Emotion Analysis of Social Media: A Study on Emotional Expression in Tweets using Sentiment Analysis Approach

Jeff Rouzel Bat-og, Zyrex Djewel Ganit, and Rainer Mayagma

University of the Philippines Visayas, Miagao, Iloilo
{jabatog, zfganit, rtmayagma}@up.edu.ph

Abstract. Social media platforms, particularly Twitter, have become vital sources of real-time public expression, providing valuable insights into people's emotional states. This study aims to analyze and classify emotions expressed in tweets using natural language processing (NLP) techniques. We explore various approaches for detecting emotions such as anger, joy, sadness, surprise, and love from a diverse dataset of tweets. By applying machine learning models, we examine how different emotions manifest in social media content and how these emotions correlate with current events, public sentiments, and user behaviors. Our findings provide a deeper understanding of emotional dynamics in online communities, contributing to the broader field of emotion analysis in social media. This research has potential applications in areas such as sentiment monitoring, mental health assessment, and social media analytics.

Keywords: Emotion analysis · Sentiment analysis · Machine learning · Natural language processing

1 Introduction

Social media platforms like Twitter are among the most popular social networks worldwide, with a significant number of monthly active users as of April 2024. It ranks 12th globally by monthly active users [1]. Twitter has emerged as an influential platform for individuals to express their thoughts, opinions, and emotions. With over 500 million tweets posted daily and 237.8 million daily active users, Twitter offers a unique opportunity to study human emotions on a large scale [2][3]. As public expression increasingly occurs in digital spaces, understanding the emotional content of social media posts becomes crucial for various fields, including sentiment analysis (SA), mental health, and public opinion research.

Emotions are very important in human communication, and social media platforms provide a rich source of emotional data. Emotion analysis has become an important area of research in the broader field of SA, where the emotional undertone of textual data is analyzed to understand human behavior [4]. It is significant in understanding the affective dimension of literature, but its applications extend to various domains, including social media, where user-generated

content offers insights into public sentiment, mental well-being, and social interactions [5]. The growth of digital platforms like Twitter has further fueled the interest in this research, with vast amounts of text-based data available for analysis, enabling the study of human emotions on an unprecedented scale.

As technology advances, the ability of computers to recognize and respond to human emotions has become increasingly important. Emotion detection, often tied closely to SA, is a key component of human-computer interaction (HCI) [6]. It enables systems to adjust their behavior based on users' emotional states, improving user experience and satisfaction. Machine learning (ML) techniques, particularly supervised learning approaches such as Support Vector Machines (SVM) and Naïve Bayes, have proven effective in detecting emotions across various data sources, including text, speech, and even biosignals. These techniques have been applied successfully in diverse domains, ranging from mental health to interactive systems design. However, challenges remain in standardizing datasets and evaluation procedures, as well as in expanding the scope beyond English-language data sources [6].

The growth of online platforms like Twitter has provided a wealth of data for emotion analysis, as users regularly express their feelings through text-based posts. This vast, unstructured data can be analyzed to gain insights into public sentiment and emotional trends. Machine learning methods, by processing this data, can uncover hidden patterns and provide valuable feedback for improving system interactions.

2 Literature Review

Sentiment analysis, also known as opinion mining, is a subfield of natural language processing (NLP) concerned with identifying the sentiment or emotional tone in a piece of text, typically classified as positive, negative, or neutral [7]. It has found widespread applications in domains such as marketing, customer service, and public opinion monitoring, making it a critical tool for deriving insights from text data [8].

Traditional lexicon-based approaches have been largely replaced and outperformed by machine learning models, which can automatically learn sentiment patterns from labeled data [9]. In an experimental result, SVM and Naive Bayes come out with higher precision, while the lexicon-based approach requires less human efforts in labeling. This section reviews key models and techniques used in sentiment analysis, focusing on classification models, feature extraction methods, and data sampling strategies.

Emotion analysis refers to the process of determining the emotional status of individuals, such as anxiety, stress, depression, and fear, using data analytics techniques like TF-IDF and Bag of Words. It is applied to understand the psychological effects of events like Covid-19 lockdowns on human psychology [10].

2.1 Machine Learning Models for Sentiment Analysis

Several machine learning models have been used in sentiment analysis. Each model offers different strengths in terms of interpretability, scalability, and performance.

Naive Bayes Naive Bayes is one of the simplest yet widely used models for sentiment analysis due to its ease of implementation and low computational cost. This classifier provides more efficient and quick results compared to other machine learning models and techniques such as SVM and Maximum Entropy [11]. It operates on the assumption that the features (words) are conditionally independent given the class label. Despite this simplification, Naive Bayes performs well in text classification tasks like sentiment analysis, especially when the number of features is large [12].

Logistic Regression Logistic Regression is often used in various applications, such as predicting the risk of developing a disease or classifying data into different categories. It estimates the probability that a given input belongs to a particular class (positive or negative sentiment) by using the sigmoid function. One of the advantages of logistic regression is its interpretability, as it provides direct insight into the importance of individual features in the classification decision [13].

Studies says that logistic regression can be employed in the emotion analysis of speech signals to identify various emotional states during verbal communication. The study utilized machine learning algorithms to select the best features influencing emotional states, aiding in the understanding of human emotions through speech signals[14].

Support Vector Machines (SVM) Support Vector Machines (SVM) are a well-established technique in sentiment analysis [9]. SVMs work by finding the hyperplane that best separates the data into different classes [15]. In sentiment analysis, SVMs have proven to be effective in handling high-dimensional data typical of text classification tasks.

[**?**] highlighted the use of SVMs in classifying sentiment from short texts like tweets, where it performed better than some probabilistic models such as Naive Bayes.

Random Forest Random Forest is an ensemble learning method that constructs multiple decision trees and outputs the class that is the majority vote of the individual trees. Random Forests are well-suited for text classification tasks as they reduce overfitting and handle a large number of features effectively.

In [**?**], Random Forest was applied to Twitter data, where its ensemble nature helped achieve better classification accuracy by considering multiple decision paths.

Artificial Neural Networks (ANN) Artificial Neural Networks (ANN) are computational models inspired by the human brain. In the context of sentiment analysis, ANNs have gained attention for their ability to capture complex patterns and relationships in data. ANNs work by passing input features through multiple layers of neurons and adjusting the weights of connections based on training data.

Although ANNs are typically more complex and computationally expensive compared to simpler models like logistic regression, their flexibility allows them to handle non-linear patterns in text data, as shown by [**?**].

XGBoost XGBoost is an advanced gradient boosting algorithm that has gained significant popularity for its efficiency and performance in a variety of machine learning tasks, including sentiment analysis. XGBoost optimizes the decision tree-based boosting technique and incorporates regularization to prevent overfitting, making it highly effective for classification tasks on large datasets.

[**?**] introduced XGBoost, and its application to sentiment analysis has led to state-of-the-art performance on several datasets due to its ability to handle both linear and non-linear decision boundaries.

2.2 Feature Extraction Techniques

Feature extraction is crucial for sentiment analysis as it transforms raw text into numerical representations that machine learning models can process. Some of the most commonly used methods are described below:

TF-IDF Vectorizer Term Frequency-Inverse Document Frequency (TF-IDF) is a widely used feature extraction technique that converts text data into numerical vectors. It assigns a weight to each word based on its frequency in a document (Term Frequency) and its inverse frequency across all documents (Inverse Document Frequency). The TF-IDF vectorizer is useful in sentiment analysis as it helps reduce the influence of common words while emphasizing rare but significant words in the text.

[**?**] described TF-IDF as a powerful method for converting text into features that can be used for classification tasks like sentiment analysis.

n-Grams n-Grams refer to contiguous sequences of words or characters in a text. In sentiment analysis, n-grams help capture context and local dependencies between words that individual words (unigrams) might miss. For example, using bigrams or trigrams (sequences of two or three words) can help capture phrases like "not good" that express sentiment beyond individual word-level analysis.

[**?**] introduced n-grams for text classification, demonstrating their effectiveness in sentiment classification tasks.

2.3 Data Sampling Techniques

Data imbalance is a common problem in sentiment analysis, where one sentiment class (e.g., positive) might be overrepresented compared to others (e.g., negative). Various sampling techniques can be used to address this issue.

Upsampling and Downsampling Upsampling involves increasing the number of samples in the minority class by replicating existing samples or generating synthetic samples. Downsampling involves reducing the number of samples in the majority class to balance the dataset. These methods help ensure that machine learning models do not become biased toward the majority class, as discussed by [**?**].

Stratified Sampling Stratified sampling ensures that the training and test sets maintain the same proportion of classes as the original dataset. This technique is important in sentiment analysis, where certain sentiments might be underrepresented. By preserving class distributions, stratified sampling helps models generalize better to unseen data.

[**?**] showed that stratified sampling improves model performance, especially when working with imbalanced datasets.

2.4 Model Optimization: Hyperparameter Tuning

Hyperparameter tuning involves optimizing the parameters of machine learning models that are not learned from the data but set prior to training (e.g., the number of trees in Random Forest, the regularization parameter in logistic regression). Tuning these parameters can significantly impact the performance of models in sentiment analysis tasks.

Common methods for hyperparameter tuning include:

Grid Search: An exhaustive search over specified hyperparameter values. Random Search: A randomized search over hyperparameter space, which is more computationally efficient than grid search. Bayesian Optimization: A probabilistic model-based optimization technique that searches for the best hyperparameters by updating its knowledge about the parameter space over time. [**?**] showed that random search and Bayesian optimization often outperform traditional grid search in terms of finding the optimal hyperparameters efficiently.

2.5 Datasets for Sentiment Analysis

Several datasets have been pivotal in advancing sentiment analysis research. These datasets provide labeled examples of text with corresponding sentiment labels (positive, negative, neutral), and are used for training, validating, and testing models.

Stanford Sentiment Treebank (SST) The Stanford Sentiment Treebank is one of the most widely used datasets in sentiment analysis research. It contains over 10,000 movie reviews labeled with sentiment polarity. Unlike other datasets, SST provides a fine-grained sentiment score for each phrase in the sentence, making it particularly useful for models that need to capture the subtleties of sentiment at different granular levels.

IMDB Movie Reviews The IMDB movie review dataset is a large collection of movie reviews with binary sentiment labels (positive or negative). This dataset has been widely used to evaluate sentiment analysis models, particularly in supervised learning scenarios.

Twitter Sentiment Datasets Due to the rise of social media, Twitter sentiment datasets have become increasingly important for analyzing public opinion. (CITE HERE)introduced a Twitter dataset where tweets were automatically labeled using emoticons as proxies for sentiment. This dataset has been used extensively for training models that can handle short, informal text with a high degree of noise.

2.6 Challenges in Sentiment Analysis

Despite the advances in machine learning and deep learning models, sentiment analysis continues to face several challenges:

Sarcasm and Irony Detection Sarcasm and irony are common in online communication, and they present a significant challenge for sentiment analysis models. Lexicon-based methods, in particular, struggle with detecting sarcasm, as words with positive sentiment may be used sarcastically to express negative emotions. Recent work by (CITE HERE) explored the use of deep learning models to improve sarcasm detection, but this remains an active area of research.

Domain Adaptation Sentiment expressions vary across domains, such as product reviews, social media, and news articles. Models trained on one domain may not perform well when applied to another. (CITE HERE) proposed domain adaptation techniques to address this issue, but the challenge of transferring knowledge between domains remains.

Multilingual Sentiment Analysis Most sentiment analysis research focuses on English text, but there is growing interest in analyzing sentiments expressed in other languages. (CITE HERE) explored sentiment analysis in multilingual settings, highlighting the challenges of linguistic differences and the lack of annotated datasets in many languages.

2.7 Applications of Sentiment Analysis

Sentiment analysis has found applications across various domains, including:

Business Intelligence: Companies use sentiment analysis to monitor customer feedback and improve their products and services. Politics: Sentiment analysis has been used to analyze public opinion on political candidates and policies by studying social media trends. Healthcare: Researchers have applied sentiment analysis to analyze patient feedback and identify areas for improvement in healthcare services.

3 Data and Methodology

Here you can describe the methods and techniques used in your research. This section should provide enough detail for others to replicate your approach if needed.

4 Results and Discussion

This is the results and discussion

5 Conclusion

This section concludes the paper and discusses the results, any future work, or open problems that remain to be explored.

References

1. Statista. Most popular social networks worldwide. `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/`, 2024. Accessed: 2024-12-08.
2. Internet Live Stats. Twitter usage statistics - internet live stats. `https://www.internetlivestats.com/twitter-statistics/`, 2024. Accessed: 2024-12-08.
3. Omnicore. Twitter by the numbers: Stats, demographics and fun facts. `https://www.omnicoreagency.com/twitter-statistics/`, 2023. Accessed: 2024-12-08.
4. Evgeny Kim and Roman Klinger. A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift für digitale Geisteswissenschaften*, 2019.
5. Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. Emotion analysis: A survey. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, pages 397–402, 2017.
6. Alaa Alslaity and Rita Orji. Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. *Behaviour & Information Technology*, 43(1):139–164, 2024.
7. Saidah Saad and Bilal Saberi. Sentiment analysis or opinion mining: A review. *International Journal on Advanced Science, Engineering and Information Technology*, 7:1660, 10 2017.

8. Margarita Rodríguez-Ibánez, Antonio Casánez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223:119862, 2023.
9. Alaa Mahmood, Siti Kamaruddin, Raed Naser, and Maslinda Mohd Nadzir. A combination of lexicon and machine learning approaches for sentiment analysis on facebook. *Journal of System and Management Sciences*, 10, 11 2020.
10. A. Chatterjee, B. Das, and A. Das. Chapter 10 - towards the analyzing of social-media data to assess the impact of long lockdowns on human psychology due to the covid-19 pandemic. In Dipankar Das, Anup Kumar Kolya, Abhishek Basu, and Soham Sarkar, editors, *Computational Intelligence Applications for Text and Sentiment Data Analysis*, Hybrid Computational Intelligence for Pattern Analysis and Understanding, pages 225–238. Academic Press, 2023.
11. Pramod Mathapati, Arati Shahapurkar, and Kavita Hanabaratti. Sentiment analysis using naïve bayes algorithm. *International Journal of Computer Sciences and Engineering*, 5:75–77, 07 2017.
12. Deepak Saini, Trilok Chand, Devendra K. Chouhan, and Mahesh Prakash. A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on x-ray images. *Biocybernetics and Biomedical Engineering*, 41(2):419–444, 2021.
13. Sameena Pathan, K. Gopalakrishna Prabhu, and P.C. Siddalingaswamy. Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—a review. *Biomedical Signal Processing and Control*, 39:237–262, 2018.
14. E Poovammal, Satyam Verma, Siddhant Sharma, and Virendra Agarwal. Emotional analysis using multinomial logistic regression. *Indian Journal of Science and Technology*, 9(39), 2016.
15. V.J. Gillet and A.R. Leach. Chemoinformatics. In John B. Taylor and David J. Triggle, editors, *Comprehensive Medicinal Chemistry II*, pages 235–264. Elsevier, Oxford, 2007.