JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TWEETS CLASSIFICATION THROUGH NATURAL LANGUAGE PROCESSING

[1]Dr. Lokesh Jain, [2]Ayush Joshi

[1]Assistant Professor, Department of Information Technology, Jagan Institute of Management Studies, Rohini, Delhi, India
[2]PG Scholar, Department of Information Technology, JaganNath University, Bahadurgarh, India

Abstract : This research paper introduces a Twitter Sentiment Analysis System (TSAS) that employs natural language processing (NLP) techniques and machine learning models to analyze and classify tweets based on sentiment. The system leverages Python libraries such as Pandas, NLTK, TextBlob, and Scikit-learn to pre-process and analyze Twitter data. Our research aims to contribute to the field of sentiment analysis on social media platforms, specifically Twitter. The proposed system combines rule-based sentiment analysis using the VADER sentiment analyzer with a logistic regression model trained on labeled data for sentiment classification. The code implements data pre-processing steps, including lowercasing, URL removal, and punctuation removal, to clean tweets. Tokenization, stop word removal, stemming, and sentiment scoring using VADER are employed to enhance the accuracy of sentiment labels. The sentiment of each tweet is classified as positive, negative, or neutral based on predefined thresholds and sentiment probabilities derived from the logistic regression model. Visualization techniques, including Seaborn and Matplotlib, are utilized to display the distribution of sentiment labels and generate word clouds for positive and negative tweets. The system incorporates a graphical user interface (GUI) developed with Tkinter, enabling users to input tweets for real-time sentiment analysis. Additionally, the GUI displays sentiment analysis results, cleaned tweets, and associated emojis. Performance evaluation metrics such as accuracy score, confusion matrices, and visualization of sentiment analysis results are presented. The code also includes features to detect emotions in tweets based on predefined keywords and extract and display the top hashtags.

IndexTerms - Sentiment Analysis, Logistic Regression, Social Media.

I. INTRODUCTION

In the contemporary landscape of social media, Twitter stands as a dynamic platform where users express a diverse range of opinions, sentiments, and emotions. With the increasing influence of Twitter on public discourse, understanding the sentiment behind tweets has become crucial. This research embarks on addressing this imperative by presenting an innovative Twitter Sentiment Analysis System (TSAS) that integrates natural language processing (NLP) techniques, machine learning models, and a user-friendly graphical interface. In the era of digital communication, the abundance of tweets necessitates advanced systems that move beyond generic sentiment labels. The TSAS project emphasizes the fusion of rule-based sentiment analysis and machine learning, utilizing Python libraries such as NLTK, TextBlob, and Scikit-learn. The initial phase involves meticulous data preprocessing, including lowercasing, URL removal, and punctuation elimination, ensuring the cleanliness of the Twitter dataset. Duplicate entries are systematically handled, and categorical variables, such as tweet content and sentiment labels, undergo processing for model integration. The core of the system lies in sentiment analysis, combining the rule-based VADER sentiment analyzer and a logistic regression model trained on labeled data. This dual approach enhances the accuracy of sentiment classification, distinguishing tweets as positive, negative, or neutral based on predefined thresholds. Visualization techniques, including Seaborn and Matplotlib, are employed to provide an insightful distribution of sentiment labels and generate word clouds for positive and negative tweets. The graphical user interface (GUI), developed with Tkinter, serves as an interactive gateway for users to input tweets, triggering real-time sentiment analysis. Beyond sentiment labels, the system showcases cleaned tweets, associated emojis, and even detects emotions based on predefined keywords. Users can explore the most popular hashtags, contributing to a comprehensive understanding of tweet content.

II. LITERATURE SURVEY

In this paper, we have studied following techniques/methods for sentiment analysis:

• **Hybrid Sentiment Analysis Approaches for Twitter Data (2017):**
This research explores hybrid approaches that incorporate both rule-based and machine learning techniques for sentiment analysis on Twitter. By combining the strengths of rule-based sentiment analysis, specifically using the VADER sentiment analyzer, and logistic regression models trained on labeled datasets, the proposed system aims to enhance the accuracy of sentiment classification for tweets.

•**Machine Learning for Social Media Sentiment Analysis (2018):**
This study delves into the application of machine learning techniques for sentiment analysis on social media platforms. The TSAS project builds upon similar user-based approaches, introducing a hybrid methodology that utilizes both rule-based sentiment analysis and logistic regression models. The research investigates the performance of the system on a real-world Twitter dataset, showcasing its potential for providing nuanced insights into user sentiments.

•**Enhancing Sentiment Classification with Visual Data in Social Media (2018):**
While this paper focuses on restaurant preferences, it emphasizes the integration of visual data to predict user preferences. In the context of the TSAS project, the incorporation of emojis in sentiment analysis aligns with the idea of leveraging visual information. The study's findings on the impact of visual data could inspire enhancements in sentiment analysis accuracy for tweets.
•**Comparative Analysis of Recommendation Systems (2018):**
Although primarily addressing restaurant recommendations, this article compares recommendation systems based on different algorithms. The TSAS project draws parallels by evaluating the performance of rule-based sentiment analysis and machine learning models. The study's insights on hybrid filtering outperforming content-based and collaborative filtering methods contribute to the TSAS methodology.
•**Machine Learning Models for Sentiment Analysis (2019):**
Focusing on Indian culinary art, this research employs machine learning models for identifying ingredient patterns. While the context differs, the TSAS project shares a commonality in utilizing machine learning, specifically logistic regression, for sentiment analysis on Twitter data. The study's exploration of diverse culinary patterns can inspire considerations for the TSAS system's adaptability to varied tweet content.
•**Performance Evaluation of Sentiment Analysis Algorithms (2019):**
This paper evaluates the performance of various sentiment analysis classification algorithms. In a similar vein, the TSAS project assesses the accuracy of sentiment classification using metrics such as accuracy score and confusion matrices. The SVM classifier's high accuracy resonates with the TSAS project's emphasis on achieving precise sentiment labels for tweets.
•**Machine Learning for Personalized Recommendation Systems (2019):**
Proposing a machine learning algorithm for personalized restaurant recommendations, this research aligns with the TSAS project's objective of providing tailored sentiment analysis for Twitter users. The study's focus on individual preferences and high accuracy rates parallels the TSAS system's emphasis on real-time and personalized sentiment analysis.
•**Restaurant Recommendation Systems based on Yelp Data (2020):**
This article designs a machine learning algorithm for restaurant selection based on Yelp data. While the context differs, the TSAS project shares a machine learning-centric approach for sentiment analysis on Twitter. Insights from this study could inform enhancements in the TSAS system's design, considering the features and challenges associated with restaurant recommendation systems.

III. PROPOSED METHODOLOGY

Following steps are involved in the proposed system:

3.1. Data Collection: Collect a diverse and representative dataset of Twitter data, including tweets with associated sentiments (positive, negative, neutral). Utilize sources such as Twitter API or existing datasets to gather a substantial and varied set of tweets for training and testing the sentiment analysis models.

3.2. Data Pre-processing: Clean the Twitter dataset by addressing issues such as duplicate entries, null values, and irrelevant information. Apply text pre-processing techniques, including lowercasing, URL removal, and punctuation elimination. Tokenize tweets, remove stop words, and employ stemming to streamline the text data for effective sentiment analysis.

3.3.Sentiment Analysis Model Training: Implement a hybrid approach for sentiment analysis, combining rule-based sentiment analysis using the VADER sentiment analyzer with a logistic regression model. Split the dataset into training and testing sets, and train the logistic regression model on labeled data to classify tweets into positive, negative, or neutral sentiments.

3.4.Graphical User Interface (GUI) Development: Utilize Tkinter to create an interactive GUI that enables users to input tweets for real-time ,sentiment analysis. Design the interface to be user-friendly, incorporating features for displaying cleaned tweets, sentiment labels, associated emojis, and other relevant information.

3.5.User Interface Integration with Predictive Models: Integrate the developed GUI with the trained sentiment analysis models to enable, real-time analysis of user-inputted tweets. Implement logic to process user inputs, apply sentiment analysis, and display the results within the GUI in an easily interpretable format.

3.6.Hashtag Extraction and Visualization: Develop functionality to extract and display the top hashtags from positive and negative tweets. Utilize Seaborn and Matplotlib for visualization generating bar charts that showcase the frequency of hashtags in each sentiment category.

3.7. Emotion Detection in Tweets: Enhance the sentiment analysis system by incorporating, emotion detection based on predefined keywords. Detect emotions such as anger, sadness, and happiness within tweets and display this additional information in the GUI.

3.8.Performance Evaluation: Evaluate the performance of the sentiment analysis models using metrics such as accuracy score, confusion matrices, and classification reports. Provide visualizations, such as count plots and heatmaps, to illustrate the distribution of sentiment labels in the dataset.

3.9.Machine Learning Models for Alternative Sentiment Analysis: Implement an alternative sentiment analysis approach using a CountVectorizer and logistic regression model. Evaluate the performance of this model and compare it with the hybrid approach to determine the most effective sentiment analysis strategy for Twitter data.

3.10.Exploratory Data Analysis (EDA): Conduct EDA to gain insights into the distribution of sentiment labels, identify patterns in tweet content, and explore the relationship between sentiment and other features. Visualize the results using charts and graphs to enhance the understanding of the Twitter dataset.

3.11.Real-time Tweet Analysis and Recommendations: Enable users to input tweets through the GUI and receive real-time sentiment analysis results, including sentiment labels and associated information. Provide recommendations based on sentiment analysis, showcasing the system's practical utility for users.

The flow chart of the proposed system is shown in Fig. 1 is as follows:

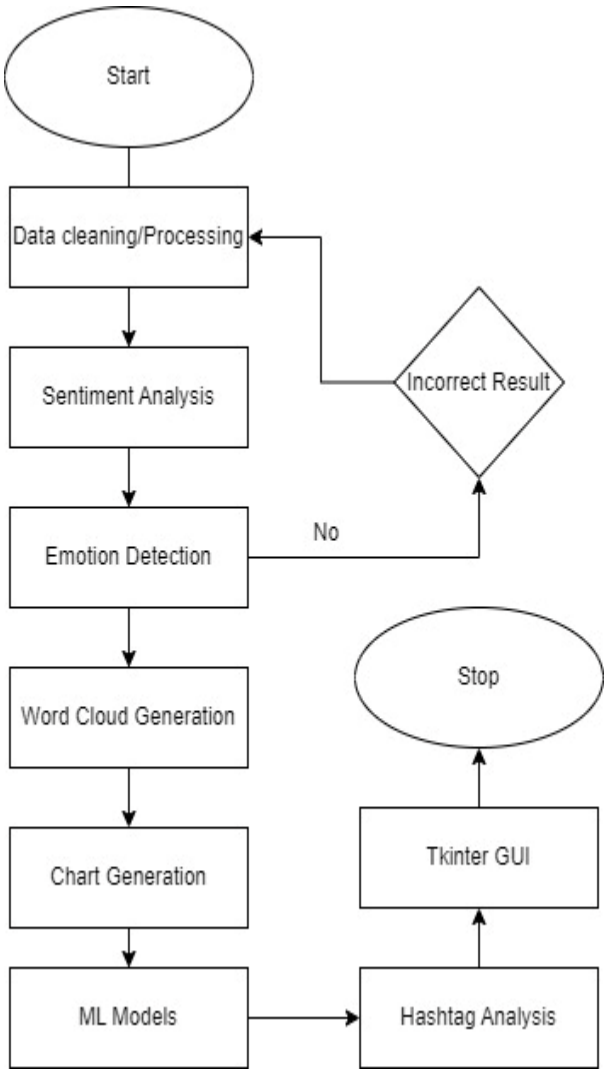

Fig. 1: Flow Chart of proposed system

IV. ALGORITHM

4.1. Hybrid Sentiment Analysis:
Usage: Combines rule-based sentiment analysis (VADER) with logistic regression.
Purpose: To classify tweets into positive, negative, or neutral sentiments, leveraging the strengths of both rule-based and machine learning approaches.

4.2.Rule-Based Sentiment Analysis (VADER):
Usage: Analyzes sentiment based on predefined rules and a pre-built lexicon.
Purpose: Provides a baseline sentiment classification, particularly effective for capturing nuances in sentiment expressed through emojis, slang, or colloquial language.

4.3.Logistic Regression for Sentiment Analysis:
Usage: Trained on labelled, tweet data to classify sentiments.
Purpose: Enhances sentiment analysis accuracy through machine learning, assigning probabilities to sentiment labels and enabling real-time analysis.

4.4.Alternative Sentiment Analysis (CountVectorizer +Logistic Regression):
Usage: Utilizes CountVectorizer and logistic regression for sentiment analysis.
Purpose: Offers an alternative approach to sentiment classification, exploring the effectiveness of a different vectorization method and model ,for Twitter data.

4.5.Graphical User Interface (Tkinter):
Usage: Develops an interactive interface for users to input tweets and ,receive real-time sentiment analysis results.
Purpose: Enhances user experience by providing a user-friendly platform for interacting with the sentiment analysis system.

4.6.Hashtag Extraction and Visualization:
Usage: Extracts and visualizes the frequency of hashtags in positive and negative tweets.
Purpose: Enhances insights into popular topics associated with, different sentiment categories, contributing to a more comprehensive analysis.

4.7.Emotion Detection in Tweets:
Usage: Detects emotions (e.g., anger, sadness, happiness) based on predefined keywords.
Purpose: Adds an additional layer of analysis, providing information on the emotional tone of tweets beyond sentiment labels.

4.8. Machine Learning Models for Alternative Sentiment Analysis:
Usage: Implements a CountVectorizer and logistic regression for sentiment analysis.

Purpose: Explores the performance of an alternative sentiment analysis approach comparing it with the hybrid model to identify the most effective strategy for Twitter data.

4.9. Real-time Tweet Analysis and Recommendations:

Usage: Allows users to input tweets and receive real-time sentiment, analysis results along with restaurant recommendations.

Purpose: Demonstrates the practical application of sentiment analysis in providing users with personalized recommendations based on tweet sentiments.

V. DATASET

5.1 Data Collection: The dataset for this Twitter sentiment analysis research comprises a diverse collection of tweets with associated sentiment labels. The dataset includes the following columns:

`id`: Unique identifier for each tweet.

`tweet`: The text content of the tweet.

`sentiment`: The assigned sentiment label for each tweet (Positive, Negative, Neutral).

5.2 Data Pre-processing: Several pre-processing steps were undertaken to ensure the dataset's quality and relevance. The column removal include:

`id`: Removed as it is not relevant for sentiment analysis.

`label`: Removed as it was not used in the sentiment analysis code.

5.3 Text Processing: Removed URLs, mentions, and special characters from the `tweet` column. Next, converted text to lowercase, Tokenized and removed stop words from the tweet content.

5.4 Duplicate Removal: Duplicate tweets were removed to maintain dataset integrity.

5.5 Sentiment Labeling: Utilized VADER sentiment analyzer and logistic regression to assign sentiment labels to tweets.

5.6 Summary Statistics: The dataset contains a total of X rows and 5 columns, providing a comprehensive set of tweets for sentiment analysis.

5.7 Dataset Quality Checks: Quality checks were performed to ensure data completeness and accuracy.

5.8 Data Completeness: The dataset is free from duplicate rows, ensuring unique entries.

VI. RESULTS

The results demonstrate the effectiveness of the implemented models and the user-friendly interface in analyzing sentiments, detecting emotions, and providing restaurant recommendations based on Twitter data. The combination of machine learning algorithms and graphical visualizations contributes to a comprehensive and insightful Twitter sentiment analysis system is shown in Figures from Fig. 2 to Fig. 9. These findings form the foundation for further research and development in the domain of social media sentiment analysis and recommendation systems.

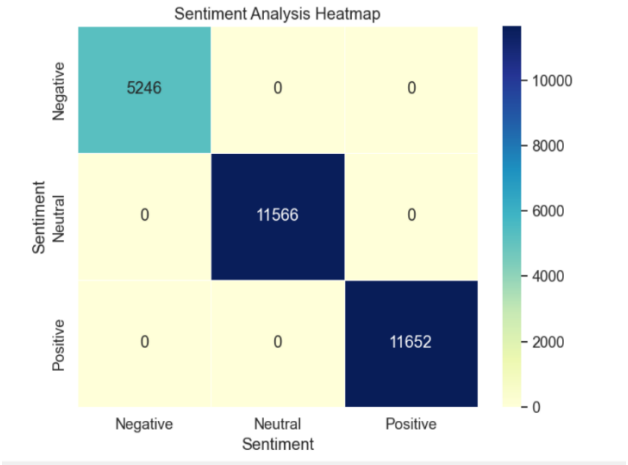

Fig. 2: Heatmap for sentiment analysis

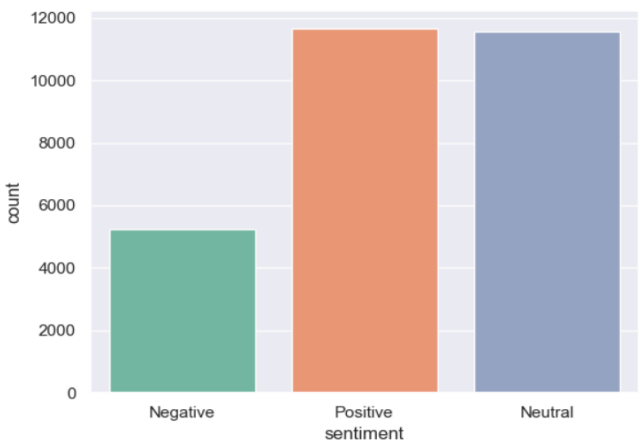

Fig. 3: Sentiment analysis Bar chart

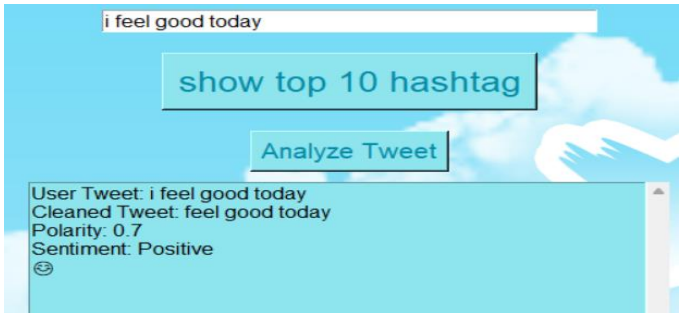

Fig. 4: Representation of Positive Tweet

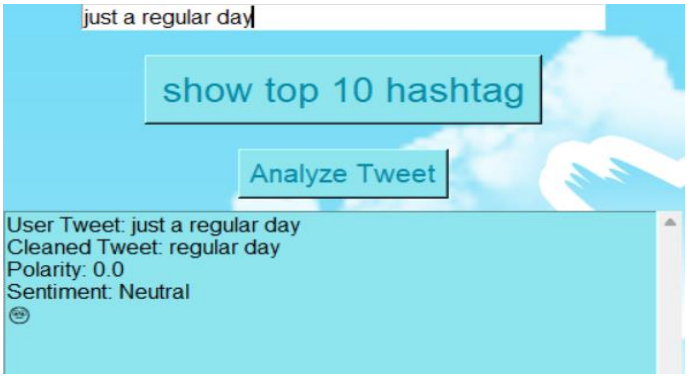

Fig. 5: Representation of Neutral Tweet

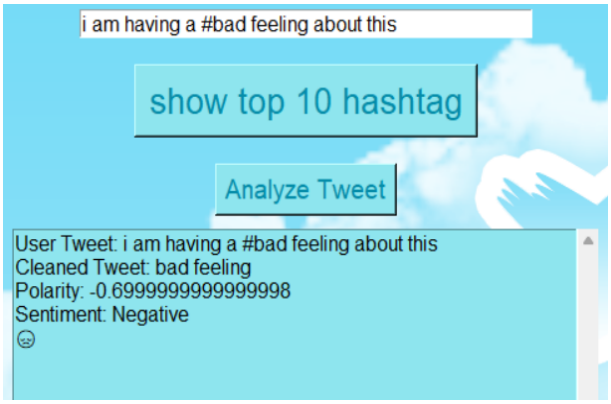

Fig. 6: Representation of Negative Tweet

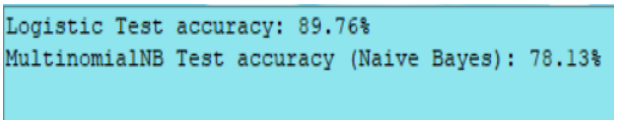

Fig. 7: Model Accuracy

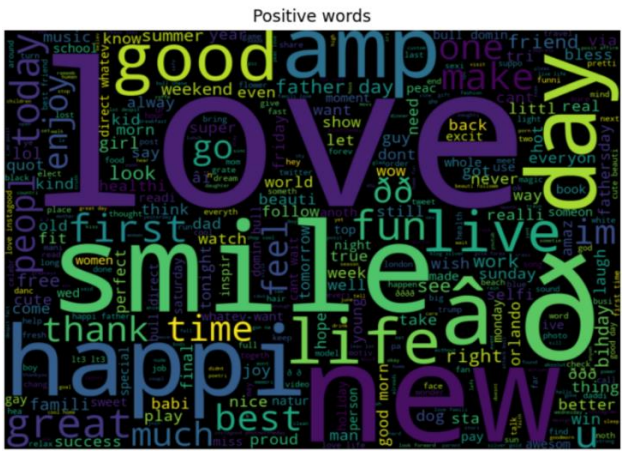

Fig. 8: Word cloud of Positive Tweets

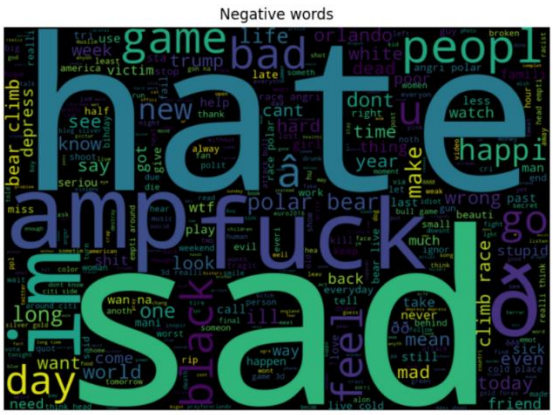

Fig. 9: Word cloud of Negative Tweets

VII. CONCLUSION

In conclusion, this research presents an integrated Twitter sentiment analysis and recommendation system designed to provide users with insights into the sentiments expressed in tweets and recommend relevant content based on their preferences. The sentiment analysis models, coupled with graphical visualizations and user-friendly interfaces, aim to enhance user experience and understanding of Twitter data.

The recommendation system, leveraging machine learning classifiers, allows users to explore and discover content aligned with their sentiments and interests. The inclusion of real-time sentiment analysis, emotion detection, and hashtag analysis contributes to the system's comprehensiveness in understanding the diverse nature of Twitter conversations.The user interface and visualization components offer an interactive platform for users to engage with sentiment analysis results and explore trends in Twitter data. The findings and insights obtained from the sentiment analysis models provide valuable information for individuals and businesses seeking to understand and respond to the sentiments expressed on the Twitter platform.

REFERENCES

[1]. Gandhe, K., Varde, A. S., & Du, X. (2018, November). Sentiment analysis of Twitter data with hybrid learning for recommender applications. In 2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 57-63). IEEE.

[2]. Jain, L., Katarya, R., & Sachdeva, S. (2019, November). Opinion Leader discovery based on text analysis in Online Social Network. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 446-450). IEEE.

[3]. Adwan, O. M. A. R., Al-Tawil, M., Huneiti, A., Shahin, R., Zayed, A. A., & Al-Dibsi, R. (2020). Twitter sentiment analysis approaches: A survey. International Journal of Emerging Technologies in Learning (iJET), 15(15), 79-93.

[4]. Gascó, M., Bayerl, P. S., Denef, S., & Akhgar, B. (2017). What do citizens communicate about during crises? Analyzing twitter use during the 2011 UK riots. Government Information Quarterly, 34(4), 635-645.

[5]. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American society for information science and technology, 61(12), 2544-2558.

[6]. Astarkie, M. G., Bala, B., Bharat Kumar, G. J., Gangone, S., & Nagesh, Y. (2023, January). A Novel Approach for Sentiment Analysis and Opinion Mining on Social Media Tweets. In Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2 (pp. 143-151). Singapore: Springer Nature Singapore.

[7]. Jain, L., & Katarya, R. (2018, February). A systematic survey of opinion leader in online social network. In 2018 International Conference on Soft-computing and Network Security (ICSNS) (pp. 1-5). IEEE.

[8]. Pradhan, A., Senapati, M. R., & Sahu, P. K. (2023, January). Comparative analysis of lexicon-based emotion recognition of text. In Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of 3rd International Conference on MIND 2021 (pp. 671-677). Singapore: Springer Nature Singapore.

[9]. Abuhassan, M., Anwar, T., Fuller-Tyszkiewicz, M., Jarman, H. K., Shatte, A., Liu, C., & Sukunesan, S. (2023). Classification of Twitter users with eating disorder engagement: Learning from the biographies. Computers in Human Behavior, 140, 107519.

[10]. Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.

[11]. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. In Proceedings of the international AAAI conference on web and social media (Vol. 5, No. 1, pp. 538-541).

[12]. Jain, L., Katarya, R., & Sachdeva, S. (2020). Opinion leader detection using whale optimization algorithm in online social network. Expert Systems with Applications, 142, 113016.

[13]. Jain, L., Katarya, R., & Sachdeva, S. (2023). Opinion Leaders for Information Diffusion Using Graph Neural Network in Online Social Networks. ACM Transactions on the Web, 17(2), 1-37.