpretation and understanding of the data.

— **Future Directions:** The project code lays the foundation for further enhancements and extensions. Future improvements could include integrating advanced machine learning models for more accurate sarcasm detection and sentiment analysis. Additional features such as user authentication, data exporting, and customization options for visualization could be incorporated to enhance usability and functionality.

— In summary, the project code provides a robust and user-friendly solution for analyzing sarcasm and sentiment in Twitter data, offering valuable insights for social media researchers, analysts, and practitioners.

3.2 Lexicon Algorithm

A lexicon algorithm is a computational method used in natural language processing (NLP) and text analysis to interpret and analyse text based on predefined dictionaries or lexicons. These algorithms rely on curated collections of words or phrases, each accompanied by attributes such as part of speech, semantic category, sentiment polarity, subjectivity, and intensity. One prominent application of lexicon algorithms is sentiment analysis, where lexicons contain sentiment scores assigned to words indicating their positive, negative, or neutral connotations. Lexicon algorithms evaluate the sentiment scores of individual words in a piece of text and aggregate them to determine the overall sentiment expressed within the text. Lexicon algorithms employ scoring mechanisms to quantify the sentiment or other attributes of text. Words in the lexicon are assigned numerical scores reflecting their sentiment polarity, and these scores are aggregated and normalized to compute an overall sentiment score for the text. Advanced lexicon algorithms consider the context in which words appear to enhance accuracy. Context-aware algorithms take into account factors such as word order, proximity to other words, syntactic structures, and semantic relationships to better interpret the meaning and sentiment of text. The effectiveness of lexicon algorithms hinges on the quality and coverage of the lexicon itself. A high-quality lexicon should be comprehensive, accurate, and contextually relevant to the domain or language being analysed. However, lexicon algorithms have limitations, particularly in handling linguistic nuances such as sarcasm, irony, ambiguity, and context-dependent meanings. to overcome limitations and enhance performance, lexicon algorithms are often combined with other NLP techniques, such as machine learning algorithms, rule-based systems, or deep learning models. Hybrid approaches leverage the strengths of both lexicon-based methods and statistical methods to achieve more robust text analysis. lexicon algorithms provide a structured framework for text analysis, particularly in sentiment analysis and other NLP tasks.

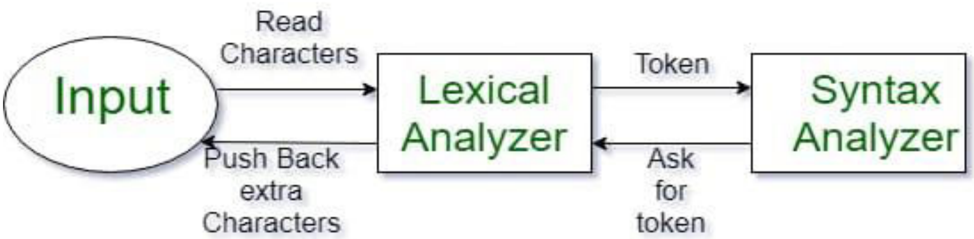

Figure 3.2: Lexicon algorithm

A lexicon algorithm, within the domain of natural language processing (NLP), employs predefined dictionaries or lexicons to analyse and decode text. Lexicons consist of words or phrases accompanied by attributes like part of speech, semantic category, sentiment polarity, subjectivity, and intensity. Among its applications, sentiment analysis stands out, where lexicons contain sentiment scores assigned to words, indicating positive, negative, or neutral sentiment. The algorithm interprets text by assessing the sentiment scores of individual words and consolidating them to determine the overall sentiment. To enhance accuracy, some lexicon algorithms consider word context, such as order, surrounding words, syntactic structure, or semantic relationships. The efficacy of lexicon algorithms heavily relies on the quality and coverage of the lexicon itself, necessitating comprehensive, accurate, and contextually relevant content. Despite their simplicity and transparency, lexicon algorithms have limitations, often struggling with nuances like sarcasm, irony, or ambiguity. They was also encounter difficulties with out-of-vocabulary words or domain-specific terminology not accounted for in the lexicon. Nonetheless, lexicon algorithms form a fundamental component of text analysis in NLP, frequently complemented by machine learning techniques for heightened precision and adaptability.

4. RESULT AND DISCUSSION

Figure 1 Present the GUI of Sarcasm Net: Extension of the Lexicon Algorithm for Emoji Based Sarcasm Detection Using Twitter Data”. It contains the below elements like Upload Social Network Data Set, Preprocess Dataset, Run First System Lexicon+ Polarity Computation, Second System Lexicon+ Sentiment Prediction, Sentiments Graph and Sarcastic Graph

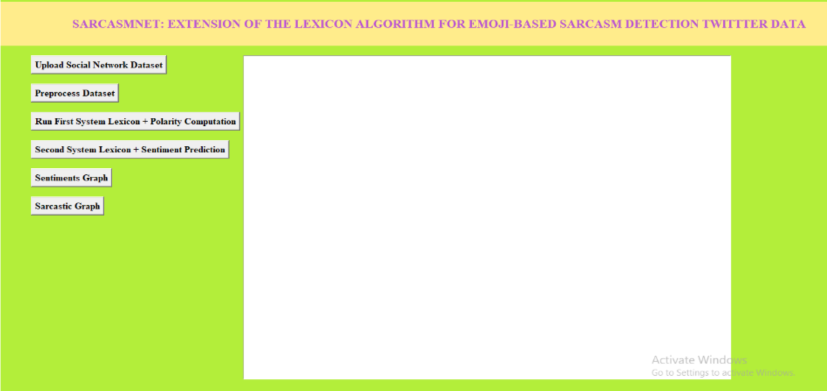

Fig 1: Provides GUI of SarcasmNet.

Fig 2 Shows the Upload Social Network Dataset: The first step in the process involves uploading a social network dataset obtained from Twitter. This dataset serves as the foundation for subsequent analyses

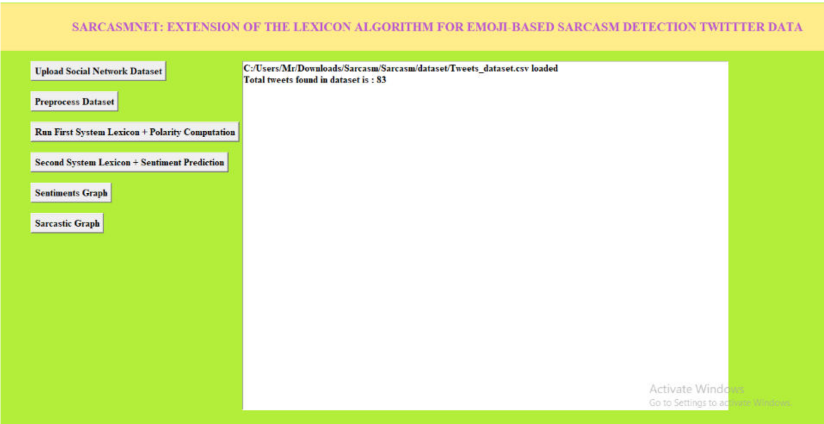

Fig 2: Upload social network dataset in the GUI

Fig 3 Preprocess Dataset: the dataset undergoes preprocessing, which includes tasks like removing irrelevant information, handling missing data, and cleaning up text to enhance the quality of the dataset.

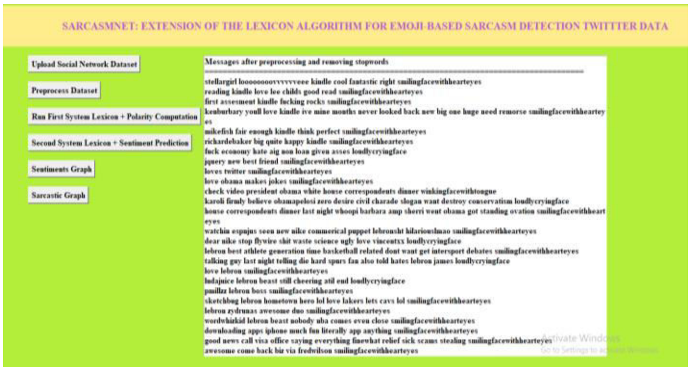

Fig 3: Run first system lexicon polarity computation

Fig 3 Run First System Lexicon Polarity Computation: involves running the first system lexicon and polarity computation. For the first system, lexicon-based polarity computation is employed. Lexicon-based approaches leverage sentiment analysis dictionaries to assign polarities (positive, negative, or neutral) to words in the dataset. This allows for a basic sentiment assessment of each tweet based on the overall sentiment of its constituent words.

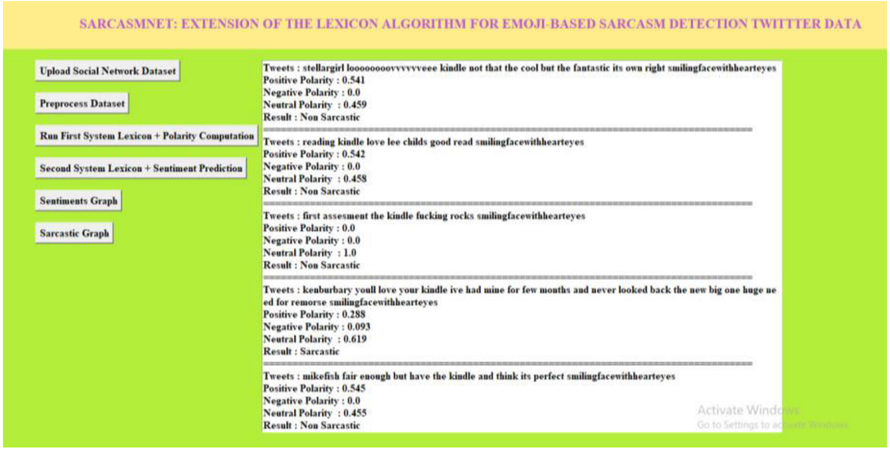

Fig 4: predicted outcomes of lexicon polarity computation.

Fig 5 Second System Lexicon Sentiment Prediction: involves second system lexicon and sentiment prediction. In the second system, a more sophisticated lexicon-based approach is employed, incorporating sentiment prediction. This involves using a lexicon with more nuanced sentiment scores and potentially integrating machine learning models to predict sentiment. This system aims to provide a more refined understanding of the sentiments expressed in the tweets.

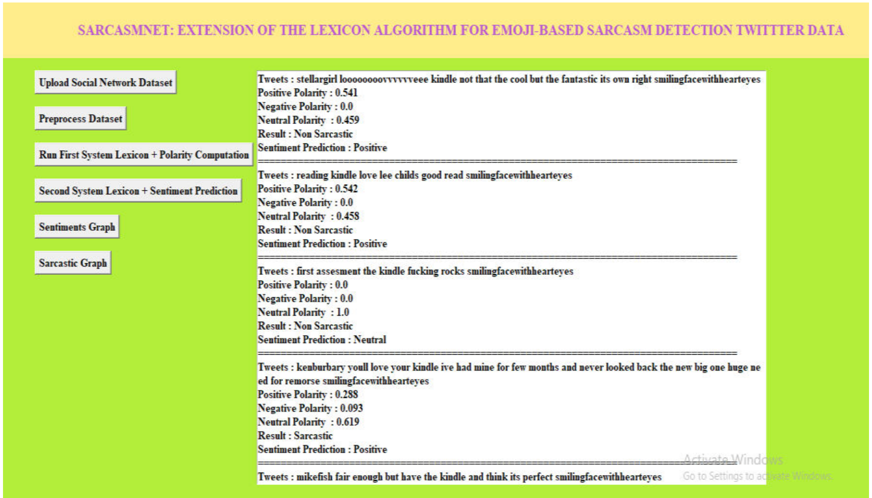

Fig 5: Proposed Lexicon Sentiment Predicted outcomes.

Sentiments Graph: The sentiments graph is a visual representation of the overall sentiments present in the dataset. It helps in identifying trends and patterns in sentiment distribution across the social network dataset

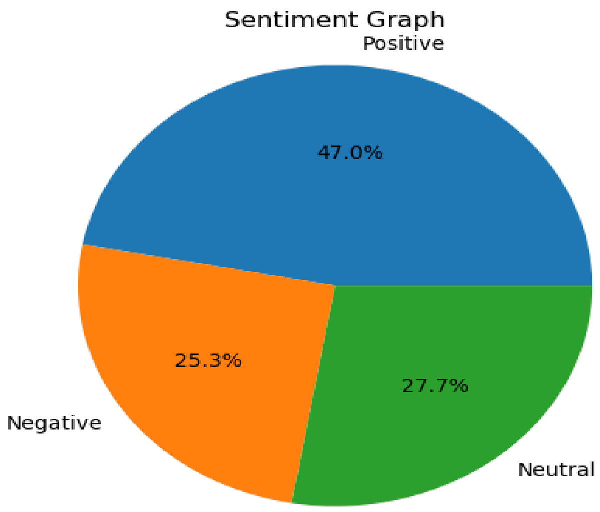

Figure 6: Pie Chart for Sentiment Visualization.

Figure 7 The sarcastic graph is created to depict instances of sarcasm within the dataset. This graph is generated based on predefined sarcastic lexicons or through the application of machine learning models trained to identify sarcastic expressions. The goal is to visualize the prevalence and distribution of sarcastic content in the social network dataset.

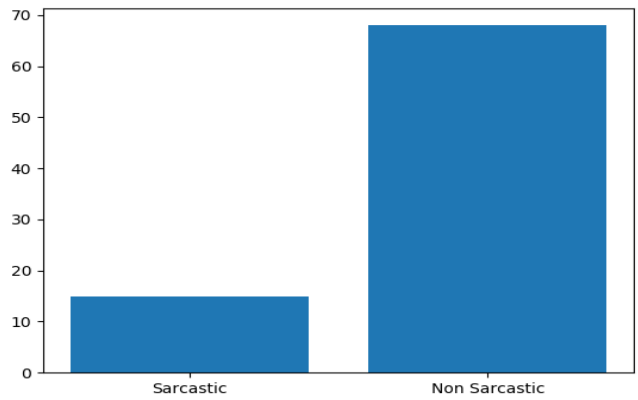

Figure 7: Count Plot of sarcastic or non-sarcastic tweets

5. CONCLUSION

The developed project code provides a functional tool for analyzing sarcasm and sentiment in Twitter data, offering a valuable resource for researchers, analysts, and social media practitioners. By leveraging lexicon-based analysis, sentiment computation, and data visualization techniques, the application enables users to gain insights into the emotional context of social media conversations.

The graphical user interface (GUI) enhances usability by providing an intuitive platform for interacting with the application, allowing users to upload datasets, preprocess data, run analysis algorithms, and visualize results seamlessly. Text preprocessing ensures that the data is cleaned and prepared for accurate analysis, while the implemented algorithms effectively detect sarcasm and predict sentiment based on polarity scores.

Data visualization techniques, including bar graphs and pie charts, aid in interpreting and understanding the analysis results, facilitating decision-making and further exploration of the data. Overall, the project code serves as a valuable tool for social media research, market analysis, sentiment monitoring, and trend analysis.

REFERENCES

[1]. Kumar, Akshi, Saurabh Raj Sangwan, Adarsh Kumar Singh, and Gandharv Wadhwa. "Hybrid deep learning model for sarcasm detection in Indian indigenous language using word-emoji embeddings." ACM Transactions on Asian and Low-Resource Language Information Processing 22, no. 5 (2023): 1-20.

[2]. Chauhan, Dushyant Singh, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. "An emoji-aware multitask framework for multimodal sarcasm detection." Knowledge-Based Systems 257 (2022): 109924.

[3]. Subramanian, Jayashree, Varun Sridharan, Kai Shu, and Huan Liu. "Exploiting emojis for sarcasm detection." In Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12, pp. 70-80. Springer International Publishing, 2019.

[4]. Prasad, Anukarsh G., S. Sanjana, Skanda M. Bhat, and B. S. Harish. "Sentiment analysis for sarcasm detection on streaming short text data." In 2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA), pp. 1-5. IEEE, 2017.

[5]. Lemmens, Jens, Ben Burtenshaw, Ehsan Lotfi, Ilia Markov, and Walter of the Daelemans. "Sarcasm detection using an ensemble approach." In proceedings second workshop on figurative language processing, pp. 264-269. 2020.

[6]. Gregory, Hunter, Steven Li, Pouya Mohammadi, Natalie Tarn, Rachel Draelos, and Cynthia Rudin. "A transformer approach to contextual sarcasm detection in twitter." In Proceedings of the second workshop on figurative language processing, pp. 270-275. 2020.

[7]. Agrawal, Ameeta, and Aijun An. "Affective representations for sarcasm detection." In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1029-1032. 2018.

[8]. Li, Yangyang, Yuelin Li, Shihuai Zhang, Guangyuan Liu, Yanqiao Chen, Ronghua Shang, and Licheng Jiao. "An attention-based, context-aware multimodal fusion method for sarcasm detection using inter-modality inconsistency." Knowledge-Based Systems (2024): 111457.

[9]. Lu, Qiang, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. "Fact-sentiment incongruity combination network for multimodal sarcasm detection." Information Fusion 104 (2024): 102203.

[10]. Prashanth, K. V. T. K. N., and Tene Ramakrishnudu. "Sarcasm-based tweet-level detection." Expert Systems (2024):e13534.

[11]. Rustagi, Aanshi, Annapurna Jonnalagadda, Aswani Kumar Cherukuri, and Amir. "Toward sarcasm detection in reviews—a dual parametric approach with emojis and ratings." Soft Computing in Interdisciplinary Sciences (2022): 245-257.

[12]. Cui, Jing, Yu R. Dandan, and Guiying Jiang. "Judging emoji by occupation: A case of emoji-based sarcasm interpretation." Acta Psychologica 234 (2023): 10387.

[13]. Sukhavasi, Vidyullatha and Venkatesulu Dondeti. ”Effective Automated Transformer Model based Sarcasm Detection Using Multilingual Data." Multimedia Tools and Applications (2023): 1-32.

[14]. Garcia, Charlotte, Alexandra Țurcan, Hannah Howman, and Ruth Filik. "Emoji as a tool to aid the comprehension of written sarcasm: Evidence from younger adults." Computers in Human Behavior 126 (2022): 106971.

[15]. Farnham, Niamh. "Unlocking the Pragmatics of Emoji: Evaluation of the Integration of Pragmatic Markers for Sarcasm Detection." (2023).