Enhanced Sentiment and Emotion Detection in Tweets Using a Hybrid Machine Learning Approach

Anoop Rao Enaganthi
A20531044
Illinois Institute of Technology
aenaganthi@hawk.iit.edu

Nachiketh Nallamaddi
A20549679
Illinois Institute of Technology
nnallamaddi@hawk.iit.edu

Ajay Babu Popuri
A20547319
Illinois Institute of Technology
apopuri1@hawk.iit.edu

April 28, 2024

1 Introduction

In today's landscape of social media, where platforms like Twitter command an increasing share of users, understanding the nuanced emotions behind what people post online has become important lately. This project points to a major stride in the right direction with the implementation of an advanced machine learning model aimed at classifying tweets into different emotional states from joy to surprise. The undertaking pursues to improve sentiment analysis methodologies through a strong methodology, which combines a voting classifier incorporating logistic regression and support vector machine techniques. By delving deeper into the emotional fabric of tweets, this approach promises to streamline the extraction of emotional data, providing a more nuanced understanding of societal sentiments.

The integration of machine learning methods in this project does not only ease the process of inferring the emotional content carried by tweets but also offers deep insights into public discourse and opinion dynamics. It uses a voting classifier of various algorithms, such as logistic regression and support vector machines. The framework ensures that it has a broad assessment of emotional states; hence, the most accurate analysis of sentiment is to be made. This development not only opens up a door to the simple extraction of emotional data from tweets but also paves the way toward a deeper understanding of societal sentiments and attitudes, therefore forming a very big leap in the area of opinion mining and social analysis.

As these media platforms continue to structure the sites of public discourses, the ability to decode emotional signals embedded within user-generated content becomes increasingly indispensable. Initiatives like this are key in helping people make their way through an evolving online communication landscape.

2 Existing Works

The sentiment analysis landscape was provided using Twitter data by several machine-learning and lexicon-based methods. The landscape is mostly occupied by methods that make use of algorithms like Support Vector Machine, Naive Bayes, Random Forest , and ensemble methods with an objective of binary sentiment classification. Besides, methods such as domain-specific lexicon generation are also used for emotion-based feature extraction to increase the accuracy of emotion recognition

Lexicon-based methods have also been researched for sentiment extraction from Twitter data, with researchers developing domain-specific lexicons to increase recognition accuracy for emotions. These methods usually rely on lists of words that have been pre-annotated with the emotions they convey and on the existence of such words within the tweet. Despite being very simple and transparent, these lexicon- based approaches might suffer from poor coverage of the wide range of emotional expressions and context-dependent interpretations.

Our proposed work presents a new methodology using a voting classifier that combines Logistic Re- gression and Support vector machine. This approach is designed to minimize errors more effectively than individual classifiers. Our model is validated across different datasets, for the first time one of them being a unique 6-Emo's dataset from Kaggle, for even wider ranges of emotions to classify, which demonstrates its versatility and better performance for emotion recognition on Twitter data.

3 Project Description

In the digital age, social media, especially Twitter, has become a significant source of data, reflecting diverse emotions and perspectives in real-time. This wealth of information, particularly evident during the COVID-19 pandemic, underscores the importance of effective sentiment analysis. Twitter's brief, informal nature, filled with slang and jargon, makes it a complex medium for accurately deciphering human emotions.

The pandemic has revealed social media's impact in the propagation of information and underlined the advanced sentiment analytics required. It is an important tool that traces the public's general sentiment and helps one navigate these murky waters, where high emotional outbursts are fueled by misinformation.

Sentiment analysis in social media presents a dual challenge technically, it demands processing vast text data using sophisticated algorithms; sociologically, it requires understanding human emotions and communication nuances. This intersection of technology and sociology is essential for effective analysis, especially during globally impactful events like the pandemic. This fusion of data analysis, technology, psychology, and sociology is vital in capturing the nuances of human sentiment in the digital realm, offering insights into the public mood as reflected in social media

Our project has set out to address this difficulty through a framework that uses state-of-the-art NLP techniques and machine-learning algorithms. One big improvement in our methodology is the integration of a Voting Classifier that can gain the advantages from Logistic Regression and Stochastic Gradient Descent. The blended model is meant to aid in greater precision in categorizing the tweets in particular emotional groupings, hence yielding deeper comprehension of public sentiment

3.1 Description of the Data

In our project, three distinct datasets from Kaggle are utilized, each serving for a specific purpose:

Twitter Sentiment Dataset: This dataset contains a set of posts from Twitter, where each of them has been annotated for sentiment: positive, negative, or neutral. This dataset would contribute to understanding the public opinions and their emotions collected over Twitter, it will be a perfect training dataset for models in sentiment and

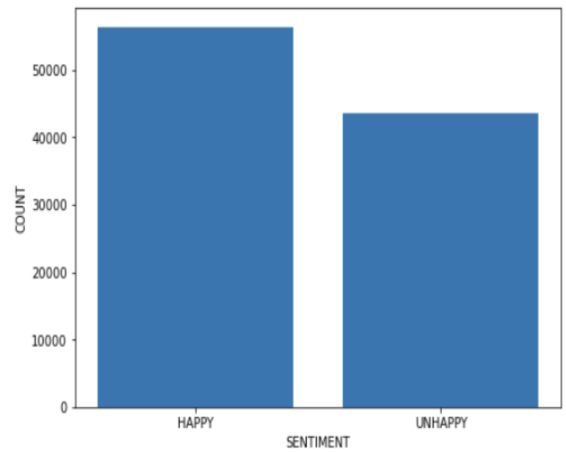


Figure 1: Visualization of dataset 1

natural language processing analysis. It has all kinds of topics and sentiments gathered from discussions on social media.

Women's E-commerce Dataset: This dataset is centered around women's e-commerce and includes detailed product descriptions, customer reviews, and ratings. It's an invaluable resource for understanding consumer behavior in the e-commerce space, specifically regarding women's products. This data, when analyzed, gives an insight into customer preference, level of satisfaction, and trending products that become highly critical inputs to market research and targeted marketing strategies in the e-commerce industry.

Sentiment Analysis Dataset on Hatred-Speech Detection from Twitter: This dataset is tailored for the detection of hate speech in Twitter posts. It contains tweets classified as containing hate speech or not containing hate speech. This is

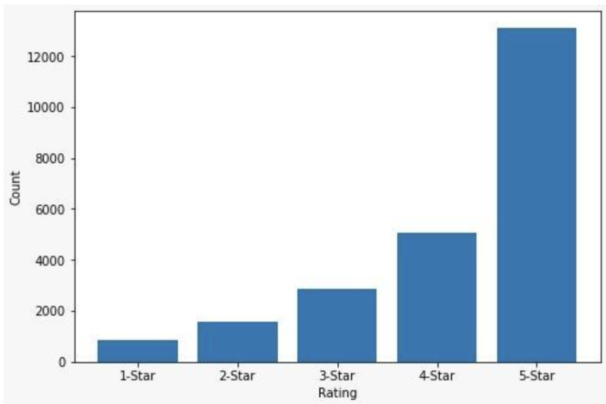


Figure 2: Visualization of dataset 2

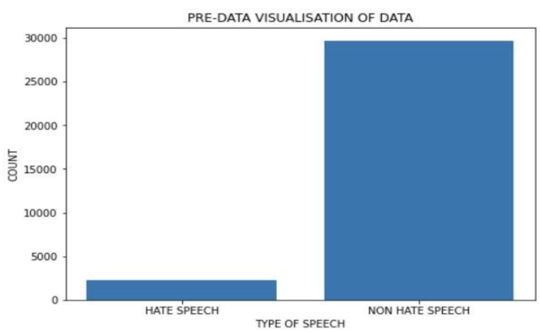


Figure 3: Visualization of dataset 3

important in developing models that are able to identify, track, and potentially flag or filter hate speech from social media platforms. This greatly boosts the safety in communication while online and promotes healthy discourse within digital spaces.

Each of these three datasets provided a unique insight from general Twitter sentiment analysis into specific consumer behavior in women's e-commerce and hate speech identification. With these, they offer a good framework for the analysis of various factors

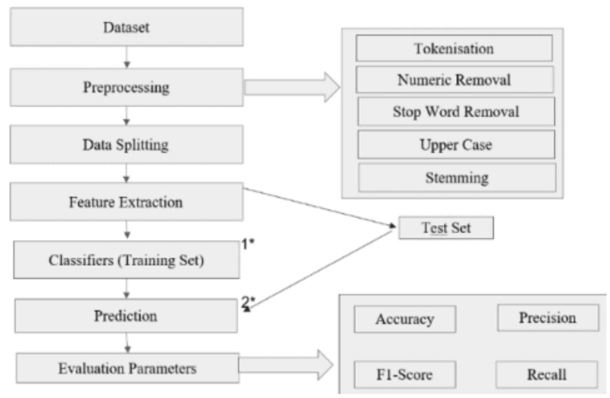
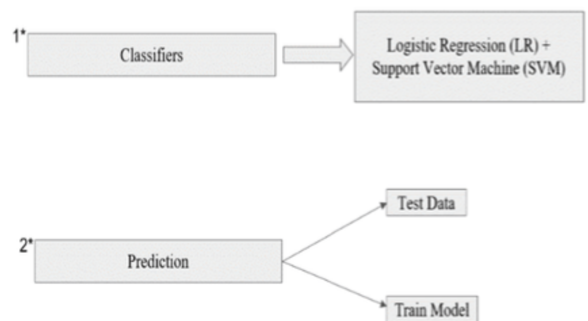


Figure 4: Methodlogy



in digital communication and consumer trend that are pivotal in the understanding and addressing of challenges currently affecting the social media dynamics and online retails

3.2 Methodology

We started with the acquisition of datasets from Kaggle. The primary dataset comprises Twitter sentiment data, which serves as the cornerstone for emotion recognition. Additionally, two supplementary datasets, including a women's e-commerce dataset and a sentiment analysis dataset focusing on hate speech detection from Twitter, are incorporated to enrich the analysis.

Once the datasets are obtained, we have done **preprocessing** to ensure data quality and consistency. This involves handling missing values, removing unnecessary characters such as emojis and symbols, eliminating numerical data as the focus is primarily on textual content, converting uppercase letters to lowercase for uniformity, removing stop words that do not contribute to the emotion expressed, and performing stemming to reduce words to their root forms. This preprocessing stage was crucial for preparing the data for further analysis and model development.

Following data preprocessing, the next step was **exploratory data analysis** (EDA) through data visualization techniques. Bar graphs, histograms, and pie charts were utilized to uncover hidden patterns and distributions within the datasets. Visualizations provided insights into the distribution of emotions and the prevalence of different sentiment classes, facilitating a deeper understanding of the data and guiding subsequent analysis.s

Feature extraction was then performed on the preprocessed data to convert textual features into a format suitable for machine learning models. Techniques such as term frequency (TF) and term frequency-inverse document frequency (TF-IDF) were employed to create a 2D feature matrix, where each row represents a tweet and each column represents a unique feature extracted from the text. Feature extraction played a crucial role in identifying important features that contribute to emotion classification.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

With the feature engineered data in hand, multiple machine learning models were developed and trained on the training set. These models include Support **Vector Machine (SVM)**, **Random Forest (RF)**, **Naive Bayes (NB)**, **Gradient Boosting Machine (GBM)**, **Logistic Regression (LR)**, **Voting Classifier (VC)**, **and optimization algorithm like (SGD)**. Each model is trained using the training set and evaluated using evaluation metrics such as accuracy, recall, precision, and F1-score to assess its performance in predicting emotions from tweets.

3.3 Challenges

The **challenges** encountered in the project was addressing the imbalance in emotional categories and ensuring that the feature selection process accurately captures sentiment nuances without overfitting.With 31,962 training records in total, Dataset 3 comprises 2,242 instances labeled as hatred speech, while the overwhelming majority, totaling 29,720 records, are categorized as non-

$$Recall = \frac{TP}{TP + FN}$$

$$F1score = 2\frac{precision.recall}{precision + recall}$$

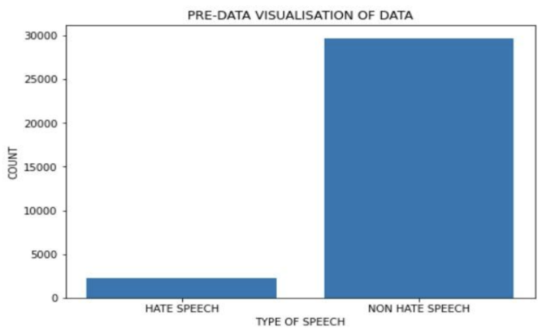


Figure 5: Before Oversampling

hatred speech. This stark class imbalance poses a significant obstacle in training machine learning models, as they tend to favor the majority class, leading to suboptimal performance in predicting the minority class.

To counteract this issue we have used oversampling technique. Specifically, instances from the minority class hatred speech were oversampled to balance the class distribution. This involved augmenting the number of instances in the minority class by adding synthetic records. To achieve this, some records from Dataset 1, the Sentiment Analysis on Twitter data, were introduced into Dataset 3.

3.4 Results

A Voting Classifier (VC) is a ensemble learning technique that harnesses the collective power of multiple machine learning algorithms to enhance predictive accuracy, generalizability, and robustness. Instead of relying on a single model, we have used VC to combine the predictions of several base classifiers through a voting mechanism, using majority voting . This approach gave more reliable outcomes and reduced the risk of overfitting.

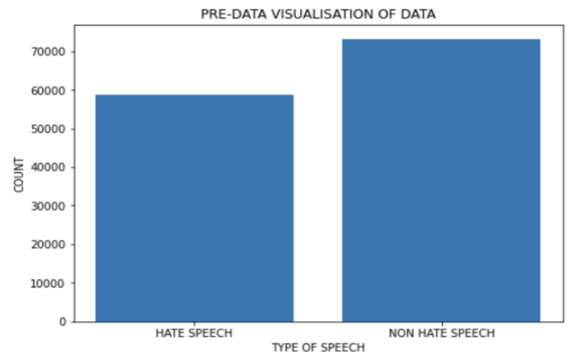


Figure 6: After Oversampling

In our project, we have used Voting Classifier, to explore its performance across three distinct datasets: Twitter sentiment data, Women's E-commerce dataset, and Sentiment Analysis for hatred speech detection from Twitter. The results obtained from these datasets illustrated the versatility and effectiveness of the Voting Classifier.For the Twitter sentiment data, VC achieves high accuracy, scoring 77.9. This demonstrates that VC can identify the correct sentiments with considerable reliability. Precision, recall, and F1 score are also notable, with 83.1, 79.0, and 81.0, respectively. The strong F1 score reflects a balanced performance, indicating that VC effectively manages the trade-offs between precision and recall, making it a robust choice for sentiment analysis.

In the Women's E-commerce dataset, VC maintainsed a similar level of accuracy, reaching 74.4. This result shows the adaptability of VC, indicating that it can work effectively with different data patterns and structures. Precision and recall scores are consistent at 79.3 and 64.4, respectively, suggesting that VC can maintain accuracy while reducing false positives. The F1 score, at 70.0, further emphasizes its balanced approach, offering a well-rounded model for predicting e-commerce trends and customer behavior.

Algorithms	Accuracy	Precision	Recall	F1_Score
NB	74.1	88.6	72.1	79.5
LR	77.4	84.1	77.8	80.8
SGD	76.1	86.1	75.2	80.3
GB	69.3	89.1	67.2	76.7
DT	68.6	72.9	71.9	72.4
RF	75.7	83.7	75.9	79.6
SVM	78.0	84.6	78.2	81.3
VC	77.9	83.1	79.0	81.0

Figure 7: Pretty Table comparing all the Algorithms with Performance Metrics (Dataset 1)

When analyzing the Sentiment Analysis dataset for hatred speech detection from Twitter, VC performs robustly, with 76.2 accuracy, 69.9 precision, 75.2 recall, and a 72.5 F1 score. This demonstrates VC's capability to identify sensitive content while minimizing false positives. The high recall rate is particularly valuable, indicating that VC can capture a high proportion of relevant instances, a critical factor in hatred speech detection.Overall, the Voting Classifier emerged as a powerful ensemble method in our project, demonstrating consistent performance across varied datasets. Its ability to combine the strengths of different classifiers allowed for improved accuracy and adaptability, making it suitable for a wide range of machine learning tasks. This flexibility and robustness underscore the importance of VC in real-world applications, where data complexity and diversity require reliable and adaptable solutions.

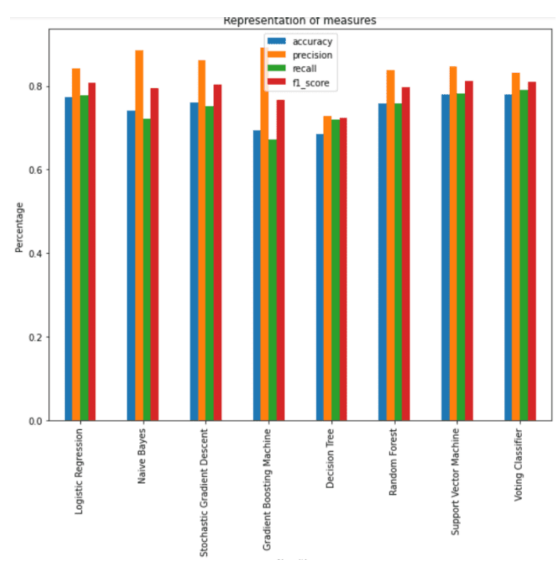


Figure 8: Bar graph comparing all the Algorithms with Performance Metrics (Dataset 1)

Models	Accuracy	Precision	Recall	F1_Score
RF	58.9	92.2	58.9	70.5
SVM	64.4	79.3	64.4	70.0
NB	56.8	98.7	56.8	56.8
GB	61.0	82.8	61.0	68.5
LR	64.4	74.3	64.4	68.2
SGD	63.4	79.5	63.4	69.3
VC	64.4	79.3	64.4	70.0

Figure 9: Pretty Table comparing all the Algorithms with Performance Metrics (Dataset 2)

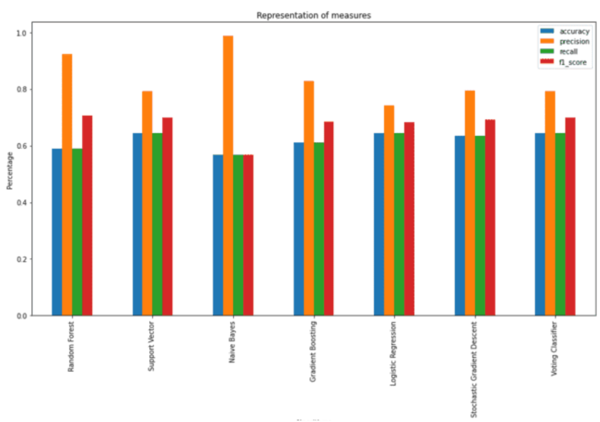


Figure 10: Bar graph comparing all the Algorithms with Performance Metrics (Dataset 2)

Models	Accuracy	Precision	Recall	F1_Score
RF	74.8	75.3	74.8	75.0
SVM	77.0	73.7	74.7	74.2
NB	73.2	54.2	79.5	64.4
GB	67.3	40.3	75.1	52.5
LR	76.2	71.7	74.4	73.0
SGD	74.0	62.4	75.4	68.3
VC	76.2	69.9	75.2	72.5

Figure 11: Pretty Table comparing all the Algorithms with Performance Metrics (Dataset 3)

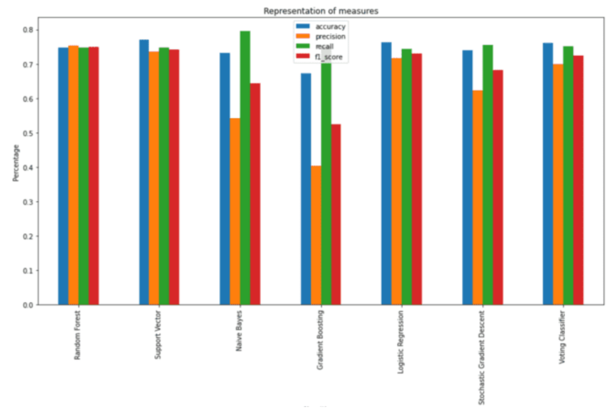


Figure 12: Bar graph comparing all the Algorithms with Performance Metrics (Dataset 3)

3.5 Conclusion

The voting classifier combines various machine learning algorithms, exploring different combinations to optimize performance. Among the combinations tested, Logistic Regression and Support Vector Machine (SVM) emerged as the most effective, yielding notable accuracy rates. Logistic Regression achieved an accuracy of 75.4, while SVM surpassed it with an accuracy of 76.93. Recognizing their superior performance, these two algorithms were selected for integration into the voting classifier. The combined model achieved a further improvement, boasting an accuracy of 80. This highlights the significance of leveraging complementary strengths from different algorithms to enhance predictive accuracy. By harnessing the power of Logistic Regression and SVM within the voting classifier, the model achieves superior performance, underscoring the importance of strategic algorithm selection and ensemble methods in machine learning optimization.

3.6 Future Scope

In the future, there is ample scope for enhancing accuracy beyond the achieved 80 percent mark by integrating deep learning algorithms. While machine learning algorithms have yielded commendable results, the complexity and non-linearity of data can be better addressed through deep learning techniques. By leveraging neural networks with their ability to capture intricate patterns, deeper insights into sentiment analysis can be attained, potentially leading to even more precise predictions.

Furthermore, the project's applicability can be expanded to encompass image and video sentiment analysis. Extending the analysis beyond textual data to include visual media opens up new avenues for understanding and interpreting human emotions expressed through different mediums. By adapting and refining existing methodologies to accommodate image and video data, the project can offer valuable insights into sentiment trends across diverse multimedia platforms, thereby broadening its impact and relevance in real-world applications.

References

[1] P. Routray, C. K. Swain, and S. P. Mishra, "A survey on sentiment analysis," *Int. J. Comput. Appl.*, vol. 76, no. 10, pp. 1–8, Aug. 2013.

[2] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (CNNL-STM)," *IEEE Access*, vol. 8, pp. 156695–156706, 2020.

[3] C. Kariya and P. Khodke, "Twitter sentiment analysis," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, pp. 212–216.

[4] A. Alsaeedi and M. Zubair, "A study on sentiment analysis techniques of Twitter data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361–374, 2019.

[5] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, and G. S. Choi, "GB-SVM: Sentiment classification from unstructured reviews using ensemble classifier," *Appl. Sci.*, vol. 10, no. 8, p. 2788, Apr. 2020.

[6] https://github.com/AnoopRao29/CS584-G23-TwitterSentiment/tree/main/Code

[7] https://github.com/AnoopRao29/CS584-G23-TwitterSentiment/tree/main/Dataset

[8] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles", *Decis. Support Syst.*, vol. 66, pp. 170–179, Oct. 2014.