Twitter Sentiment Analysis Using Machine Learning Techniques

Avi Ajmera, Steven Chang, Brahma Chilumula, Hasan Mhowwala
San Jose State University

1. Abstract

This study investigates sentiment analysis on Twitter using advanced machine learning algorithms to classify tweets into negative, neutral, or positive sentiments. Our goal was to develop a high-precision model for accurately gauging public opinion, crucial for businesses and researchers. We adopted the CRISP-DM methodology, starting with data understanding and preprocessing a dataset of 1.6 million tweets using text normalization, stopwords removal, and TF-IDF vectorization.

We explored three machine learning algorithms: Bernoulli Naive Bayes, Support Vector Machine, and Logistic Regression, assessing them based on precision, recall, accuracy, F1-score, and ROC-AUC. Logistic Regression proved the most effective, balancing accuracy with precision and recall. The project also featured a sentiment analysis application for real-time analysis, showcasing the practical utility of our research.

2. Introduction

In the digital age, social media platforms, particularly Twitter, have become a hub for public expression, making sentiment analysis increasingly vital for understanding public opinion. This project focuses on analyzing sentiments expressed in Twitter data, a challenging task due to the dynamic and nuanced nature of natural language used in social media. The importance of this analysis lies in its potential to provide businesses, researchers, and policymakers with critical insights into public sentiments, trends, and reactions to various topics or events.

Overview of Results

Our approach resulted in the successful development and deployment of a sophisticated machine learning model capable of classifying Twitter sentiments. Utilizing a dataset of 1.6 million tweets, we applied advanced

data preprocessing techniques, followed by the implementation of three machine learning algorithms. The Logistic Regression model, in particular, demonstrated outstanding performance with the highest accuracy and a robust balance between precision and recall. Additionally, the project's achievement includes the deployment of a sentiment analysis application, offering real-time insights into Twitter data.

Contribution to the Subject

This study uses a large dataset and sophisticated machine learning techniques, which makes a substantial contribution to the field of sentiment analysis. Our methodology, which combines rigorous data preprocessing with the application of multiple machine learning models, offers a replicable framework for similar sentiment analysis tasks.

Moreover, the deployment of a sentiment analysis application demonstrates the practical application of our findings, bridging the gap between theoretical research and real-world utility.

3. Related Work

The domain of sentiment analysis on social media data has been enriched by several pivotal studies, each contributing unique insights and methodologies:

Pang and Lee (2008) Their seminal work, "Opinion Mining and Sentiment Analysis," laid the groundwork for understanding sentiment analysis algorithms and applications. They explored various machine learning techniques, offering a comparative analysis that remains relevant to our approach.

Go, Bhayani, and Huang (2009) In their paper, "Twitter Sentiment Classification using Distant Supervision," they applied Naive Bayes, Maximum Entropy, and Support Vector Machines to sentiment analysis on Twitter data. Because it uses similar classification techniques and focuses on the same social networking platform, this study is especially relevant to our project.

Agarwal et al. (2011): The goal of their work, "Sentiment Analysis of Twitter Data," was to improve sentiment categorization in tweets by utilising a range of natural language processing (NLP) approaches, such as POS-tagging and N-grams. Our method for text

preparation and feature extraction was influenced by this research.

Comparison with Current Approach

Our project builds upon these foundational studies:

Data Preprocessing and NLP Techniques

Following Agarwal et al.'s approach, we implemented advanced NLP techniques like TF-IDF vectorization and lemmatization. However, we adapted these techniques to handle our large dataset, demonstrating their scalability and effectiveness.

Machine Learning Models

Echoing Go, Bhayani, and Huang's methodology, we employed Naive Bayes and SVM, but we further extended our analysis by including Logistic Regression and conducting a comparative performance evaluation, adding depth to the existing understanding of these models in Twitter sentiment analysis.

Our project contributes to the field in several unique ways:

Scale and Diversity of Dataset

We utilized a dataset of 1.6 million tweets, considerably larger than those typically used in previous studies, offering insights into sentiment analysis at scale.

Practical Application

The deployment of a sentiment analysis application underscores our project's practical relevance, transforming theoretical approaches into a usable tool, a facet often overlooked in academic research.

4. Data

Dataset Description

The dataset used in this project plays a pivotal role in the analysis and modeling of Twitter sentiment. Key characteristics of the dataset include:

Source and Composition

The dataset was sourced as a .csv file, a standard format for handling tabular data in machine learning projects. There are 1.6 million tweets in all, and each one has an emotion polarity designation (positive or negative).

Dataset Columns

target: Denotes the tweet's polarity of feeling.

ids: Tweet's identification.

date: This is the tweet's date.

flag: The query flag that was used to get the tweet.

user: The person who posted the tweet.

text: The real content of the tweet that is utilized to analyze emotion.

Data Acquisition and Preprocessing

A critical aspect of the project was the acquisition and preprocessing of the dataset:

Data Integrity

The dataset was carefully checked for completeness and consistency, with no missing values detected, ensuring a robust foundation for analysis.

Preprocessing Techniques Applied

Text normalization (converting to lowercase, removing punctuations and URLs).

Elimination of stopwords to focus on more meaningful words in the text.

Application of NLP techniques like tokenization, stemming, and lemmatization.

Transformation of text data using TF-IDF vectorization.

Data Analysis and Insights

Preliminary analysis provided key insights:

Sentiment Distribution

The dataset is evenly balanced between negative and positive sentiments, which negates the need for balancing techniques in model training.

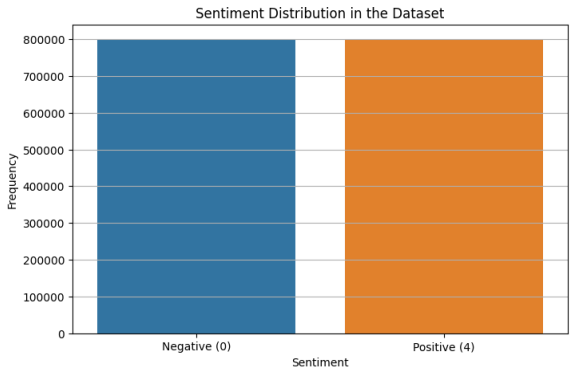

Fig. 1 : Sentiment distribution of tweets

Textual Properties: Analysis of tweet lengths and common phrases (using methods like word clouds) revealed significant variations in text length and frequent use of specific terms, shaping the approach to feature extraction.

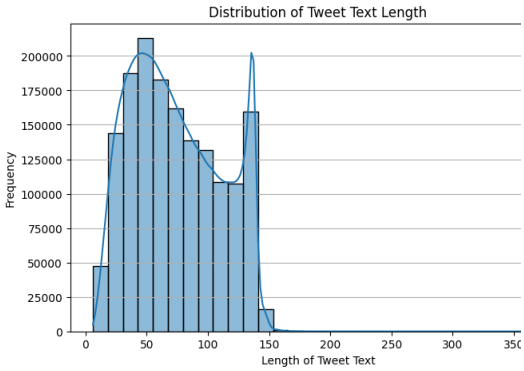

Fig.2: Distribution of tweet text length

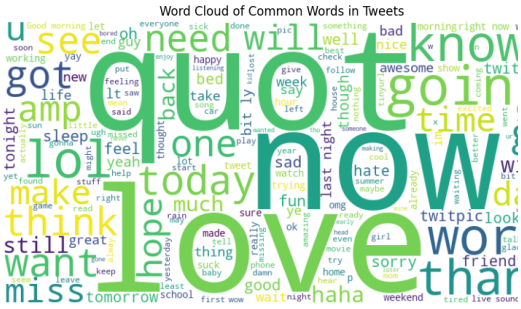

Fig. 3: Word Cloud of Common words in tweets

Implications for Modeling: The diverse and large-scale nature of the dataset presented both challenges and opportunities, ensuring that the sentiment analysis model developed would be robust and generalizable across different types of tweets.

5. Methods

Our methodological approach was grounded in the CRISP-DM methodology, ensuring a systematic and comprehensive process from data understanding to model deployment. Key aspects of our methods are outlined below:

Machine Learning Models

Bernoulli Naive Bayes: Chosen for its efficiency in handling binary classification problems and its robustness to irrelevant features.

Support Vector Machine (SVM): Selected for its effectiveness in high-dimensional spaces, which is characteristic of text data.

Logistic Regression: Employed for its ability to provide probabilities for outcomes, allowing for a more nuanced understanding of sentiment classifications.

These models were chosen due to their proven effectiveness in sentiment analysis tasks, as evidenced by previous studies. Additionally, the variety of models allowed for a comparative analysis to determine the best fit for our dataset.

Data Preprocessing and Feature Engineering

Text Preprocessing: Included converting text to lowercase, removing punctuation, URLs, and numbers, and normalizing characters.

Feature Extraction: Utilized TF-IDF vectorization to transform text data into a format suitable for machine learning algorithms. This method was chosen for its ability to reflect the importance of words in relation to the dataset as a whole.

Model Training and Validation

Training Process: The models were trained on a subset of the dataset, with rigorous tuning of parameters to optimize performance.

Validation Strategy:Used cross-validation methods to guarantee the robustness against overfitting and generalizability of the model.

Alternative Approaches Considered

Deep Learning Models: While deep learning models like LSTM and CNN were considered, they were not pursued due to the increased computational resources and complexity they entail.

Feature Selection Techniques: Alternative feature selection methods, such as word embeddings, were explored but not implemented in the final model due to the satisfactory performance of TF-IDF vectorization.

The project saw the introduction of several innovative approaches:

Balanced Dataset Utilization: Leveraging an evenly distributed dataset in terms of sentiment polarity, which is relatively rare in sentiment analysis projects.

Feature Engineering: The combination of various text preprocessing and feature extraction techniques provided a rich set of features for the models to learn from.

Experimental Design

Our experimental approach was methodically structured to validate the effectiveness of the chosen machine learning models in sentiment analysis of Twitter data:

Dataset Splitting: The dataset was split into training (95%) and testing (5%) sets, ensuring a substantial amount of data for model training while retaining an adequate portion for unbiased evaluation.

Model Training: Each model (Bernoulli Naive Bayes, SVM, Logistic Regression) was trained on the same dataset, allowing for a direct comparison of their performance.

Performance Metrics: The models were assessed and compared using important

metrics such accuracy, precision, ROC-AUC, recall, and F1-score.

Key Experiments Conducted

Model Comparison: The primary experiment involved comparing the three models on the same dataset to identify which model performed best in terms of accuracy and other metrics.

Feature Impact Analysis: We examined the impact of different feature engineering techniques, such as TF-IDF vectorization, on model performance.

Parameter Tuning: Each model was fine-tuned to find the optimal set of parameters for our specific dataset and task.

Bernoulli Naive Bayes showed good efficiency but was slightly less accurate.

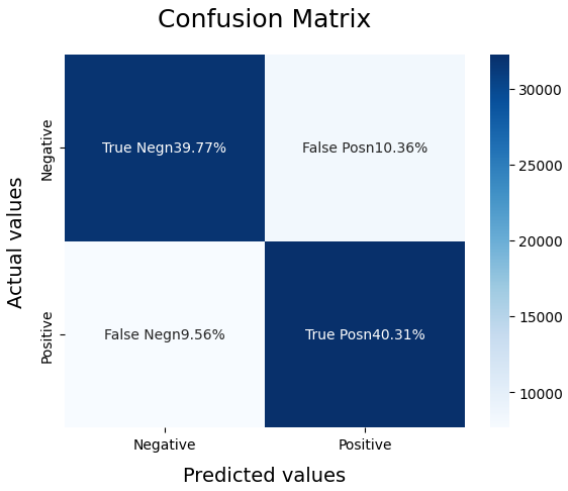

Fig. 4: Confusion Matrix -BNB

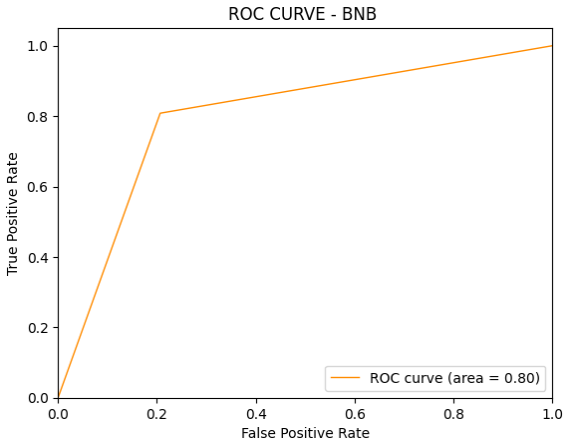

Fig. 5 : ROC Curve -BNB

6. Results

SVM provided a balance between accuracy and computational efficiency.

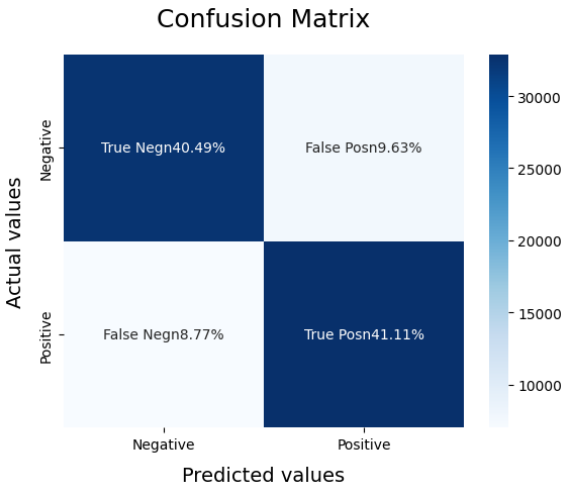

Fig 6 : Confusion Matrix-SVM

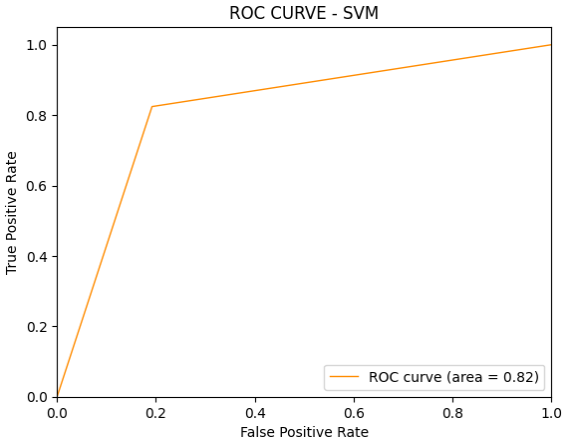

Fig. 7 : ROC Curve -SVM

Logistic Regression achieved the highest accuracy and a strong balance in precision and recall but required the longest computation time.

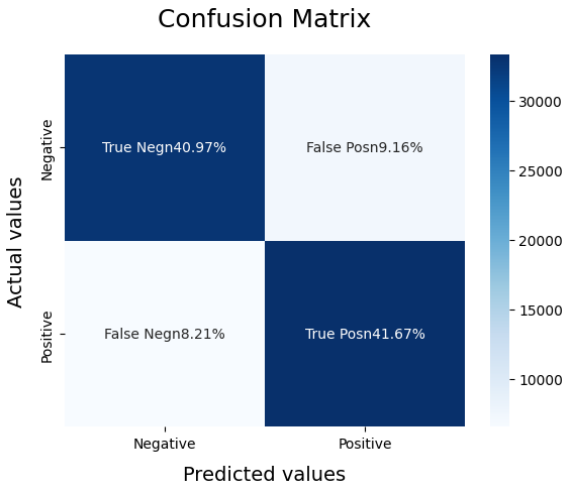

Fig. 8 : Confusion Matrix- LR

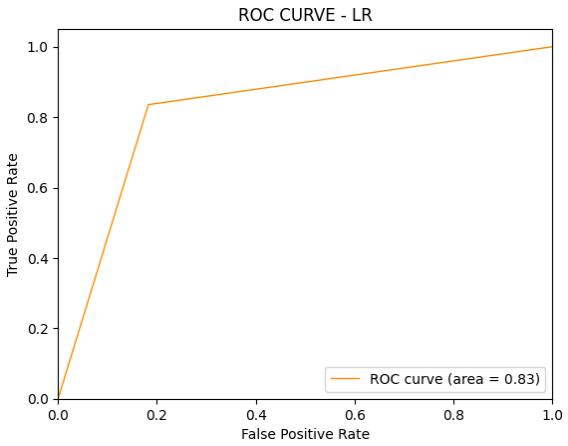

Fig. 9 : ROC Curve -LR

Insights Gained

Feature engineering played a crucial role in the performance of all models.

Parameter tuning was essential in achieving the highest possible accuracy.

Visualizations: Performance metrics were visualized using confusion matrices, ROC curves, and bar charts

for an intuitive understanding of each model's strengths and weaknesses.

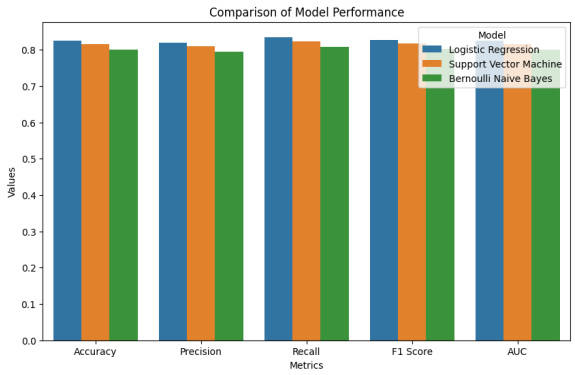

Fig. 10 : Model Metrics Comparison

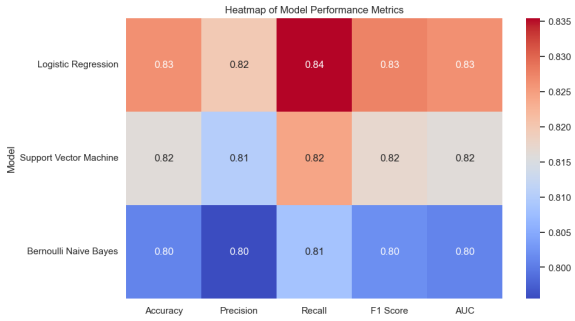

Fig. 11 : HeatMap of Model Performance Metrics

Model Selection and Practical Implications

Bernoulli Naive Bayes is recommended for scenarios where speed is a priority over slight decreases in accuracy.

Logistic Regression is ideal for applications where the highest accuracy is needed, despite longer processing times.

SVM strikes a balance and is suitable for cases where a compromise between speed and accuracy is acceptable.

7. Conclusion

Summary of Key Findings

This project embarked on an ambitious journey to explore sentiment analysis of Twitter data using advanced machine learning techniques. Key findings from this study include:

Model Efficacy: Among the models tested – Bernoulli Naive Bayes, SVM, and Logistic Regression – the Logistic Regression model demonstrated the highest accuracy and a balanced precision-recall profile, making it the most effective model for our sentiment analysis task.

	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)
0	Logistic Regression	82.63	81.98	83.54	82.75	82.63
1	Support Vector Machine	81.60	81.02	82.42	81.72	81.60
2	Bernoulli Naive Bayes	80.08	79.56	80.83	80.19	80.08

Table 1: Model Metrics Summary

Importance of Preprocessing: The extensive data preprocessing and feature engineering phases, including text normalization, stopwords removal, and TF-IDF vectorization, were

instrumental in enhancing model performance.

Practical Application: The deployment of a sentiment analysis application illustrated the practical utility of our models, showcasing their potential in real-world scenarios.

Lessons Learned

Throughout the course of this project, several key lessons were learned:

Data Quality and Preparation: The significance of a well-prepared and balanced dataset cannot be overstated. It forms the backbone of any successful machine learning project.

Model Selection: Different models offer different trade-offs between speed, accuracy, and complexity. Understanding these trade-offs is crucial in selecting the right model for specific requirements.

Application of Theoretical Knowledge: The transition from theoretical models to practical applications presents its own set of challenges and learnings, emphasizing the importance of adaptability and continuous testing.

Future Work and Extensions

Looking forward, there are several potential avenues for extending this project:

Exploring Deep Learning: Implementing deep learning models like CNNs or LSTMs could further improve the accuracy of sentiment analysis, especially in capturing the nuances of natural language.

Real-Time Analysis: Enhancing the sentiment analysis application for real-time data processing and visualization could offer more immediate insights into public opinion trends.

Cross-Platform Analysis: Incorporating information from additional social media sites into the investigation would yield a more all-encompassing perspective of popular opinion.

Final Thoughts

In summary, this research constitutes a noteworthy advancement in the comprehension and evaluation of public opinion on social media platforms. The results and approaches described in this paper open up new avenues for practical applications and further study in the dynamic and developing field of sentiment analysis. They also provide insightful information to the field.

8. References

1. Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Oxford University Press.
2. Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision*. Stanford University.
3. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). *Sentiment Analysis of Twitter Data*. Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics.
4. Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC'10).
5. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1), 1-8.
6. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.