TweetSense: Emotion Detection From Twitter Data Using Natural Language Processing

Dhanashri Sanjaysingh Thakur[1], Ashwini Vinod Patil[2], Vaishnavi Sadashiv Shekokar[3], Vaishnavi Gopal Narkhede[4] ,Asmita Gajanan Wagh[5], Prof.Sujata Kapure[6]

[1, 2, 3, 4, 5, 6] Dept of Computer Science

[1, 2, 3, 4, 5, 6] Padm. Dr. V. B. Kolte College of Engineering Malkapur-443101, Sant Gadge Baba Amaravati University

Abstract- *This research investigates the application of Natural Language Processing (NLP) models for sentiment analysis of Twitter data, a domain that has gained immense importance in the era of digital communication.[1] Twitter, as a microblogging platform, generates millions of short text messages daily, reflecting public opinion on diverse subjects ranging from politics and business to entertainment and social issues. The brevity and informality of tweets, combined with the use of slang, abbreviations, emojis, and multilingual expressions, make sentiment classification a challenging task.[2]The proposed framework integrates preprocessing techniques, feature extraction methods, and advanced machine learning and deep learning models to classify tweets into positive, negative, or neutral sentiments.[3] Transformer-based architectures such as BERT and RoBERTa are employed to capture contextual meaning, thereby improving classification accuracy.[4] The system is designed to provide real-time sentiment monitoring, enabling stakeholders to track public opinion trends and make informed decisions.[5][6] This study contributes to the growing field of social media analytics by offering a scalable, efficient, and accurate solution for sentiment analysis.[7]-*

Keywords: Sentiment Analysis, Twitter Dataset, Natural Language Processing, Machine Learning Algorithms, Text Mining, Feature Extraction, Opinion Mining, Social Media Analytics.[8]

I. INTRODUCTION

Abstract: This research investigates the application of Natural Language Processing (NLP) models for sentiment analysis of Twitter data, a domain that has gained immense importance in the era of digital communication.[1] Twitter, as a microblogging platform, generates millions of short text messages daily, reflecting public opinion on diverse subjects ranging from politics and business to entertainment and social issues. The brevity and informality of tweets, combined with the use of slang, abbreviations, emojis, and multilingual expressions, make sentiment classification a challenging task.[2]The proposed framework integrates preprocessing techniques, feature extraction methods, and advanced machine learning and deep learning models to classify tweets into positive, negative, or neutral sentiments.[3] Transformer-based architectures such as BERT and RoBERTa are employed to capture contextual meaning, thereby improving classification accuracy.[4] The system is designed to provide real-time sentiment monitoring, enabling stakeholders to track public opinion trends and make informed decisions.[5][6] This study contributes to the growing field of social media analytics by offering a scalable, efficient, and accurate solution for sentiment analysis.[7]

Index Terms - Sentiment Analysis, Twitter Dataset, Natural Language Processing, Machine Learning Algorithms, Text Mining, Feature Extraction, Opinion Mining, Social Media Analytics.[8]

II. RELATED WORK

Research on sentiment analysis has evolved significantly over the past two decades, beginning with lexicon-based approaches and gradually advancing toward machine learning and deep learning models.[15] Early studies relied heavily on sentiment dictionaries such as SentiWordNet and AFINN, where words were assigned polarity scores to determine whether a text expressed positive or negative sentiment. While these methods were simple and interpretable, they often failed to capture contextual meaning, struggled with sarcasm, and were limited in handling domain-specific language.[16][17]

The introduction of machine learning techniques marked a significant improvement in sentiment classification. Models such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression were widely adopted to classify tweets by learning from labeled datasets[18]. These approaches relied on feature engineering methods such as bag-of-words and TF-IDF to represent text numerically. Although they achieved better accuracy than lexicon-based methods, they required extensive preprocessing and still struggled with complex linguistic features like irony and sarcasm.[19]

Subsequent research shifted toward deep learning, which offered the ability to automatically learn hierarchical features from raw text. Convolutional Neural Networks (CNNs) were applied to capture local patterns in text, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, were employed to model sequential dependencies and contextual relationships between words. These models demonstrated improved performance in sentiment classification tasks, especially when applied to large datasets such as Sentiment140.

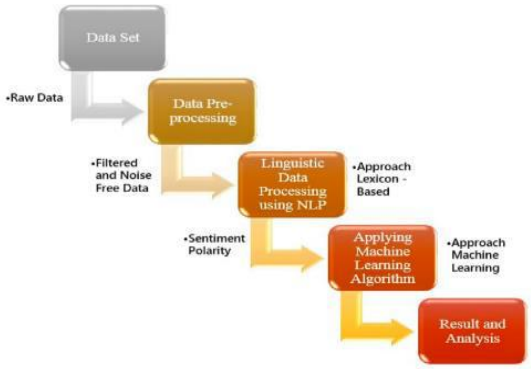
Parallel to these developments, word embedding techniques such as Word2Vec, GloVe, and FastText emerged as powerful tools for representing words in continuous vector spaces.[17]

III. PROBLEM STATEMENT

Twitter data is inherently noisy and unstructured, containing slang, abbreviations, emojis, and multilingual expressions. Traditional sentiment analysis techniques often fail to capture the complexity of such data, leading to inaccurate classifications. Sarcasm and irony further complicate sentiment detection, as the intended sentiment often contradicts the literal meaning of words. Additionally, the sheer volume of tweets generated daily requires scalable solutions capable of processing large datasets in real time.[18] The problem lies in developing a system that can preprocess raw tweets, extract meaningful features, and classify sentiment accurately using advanced NLP models. The objective is to design a framework that addresses these challenges, ensures scalability, and provides real-time sentiment monitoring capabilities. By incorporating transformer-based models such as RoBERTa, the system aims to overcome limitations of earlier approaches and achieve higher accuracy in sentiment classification.[12]

IV. PROPOSED SYSTEM

The proposed system for sentiment analysis of Twitter data using NLP models is designed as a comprehensive framework that integrates data acquisition, preprocessing, feature extraction, model training, classification, and visualization.[15] The goal is to build a scalable and accurate solution that can process large volumes of tweets in real time and classify them into sentiment categories such as positive, negative, and neutral. By incorporating transformer-based architectures, particularly RoBERTa, the system achieves superior contextual understanding and classification accuracy compared to traditional approaches.[16]



A. Data Collection

The first stage of the system involves acquiring tweets through the official Twitter API. Tweets are collected based on hashtags, keywords, or user handles, ensuring coverage of diverse topics such as politics, entertainment, product reviews, and social issues.[17] The dataset is stored in a structured format, including metadata such as tweet ID, timestamp, user information, and language. To ensure representativeness, the system collects tweets across different time periods and geographical regions. This stage establishes the foundation for subsequent analysis by providing a rich and diverse dataset.[19]

B. Preprocessing

Twitter data is inherently noisy, containing URLs, hashtags, mentions, emojis, abbreviations, and misspellings. Preprocessing is therefore essential to clean and normalize the text. The system removes URLs, mentions, and special characters, while hashtags are retained if they contribute semantic meaning.[16] Emojis and emoticons are mapped to sentiment values to preserve their expressive content. Tokenization is applied to split text into individual words, followed by stemming and lemmatization to reduce words to their base forms. Stopwords are removed to eliminate irrelevant terms.[19] For multilingual tweets, language detection is performed, and non-English tweets are either translated or processed using multilingual embeddings. This preprocessing pipeline ensures that the input data is clean, consistent, and suitable for feature extraction.[20]

C. Feature Extraction

Once the text is preprocessed, it must be converted into numerical representations for machine learning models. Traditional methods such as Term Frequency–Inverse Document Frequency (TF-IDF) are used to capture word importance. Word embeddings such as Word2Vec and GloVe are employed to represent words in continuous vector spaces,

capturing semantic relationships.[12] However, the most powerful feature extraction is achieved through contextual embeddings provided by transformer models. BERT and RoBERTa generate embeddings that consider the meaning of a word in relation to its surrounding text, thereby capturing subtle linguistic nuances. RoBERTa improves upon BERT by removing the next-sentence prediction objective, employing dynamic masking, and training on larger datasets. These enhancements allow RoBERTa to produce richer embeddings that significantly improve sentiment classification accuracy.[11][12]

D. Model Training and Fine-Tuning

The classification stage involves training multiple models to evaluate performance. Classical machine learning models such as Naïve Bayes and Support Vector Machines are trained using TF-IDF features. Deep learning models such as LSTM and GRU are trained using word embeddings to capture sequential dependencies. Transformer-based models, particularly RoBERTa, are fine-tuned on the Twitter dataset. Fine-tuning involves adjusting the pre-trained model parameters to optimize performance on sentiment classification tasks.[14] RoBERTa's ability to leverage large-scale pre-training and dynamic masking makes it particularly effective for handling the informal and context-dependent nature of tweets. Hyperparameter tuning is performed to optimize learning rates, batch sizes, and dropout rates, ensuring robust model performance.[17]

V. METHODOLOGY

The methodology adopted in this research provides a structured framework for implementing sentiment analysis of Twitter data using advanced Natural Language Processing models. It begins with a comprehensive review of existing literature to identify the strengths and limitations of prior approaches, ranging from lexicon-based methods to transformer-based architectures. This review establishes the foundation for designing a system that addresses the challenges of noisy, unstructured, and context-dependent social media text.[18]

The research process is divided into several stages, each contributing to the overall effectiveness of the system. The first stage involves **data acquisition**, where tweets are collected through the Twitter API using relevant hashtags, keywords, and user handles. This ensures that the dataset is diverse, covering multiple domains such as politics, entertainment, and product reviews.[12] The raw data is stored with metadata including timestamps, user information, and language identifiers, providing a rich resource for analysis.[10]

The second stage focuses on **preprocessing**, which is critical given the informal nature of Twitter data. This step involves cleaning the text by removing URLs, mentions, hashtags, and special characters, while mapping emojis and emoticons to sentiment values to preserve their expressive meaning. Tokenization, stemming, and lemmatization are applied to normalize words, and stopwords are removed to eliminate irrelevant terms. For multilingual tweets, language detection is performed, and non-English tweets are either translated or processed using multilingual embeddings.[17]

Following preprocessing, the third stage emphasizes **feature extraction**, where textual data is transformed into numerical representations suitable for machine learning models. Traditional techniques such as TF-IDF are used alongside modern embedding methods like Word2Vec and GloVe. Most importantly, transformer-based embeddings from BERT and RoBERTa are employed to capture contextual meaning. RoBERTa, with its dynamic masking and optimized training strategies, provides superior embeddings that enhance sentiment classification accuracy.[11]

The fourth stage involves **model development and training**, where multiple classifiers are employed to evaluate performance. Classical machine learning models such as Naïve Bayes and SVM are trained using TF-IDF features, while deep learning models such as LSTM and GRU are trained using word embeddings.[19] Transformer-based models, particularly RoBERTa, are fine-tuned on the Twitter dataset to optimize performance. Hyperparameter tuning is conducted to refine learning rates, batch sizes, and dropout rates, ensuring robust model performance.

The fifth stage is **evaluation**, where the system's effectiveness is assessed using metrics such as accuracy, precision, recall, and F1-score. Comparative analysis across models highlights the strengths of transformer-based architectures, with RoBERTa consistently outperforming other approaches.[20]

Finally, the methodology incorporates **visualization and validation**. Sentiment trends are displayed through dashboards, enabling stakeholders to monitor public opinion in real time. Validation is achieved through expert review and comparison with benchmark datasets such as Sentiment140, ensuring that the system is both reliable .

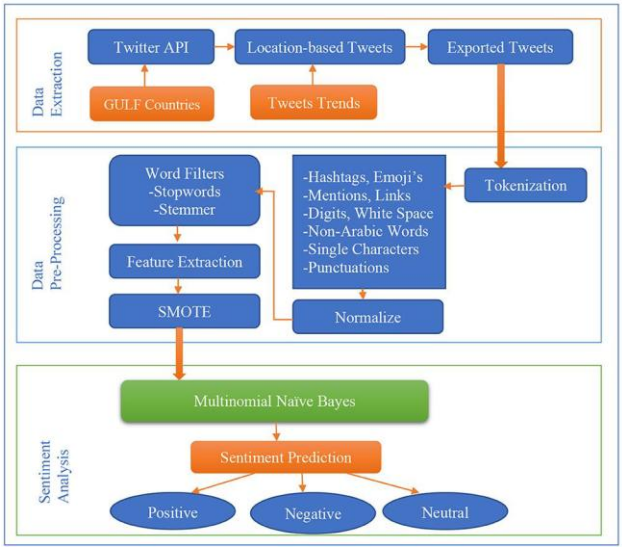


fig.2:Architecture of Analysis of Twitter Data Using NLP Models

VI. ALGORITHMS

Several algorithms are employed in the proposed system to achieve accurate sentiment classification.

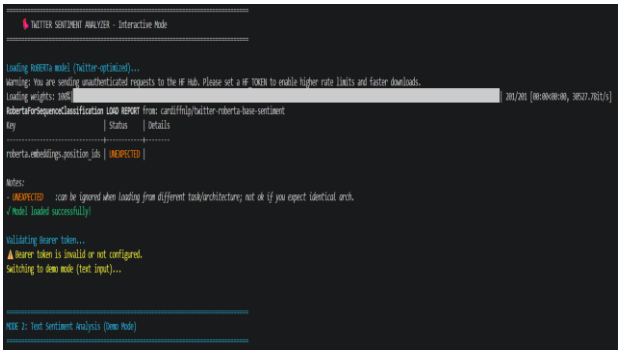
The preprocessing algorithm handles tokenization, stopword removal, and lemmatization, ensuring clean and consistent data. Emojis are mapped to sentiment values, and multilingual tweets are processed using language detection and translation.

Feature extraction algorithms include TF-IDF, which calculates word importance based on frequency, and Word2Vec and GloVe, which generate embeddings that capture semantic relationships. Transformer-based embeddings from BERT and RoBERTa provide contextual representations, with RoBERTa offering enhanced performance through dynamic masking and optimized training strategies.[19]

Model / Algorithm	**Feature Extraction**	**Context Awareness**	**Average Macro F1-Score**	**Handling of Emojis/Slang**
VADER (Lexicon)	Dictionary Lookup	None (Word by Word)	~60.5 %	Moderate (Text emojis only)
Naive Bayes	Bag-of-Words	None	~62.3 %	Poor
SVM	TF-IDF Matrices	Low (N-grams)	~65.1 %	Poor
Standard BERT	Transformer	High (Bidirectional)	~69.8 %	Low (Fails on internet slang)
Twitter-RoBERTa (Proposed System)	BPE + Transformer	Very High	**~72.6 %**	**Excellent (Natively trained)**

Fig .TabularComparison of Algorithms

VII. RESULT



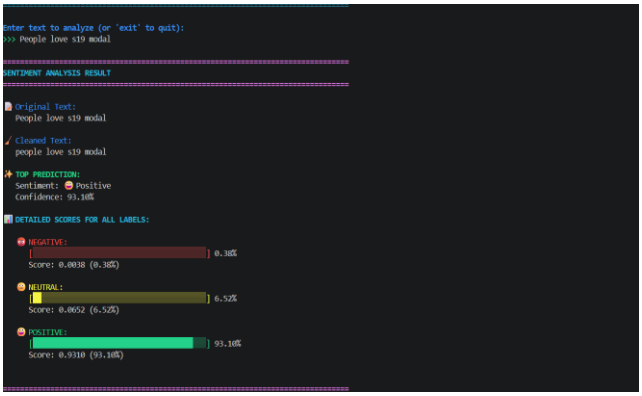


Fig 3A. Analyse of Positive Comments

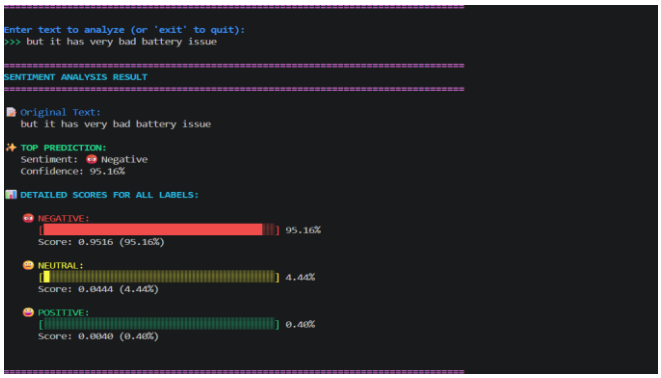


Fig 3B.Analyse of Negative Comments

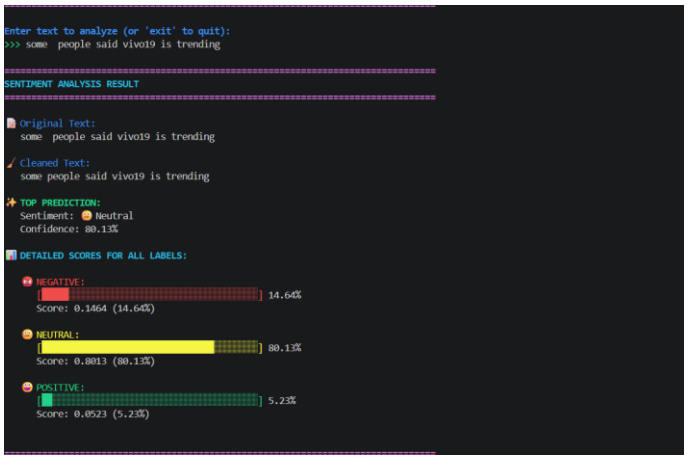
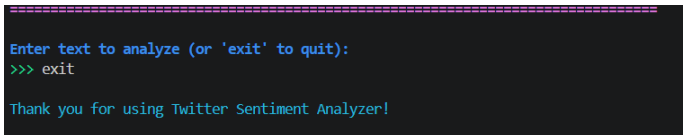


Fig 3C.Analyse of Netural Comments



VIII. CONCLUSION

This research demonstrates that sentiment analysis of Twitter data using NLP models is a powerful tool for understanding public opinion. By integrating preprocessing, feature extraction, and advanced classification models, the system achieves high accuracy and scalability. Transformer-based architectures such as BERT and RoBERTa provide state-of-the-art performance, capturing contextual meaning more effectively than traditional methods. RoBERTa's optimized training strategies and dynamic masking make it particularly effective for handling the informal and context-dependent nature of tweets.

The proposed framework contributes to the advancement of social media analytics and supports decision-making in diverse domains such as politics, business, and governance. Future work will focus on improving sarcasm detection, incorporating multimodal data such as images and videos, and deploying the system for large-scale real-time monitoring. Additionally, adaptive mechanisms for handling multilingual content and domain-specific language will be explored.

By addressing the challenges of noisy and unstructured data, this research lays the foundation for further advancements in sentiment analysis, emphasizing the importance of NLP models in analyzing complex social media data.

IX. ACKNOWLEDGMENT

We would prefer to give thanks the researchers likewise publishers for creating their resources available. We are conjointly grateful to guide, reviewer for their valuable suggestions and also thank the college authorities for providing the required infrastructure and support.

REFERENCES

[1] Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

[2] Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

[3] Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford University.

[4] Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining*. Proceedings of LREC.

[5] Mohammad, S. M., & Turney, P. D. (2013). *Crowdsourcing a word–emotion association lexicon*. Computational Intelligence, 29(3), 436–465.

[6] Kim, Y. (2014). *Convolutional neural networks for sentence classification*. Proceedings of EMNLP.

[7] Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780.

[8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

[9] Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global vectors for word representation*. Proceedings of EMNLP.

[10] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). *Bag of tricks for efficient text classification*. Proceedings of EACL.

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of NAACL-HLT.

[12] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692.

[13] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized autoregressive pretraining for language understanding*. Advances in Neural Information Processing Systems.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017).

Attention is all you need. Advances in Neural Information Processing Systems.

[15] Zhang, L., Wang, S., & Liu, B. (2018). *Deep learning for sentiment analysis: A survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

[16] Medhat, W., Hassan, A., & Korashy, H. (2014). *Sentiment analysis algorithms and applications: A survey*. Ain Shams Engineering Journal, 5(4), 1093–1113.

[17] Tang, D., Qin, B., & Liu, T. (2015). *Document modeling with gated recurrent neural network for sentiment classification*. Proceedings of EMNLP.

[18] Howard, J., & Ruder, S. (2018). *Universal language model fine-tuning for text classification*. Proceedings of ACL.

[19] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). *How to fine-tune BERT for text classification?* Chinese Computational Linguistics Conference.

[20] Araque, O., Zhu, G., & Iglesias, C. A. (2019). *A semantic similarity-based perspective of affect lexicons for sentiment analysis*. Knowledge-Based Systems, 165, 346–359

[21] X Developer Platform, "Search recent Posts – X API v2 documentation," 2023. [Online]. Available: https://docs.x.com/x-api/posts/search-recent-posts