MACHINE LEARNING AND NLP-BASED SENTIMENT ANALYSIS OF SOCIAL MEDIA POSTS: UNCOVERING PUBLIC EMOTIONS

Swetha D, III B. Sc. Computer Science (Data Science), Dept. Of Computer Science and Data Science, Nehru Arts and Science College, Coimbatore-641105.
Ms. S. Shanmugapriya, Asst. Prof., Dept. Of Computer Science and Data Science, Nehru Arts and Science College, Coimbatore-641105.

ABSTRACT:

Social media serves as a crucial platform for sharing opinions and feelings, making it essential for sentiment analysis[1]. This paper explores the use of machine learning (ML) and natural language processing (NLP) techniques to categorize sentiments in social media posts as positive, negative, or neutral. It merges lexicon-based approaches, such as VADER[2], with more sophisticated models like RoBERTa[3] to tackle issues such as sarcasm, ambiguity, and linguistic variety. The research compares traditional and contemporary methods, emphasizing the greater accuracy of transformer-based model[4]. Potential applications include monitoring public opinion, managing brands, and tracking sentiment in real-time[5]. By combining ML advancements with specialized knowledge, this study illustrates the capability of sentiment analysis to reveal public emotions and enhance social media analytics.

Key-words: Sentiment Analysis; Public Emotions; Deep Learning Models; RoBERTa; VADER.

INTRODUCTION:

Social media has changed how people interact, allowing for immediate sharing of thoughts, feelings, and ideas on a wide range of subjects[6]. Platforms such as Twitter, Facebook, and Instagram have become influential venues for public discussion, producing a large volume of text data every day[7]. This data provides valuable insights into public opinions and sentiments, making it useful for businesses, policymakers, and researchers[8]. Sentiment analysis, a branch of natural language processing (NLP), has become an essential tool for extracting and interpreting these sentiments, helping to understand the emotions present in social media content[9].

However, social media data poses significant challenges. The text is often noisy, unstructured, and reliant on context[10], incorporating slang, emojis, abbreviations, and various writing styles. Capturing the subtle nuances of emotions, particularly sarcasm, irony, and cultural differences, is a difficult task. To tackle these issues, this paper investigates a hybrid framework that combines traditional lexicon-based techniques, like VADER[2], with advanced machine learning (ML) models, such as the transformer-based RoBERTa[3]. These methods are evaluated for their ability to accurately classify sentiments as positive, negative, or neutral while addressing contextual and linguistic complexities.

The research also highlights the need for scalability and robustness in sentiment analysis applications. Comparative assessments reveal the benefits of deep learning methods over traditional approaches, especially in managing intricate language patterns. The results have practical implications for monitoring public opinion, managing brand reputation, and tracking sentiment in real-time during crises. By integrating domain knowledge with state-of-the-art ML models, this study contributes to the expanding field of social media analytics, providing avenues for a deeper understanding and actionable insights into global public emotions.

OBJECTIVES:

The objective is to utilize machine learning and natural language processing (NLP) methods for conducting sentiment analysis on social media posts, aiming to identify, measure, and comprehend public emotions. The study's goals include:

i. Model Development and Assessment: Create and assess strong machine learning models capable of accurately and efficiently classifying social media content by sentiment[11].

https://doi.org/10.36893/JK.2025.V15I2.042

ii. Public Sentiment Analysis: Employ NLP techniques to derive insights regarding public emotions, opinions, and trends from unstructured text data on platforms like Twitter, Facebook, and Instagram[12].

iii. Identifying Influential Factors: Determine significant topics, hashtags, or events that influence changes in public sentiment[13].

iv. Practical Applications: Investigate how sentiment analysis can be utilized in areas like marketing, politics, disaster management, and social behavior research to aid in decision-making and policy development[14].

v. Addressing Social Media Data Challenges: Tackle the intricacies of social media language, including slang, abbreviations, emojis, and multilingual content, to enhance the reliability and scalability of sentiment analysis systems[15].

LITERATURE REVIEW:

Sentiment analysis, also known as opinion mining, has been a prominent research focus for the past two decades. Early research primarily utilized lexicon-based methods, such as the VADER (Valence Aware Dictionary and Sentiment Reasoner) approach created by Hutto and Gilbert in 2014. VADER uses a rule-based framework to assign sentiment scores to words, making it particularly effective for analyzing social media content, which often features emojis, abbreviations, and slang.

The advent of deep learning revolutionized sentiment analysis with the introduction of models like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), which are adept at capturing long-term dependencies in text. Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), further advanced the field by offering a bidirectional understanding of context, as highlighted by Devlin et al. in 2019. RoBERTa, an enhanced version of BERT introduced by Liu et al. in 2019, builds on BERT's strengths by leveraging larger datasets and longer training times, thus improving its ability to grasp contextual nuances and implicit emotions.

Sentiment analysis has diverse applications across multiple sectors. In marketing, it allows brands to understand consumer perceptions of their products, while in politics, it aids in evaluating public sentiment toward policies and candidates. The evolution of sentiment analysis techniques, from lexicon-based methods to deep learning approaches, underscores the growing need for advanced models capable of addressing the intricacies of natural language.

METHODOLOGY:

Data Collection: Social media data was obtained using Twitter and Facebook APIs[8], as well as web scraping, targeting posts that included specific keywords, hashtags, and phrases relevant to the research topic. Filters were applied to ensure a diverse range of posts in various languages and from different geographical areas.

Data Preprocessing: The raw data underwent cleaning to eliminate stop words, irrelevant symbols, special characters, URLs, and other unnecessary elements. Tokenization broke the text into smaller units (tokens), while normalization standardized formats and converted text to lowercase for consistency[9]. Emoticons and emojis were assigned sentiment values to capture emotional expressions beyond the text.

Feature Extraction: The textual data was transformed into numerical formats using advanced feature extraction methods like TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings (such as GloVe and Word2Vec)[10]. These methods helped to capture semantic relationships between words and enhanced model performance by representing the text in a machine learning-friendly format.

Sentiment Analysis: Two approaches were utilized for sentiment analysis:
Lexicon-Based Method: VADER[2] (Valence Aware Dictionary and Sentiment Reasoner), a lexicon-based tool, was employed to classify sentiments using a predefined dictionary, allowing for quick categorization into positive, negative, or neutral sentiments.

Deep Learning Models: A transformer-based model, RoBERTa, was fine-tuned for sentiment classification tasks, effectively capturing deeper contextual nuances and demonstrating superior performance with complex sentence structures.

Model Training and Testing: The dataset was divided into training and testing sets to ensure that models were trained on a diverse array of data and evaluated on unseen data. Performance metrics such as accuracy, precision, recall, and F1-score[11] were utilized to measure the effectiveness of the sentiment analysis models. Cross-validation techniques were also implemented to enhance generalization and minimize overfitting.

Comparison and Evaluation: The traditional lexicon-based sentiment analysis (VADER) was compared with modern deep learning techniques (RoBERTa) to assess their effectiveness in real-world scenarios. Particular focus was given to the handling of sarcasm, irony, and ambiguous language, with deep learning models showing better performance in these areas[3].

Applications and Insights: The outcomes of the sentiment analysis were leveraged to generate actionable insights for businesses and organizations. These insights were applied in: Brand Reputation Management: Tracking public sentiment towards brands and identifying potential issues before they escalate.

Public Opinion Monitoring: Gaining insights into societal trends, political sentiment, and public reactions to significant events.

Crisis Management: Identifying emerging crises through sentiment analysis of social media posts, enabling timely intervention.

IMPLEMENTATION: TOOLS AND TECHNIQUES:

The project is named "Machine Learning and NLP-Based Sentiment Analysis of Social Media Posts: Uncovering Public Emotions" and utilizes HTML, CSS, and JavaScript for the front-end development, while Python is employed for the back end. The subsequent sections outline the tools and frameworks utilized in the implementation process.

Languages and Frameworks

- Front End: The interface is built using **HTML**, **CSS**, and **JavaScript**, which provide a user-friendly, responsive, and interactive design for users to input and visualize sentiment analysis results.
- Back End: The back end is powered by **Python**, leveraging its robust libraries and frameworks to process data, perform sentiment analysis, and return results efficiently.

TOOLS COMPARISON: VADER VS. ROBERTA :

VADER (Valence Aware Dictionary and sentiment Reasoner) is a sentiment analysis tool that relies on a predefined set of word sentiment scores to assess text. It is ideal for obtaining quick and easily interpretable results, making it a practical option for basic sentiment analysis tasks. Using straightforward rule-based algorithms, VADER produces sentiment scores for positive, negative, and neutral sentiments, offering a clear and comprehensible sentiment overview.

In contrast, RoBERTa (Robustly Optimized BERT Pretraining Approach) employs sophisticated deep learning methods to analyze text. It is trained on extensive datasets, which allows it to gain a contextual understanding of language. This ability enables RoBERTa to recognize subtle shifts in sentiment and identify implicit emotions, making it a robust tool for more intricate and context-dependent sentiment analysis tasks.

MODULES DESCRIPTION:

A. Data Collection and Preprocessing

The initial phase of any sentiment analysis project involves collecting a varied and thorough dataset. This process includes gathering text samples from various sources such as social media sites like Twitter, Facebook, and Instagram, as well as review sites, forums, and blogs. The dataset should encompass a wide range of sentiments across different subjects and contexts. After the data is collected, it goes through a preprocessing stage to ensure uniformity and eliminate any irrelevant elements that could distort the analysis. Important preprocessing steps include:

1. Removing Special Characters and Punctuation: Non-textual elements like emojis, hashtags, or extra symbols are eliminated to keep the focus on the main text.
2. Eliminating Stop words: Common words such as "the," "and," or "is" are discarded as they do not convey significant sentiment and can hinder the model's effectiveness[5].
3. Lowercasing: All text is converted to lowercase to standardize it and prevent the model from interpreting the same word differently due to case variations.
4. Tokenization: The text is divided into individual words or phrases (tokens) for analysis.

B. Sentiment Analysis using VADER

VADER (Valence Aware Dictionary and Sentiment Reasoner) is a sentiment analysis tool that operates on a rule-based system and is included in the NLTK (Natural Language Toolkit) library. This module utilizes VADER to evaluate each preprocessed text sample and determine sentiment scores. The procedure includes:

1. Utilizing VADER for Sentiment Scoring: VADER employs a set lexicon of sentiment-related words and applies specific rules to ascertain the sentiment of a sentence or text sample.
2. Assigning Sentiment Labels: VADER's sentiment analysis provides a compound score for each text sample, which serves as a normalized indicator of the overall sentiment. Sentiment labels are assigned based on the compound score:
   - Positive: Scores exceeding a certain threshold (e.g., > 0.05)
   - Negative: Scores falling below a certain threshold (e.g., < -0.05)
   - Neutral: Scores that are near zero (e.g., between -0.05 and 0.05)

C. Sentiment Analysis using RoBERTa

RoBERTa (Robustly Optimized BERT Pretraining Approach) is an advanced pre-trained deep learning model designed for natural language understanding. This module utilizes the Transformers library from Hugging Face to implement RoBERTa on text samples. RoBERTa thoroughly examines each text by taking into account the context and the relationships among words within a sentence. The procedure involves:

1. Setting Up Transformers: The required libraries are installed, and the pre-trained RoBERTa model is accessed using the pipeline feature from the Hugging Face library.
2. Sentiment Analysis: RoBERTa delivers sentiment predictions (positive, negative, or neutral) along with a confidence score for each text sample, providing more detailed insights into sentiment.

D. Comparison and Evaluation:

After analyzing the text samples with both VADER and RoBERTa, the results are compiled and compared. This comparison involves examining the distribution of sentiment labels produced by each method, particularly how they perform in various scenarios. The evaluation process consists of the following components:

1. Sentiment Label Distribution: By combining the predictions from both tools, we can assess the occurrence of positive, negative, and neutral sentiment labels assigned by VADER and RoBERTa.
2. Performance Metrics: The accuracy of sentiment predictions from both methods is assessed using standard machine learning metrics, including:
   - Precision: The ratio of true positive sentiment predictions to the total number of positive predictions made.

https://doi.org/10.36893/JK.2025.V15I2.042

- Recall: The ratio of true positive sentiment predictions to the total number of actual positive sentiment instances in the dataset.
- F1-score: The harmonic mean of precision and recall, offering a balanced measure for evaluating the effectiveness of sentiment analysis.

RESULTS AND DISCUSSION:

The analysis highlights several distinctions between VADER and RoBERTa, especially regarding their performance, accuracy, interpretability, and computational efficiency. The main findings are as follows:

Accuracy: RoBERTa surpasses VADER in performance, especially when dealing with more intricate and context-sensitive text. Its deep learning framework enables it to grasp nuances such as sarcasm, implicit meanings, and complex sentence structures, while VADER is constrained by its fixed lexicon and rule-based methodology.

Interpretability: VADER offers clear and easily understandable sentiment labels with straightforward explanations, whereas RoBERTa's outcomes are less transparent. The predictions made by the deep learning model are often difficult to interpret, necessitating a higher level of expertise to comprehend the decision-making process behind them.

Speed and Efficiency: VADER, as a lightweight rule-based tool, operates much more quickly and efficiently than RoBERTa. It can handle large datasets in real-time, making it ideal for situations where speed is crucial. In contrast, RoBERTa is resource-intensive, requiring significant processing power and time to produce predictions.

APPLICATIONS:

The sentiment analysis system created in this project has numerous applications across different fields:

Marketing: Companies can utilize sentiment analysis to assess public perception of their products, services, or brands. By examining customer reviews, social media interactions, and feedback, businesses can determine customer sentiment and make well-informed choices.

Politics: Policymakers and analysts can leverage sentiment analysis to assess public feelings about policies, political figures, or major events. By grasping public opinion, politicians can tailor their messaging and strategies accordingly.

Customer Insights: Organizations can choose the sentiment analysis tool that best fits their unique requirements. For instance, VADER may be preferred for quick, real-time analysis of social media data, while RoBERTa might be selected for more in-depth and accurate insights into complex customer feedback.

CONCLUSION:

This project showcases the effectiveness of merging machine learning and natural language processing methods to analyze public sentiment. By evaluating VADER and RoBERTa, businesses and researchers can select the most appropriate tool for their particular needs. VADER is quick and easy to use, making it ideal for simple tasks, whereas RoBERTa offers more sophisticated features for examining intricate and subtle sentiments. Future efforts will aim to incorporate multimodal data, including images and videos, to gain a deeper insight into public feelings and viewpoints. This development will improve sentiment analysis capabilities and broaden its use in various sectors.

REFERENCES:

[1] Asghar, M. Z., Habib, A., Habib, A., & Khan, A. (2019). Lexicon-enhanced sentiment analysis framework using hybrid approaches.

[2] Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 5: Discovering conceptual primitives for sentiment analysis.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[4] He, S., Lin, C., & Al, H. (2020). Sentiment analysis in marketing: Applications and trends.

[5] Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.

[6] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space.

[8] Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states.

[9] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis.

[10] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis.

[11] Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2018). Social media analytics for crisis communication.

[12] Thelwall, M. (2018). Social media analytics for sentiment analysis in politics.

[13] Twitter API Documentation (2021). Retrieved from https://developer.twitter.com/

[14] Yang, Z., Dai, Z., Yang, Y., et al. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding.

[15] Zadeh, A., Chen, M., Poria, S., & Morency, L. P. (2020). Multimodal sentiment analysis: A survey and taxonomy.