Optimized Sentiment Analysis of Twitter Reviews Using GOSM and Deep Sentiment Metrics

Jayagopi G[1], Sumathi C B[2]
[1]*Department of Computer Science and Engineering, Mother Theresa Institute of Engineering and Technology, Palamanar, India*
[2]*PG and Research Department of Mathematics, Marudhar Kesari Jain College for Women, Tamil Nadu, India*

Abstract: Sentiment analysis involves identifying and classifying opinions or sentiments expressed within textual data. Social media platforms, especially Twitter, serve as significant sources of sentiment-rich content through tweets, status updates, and blog posts. Extracting meaningful insights from such data can help understand public opinion trends. However, sentiment analysis on Twitter presents unique challenges due to informal language, frequent use of slang, misspellings, and the platform's 140-character limit. Two widely adopted approaches in sentiment analysis are the knowledge-based and machine learning methods. This study focuses on analysing Twitter posts related to electronic products such as mobile phones and laptops using a machine learning approach. Specifically, it utilizes the Grid Optimized Search Machine (GOSM) algorithm to assess sentiments within tweets. To enhance the evaluation, a deep sentiment difference metric is introduced to measure the impact of reviews more effectively. The proposed system demonstrates promising performance, achieving an accuracy of 85%.

Keywords— Sentiment Analysis, Machine Learning, Opinion Mining, Neural Networks, Optimization Algorithms.

I. INTRODUCTION

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence that focuses on enabling effective communication between computers and human languages. It involves the design and development of algorithms and computational models that can analyse, interpret, and generate natural language in both text and speech formats.

Some of the applications where NLP is used include Sentiment Analysis: Analysing the sentiment behind text or speech to determine whether it expresses a positive, negative, or neutral opinion. *Example:* Understanding customer satisfaction from product reviews. Machine Translation: Automatically translating text or spoken language from one language to another. *Example:* Google Translate converting English text into Spanish. Named Entity Recognition (NER): Identifying and classifying key entities in text, such as names of people, places, dates, and organizations. *Example:* Extracting "Barack Obama," "White House," or "Microsoft" from a news article. Text Summarization: Producing a concise summary of a larger text document while preserving its key points. *Example:* Summarizing news articles or research papers. Chatbots and Conversational Agents: Creating systems that can understand and respond to human language in a natural way. *Example:* Customer support bots on e-commerce websites. Question Answering: Answering questions posed in natural language by extracting or generating answers based on text data. *Example:* A virtual assistant answering, "Who was the first president of the United States?"

Some of the most important techniques used in Natural Language Processing (NLP) are:

Tokenization: Breaking down text into individual words or tokens. Part-of-Speech Tagging: Assigning each token a grammatical category, such as noun, verb, or adjective. Dependency Parsing: Analysing sentence structure to identify how words relate to each other. Named Entity Recognition (NER): Detecting and extracting names of people, places, organizations, and other entities from text. Sentiment Analysis: Determining the emotional tone of text using machine learning or rule-based approaches. Language Modelling: Creating text that resembles a given input by using statistical or neural network models.

Natural Language Processing (NLP) supports various sectors including healthcare, finance, education, and social media analysis. Recent advances in deep learning have fuelled the development of sophisticated NLP models like transformer-based architectures—BERT, GPT-2,

and GPT-3—which greatly improve performance across a wide range of NLP tasks.

- This system has been developed as a social initiative to raise public awareness and provide users with a self-assessment tool.
- To gather information about users' symptoms, concerns, and timelines related to their health, a software-integrated survey is used. The survey questions are linked to a backend database of frequently asked questions, while the Opinion Lexicon model analyses user responses.
- The interactive interface collects key data on persistent symptoms and allows users to update their inputs, which are then converted into numerical data for backend analysis.

II. BACKGROUND STUDY

A substantial number of studies have focused on developing NLP techniques for real-time applications in the healthcare domain. Below are some significant contributions:

[7] This study employed six algorithms, including N-grams and TF-IDF (Term Frequency-Inverse Document Frequency), to categorize sentiments using the SS-Tweet dataset. The authors found that machine learning algorithms are highly effective, with logistic regression combined with TF-IDF features outperforming other methods across multiple evaluation metrics.

[8], [9] These papers conducted detailed research on applying NLP methods to various Twitter datasets. However, their work was limited to the English language, which restricts access to Twitter's rich multilingual data, as the platform supports 35 languages. This limitation highlights the need for developing NLP approaches that encompass other languages to fully utilize Twitter's data.

[10] The authors introduced three levels of feature extraction and applied classifiers such as SVM, J48, and Naïve Bayes. Among these, SVM achieved the highest precision, while K-Nearest Neighbor (KNN with k=10) demonstrated superior recall on the evaluated datasets.

[11] Sohan et al. presented a comprehensive survey of datasets relevant to COVID-19 research. Their work compiles various resources including pathological reports, radiographic images, final test results, and patient summaries, providing valuable datasets for academic and clinical research.

[12] Barbara Calabrese et al. emphasized the importance of social media as a data source. Their review discusses methods for data extraction, feature extraction, normalization, and modeling for contextual and emotion detection. Additionally, they identify future research directions such as behavioral analysis and computational tool development.

[13] Chiara Zucco et al. provided a detailed review of different NLP techniques for sentiment analysis. They categorized methods such as lexicon-based approaches and bag-of-words models, highlighting the effectiveness of these explainable methods in sentiment classification tasks.

[14] This study focused on sentiment analysis in the medical domain, specifically analysing drug reviews. The authors employed TF-IDF for feature extraction along with Fast Text embeddings and linguistic preprocessing. Their approach aimed to capture the contextual meaning of reviews and demonstrated superior performance compared to baseline models.

[15] Yue Han et al. introduced the "Sent Drugs" dataset, designed for sentiment extraction from drug reviews. They proposed a BiGRU model that leverages pretrained embeddings to capture semantic context bidirectionally, achieving improved sentiment analysis performance. The paper also includes several benchmark datasets for evaluation.

[16] Hasan et al. utilized character n-grams and applied a Linear Support Vector Machine (Linear SVM) on Twitter data. Their study demonstrated the effectiveness of n-gram features in enhancing text classification accuracy.

III. SYSTEM DESIGN

The system utilizes a pre-trained model built using the PANACEA Twitter dataset, which is specifically designed for sentiment analysis of Twitter conversations during the COVID-19 period. Multiple feature extraction techniques and NLP methods were applied during training to develop an effective sentiment classification model. This pre-

trained model is incorporated into a processing pipeline that categorizes incoming data. For evaluation purposes, medical test data and current patient symptoms, collected through surveys, are combined and used as input. The input data is pre-processed using the same techniques employed during training, including lemmatization and other feature extraction methods. The model then analyses the pre-processed data to determine sentiment. Through iterative training and fine-tuning, the system achieves improved accuracy. The resulting outputs provide final sentiment assessments along with actionable insights and recommendations.

IV. METHODOLOGY

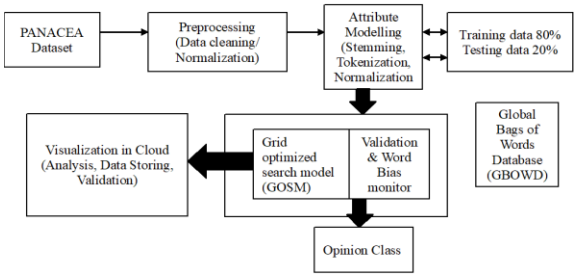

Fig 1. system architecture of proposed opinion mining framework.

Data Pre-processing

Data pre-processing is a foundational and essential step in building accurate and objective models. It involves a set of techniques aimed at cleaning, transforming, and standardizing raw textual data to enhance its quality and usability for analysis. Effective pre-processing ensures that the data used in modelling is free from bias and inconsistencies, allowing the system to extract meaningful patterns and make reliable predictions.

Tokenization

Tokenization is the initial step in processing text, where long sentences or texts are broken down into smaller, manageable units called *tokens*. These tokens can range from individual words to phrases or characters. This process enables detailed lexical and semantic analysis by algorithms, forming the basis for further natural language understanding.

Normalization

Normalization is the process of standardizing textual data to ensure consistency across the dataset. It typically includes:

- Converting all text to lowercase
- Removing punctuation
- Replacing numeric values with their word equivalents

These transformations help ensure that the text is uniformly processed, preventing case sensitivity or special characters from skewing the analysis.

Stemming

Stemming reduces words to their root forms by removing suffixes. For example:

- "Running" and "ran" become "run"
- "Achieved" becomes "achieve"

This technique minimizes the diversity of word forms, which simplifies the data and improves the model's focus on core semantics.

Lemmatization

Lemmatization, like stemming, converts words to their base forms but does so using a vocabulary and grammatical context. Unlike stemming, it considers the part of speech and ensures that the resulting lemma is a valid word. For example:

- "Better" is lemmatized too "good"
- "Running" to "run" (when used as a verb)

According to the Stanford NLP Group, lemmatization aims to remove inflectional endings while returning the dictionary form of a word. It enhances the contextual understanding of language elements, particularly in distinguishing between nouns, verbs, and other parts of speech.

Additional Preprocessing Techniques

Several supplementary techniques are also employed to further clean and optimize the text data:

- Lowercasing: Ensures uniformity by converting all characters to lowercase.
- Punctuation Removal: Eliminates unnecessary symbols that do not add semantic value.
- Stop Word Removal: Filters out frequently used words (e.g., "the", "is", "and") that contribute little to the meaning.
- Filler Word Removal: Removes non-essential verbal fillers like "uh" or "um" to enhance clarity.

Noise Removal

Noise refers to irrelevant or extraneous content within the raw data, such as special characters, irregular formatting, or web links. Techniques like regular expressions (regex) are employed to detect and eliminate these patterns. This step is critical in

ensuring that the dataset used for training is clean and relevant.

Feature Extraction and Modelling

N-gram Approach

The n-gram model is a widely used method for extracting features in text analysis. An n-gram is a continuous sequence of *n* items (words or characters) from a text. For instance:

For the phrase "The coronavirus pandemic":

- Unigrams (n=1): ["The", "coronavirus", "pandemic"]
- Bigrams (n=2): ["The coronavirus", "coronavirus pandemic"]

  This method is both flexible and efficient. It requires less memory, automatically captures stemming effects, and is effective in handling variations in text structure, making it suitable for classification tasks.

Opinion Mining

The core objective of the proposed system is to implement a robust framework for opinion mining. Leveraging the PANACEA database and a standardized dataset, the system extracts meaningful sentiments and opinions through a lexicon-based model. This model analyses textual input in relation to pre-defined sentiment-rich word lists. Opinion mining involves examining the relationships between terms, sentiment polarity, and conversational patterns to derive insights. It identifies subjective information and generates actionable suggestions by analysing similarities across conversations. This approach supports decision-making by recognizing trends in user-generated content.

V. RESULTS AND DISCUSSIONS

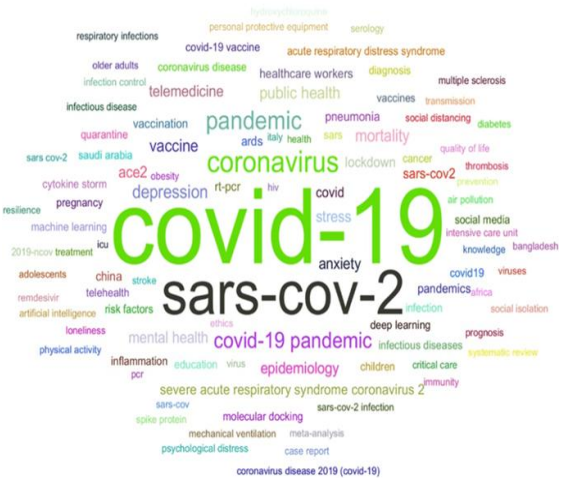

Fig. 2. Word cloud of input Dataset

Opinion mining using GOSM

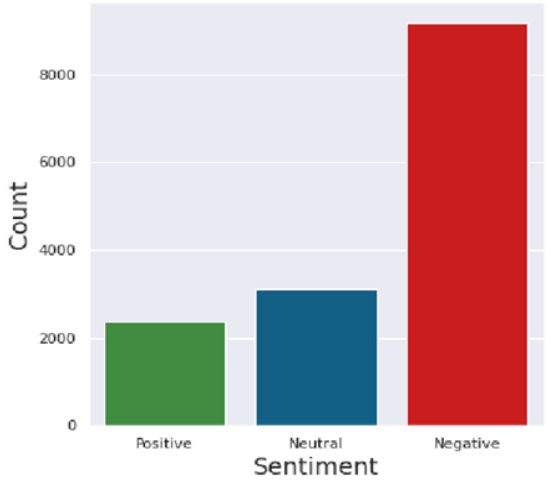

Figure 3 illustrates the sentiment analysis results generated using the GOSM (Global Opinion Sentiment Modelling) technique. To enhance the accuracy of COVID-19 sentiment analysis, the model has been further developed using an improved algorithm based on logistic regression. Moreover, the process of recommending suitable medications relies on continually updated information from global healthcare authorities.

VI. CONCLUSION

Sentiment analysis involves detecting and categorizing opinions or emotions expressed in textual data. Social media platforms, such as Twitter, provide a vast amount of sentiment-rich content in the form of tweets, status updates, and blog posts. Analysing this user-generated content enables the extraction of meaningful insights into public opinion. However, sentiment analysis on Twitter is particularly challenging due to the informal language, use of slang, frequent misspellings, and the character limit of 140 per tweet. The proposed system addresses these challenges by employing an opinion mining framework based on a Grid Optimized Search Model (GOSM). This system classifies the sentiment expressed in the text and determines opinion bias with an achieved accuracy of 85%. To further enhance the system's performance and improve result optimization, future work will focus on integrating more advanced hybrid techniques.

REFERENCE

[1] Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2020). COVID-19 epidemic analysis using

machine learning and deep learning algorithms. *medRxiv*. https://doi.org/10.1101/2020.04.08.20057679

[2] Horry, M. J., et al. (2020). COVID-19 detection through transfer learning using multimodal imaging data. *IEEE Access*, 8, 149808–149824. https://doi.org/10.1109/ACCESS.2020.3016780

[3] Wang, S., et al. (2020). A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *European Respiratory Journal*.https://doi.org/10.1183/13993003.00775-2020

[4] Zhang, H.-T., Zhang, J.-S., Zhang, H.-H., et al. (2020). Automated detection and quantification of COVID-19 pneumonia: CT imaging analysis by a deep learning-based software. *European Journal of Nuclear Medicine and Molecular Imaging*, 47. https://doi.org/10.1007/s00259-020-04953-1

[5] Samson, P., Navale, G., & Dharne, M. (2020). Biosensors: Frontiers in rapid detection of COVID-19. *3 Biotech*, 10. https://doi.org/10.1007/s13205-020-02369-0

[6] Zebin, T., & Rezvy, S. (2020). COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization.

[7] Kumar, R., Nagpal, S., Kaushik, S., et al. (2020). COVID-19 diagnostic approaches: Different roads to the same destination. *VirusDisease*, 31, 97–105. https://doi.org/10.1007/s13337-020-00599-7

[8] Ghafoor, K. (2020). COVID-19 pneumonia level detection using deep learning algorithm. *TechRxiv Preprint*. https://doi.org/10.36227/techrxiv.12619193.v1

[9] Ahuja, S., Panigrahi, B. K., Dey, N., et al. (2020). Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices. *Applied Intelligence*. https://doi.org/10.1007/s10489-020-01826-w

[10] Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. *Journal of Medical Systems*, 44(8), 135. https://doi.org/10.1007/s10916-020-01597-4

[11] Yoo, S. H., Geng, H., Chiu, T. L., et al. (2020). Deep learning-based decision tree classifier for COVID-19 diagnosis from chest X-ray Imaging.DOI: 10.1183/13993003.00775-2020

[12] Alazab, M., Awajan, A., Mesleh, A., Abraham, A., Jatana, V., & Alhyari, S. (2020). COVID-19 prediction and detection using deep learning. *International Journal of Computer Information Systems and Industrial Management Applications*, 12, 168–181.

[13] Narinder Singh Punn, Sanjay Kumar Sonbhadra, Sonali Agarwal COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms, medRxiv 2020.04.08.20057679; doi:https://doi.org/10.1101/2020.04.08.20057679

[14] Elaziz, M. A., Hosny, K. M., Salah, A Darwish, M. M., Lu, S., & Sahlol, A. T. (2020). New machine learning method for image-based diagnosis of COVID-19. *PLOS ONE*, 15(6), e0235187. https://doi.org/10.1371/journal.pone.0235187

[15] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., et al. (2020). Machine learning-based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*, 12, 731–739. https://doi.org/10.1007/s41870-020-00495-9

[16] Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Acharya, U. R. (2020). Automated detection of COVID-19 cases Using deep neural networks with X-ray images. *Computers in Biology and Medicine*, 121,103792. https: //doi.org/10.1016/j.compbiomed. 2020.103792