VibeCheck: A Machine Learning Approach to Emotion Detection in Twitter Data

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 social media platforms, 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 emotion 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 social media 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 large 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). 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 Naive Bayes, have proven effective in detecting emotions across various data sources, including text, speech, and even biosignals. These techniques have been applied successfully in different 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].

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].

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 [9].

The distinction between these two processes can sometimes blur, as both methods rely on similar techniques, such as natural language processing (NLP) and machine learning models, to assess the emotional tone of the text. In many cases, sentiment analysis systems may also be trained to recognize particular emotions, effectively overlapping with emotion detection. While sentiment analysis tends to provide a broader understanding of a text's emotional polarity, emotion detection offers more nuanced insights into specific emotional states [10].

2.1 Machine Learning Models for Emotion and Sentiment Analysis

Emotion detection has gained traction in recent research. Studies like the one conducted by Tanwar and Vishesh explored emotion detection using Twitter data, focusing on six primary emotions: joy, anger, sadness, fear, love, and surprise [11]. The study employed deep learning models like BiLSTM and CNN, demonstrating the superiority of these approaches in capturing the complexities of emotional expressions in tweets. Similarly, Lora et al. compared traditional machine learning models like SVM with deep learning methods such as Stacked LSTM and CNN, underscoring the importance of data preprocessing and utilizing large datasets for improved accuracy in emotion detection tasks [12].

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

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 [13]. 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 [14].

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 [15].

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[16].

Support Vector Machines (SVM) are a well-established technique in sentiment analysis [17]. SVMs work by finding the hyperplane that best separates the data into different classes [18]. In sentiment analysis, SVMs have proven to be effective in handling high-dimensional data typical of text classification tasks.

A Random Forest classifier is a machine learning algorithm that uses a collection of decision trees to classify data into different classes. It performs well in predicting most classes, but may struggle with classes that have similar characteristics in their data [19].

Artificial Neural Networks (ANN) are computational models inspired by the human brain. ANNs work by passing input features through multiple layers of neurons and adjusting the weights of connections based on training data [20].

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 [21].

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 [22].

2.2 Feature Extraction Techniques

Feature extraction is crucial for emotion and sentiment analysis as it transforms raw text into numerical representations that machine learning models can process.

The Term Frequency-Inverse Document Frequency (TF-IDF) is a widely used feature extraction technique that converts text data into numerical vectors [23]. 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 both analysis as it helps reduce the influence of common words while emphasizing rare but significant words in the text [24].

The n-Grams refer to contiguous sequences of words or characters in a text. In both 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 [25].

Bhardwaj and Pant introduced n-grams for text classification, demonstrating their effectiveness in sentiment classification tasks. By combining n-gram feature extraction with the KNN classifier, their approach effectively captured sentiment patterns in Twitter data [26]. The experiments showed that using n-grams improved performance in terms of precision, recall, and accuracy, outperforming traditional methods such as SVM classifiers. This highlights the value of n-grams in capturing contextual relationships in sentiment analysis tasks.

2.3 Data Sampling Techniques

Data imbalance is a common problem in emotion analysis, where one sentiment class (e.g., joy) might be overrepresented compared to others (e.g., surprise) [27]. This can lead to difficulties in training models and lower accuracy in object detection. Various sampling techniques can be used to address this issue.

Upsampling involves increasing the number of samples in the minority class through replication or synthetic sample generation. Downsampling involves reducing the number of samples in the majority class.

The choice between upsampling and downsampling depends on factors like dataset size, class imbalance ratio, and computational resources. Additionally, ensuring data quality, selecting appropriate machine learning algorithms, and using suitable evaluation metrics are crucial for improving model performance on imbalanced datasets [28].

Stratified sampling ensures that after splitting the datasets, it maintains the same proportion of classes as in 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 [29].

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.

Grid Search is an exhaustive search over specified hyperparameter values. Random Search randomized search over hyperparameter space, which is more computationally efficient than grid search [30]. Bayesian Optimization, a probabilistic model-based optimization technique that searches for the best hyperparameters by updating its knowledge about the parameter space over time [31].

2.5 Challenges in Emotion and Sentiment Analysis

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

Sarcasm and irony are common in online communication, and they present a significant challenge for sentiment analysis models [32]. Lexicon-based methods, in particular, struggle with detecting sarcasm, as words with positive sentiment may be used sarcastically to express negative emotions. Researchers explored the use of deep learning models to improve sarcasm detection, but this remains an active area of research [33].

Most sentiment analysis research focuses on English text, but there is growing interest in analyzing sentiments expressed in other languages. Dashtipour et al. explored sentiment analysis in multilingual settings, highlighting the challenges of linguistic differences and the lack of annotated datasets in many languages [34].

2.6 Applications of Emotion Analysis

Emotion analysis is widely used across industries to gain insights from customer feedback, enhance product development, and improve marketing strategies. It helps businesses analyze online reviews, social media posts, and customer surveys to understand public opinion. It is important in monitoring brand reputation and tracking trends. Other applications include political sentiment tracking, financial market analysis, and providing insights into public health sentiment during crises. These real-world applications allow organizations to make data-driven decisions based on customer emotions and opinions [35].

3 Data and Methodology

3.1 Data

The dataset used in this study is sourced from Kaggle and is titled "Emotions Dataset for NLP." This dataset contains 16,000 labeled samples, each expressing one of five distinct emotions: joy, love, surprise, sadness, and anger. The data is structured into two columns: the first representing the text (tweet or statement) and the second representing the associated emotion label. The dataset is highly suitable for training machine learning models designed for emotion detection tasks, as it includes diverse samples of short, real-world text data. The dataset can be accessed via the following link: Kaggle Dataset.

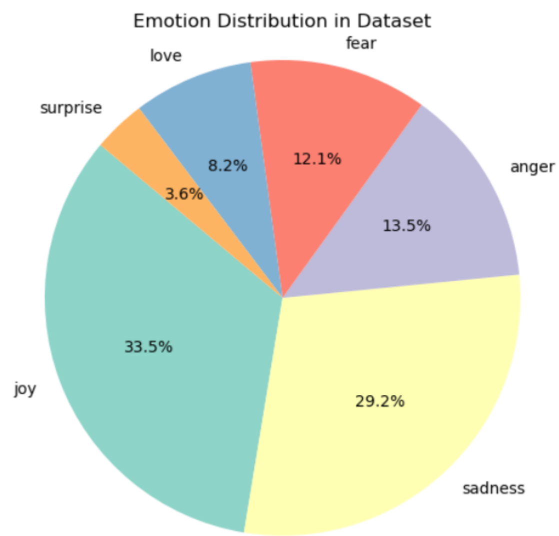


Fig. 1: Percentage distribution of emotions in the dataset.

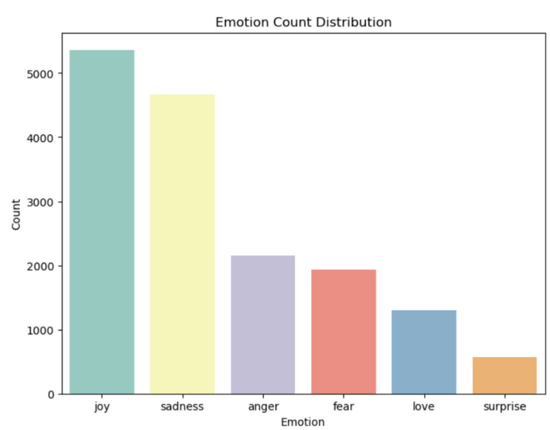


Fig. 2: Count of each emotion in the dataset.

The dataset contains six distinct emotions: *joy*, *sadness*, *anger*, *fear*, *love*, and *surprise*. The distribution of these emotions can be visualized in Figure 1, which shows that *joy* is the most frequent emotion, comprising 33.5% of the dataset. This is followed by *sadness*, which accounts for 29.2%, reflecting a notable representation of negative sentiments. Emotions such as *anger* (13.5%) and *fear* (12.1%) also have a significant presence, while *love* (8.2%) is moderately represented. Finally, *surprise* is the least frequent emotion, making up only 3.6% of the dataset. The bar chart in Figure 2 further highlights the exact counts for each emotion, reinforcing the dominant presence of *joy* and *sadness*, and the relatively rare occurrence of *surprise*.

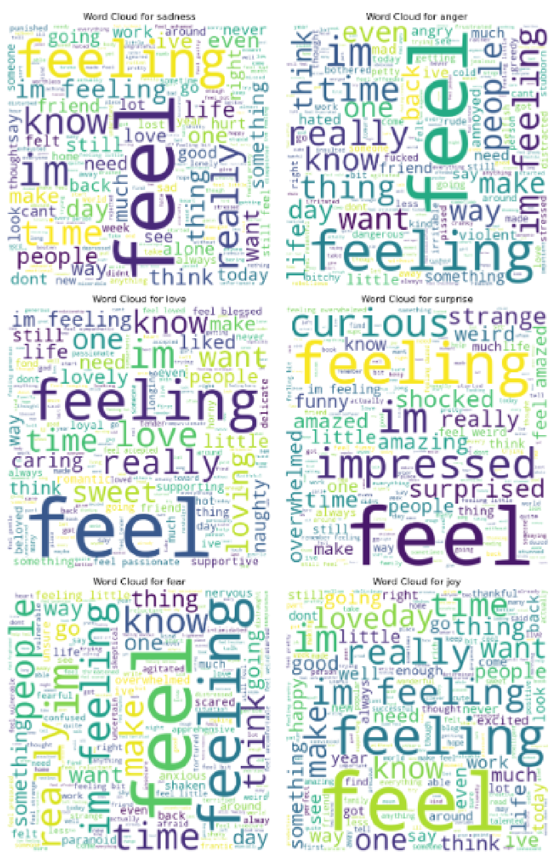


Fig. 3: Word cloud for the most frequent terms in each emotion.

In addition, Figure 3 presents a word cloud visualization that showcases the most frequently occurring words in each emotion category. For example, in the *joy* category, words such as *happy*, *excited*, and *great* appear frequently, emphasizing themes of positivity. On the other hand, in the *sadness* category, terms like *loss*, *hurt*, and *tears* dominate, reflecting themes of grief. The *anger* category includes terms such as *mad*, *frustrated*, and *hate*, while the *fear* category features words like *scared* and *nervous*. In the *love* category, words such as *care*, *heart*, and *dear* are more prominent, highlighting affection and compassion. Finally, the *surprise* category shows terms like *shock* and *unexpected*, which represent reactions to unexpected events. Additionally, common words such as *feel* and *feeling* appear across all emotion categories, indicating the strong emotional undertones in the dataset and emphasizing how people often associate emotions with personal feelings.

These visualizations provide crucial insights into the distribution and underlying themes of each emotion in the dataset, offering a comprehensive view of the text data's emotional content.

4 Methodology

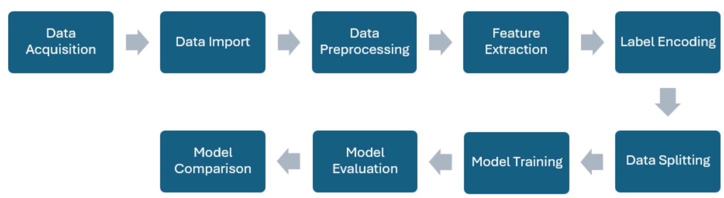


Fig. 4: Flowchart representing the methodology for sentiment analysis.

The methodology for this sentiment analysis project is outlined in the flowchart shown in Figure 4.

The first step in the methodology is data acquisition, where the dataset is obtained from Kaggle. The dataset used in this study is titled "Emotions Dataset for NLP," and it contains textual data labeled with six distinct emotions: joy, sadness, anger, fear, love, and surprise. This dataset serves as the foundation for the sentiment analysis task. Once the dataset is acquired, it is imported into the system using Pandas. The 'train.csv' file, which contains 16,000 sentiment samples, is read into a Pandas DataFrame for further processing and analysis.

Data preprocessing is a crucial step to prepare the dataset for analysis. This involves cleaning the text by removing HTML tags, URLs, numbers, special characters, and unwanted patterns using regular expressions (regex). Additionally, stopwords—common words that do not contribute significant meaning, such as "the," "a," "an," etc.—are removed using the NLTK library [36].

After preprocessing, the text data is converted into numerical features using the TF-IDF (Term Frequency-Inverse Document Frequency) technique. The

'TfidfVectorizer' is used to transform the text data into a matrix of TF-IDF features. This method also includes n-grams, specifically unigrams, bigrams, and trigrams, to capture not only individual words but also word pairs and triples, which can provide more context and improve the model's understanding of the relationships between words in the text.

Next, label encoding is performed. Since machine learning models cannot process categorical data directly, the emotion labels (joy, sadness, anger, etc.) are encoded into numerical values using 'LabelEncoder', allowing the model to understand and classify the emotions.

The dataset is then split into training and testing sets. Typically, 80% of the data is used for training the model, while the remaining 20% is reserved for testing and evaluation. This ensures that the model is trained on a portion of the data and evaluated on unseen data.

Once the data is preprocessed and ready, several machine learning models are trained on the dataset, including Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, Artificial Neural Networks (ANN), and XGBoost. Each model is trained using the training set, and its performance is evaluated using the test set.

After training the models, they are evaluated using various performance metrics. These include a classification report showing precision, recall, and F1-score for each emotion class, a confusion matrix illustrating the true positive, true negative, false positive, and false negative counts, and an overall accuracy score for each model on the test data.

The performance of all trained models is then compared. The best performing model is selected based on its accuracy and other evaluation metrics, and it is used for further analysis and potential deployment. Finally, the results are summarized, highlighting the best performing model and its potential application for real-world emotion detection tasks. The selected model is further analyzed to understand its strengths and weaknesses in classifying the emotions.

5 Results and Discussion

This chapter present the performance evaluation of various machine learning models for emotion classification and discuss their strengths and limitations. Six models—Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine (SVM), Artificial Neural Network (ANN), and XGBoost(Extreme Gradient Boosting)—were trained and tested on the dataset to identify the emotions expressed in textual data. Their performance was evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

Overall, the results highlight the varying capabilities of the models in distinguishing between six emotions: anger, fear, joy, love, sadness, and surprise. While all models performed reasonably well in identifying the more frequent emotions like joy and sadness, challenges emerged in classifying nuanced or underrepresented emotions like love and surprise. Each model's strengths and weaknesses are discussed in detail, emphasizing the trade-offs between simplicity, computational efficiency, and classification accuracy.

5.1 Initial Results

This section presents the initial results of the models before implementing optimizations such as stratified sampling, upsampling, downsampling, hyperparameter tuning, and adjustments to the Random Forest parameters. Notably, the previous results did not include the application of XGBoost and Artificial Neural Networks (ANN) to the dataset, which are addressed in this analysis.

Accuracy: 0.629375
Classification Report

	precision	recall	f1-score	support
anger	0.93	0.19	0.31	427
fear	0.93	0.14	0.24	397
joy	0.54	0.99	0.70	1021
love	1.00	0.01	0.01	296
sadness	0.73	0.91	0.81	946
surprise	0.00	0.00	0.00	113
accuracy			0.63	3200
macro avg	0.69	0.37	0.35	3200
weighted avg	0.72	0.63	0.54	3200

Fig. 5: Naive Bayes Classification Report

Accuracy: 0.826875
Classification Report

	precision	recall	f1-score	support
anger	0.92	0.71	0.81	427
fear	0.86	0.69	0.77	397
joy	0.77	0.97	0.86	1021
love	0.91	0.47	0.62	296
sadness	0.85	0.95	0.90	946
surprise	0.86	0.33	0.47	113
accuracy			0.83	3200
macro avg	0.86	0.69	0.74	3200
weighted avg	0.84	0.83	0.81	3200

Fig. 6: Logistic Regression Classification Report

Accuracy: 0.8825
Classification Report

	precision	recall	f1-score	support
anger	0.88	0.87	0.87	427
fear	0.88	0.77	0.82	397
joy	0.87	0.95	0.91	1021
love	0.88	0.71	0.79	296
sadness	0.91	0.94	0.93	946
surprise	0.79	0.68	0.73	113
accuracy			0.88	3200
macro avg	0.87	0.82	0.84	3200
weighted avg	0.88	0.88	0.88	3200

Fig. 7: Random Forest Classification Report

```
Accuracy: 0.7840625
Classification Report
              precision    recall  f1-score   support

       anger       0.93      0.62      0.74       427
        fear       0.86      0.60      0.71       397
         joy       0.70      0.97      0.82      1021
        love       0.91      0.29      0.44       296
     sadness       0.82      0.95      0.88       946
    surprise       0.76      0.22      0.34       113

    accuracy                           0.78      3200
   macro avg       0.83      0.61      0.66      3200
weighted avg       0.81      0.78      0.76      3200
```

Fig. 8: SVM Classification Report

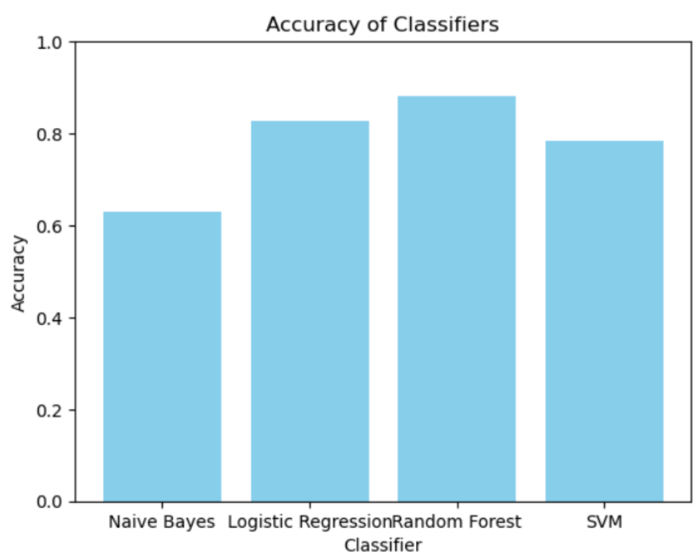
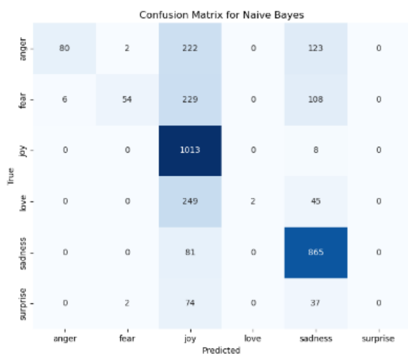
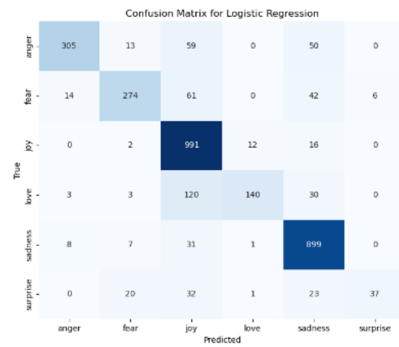


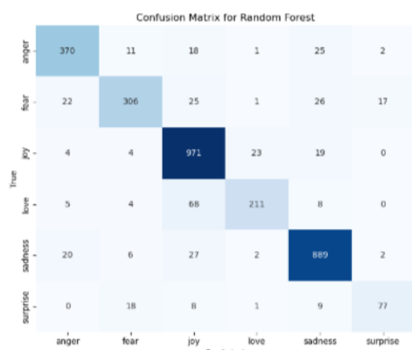
Fig. 9: Initial Accuracy Model



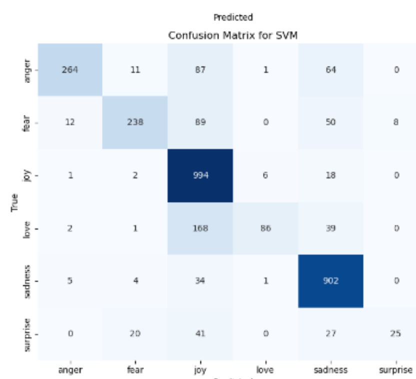
(a) Confusion Matrix for Naive Bayes



(b) Confusion Matrix for Logistic Regression



(c) Confusion Matrix for Random Forest



(d) Confusion Matrix for SVM

Fig. 10: Initial Results Confusion Matrix

5.2 Naive Bayes

The Naive Bayes model achieved an accuracy of 0.63 (62.93%) 5. It demonstrated strong performance in recognizing "joy" and "sadness," with high recall values for these classes. Notably, the recall for "joy" was 0.99, indicating that nearly all joyful instances were correctly identified. However, the model struggled significantly with other emotions such as "anger," "fear," "love," and particularly "surprise," where the recall was extremely low at 0.00. This suggests that Naive Bayes had difficulty distinguishing between more nuanced emotions. The F1-scores were moderate for "joy" (0.70) and "sadness" (0.81), but very low for other categories, indicating poor balance between precision and recall. The confusion matrix showed that Naive Bayes experienced the most misclassifications, particularly between "joy" and other emotions like "anger" and "fear 10a."

5.3 Logistic Regression

Logistic Regression significantly improved accuracy to 0.83 , demonstrating strong performance across most emotions(82.68%) 6. The model was especially effective in identifying "joy" (F1-score: 0.86) and "sadness" (F1-score: 0.90), with high recall values of 0.97 and 0.95, respectively, indicating reliable identification of these emotions. It also performed well for "anger" and "fear," with good precision and recall metrics. However, the emotions "love" and "surprise" showed slightly lower recall values of 0.47 and 0.33, respectively, suggesting that these emotions were harder to predict accurately. While Logistic Regression reduced misclassifications overall, "love" and "surprise" remained challenging for the model.

5.4 Random Forest

The Random Forest model achieved the highest accuracy at 0.88 (88.25%), outperforming both Logistic Regression and Naive Bayes 7. It performed very well on "joy," "sadness," "fear," and "anger," with F1-scores around or exceeding 0.80. While the recall for "love" and "surprise" was still lower at 0.79 and 0.73, respectively, this was an improvement over the other models. Random Forest showed a balanced performance between precision and recall across most emotion categories, making it the best performer overall in this context. This model's ability to maintain high scores for harder-to-classify emotions like "love" and "surprise" highlighted its robustness.

5.5 Support Vector Machine (SVM)

The SVM model achieved an accuracy of 0.85 (85.28%), performing comparably to Logistic Regression. It exhibited strong precision and recall for "anger," "fear," "joy," and "sadness. 8" However, similar to Logistic Regression, "love" and "surprise" had lower recall values of 0.29 and 0.22, respectively, suggesting these emotions posed a challenge for the model. Despite these challenges, SVM achieved strong F1-scores for major emotion categories such as "joy" (0.82) and "sadness" (0.88). While its overall performance was commendable, SVM also faced difficulties with "love" and "surprise," similar to other models.

5.6 Overall Insights for Initial Results

The Random Forest model emerged as the best-performing model, achieving the highest accuracy (88.25%) and demonstrating a well-balanced performance across most emotions. However, all models faced challenges with "love" and "surprise," which showed lower recall and F1-scores. These difficulties could be attributed to the lower frequency of these emotions in the dataset or their nuanced nature. The Naive Bayes model, while effective for simpler tasks, struggled significantly with complex emotions and exhibited poor performance on less common categories like "surprise" and "love." This analysis underscores the importance of using more robust models like Random Forest for achieving better balance and accuracy in emotion classification tasks.

5.7 Final Results

The following are the precision, recall, F1-score, and support metrics for each classification model, evaluating its performance across six emotion categories: Sadness, Joy, Love, Anger, Fear, and Surprise.

Accuracy for Naive Bayes: 0.785
Classification report for Naive Bayes:

	precision	recall	f1-score	support
anger	0.86	0.74	0.79	432
fear	0.76	0.79	0.78	387
joy	0.89	0.78	0.83	1072
love	0.56	0.89	0.69	261
sadness	0.92	0.77	0.84	933
surprise	0.34	0.90	0.50	115
accuracy			0.79	3200
macro avg	0.72	0.81	0.74	3200
weighted avg	0.83	0.79	0.80	3200

Fig. 11: Naive Bayes Classification Report

Accuracy for Logistic Regression: 0.8584375
Classification report for Logistic Regression:

	precision	recall	f1-score	support
anger	0.85	0.87	0.86	432
fear	0.83	0.84	0.84	387
joy	0.92	0.84	0.88	1072
love	0.63	0.97	0.76	261
sadness	0.94	0.85	0.89	933
surprise	0.70	0.89	0.78	115
accuracy			0.86	3200
macro avg	0.81	0.88	0.84	3200
weighted avg	0.87	0.86	0.86	3200

Fig. 12: Logistic Regression Classification Report

Accuracy for Random Forest: 0.8559375
Classification report for Random Forest:

	precision	recall	f1-score	support
anger	0.87	0.84	0.86	432
fear	0.87	0.85	0.86	387
joy	0.90	0.83	0.86	1072
love	0.67	0.97	0.79	261
sadness	0.92	0.85	0.88	933
surprise	0.64	0.96	0.77	115
accuracy			0.86	3200
macro avg	0.81	0.88	0.84	3200
weighted avg	0.87	0.86	0.86	3200

Fig. 13: Random Forest Classification Report

```
Accuracy for SVM: 0.854375
Classification report for SVM:
              precision    recall  f1-score   support

       anger       0.85      0.86      0.85       432
        fear       0.82      0.84      0.83       387
         joy       0.91      0.84      0.87      1072
        love       0.64      0.95      0.77       261
     sadness       0.93      0.85      0.89       933
    surprise       0.71      0.87      0.78       115

    accuracy                           0.85      3200
   macro avg       0.81      0.87      0.83      3200
weighted avg       0.87      0.85      0.86      3200
```

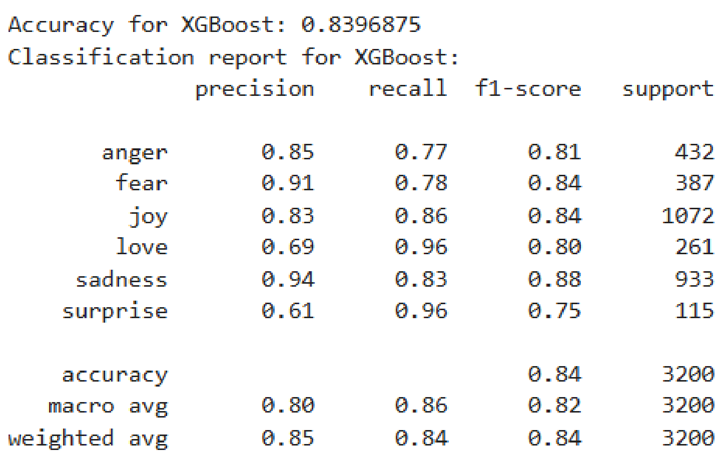
Fig. 14: SVM Classification Report

```
Accuracy for ANN: 0.835
Classification report for ANN:
              precision    recall  f1-score   support

       anger       0.78      0.84      0.81       432
        fear       0.78      0.83      0.81       387
         joy       0.92      0.82      0.87      1072
        love       0.59      0.92      0.72       261
     sadness       0.93      0.84      0.88       933
    surprise       0.72      0.75      0.74       115

    accuracy                           0.83      3200
   macro avg       0.79      0.83      0.80      3200
weighted avg       0.85      0.83      0.84      3200
```

Fig. 15: ANN Classification Report



Accuracy for XGBoost: 0.8396875
Classification report for XGBoost:

	precision	recall	f1-score	support
anger	0.85	0.77	0.81	432
fear	0.91	0.78	0.84	387
joy	0.83	0.86	0.84	1072
love	0.69	0.96	0.80	261
sadness	0.94	0.83	0.88	933
surprise	0.61	0.96	0.75	115
accuracy			0.84	3200
macro avg	0.80	0.86	0.82	3200
weighted avg	0.85	0.84	0.84	3200

Fig. 16: XGBoost Classification Report

5.8 Naive Bayes Result

The Naive Bayes model achieved an accuracy of **0.785**. It performed reasonably well across several emotions, with **Sadness** showing the highest recall (0.77) and **Surprise** the lowest precision (0.34) as shown in Figure 11. The weighted average F1-score of 0.80 indicates balanced performance across emotions, though some areas like **Surprise** and **Love** could be improved. In terms of precision, the model had the best result for **Sadness** (0.92).

5.9 Logistic Regression Result

Logistic Regression showed the highest accuracy at **0.8584**, as displayed in Figure 12. It achieved solid F1-scores across emotions, with **Sadness** and **Joy** showing excellent results (F1-scores of 0.89 and 0.88, respectively). The model performed particularly well in **Love** with a precision of 0.63 and recall of 0.97, which led to a high F1-score of 0.76. The overall weighted average F1-score of 0.86 reflects its strong performance, particularly in **Anger**, **Joy**, and **Sadness**.

5.10 Random Forest Result

The Random Forest model delivered an accuracy of **0.8559**, close to that of Logistic Regression, as shown in Figure 13. The model showed good precision and recall across emotions, with **Sadness** again being a strong performer (precision 0.92, recall 0.85). Notably, **Love** and **Surprise** had slightly lower precision scores, though the weighted average F1-score of 0.86 suggests robust performance overall, with good consistency between precision and recall.

5.11 Support Vector Machine (SVM) Result

SVM achieved an accuracy of **0.8544**, performing similarly to Random Forest, as depicted in Figure 14. The **Anger** class had a precision of 0.85 and recall of 0.86, contributing to a high F1-score. **Surprise** had relatively high precision (0.71) and recall (0.87), leading to a balanced F1-score. This indicates SVM's ability to generalize well, even with some misclassifications in less frequent emotions like **Surprise**.

5.12 Artificial Neural Network (ANN) Result

ANN had an accuracy of **0.835**, which was the lowest among the models tested, as shown in Figure 15. Despite this, it achieved impressive recall in **Love** (0.92), showing its strength in detecting emotional subtleties in such cases. However, it struggled with **Surprise**, as reflected in the F1-score of 0.74. The overall macro average F1-score of 0.80 highlights ANN's potential, although it might need additional tuning to improve accuracy across all emotions.

5.13 XGBoost Result

XGBoost showed an accuracy of **0.8397** and demonstrated good overall performance, particularly with **Fear** (precision 0.91, recall 0.78), as shown in Figure 16. The model performed reasonably well with **Sadness** and **Love**, achieving balanced results in both precision and recall. However, **Surprise** showed lower precision (0.61), impacting the F1-score, which was still relatively high at 0.75. The overall model's weighted F1-score of 0.84 indicates a solid all-around performance.

5.14 Model Accuracy Comparison

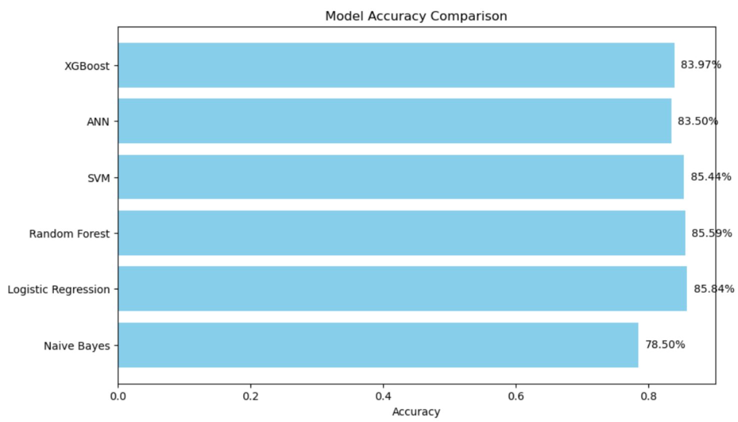


Fig. 17: Model Accuracy Comparison

As shown in Figure 17, Logistic Regression emerged as the top-performing model with an accuracy of **0.8584**. It was closely followed by Random Forest (**0.8559**) and SVM (**0.8544**). ANN achieved the lowest accuracy at **0.835**, highlighting the potential for further improvements in performance. XGBoost showed a solid accuracy of **0.8397**, rounding out the list of top models.

Logistic Regression performed exceptionally well across all emotion categories, with Random Forest and SVM also demonstrating strong results. Although ANN had the lowest accuracy, its strong recall in detecting emotions such as **Love** suggests it could be improved with further tuning. XGBoost demonstrated competitive performance but could benefit from addressing lower precision in emotions like **Surprise**. The results indicate that further optimization of hyperparameters for each model could enhance their performance, especially for less represented emotions.

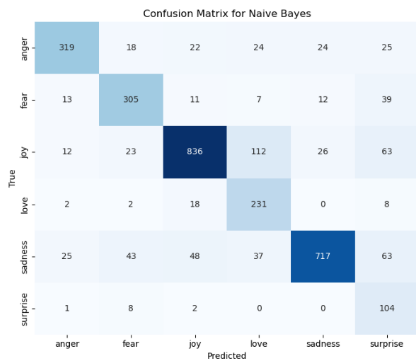
5.15 Confusion Matrices

The confusion matrices for each model show significant improvements in true positive results after applying optimizations. Among all emotion categories, "Joy" consistently achieves the highest true positive counts across all models, followed by "Sadness." These results suggest that the models are particularly effective at identifying these two emotions, likely due to their distinct features and higher frequency in the dataset.

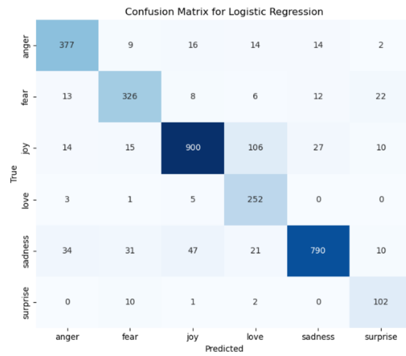
There is a notable improvement in the new results compared to the previous ones, where the true positive counts for some categories, such as "Love" and "Surprise," were effectively zero. While these values remain low, they have significantly improved. For instance, Naive Bayes previously had zero true positive counts for "Love" and "Surprise." After the distribution modifications, these values increased to 231 and 104, respectively. The fact that they are no longer zero indicates that the applied optimizations, such as stratified sampling and hyperparameter tuning, had a meaningful impact, even on the weakest-performing model.

For the other models, the improvements are also evident in the previously challenging categories of "Love" and "Surprise." After optimization, these models now demonstrate significantly higher true positive counts for these emotions. "Love" has achieved a true positive count of around 200 in most models, while "Surprise" has improved to approximately 80 to 100. Despite these gains, ANN has the lowest true positive count for "Surprise," with only 88, indicating it still struggles with this emotion. On the other hand, Random Forest demonstrates exceptional performance for "Love" and "Surprise," achieving a true positive count of 252 and 110 respectively, the highest among all models.

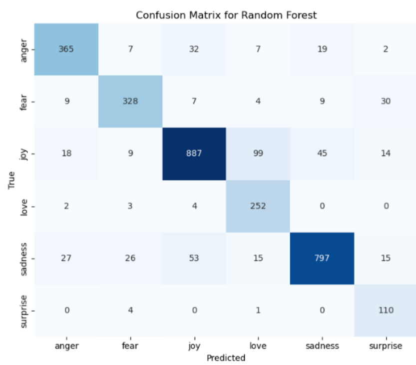
These results highlight the effectiveness of the optimizations in addressing imbalanced and nuanced emotion categories while underscoring the strengths and limitations of each model.



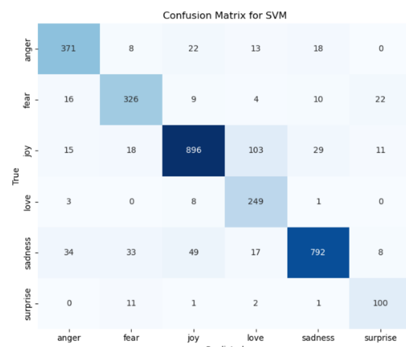
(a) Confusion Matrix for Naive Bayes



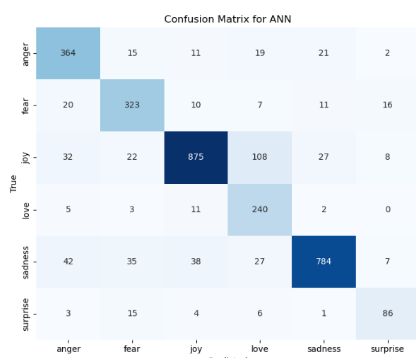
(b) Confusion Matrix for Logistic Regression



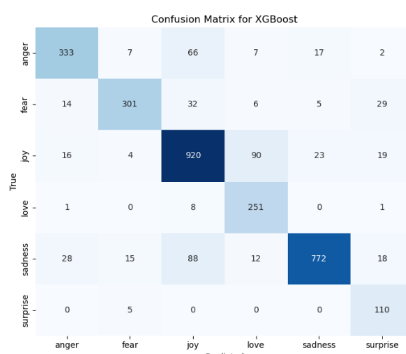
(c) Confusion Matrix for Random Forest



(d) Confusion Matrix for SVM

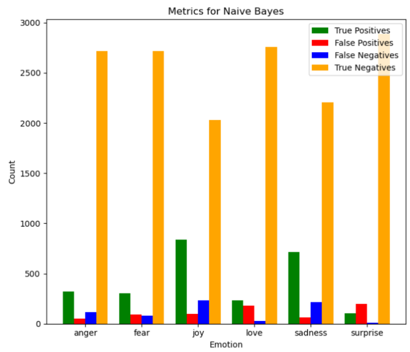


(e) Confusion Matrix for ANN

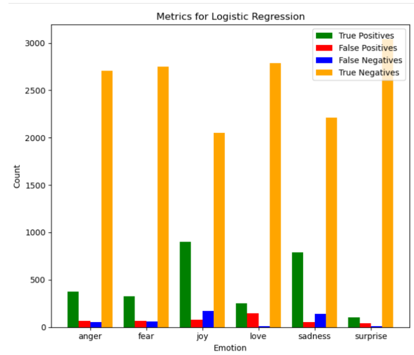


(f) Confusion Matrix for XGBoost

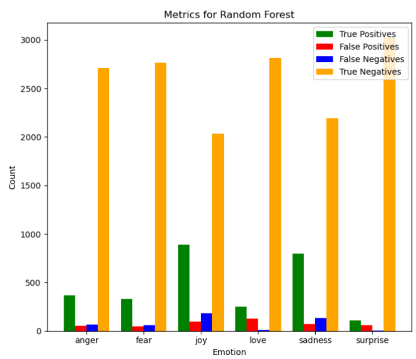
Fig. 18: Final Results Confusion Matrix



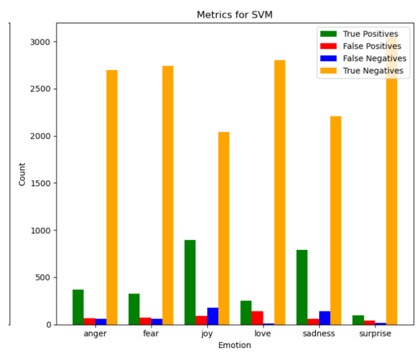
(a) Metrics for Naive Bayes



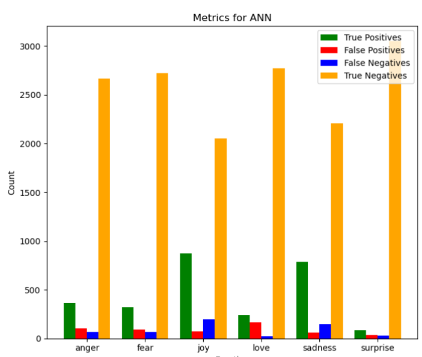
(b) Metrics for Logistic Regression



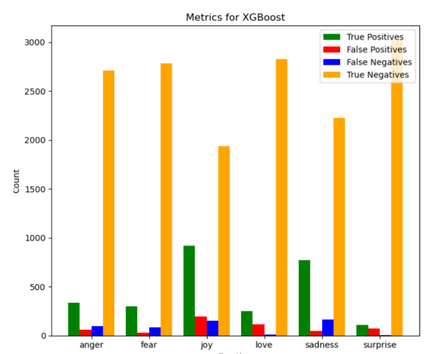
(c) Metrics for Random Forest



(d) Metrics for SVM



(e) Metrics for ANN

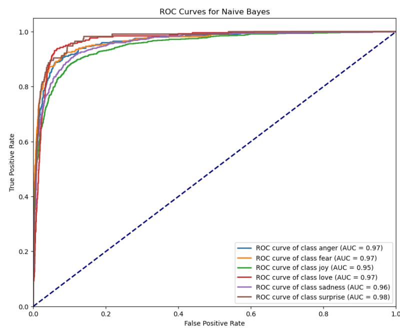


(f) Metrics for XGBoost

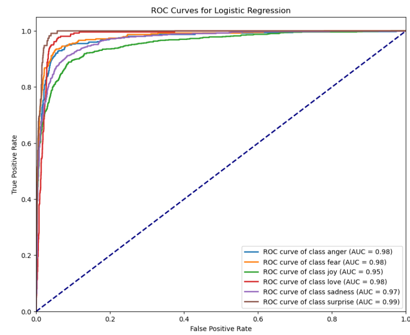
Fig. 19: Metrics for Models

For the metrics evaluation across all models, "Surprise" consistently exhibits the highest number of True Negatives, followed by "Love," "Fear," and "Anger."

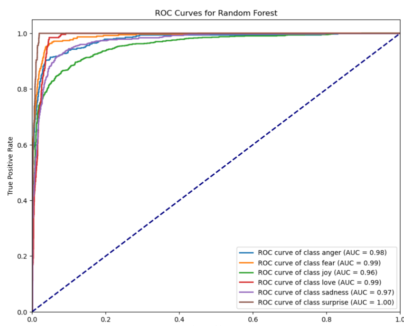
On the other hand, "Joy" achieves the highest True Positive count, closely followed by "Sadness." The distribution of False Positives and False Negatives varies across the models, highlighting differences in their misclassification patterns. These trends indicate that while some emotions are more easily distinguished by the models, others, like "Surprise" and "Love," may still pose challenges, depending on the specific classification algorithm.



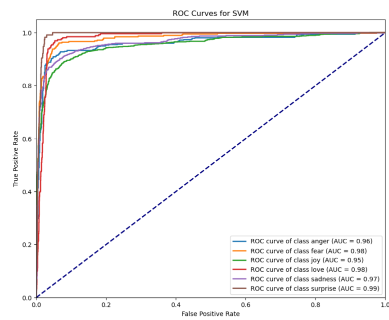
(a) ROC for Naive Bayes



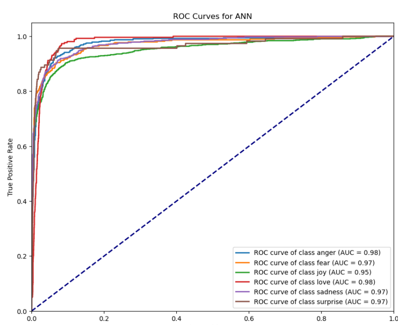
(b) ROC for Logistic Regression



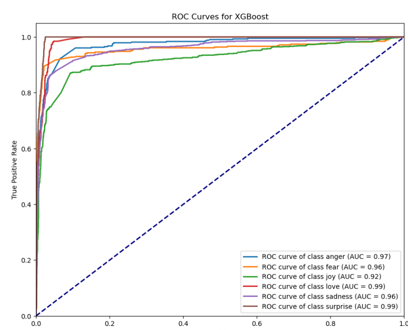
(c) ROC for Random Forest



(d) ROC for SVM



(e) ROC for ANN



(f) ROC for XGBoost

Fig. 20: RO Curve for Models

The ROC curves for the models show a high degree of classification performance across all emotion categories, as indicated by the Area Under the Curve (AUC) values. Random Forest demonstrate an AUC of 1.00 for "Surprise" and consistently high scores across all categories, indicating excellent discrimination.

However, since an AUC of 1.00 is rare in practice, it might indicate potential overfitting, data leakage, or unusually strong signals in the dataset for this specific class. Similarly, SVM achieves near-perfect AUC values for "Fear," "Love," and "Surprise."

Naive Bayes, while generally lower in classification accuracy compared to other models, shows respectable AUC values ranging from 0.95 to 0.98, indicating improved ability to separate classes after optimization. Logistic Regression and ANN also perform well, with AUC values typically above 0.95, showcasing their robustness in emotion detection.

XGBoost shows competitive performance with high AUC values, particularly excelling in "Love" (AUC = 0.99) and "Surprise" (AUC = 0.99). However, its AUC for "Joy" (0.92) is slightly lower compared to other models.

6 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.
2. Internet Live Stats. Twitter usage statistics - internet live stats. `https://www.internetlivestats.com/twitter-statistics/`, 2024.
3. Omnicore. Twitter by the numbers: Stats, demographics and fun facts. `https://www.omnicoreagency.com/twitter-statistics/`, 2023.
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. 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.
10. Olga Miroshnyk. Sentiment analysis vs. emotion detection. `https://oneai.com/learn/sentiment-analysis-vs-emotion-detection`, 2022.
11. Vishesh Tanwar. Deep learning-based emotion detection on twitter: A study of six primary emotions. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, pages 1232–1237, 2024.
12. Sanzana Lora, Nazmus Sakib, Shahana Antora, and Nusrat Jahan. A comparative study to detect emotions from tweets analyzing machine learning and deep learning techniques. 12:6–12, 06 2020.
13. 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.
14. 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.

15. 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.
16. E Poovammal, Satyam Verma, Siddhant Sharma, and Virendra Agarwal. Emotional analysis using multinomial logistic regression. *Indian Journal of Science and Technology*, 9(39), 2016.
17. 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.
18. 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.
19. Umit Senturk, Kemal Polat, Ibrahim Yucedag, and Fayadh Alenezi. Chapter 4 - arrhythmia diagnosis from ecg signal pulses with one-dimensional convolutional neural networks. In Kemal Polat and Saban Öztürk, editors, *Diagnostic Biomedical Signal and Image Processing Applications with Deep Learning Methods*, Intelligent Data-Centric Systems, pages 83–101. Academic Press, 2023.
20. Steven Walczak and Narciso Cerpa. Artificial neural networks. In Robert A. Meyers, editor, *Encyclopedia of Physical Science and Technology (Third Edition)*, pages 631–645. Academic Press, New York, third edition edition, 2003.
21. Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 02 2011.
22. Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of xgboost, 11 2019.
23. Noura Semary, Wesam Ahmed, Khalid Amin, Paweł Pławiak, and Mohamed Hammad. Enhancing machine learning-based sentiment analysis through feature extraction techniques. *PLOS ONE*, 19:e0294968, 02 2024.
24. Eric Nguyen. Chapter 4 - text mining and network analysis of digital libraries in r. In Yanchang Zhao and Yonghua Cen, editors, *Data Mining Applications with R*, pages 95–115. Academic Press, Boston, 2014.
25. Olumide Ojo, Alexander Gelbukh, Hiram Calvo, and Olaronke Adebanji. Performance study of n-grams in the analysis of sentiments. *Journal of the Nigerian Society of Physical Sciences*, 3:477–483, 11 2021.
26. Shailja Bhardwaj and Janmejay Pant. Sentiment analysis approach based n-gram and knn classifier. 08 2019.
27. Ravpreet Kaur and Sarbjeet Singh. A comprehensive review of object detection with deep learning. *Digital Signal Processing*, 132:103812, 2023.
28. Saikat. 5 techniques to handle imbalanced data for a classification problem, 2021. Accessed: 2024-12-08.
29. Fei Shi. Study on a stratified sampling investigation method for resident travel and the sampling rate. *Discrete Dynamics in Nature and Society*, 2015:1–7, 03 2015.
30. Avi Chawla. Grid search vs. random search vs. bayesian optimization, 2024.
31. Avi Chawla. Bayesian optimization for hyperparameter tuning, 2024.
32. Amina Ben Meriem, Lobna Hlaoua, and Lotfi Ben Romdhane. A fuzzy approach for sarcasm detection in social networks. *Procedia Computer Science*, 192:602–611, 2021.
33. D.I. Hernández Farias and P. Rosso. Chapter 7 - irony, sarcasm, and sentiment analysis. In Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu, editors, *Sentiment Analysis in Social Networks*, pages 113–128. Morgan Kaufmann, Boston, 2017.

34. Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad Hawalah, Alexander Gelbukh, and Qiang Zhou. Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation*, 8, 08 2016.
35. Noble Desktop. Sentiment analysis: Real world applications, 2024.
36. Opinosis Analytics. Stop words explained. `https://www.opinosis-analytics.com/knowledge-base/stop-words-explained/#:~:text=Stop%20words%20are%20a%20set,carry%20very%20little%20useful%20information.`, 2024.