A Novel Approach for Sentiment Analysis on Social Media Data

***Sanjana Dehariya**
****Prof. Rakesh Shivhare**

ABSTRACT

The use of social media is increasing rapidly, significantly influencing societal changes as public opinion shared on these platforms gains more importance. Among social media platforms, Twitter has garnered substantial attention due to its real-time nature. In this study, we examine recent societal shifts associated with the many movement by developing a tool called SocialAnalyzer.
Our implementation of SocialAnalyzer follows a four-phase approach, analyzing a dataset of 393,869 static and streamed entries collected from the Data World website. Using a classifier, we categorize the data into sentiment score. Results indicate that most opinions fall under the 2-3 score, with contrary opinions constituting the second-largest group.
To validate our findings, we compared SocialAnalyzer performance against TextBlob on a subset of 765 tweets. When treating neutral tweets as positive, the precision scores for SocialAnalyzer and TextBlob were 70.74% and 72.92%, respectively. These results demonstrate the potential of SocialAnalyzer in analyzing societal trends and public sentiment.

Keywords:- *Natural Language Processing (NLP), Tweet, Social Media, Machine Learning.*

*Prabhati Bharti, Research Scholar, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal , sanjanadehariya13@gmail.com
**Prof. Rakesh Shivhare, HOD, Department of Computer Science & Engineering, Radharaman Engineering College, Bhopal, India

I. INTRODUCTION

As humans, we are naturally drawn to like-minded individuals. Studies have shown that we feel more comfortable socializing with people who share similar beliefs, earn our trust, and support us in achieving our goals. Historically, this tendency to associate with like-minded groups has shaped community dynamics. Communities are often composed of multiple interconnected clusters, and modularity is a key measure used to quantify these communities [1].

By closely analyzing the characteristics of these clusters, we can gain valuable insights into the distinct traits of individual groups or like-minded communities. In essence, the presence of a shared connection among a group of individuals reflects common principles and purposes within that group.

Broadly speaking, social media platforms can be categorized into two types: social networks and online communities, each serving unique roles in fostering connections and interactions among users.

Social networks consist of individuals connected through pre-existing personal relationships, maintaining those ties while seeking new connections to expand their circle. These networks link people with direct relationships, fostering an organized structure. In contrast, communities bring together individuals from diverse backgrounds who may have little or no prior connection. What unites members of a community is a shared interest or common goal. People often engage with communities for various reasons—shared passions, a sense of belonging, or the opportunity to achieve personal objectives. Unlike the structured arrangements of social networks, communities often exhibit overlapping and nested configurations.

Social media serves as a platform for sharing information with vast audiences, acting as an interface to disseminate content widely. When paired with social networks, social media enables individuals to reach larger audiences for sharing ideas, promoting initiatives, or facilitating discussions [2].

Sentiment analysis is the process of categorizing opinions expressed about a specific subject. With advancements in technology, sentiment analysis has become essential for understanding public attitudes toward products, services, or general preferences. By analyzing the emotions behind social media posts, organizations can contextualize user reactions and predict behavior.

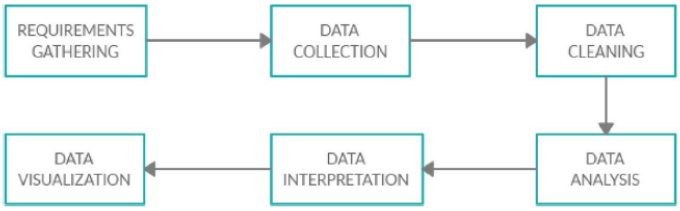

Figure 1. Data analysis steps

Natural Language Processing

Natural Language Processing (NLP) is a technique that enables intelligent systems to communicate using natural languages, such as English. It facilitates a wide range of tasks on these systems.

- **Lexical Analysis** involves identifying and analyzing the structure of words within a sentence, focusing on their formation and composition.
- **Syntactic Analysis** examines the grammatical structure and relationships between words, ensuring proper sentence formation.

- **Semantic Analysis** extracts the precise or dictionary meaning of the text, focusing on understanding its true context and significance.

Sentiment Analysis

Sentiment Analysis is the process of classifying a block of text as positive, negative, or neutral. Its primary goal is to understand people's interests and preferences in a way that can aid businesses in growth and decision-making. Beyond polarity (positive, negative, neutral), sentiment analysis also identifies emotions such as happiness, sadness, or anger.

This process leverages various Natural Language Processing (NLP) algorithms and involves contextual mining of words to reveal the social sentiment surrounding a brand or topic. It provides insights that help businesses determine whether their products are likely to meet market demand. Figure 2 illustrates the key steps involved in sentiment analysis.

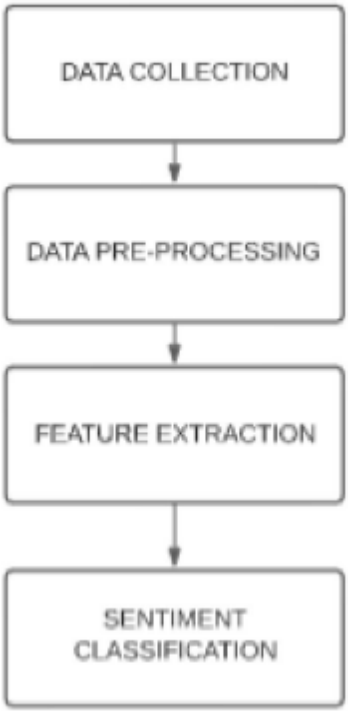

Figure 2. Sentiment analysis steps

- **Rule-based Sentiment Analysis:** This approach relies on predefined rules and a lexicon of words labeled by polarity to determine the sentiment of a text. However, accurately interpreting sentences with sarcasm, negations, or dependent clauses often requires additional processing to capture their true sentiment.

- **Machine Learning-based Sentiment Analysis:** This method involves training a Machine Learning model to identify sentiment polarity by analyzing the sequence and context of words, using a sentiment-labeled training dataset.

II. LITERATURE REVIEW

Joshi et al. [3] explore sentiment analysis from social media data in code-mixed Indian languages, employing machine learning classifiers with TF-IDF and weighted word features. Their work highlights the importance of considering linguistic diversity in sentiment analysis tasks.

Molenaar et al. [4] investigate social media opinions on food security using natural language processing techniques. Through sentiment analysis and topic modeling, they uncover valuable insights into public perceptions and attitudes towards food security issues.

He et al. [5] evaluate the use of large language models for sentiment analysis of health-related social media data. Their empirical study provides practical tips and insights into leveraging advanced language models for sentiment analysis tasks.

Almuhaya et al. [6] conduct a comparative analysis of machine learning algorithms for Arabic sentiment analysis on imbalanced social media data. Their study contributes to the understanding of sentiment analysis challenges in linguistic-specific contexts.

Paulraj et al. [7] propose a deep learning modified neural network (DLMNN) for proficient sentiment analysis on Twitter data. Their deep learning approach demonstrates promising performance in capturing nuanced sentiment patterns.

Dean and Porter [8] focus on sentiment analysis of Russian-language social media posts discussing the 2022 Russian invasion of Ukraine. Their study sheds light on public sentiment dynamics in the context of geopolitical events.

Srivastava et al. [9] explore sentiment analysis of Twitter data from a COVID-19 perspective, highlighting the role of social media in shaping public discourse during global crises.

Bukhari and Ramzan [10] conduct a systematic literature review on text mining on social media data, providing insights into the methodologies and applications of text mining techniques in sentiment analysis tasks.

Xinwei et al. [11] investigate the potential of social media data in business decision-making processes through an exploratory study. Their research emphasizes the utility of social media data for informing strategic decisions in various business domains.

Poudel and Weninger [12] examine biases in search engine results pages (SERPs) and their impact on social media data analysis. Their work highlights the importance of considering biases and limitations when analyzing social media data sourced from search engines.

Gothane et al. [13] propose deep learning techniques for sentiment analysis in social media, contributing to the advancement of sentiment analysis methodologies.

Tejaswini et al. [14] focus on depression detection from social media text analysis using natural language processing techniques and hybrid deep learning models. Their research addresses the critical issue of mental health monitoring through social media data analysis.

III. PROPOSED WORK

Our workflow for collecting and analyzing social media data consists of four distinct phases:

A. Data Collection

B. Preprocessing and Storage

C. Sentiment Analysis

D. Result Validation

These phases are shown in Figure 3.

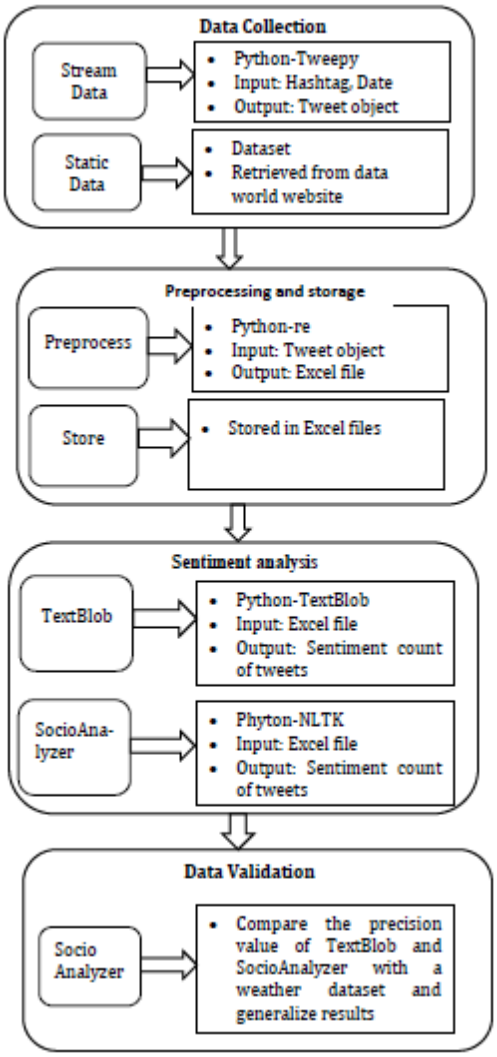

Figure 3. Methodology

A. Data Collection

We categorized our data into two types: static and stream data. The static data, consisting of content, was sourced from the Data World website. Stream data, also focused on the hashtag, was periodically scraped from social media using the Python Tweepy library. The hashtag and a specified date served as inputs to the Tweepy library, which generated social media objects for a given hashtag over a one-week period based on the input date.

To avoid redundancy, only the static data from the Data World website was used for analysis. A total of 393,869 post with the hashtag were retrieved from this source. Each post, with a maximum length of 140 characters, included elements such as the message, hashtags and emoticons, which help convey the user's emotions [15].

B. Preprocessing and Storage

Data preprocessing is a critical step in any Natural Language Processing (NLP) task [16]. We utilized Python's regular expression library to remove random patterns in posts, such as emoticons, embedded URLs and hashtags. The cleaned posts were then stored in an Excel file. The cleaned dataset includes the following fields: postId, dateOfPost and text (with URLs and hashtags removed).

C. Sentiment Analysis

We developed the SocialAnalyzer to assess whether people's reactions to the posts are positive, negative, or neutral. Key steps in creating the SocialAnalyzer include building a sentiment dictionary using Natural Language Processing (NLP) techniques with NLTK, and then classifying the input data into one of three score categories: 1, 2, 3. Figure 4 illustrates the process of data analysis and categorization.

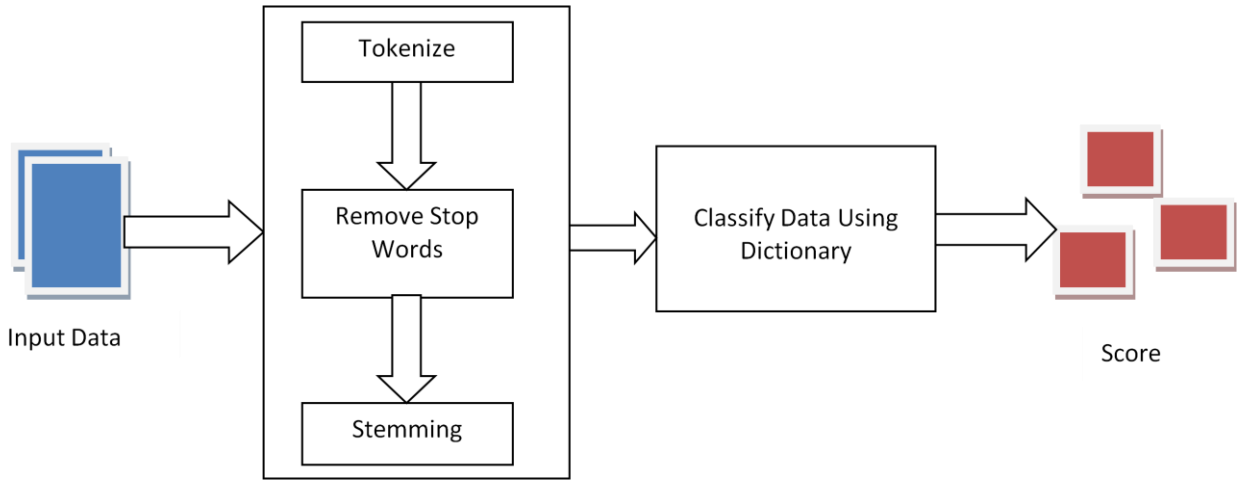

Figure 4. Architecture of proposed work

The input data consists of a set of posts. Each post is tokenized, separating the words using spaces as delimiters. Next, we remove stop words, which are predefined in NLTK. The resulting set of tokens contains various forms of words, which are then unified by stemming. The sentiment extraction process from the posts follows this approach using the dictionary.

1. Start downloading and caching the sentiment dictionary
2. Download twitter testing data sets and input it into the program

   Input: It is beautiful day to go for _shing

3. Tokenize each word in the data set and feed in to the program

   ['it', 'is', 'beautiful', 'day', 'to', 'go', 'for', '_shing']

4. Clean the tweets by removing the stop words

   ['beautiful', 'day', 'go', '_shing']

5. The multiple forms of each word are counted as one word using stemming.

   ['beautiful', 'day', 'go', '_shing']

   ['beauti', 'day', 'go', '_sh']

6. Now, for each word, compare it with positive and negative sentiments word in the dictionary. If matches, then increment positive count or negative count.

D. Data Validation

The Data validation is a critical phase in any data science project, but validating 393,869 records is a significant challenge. Manually verifying all the posts is impractical, so we selected a benchmark dataset of 765 posts related to weather data for our project. These posts were manually categorized into sentiments score and the results were verified using both TextBlob and our SocialAnalyzer.

For validation, we compared the precision of TextBlob with that of SocialAnalyzer.

Although numerous sentiment analysis tools are available, we selected TextBlob for validation after conducting a pilot study. In this study, we randomly chose ten posts and evaluated them using three tools: TextBlob and SocialAnalyzer. Among these, SocialAnalyzer provided the most accurate results for all ten posts. The outcomes of this evaluation are presented in Table 1.

Table 1. Comparisons of proposed and previous method

Text	SocialAnalyzer	TextBlob
love this sandwich	3	2
This is an amazing place	4	1
I feel very good about these	3	3

beers		
This is my best work	3	2
What an awesome view?	4	3
Tomorrow is Wednesday	0	0
I am tired of this stuff	2	0
I can't deal with this	0	1
He is my sworn enemy!	2	1
I am here	0	1

IV. RESULTS AND ANALYSIS

The primary objective of our research is to develop SocialAnalyzer, a tool for analyzing social media data through sentiment analysis. We specifically focused on posts related to the sentiment of user.

We evaluated the precision of both SocialAnalyzer and TextBlob using the weather dataset. Precision, which measures the accuracy of positive predictions, is calculated using the formula:

$$Precision = \frac{\text{TruePositive}}{(\text{TruePositive} + \text{FalseP ositive})} \qquad 1$$

Since our dataset includes three categories—positive, negative and neutral. we computed precision under two scenarios:

1. Treating neutral posts as positive.
2. Treating neutral posts as negative.

The precision values for SocialAnalyzer and TextBlob under these conditions are presented in Table 2. When treating neutral posts as positive, SocialAnalyzer achieved a precision of 73.72%, while TextBlob achieved 70.74%.

Table 2. Precision Values

Tools	TP	TF	Precision
SocialAnalyzer	449	160	0.7372
TextBLob	428	117	0.7074

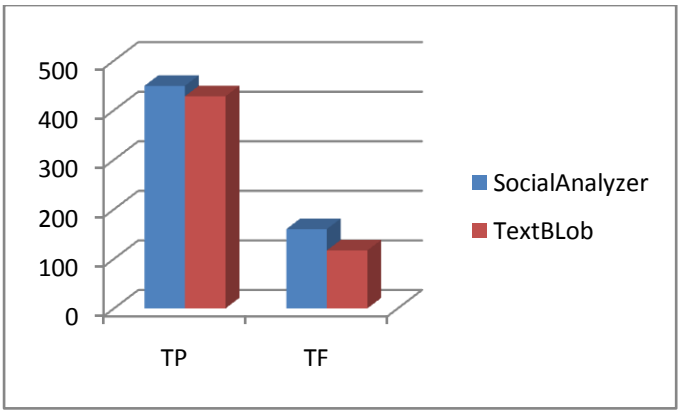

Figure 3. Result Comparison Graph

V. CONCLUSION

As detailed in the manuscript, we analyzed posts containing the sentiment. The results indicate that the sentiment analysis of posts uing proposed method is efficient compare to old method. In this paper we show that we achieve our goal.

We compared the performance of our SocialAnalyzer with TextBlob. When treating neutral posts as positive, the precision values for SocialAnalyzer and TextBlob were 73.72% and 70.74%, respectively, demonstrating that SocialAnalyzer achieves results comparable to TextBlob.

However, our study has certain limitations. We were unable to extract demographic information (such as age, geographic location, and gender) of users from the posts. Additionally, posts collection was restricted to one-week intervals, making it labor-intensive and costly to gather data for every week within the study period.

REFERENCES

1. M. Hoffman, D. Steinley, K. M. Gates, M. J. Prinstein, and M. J. Brusco, “Detecting clusters/communities in social networks,” Multivariate Behav. Res., vol. 53, no. 1, pp. 57–73, 2018, doi: 10.1080/00273171.2017.1391682.

2. J. Leskovec, “Social media analytics: Tracking, modeling and predicting the flow of information through networks,” in Proc. 20th Int. Conf. Companion World Wide Web, Mar. 2011, pp. 277–278.
3. Joshi, Prasad A., Varsha M. Pathak, and Manish R. Joshi. "Sentiment Analysis from Social Media Data in Code-Mixed Indian Languages Using Machine Learning Classifiers with TF-IDF and Weighted Word Features." In International Conference on Data Science and Big Data Analysis, pp. 203-222. Springer, Singapore, 2024.
4. Molenaar, Annika, Dickson Lukose, Linda Brennan, Eva L. Jenkins, and Tracy A. McCaffrey. "Using Natural Language Processing to Explore Social Media Opinions on Food Security: Sentiment Analysis and Topic Modeling Study." Journal of Medical Internet Research 26 (2024): e47826.
5. He, Lu, Samaneh Omranian, Susan McRoy, and Kai Zheng. "Using Large Language Models for sentiment analysis of health-related social media data: empirical evaluation and practical tips." medRxiv (2024): 2024-03.
6. Almuhaya, Basheer, Bishal Saha, Manbir Kaur, Mahmood A. Bazel, and Rehab Mohammed. "Comparative Analysis of Machine Learning Algorithms for Arabic Sentiment Analysis on Imbalanced Social Media Data." In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), pp. 1362-1367. IEEE, 2024.
7. Paulraj, D., P. Ezhumalai, and Mohan Prakash. "A Deep Learning Modified Neural Network (DLMNN) based proficient sentiment analysis technique on Twitter data."Journal of Experimental & Theoretical Artificial Intelligence 36, no. 3 (2024): 415-434.
8. Dean, Matthew C., and Ben Porter. "Sentiment Analysis of Russian-Language Social Media Posts Discussing the 2022 Russian Invasion of Ukraine." Armed Forces & Society (2024): 0095327X241235987.
9. Srivastava, Shobhit, Mrinal Kanti Sarkar, and Chinmay Chakraborty. "Sentiment analysis of Twitter data using machine learning: COVID-19 perspective."International Journal of Data Analysis Techniques and Strategies 16, no. 1 (2024): 1-16.
10. Bukhari, Sarah, and Muhammad Ramzan. "Text mining on social media data: a systematic literature review." International Journal of Data Analysis Techniques and Strategies 16, no. 1 (2024): 82-104.
11. Xinwei, Li, Ying Kei Tse, and Fernando Fastoso. "Unleashing the power of social media data in business decision making: an exploratory study." Enterprise Information Systems 18, no. 1 (2024): 2243603.
12. Poudel, Amrit, and Tim Weninger. "Navigating the Post-API Dilemma Search Engine Results Pages Present a Biased View of Social Media Data." arXiv preprint arXiv:2401.15479 (2024).
13. Gothane, Suwarna, G. Vinoda Reddy, K. Praveen Kumar, D. Baswaraj, Gumma Parvathi Devi, Sruthi Thanugundala, and Ravindra Changala. "Sentiment Analysis in Social Media Using Deep Learning Techniques."
14. Tejaswini, Vankayala, Korra Sathya Babu, and Bibhudatta Sahoo. "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model." ACM Transactions on Asian and Low-Resource Language Information Processing 23, no. 1 (2024): 1-20.
15. Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1528{1531. ACM, 2012.
16. Bryan Pratama, Dedi Dwi Saputra, Deny Novianti, Endah Putri Purnamasari, Antonius Yadi Kuntoro, Windu Gata, Nia K Wardhani, Sfenrianto Sfenrianto, Sularso Budilaksono, et al. Sen- timent analysis of the indonesian police mobile brigade corps based on twitter posts

using the svm and nb methods. In Journal of Physics: Conference Series, volume 1201, page 012038. IOP Publishing, 019.