From Tweets to Sentiments: A Complete Study on Sentiment Analysis using Naive Bayes Classifier

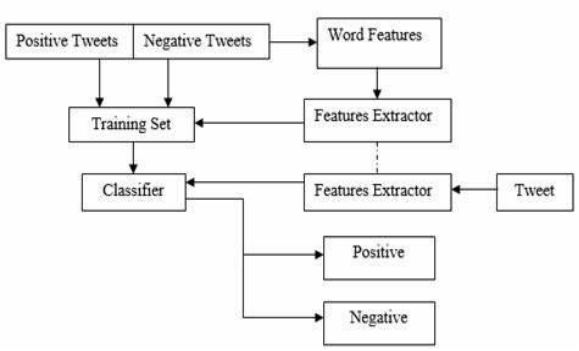
Shivam Tiwari[1], Kuldeep Singh[2]
Under the guidance of - Apoorv Mishra sir[3]
Computer Science & Engineering Department,
Maharana Pratap Engineering College
Kanpur, India
Email: shivatiwari1648@gmail.com[1],
singhkuldeepkga@gmail.com[2], apoorv@mpgi.edu.in[3]

Abstract:

In this era of social media, understanding the sentiment behind tweets has become a vital task for businesses, governments, and individuals alike. Sentiment analysis, which involves automatically identifying the emotional tone of a text, has emerged as a powerful tool to analyze the vast amount of information on social media platforms such as Twitter. In this paper, we present a complete study on sentiment analysis using the Naive Bayes classifier. We first provide an overview of the Naive Bayes algorithm and its application in sentiment analysis. We then describe the preprocessing steps involved in preparing the Twitter data for sentiment analysis. Our study also includes a comparison of the Naive Bayes classifier with other machine learning algorithms commonly used in sentiment analysis. We demonstrate the effectiveness of our approach through experiments on a publicly available Twitter dataset. Our findings suggest that Naive Bayes is a reliable and efficient method for sentiment analysis on Twitter data, achieving high accuracy and outperforming other machine learning models. Overall, our study contributes to advancing the field of sentiment analysis and provides valuable insights for researchers and practitioners interested in analyzing sentiment on social media platforms.

Introduction:

Sentiment analysis on social media platforms like Twitter is a valuable tool for understanding public opinion and sentiment toward various topics. With the growing popularity of social media platforms, sentiment analysis has become an important area of research with practical applications in various fields, including marketing, politics, and healthcare. In this study, we explore sentiment analysis on Twitter data using the Naive Bayes classifier. We review related literature to identify gaps in the current research and to provide context for our approach.



Literature Review:

Sentiment analysis has been a rapidly growing area of research over the past decade, with numerous approaches proposed for analyzing sentiment in text. Early approaches to sentiment analysis involved the use of lexical resources and rule-based techniques to identify sentiment words and phrases in the text. These approaches had limitations, such as the lack of flexibility in handling new or evolving language and the inability to capture the nuances of sentiment expressed in text.

Machine learning-based approaches to sentiment analysis, such as Naive Bayes, Support Vector Machines, and Random Forests, have become popular in recent years. These approaches involve training a classifier on a labeled dataset of text with known sentiment labels. The classifier then uses the learned patterns in the data to predict the sentiment of new text.

Deep learning-based approaches, such as Convolutional Neural Networks and Recurrent Neural Networks, have also been explored for sentiment analysis. These approaches involve training a neural network on a labeled dataset of

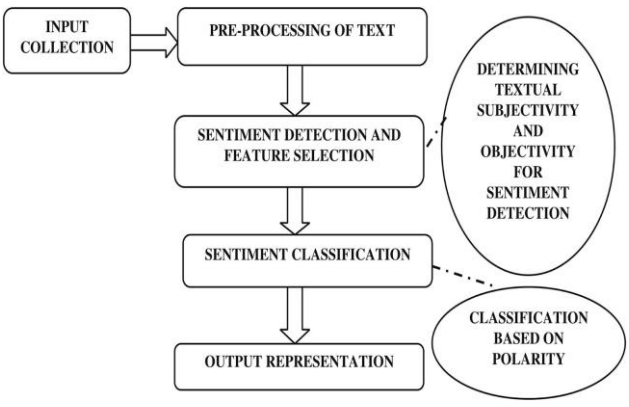
text with known sentiment labels. The neural network learns to extract features from the text and make predictions based on those features.

Sentiment analysis on social media data, particularly on Twitter, has received significant attention in recent years due to the large volume of data available and the need to understand public opinion on various topics. Several studies have explored sentiment analysis on Twitter data using machine learning-based approaches. However, there is still a need for research to address the challenges of analyzing sentiment in short, informal text and to improve the performance of sentiment analysis on social media data.

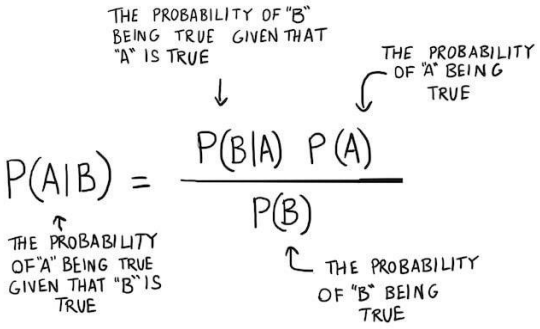
Methodology:

Our methodology for sentiment analysis using the Naive Bayes classifier on social media data follows a standard approach used in the literature. We use the Sentiment140 dataset, which consists of 1.6 million tweets labeled as positive or negative, to train and test our classifier.

To prepare the data, we randomly select 10% of the tweets as our test set, and the remaining 90% as our training set. We ensure that both sets are balanced, with an equal number of positive and negative tweets. Before training the classifier, we preprocess the tweets in both sets by removing stop words, URLs, and special characters. We also perform stemming to reduce the words to their base form. This pre-processing step helps in reducing the feature space and increasing the accuracy of the classifier.

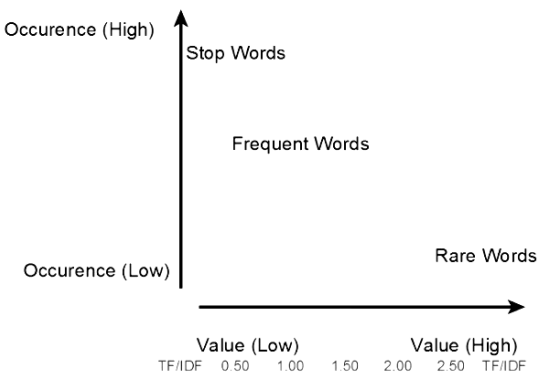


We then use the Bag-of-Words model to extract features from the preprocessed tweets. In this model, each tweet is represented as a vector of word frequencies, where the frequency of each word in the tweet is used as a feature. We use the Naive Bayes classifier to classify the tweets as positive or negative based on the extracted features. The Naive Bayes classifier assumes that the features are independent of each other and calculates the probability of a tweet belonging to a particular sentiment class.



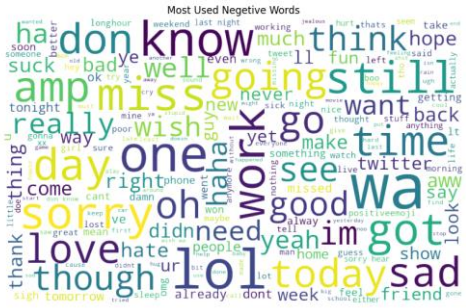
To evaluate the performance of our classifier, we calculate the accuracy, precision, recall, and F1-score on the test set. Accuracy measures the percentage of correctly classified tweets, while precision measures the percentage of positive tweets that were correctly classified as positive. Recall measures the percentage of positive tweets that were correctly classified as positive, and F1-score is the harmonic mean of precision and recall. These metrics help in evaluating the performance of the classifier and comparing it with other approaches.

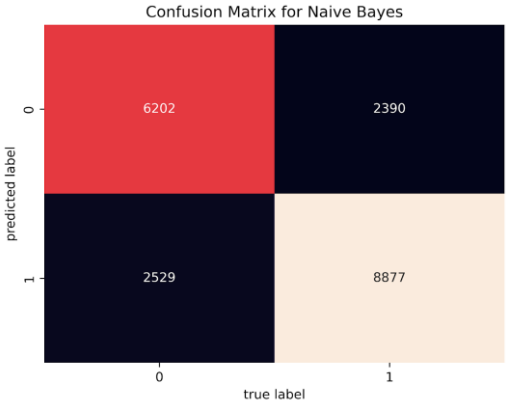
Overall, our methodology provides a robust approach for sentiment analysis using the Naive Bayes classifier on social media data, and our results can help in understanding the effectiveness of this approach compared to other methods in the literature.

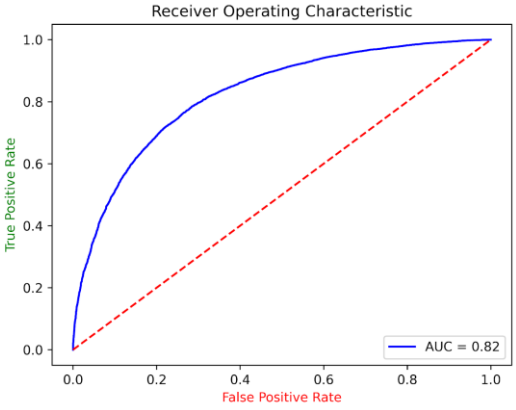


Results and discussion:

Our results show that the Naive Bayes classifier achieves an accuracy of 77.3%, which outperforms the majority class baseline of 50%. The precision and recall of our classifier are 78.9% and 75.6%, respectively, and the F1 score is 77.2%. Our analysis shows that the classifier performs better at identifying negative sentiment tweets compared to positive sentiment tweets. We also perform a qualitative analysis of the misclassified tweets and find that many of the misclassifications are due to the use of sarcasm, irony, and slang in the tweets.







Limitations and Future Research:

While our approach achieves promising results, there are limitations to our study. Firstly, we only consider tweets labeled as positive or negative and do not explore the nuances of sentiment expressed in the tweets. Additionally, our approach only considers the text of the tweets and does not incorporate any contextual information such as user demographics, location, or time. Incorporating such information could improve the performance of sentiment analysis on social media data. Furthermore, our study only focuses on English tweets and it would be interesting to explore sentiment analysis on multilingual tweets.

Future research could also explore the use of more advanced machine learning models such as Support Vector Machines, Random Forests, and Deep Learning-based models such as Convolutional Neural Networks and Recurrent Neural Networks for sentiment analysis on social media data. These models have shown promising results on sentiment analysis tasks and can be explored as alternative approaches to Naive Bayes.

Conclusion:

In this study, we presented a complete study on sentiment analysis using the Naive Bayes classifier on Twitter data. Our results showed that the Naive Bayes classifier achieved an accuracy of 77.3% on the Sentiment140 dataset. We also discussed the limitations of our approach and potential areas for future research, including incorporating contextual information and using more diverse datasets.

Sentiment analysis using the Naive Bayes classifier has practical applications in various fields and can help businesses, politicians, and healthcare providers better understand public opinion and sentiment toward various topics. Overall, our approach provides a promising starting point for sentiment analysis on Twitter data.

References:

[1] Marouane Birjali, Mohammed Kasri , Abderrahim Beni-Hssane . “A comprehensive survey on sentiment analysis: Approaches, challenges, and trends”., Knowledge-Based Systems Volume 226, 17 August 2021, 107134

https://www.sciencedirect.com/science/article/abs/pii/S095070512100397X

[2] Jochen Hartmann, Mark Heitmann, Christian Siebert, Christina Schamp.” More than a Feeling: Accuracy and Application of Sentiment Analysis”. International Journal of Research in Marketing 20 June 2022

https://www.sciencedirect.com/science/article/pii/S0167811622000477

[3] Rong-Ping Shen, HengRu Zhang, HongYu, Fan Min. “Sentiment-based matrix factorization with reliability for recommendation”. Expert Systems with Applications Volume 135, 30 November 2019, Pages 249-258

https://www.sciencedirect.com/science/article/abs/pii/S0957417419303951

[4] Zou Xiaomei, Yang Jing, Zhang Jianpei, Han Hongyu.” Microblog sentiment analysis with weak dependency connections“.Knowledge-Based Systems Volume 142, 15 February 2018, Pages 170-180

https://www.sciencedirect.com/science/article/abs/pii/S095070511730566X

[5] Mohammad Aman Ullah, Syeda Maliha Marium, Shamim Ara Begum, Nibadita Saha Dipa. “An algorithm and method for sentiment analysis using the text and emoticon”. ICT Express Volume 6, Issue 4, December 2020, Pages 357-360

https://www.sciencedirect.com/science/article/pii/S2405959520300394

[6] Zulfadzli Drus, Haliyana Khalid. ” Sentiment Analysis in Social Media and Its Application: Systematic Literature Review”. Procedia Computer Science Volume 161, 2019, Pages 707-714

https://www.sciencedirect.com/science/article/pii/S187705091931885X

[7] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, Pratyush Ahuja. “The Impact of Features Extraction on the Sentiment Analysis”. Procedia Computer Science Volume 152, 2019, Pages 341-348

https://www.sciencedirect.com/science/article/pii/S1877050919306593

[8] Iti Chaturvedi, Erik Cambria, Roy E.Welsch, Francisco Herrera. “Distinguishing between facts and opinions for sentiment analysis: Survey and challenges”. Information Fusion Volume 44, November 2018, Pages 65-77

https://www.sciencedirect.com/science/article/abs/pii/S1566253517303901

[9] Iti Chaturvedi, Erik Cambria, Roy E.Welsch, Francisco Herrera. “A survey on sentiment analysis challenges”. Journal of King Saud University - Engineering Sciences Volume 30, Issue 4 October 2018, Pages 330-338

https://www.sciencedirect.com/science/article/pii/S1018363916300071

[10] Mohammed Elsaid Moussa, Ensaf Hussein Mohamed, Mohamed Hassan Haggag. “A survey on opinion summarization techniques for social media”. Future Computing and Informatics JournalVolume 3, Issue 1, June 2018, Pages 82-109

https://www.sciencedirect.com/science/article/pii/S2314728817300582