WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

TWEET SENTIMENT ANALYSIS USING SVC, LOGISTIC REGRESSION AND BERNOULLI NB

Rohit Vijay[a], Aditya Singh Rawat[b], Manpreet Singh[c]

[a,b] UG Student, Department of Computer Applications, Maharaja Surajmal Institute, New Delhi
[c]Assistant Professor, Department of Computer Applications, Maharaja Surajmal Institute, New Delhi

ABSTRACT

This research paper focuses on tweet sentiment analysis, the process of automatically categorizing tweets as positive, negative. It explores the techniques, challenges, and applications of analyzing sentiments expressed in short-form text on social media platforms like Twitter. By leveraging machine learning businesses can extract valuable insights from tweets to make informed decisions, enhance customer engagement, and manage online reputation effectively. This study aims to provide insights into tweet sentiment analysis methodologies and their implications for businesses and individuals in the dynamic realm of social media.

Keywords: Tweet , sentiment analysis, logistic regression, SVC, Bernoulli NB

Introduction

Social media platforms have become an integral part of modern communication, providing a vast and dynamic landscape for individuals and businesses to express opinions, share experiences, and engage with a global audience. The proliferation of user-generated content on these platforms has led to an exponential growth in the amount of data available for analysis, presenting both opportunities and challenges for organizations seeking to understand the sentiments and preferences of their target audience . Sentiment analysis, the process of extracting and categorizing the emotions and opinions expressed by users, has emerged as a crucial tool in this digital age. By analysing the sentiments expressed on social media, businesses can gain valuable insights into customer feedback, brand perception, and emerging trends, enabling them to make more informed decisions and enhance their overall strategy .This research paper delves into the multifaceted world of sentiment analysis on social media platforms, exploring the various techniques, challenges, and applications of this dynamic field. Through a comprehensive review of existing literature and a detailed discussion of the research findings, this paper aims to provide a holistic understanding of the significance of sentiment analysis in the context of social media and its implications for businesses and individuals navigating the digital landscape.

Research Objective

1. Explore the importance of sentiment analysis on social media platforms. This study aims to explore the importance of sentiment analysis in understanding user sentiment, monitoring brand awareness, and making informed decisions in the digital age.
2. Learn about the methods and tools used in sentiment analysis. The goal of this study is to explore a wide range of technologies , machine learning, and, and tools used for sentiment analysis on social media platforms.
3. Identify problems when interpreting user sentiment. The goal is to identify and analyze the challenges faced in accurately interpreting user sentiment in social media, including linguistic nuances, cultural differences, and understanding context.
4. Discuss ethical considerations when performing sentiment analysis. This study aims to ensure responsible and ethical use of sentiment analysis techniques by addressing ethical considerations associated with sentiment analysis, including user privacy, data security, and bias mitigation.
5. We highlight future research directions in sentiment analysis. The goal is to suggest potential future research directions, including advanced sentiment analysis techniques, multilingual sentiment analysis, and integration with other data.

Research Significance

Sentiment analysis of Twitter tweets is important in various fields such as social media analysis, market research, brand monitoring, and public opinion analysis. This study contributes to the advancement of Twitter sentiment analysis methodology and provides several key findings and implications:

1) Information for decision making: Inan age of widespread social media influence, understanding public sentiment and opinions is paramount to making informed decisions across a variety of sectors. Sentiment analysis provides valuable insights into consumer attitudes, market trends and brand awareness, allowing businesses to tailor strategies, products and services to customer needs and preferences.

2) Improved social media analytics: Social media platforms like Twitter act as virtual echo chambers where people express their thoughts, feelings, and opinions on a variety of topics. Sentiment analysis makes it easier to extract actionable information from vast amounts of user-generated content, allowing researchers, analysts, and organizations to decipher sentiment trends, identify influencers, and gauge public reactions to events and phenomena in real time. You can measure it.

3) Monitor public discussion: Twitter has become an important forum for public discourse, serving as a platform for political debate, social movements, and cultural debate. Sentiment analysis allows politicians, government agencies, and advocacy groups to monitor public sentiment, gauge public opinion on political issues, and proactively respond to emerging issues and sentiments in society.

4) Advances in computational linguistics: Sentiment analysis presents unique challenges and opportunities for advancement in the field of computational linguistics. By developing and evaluating machine and deep learning models for sentiment classification, researchers are contributing to the development of powerful, scalable, and interpretable algorithms for text data analysis. Sentiment analysis also serves as a testing ground for exploring new technologies, methodologies, and models in natural language processing and artificial intelligence research.

5) Promote research and innovation: The results and methodology presented in this research article provide a foundation for future research and innovation in sentiment analysis research. Based on our work, researchers can explore new research questions, data sets, and methodologies and further expand our understanding of the emotional dynamics of social media discourse. Additionally, open access to this research promotes collaboration, replication, and dissemination of results within the scientific community, thereby fostering interdisciplinary research and innovation in sentiment analysis

Literature Review

Sentiment analysis, also known as sentiment mining, has received significant attention in natural language processing (NLP) and speech processing. With the growth of social media platforms such as Twitter, sentiment analysis has become an important tool to disseminate valuable information from a wealth of user-generated content. In this section, we review the existing literature on the sentiment of Twitter tweets, highlighting the main methods, methods, and results of previous studies. Support vector machines (SVM), Naive Bayes, logistic regression, and decision trees are popular machine learning algorithms for problem classification.

1. Researchers such as Aliza and Shuib (2014) used NLP, CBR, SVM and ANN to analyze sentiment on Twitter and achieved good results in classifying tweets as positive, negative and neutral. This article discusses Twitter sentiment analysis and highlights the importance of understanding consumer sentiment through social media. It presents challenges such as informal speech and expressions. This study aims to develop a sentiment analysis system that separates customer comments into positive and negative sentiments. The process involves using Python to extract tweets, provide emotional results, and classify them. This research focuses on sentiment analysis on Weibo, with the aim of creating a project to analyze positive customer feedback
2. Arwa alshamsi and Reem Bayari (2020) investigate decision trees, naive Bayes, Random Forest, Sensitivity Analysis, K-NN for ID3 and Random Trees. Yes, Random Forest, K-NN, ID3, Random Tree) are used to classify tweets as good, bad or neutral1. Achieve up to 35% accuracy at best. On non-equivalent data, Naive Bayes and ID3 perform best with 82.7% accuracy 1. This article concludes that hypothesis analysis is an important tool for understanding public sentiment on social media
3. Munir Ahmad, Shabib Aftab and Iftikhar Ali (2017) explore SVM for sentiment discuss the importance of sentiment analysis in obtaining useful insights from social media data. It focuses on sentiment analysis using support vector machines (SVMs). and compared its performance using the previously mentioned tweet data. This study highlights the importance of sentiment analysis to develop marketing strategies and improve products/services based on the evaluation accuracy of various machine learning algorithms such as SVM, Naive Bayes. and Maximum probability are compared. The article also discusses the prioritization process, classification, and evaluation results, focusing on the measurement of Precision, Recall, and F-Measure. In addition, it touches on sentiment analysis studies and shows the progress of sentiment classification using machine learning algorithms.

4. The development of distribution theory for analyzing Twitter profiles using KNN and SVM algorithms is discussed. The authors present a strategy to identify sentiment signals in tweets using popular text and descriptive text. Their goal is to improve the model by adding more features, working with other languages, and incorporating emojis. This study focused on separating tweets into positive and negative sentiment groups, and the Sentiment Classifier Algorithm (SCA) outperformed SVM. The paper also introduces a two-step analysis classification method for Twitter using popularity data as reference data.
5. The authors emphasize the importance of analysis and meta-feature to capture the abstract representation of tweets. The power of thought comes from Twitter data. This study compares the performance of algorithms such as Bernoulli Naive Bayes, SVM and LSTM in analyzing user sentiment. He talks about the process of scraping data from Twitter, cleaning foreign words, translating Nepali sentences into English before doing steps like tokenization and lemmatization. The research also addresses the use of NLTK and iNLTK libraries for tokenization and stopping message deletion. It is based on SVM and Bernoulli Naive Bayes model. The paper shows that the LSTM model performs best across the entire set of data constraints. It also shows the potential for future improvement by testing the algorithm with larger datasets to increase accuracy. This study highlights the importance of emotional analysis in understanding customer evaluations and improving the organization's customer satisfaction.

Methodology

The methodology section describes the procedures and methods used to conduct the study, including data collection, preprocessing, feature extraction, model selection, and evaluation.

2.1 Data collection:

- We used Kaggle to collect a dataset of Twitter tweets, focusing on tweets related to specific topics or hashtags relevant to our research objectives. The dataset contains a wide range of tweets from different domains and topics, allowing a comprehensive coverage of emotional expressions.

2.2 Data preprocessing:

- Prior to analysis, we preprocessed the collected Twitter dataset to clean and standardize the text data. This involved several steps, including: - Noise Removal: Removes unnecessary information such as URLs, mentions, special characters, and punctuation marks. - Tokenization: Dividing text into individual tokens (words or phrases) for further processing. - Normalization: Normalize the text by converting all characters to lowercase to ensure uniform feature extraction. - Emoticon and slang processing: Encode emoji and slang expressions to preserve semantic meaning in sentiment analysis. These preprocessing steps were necessary to ensure the quality and consistency of the text data for subsequent analysis.

2.3 Feature extraction:

- Feature extraction is a critical step in sentiment analysis as it involves converting raw text data into a numerical representation that can be used as input to machine learning and deep learning models. In this study, we explored various feature extraction methods, including: - Bag of Words (BoW): Represents each tweet as a vector of word frequencies, regardless of word order or context. - TF-IDF (Term Frequency, Inverse Document Frequency): Weights words based on their frequency within the document and their inverse frequency in the entire corpus. - Word embeddings: Create dense, low-dimensional representations of words using methods such as Word2Vec or GloVe that capture the semantic relationships between words. These feature extraction methods capture various aspects of text information and facilitate sentiment classification tasks.

2.4 Model selection:

- For sentiment classification, we experimented with both traditional machine learning algorithms. The following models were considered: Traditional Machine Learning: Support Vector Machines (SVM), Bernoulli N.B., Logistic Regression. These models were selected based on their suitability for sentiment analysis tasks and their performance in previous studies.

2.5 Model training and evaluation:

- The selected models were trained on the preprocessed Twitter dataset using appropriate training and validation splits. To optimize model performance, hyperparameter tuning was performed using methods such as grid search or random search. The trained model was then evaluated on a separate test set using standard evaluation metrics including precision, precision, recall, F1 score, and confusion matrix. Computational efficiency and scalability were also considered during the evaluation process.

2.6 Experimental setup:

- The experimental setup included setting up computing resources, software environment, and data pipeline for data processing, model training, and evaluation. To implement the sentiment analysis pipeline, the Python programming language and popular libraries such as scikit-learn, TensorFlow, and were used.
- 2.7 Ethical considerations: Ethical considerations, including data confidentiality, informed consent, and responsible use of social media data, were observed throughout the research process. The Twitter dataset used in this study was anonymized and aggregated to protect user privacy and confidentiality. This methodology section provides a detailed overview of the procedures and methods used to perform sentiment analysis of tweets on Twitter and lays the foundation for subsequent analysis and discussion in the research work.

3.Results

We present the results of sentiment analysis experiments conducted using logistic regression, Bernoulli Naive Bayes (NB), and Support Vector Machines (SVM) on Twitter tweets dataset. The performance of each model was evaluated based on standard evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrix.

Performance metrics:

Model name	**accuracy**	**precision**	**recall**	**F1 score**
Logistic regression	**0.79**	**0.77**	**0.80**	**0.79**
SVM	**0.77**	**0.76**	**0.78**	**0.77**
Bernoulli NB	**0.77**	**0.77**	**0.78**	**0.77**

Confusion matrix:

Logistic regression

184653	**45509**
55224	**194614**

SVC

182119	**50554**
57758	**189569**

Bernoulli NB

184427	**52712**
55450	**187411**

Upon analyzing confusion matrix, logistic regression performs the best

4.Conclusion:

In this research work, we conducted a comprehensive study on sentiment analysis of Twitter tweets to investigate the effectiveness of machine learning and deep learning approaches. Through rigorous experimentation, rigorous evaluation, and in-depth analysis, we have gained valuable insights into Twitter sentiment analysis methodology, techniques, and models.

1) Summary of results: Our results showed that logistic regression was the best performing model among the evaluated algorithms. It achieved the highest accuracy, precision, recall, and F1 rank on a sentiment classification task on the Twitter dataset. Logistic regression performed well at capturing the nuances of sentiment expressed in Twitter tweets, outperforming the results of Naive Bayes Bernoulli and Support Vector Machines.

2) Implicationsand contributions: Our findings have important implicationsfor research in sentiment analysis, social network analysis,

and computational linguistics. By demonstrating the effectiveness of logistic regression in recognizing the sentiment of Twitter tweets, we provide practitioners and researchers with a powerful and interpretable model for analyzing public sentiment on social media platforms. Additionally, our study contributes to the development of sentiment analysis methodologies by revealing the pros and cons of machine learning and deep learning approaches when analyzing emotionally rich text data.

3) Future directions: Although our study has provided valuable information, severalavenues are open forfuture research and research. Fut ure research could explore ensemble methods, transfer learning methods, and domain adaptation strategies to improve the robustness and generalizability of Twitter sentiment analysis models. Moreover, the integration of multimodal data sources such as images, videos, and emoticons presents an interesting direction for extending sentiment analysis beyond textual content.

4) Ethical considerations: While conducting this study, we adhered to ethical principles and guidelines regarding the responsible use of social media data. We ensure confidentiality, anonymity and informed consent at all stages of data collection, pre- processing and analysis, and give priority to users' privacy and confidentiality.

5) Conclusion:In conclusion, our research work presents a comprehensive study on sentiment analysis of Twitter tweets and highlights the effectiveness of logistic regression as a preferred model for sentiment classification tasks. Using machine learning and deep learning approaches, we provide valuable insights into the dynamics of sentiment expressions on Twitter and contribute to the development of sentiment analysis methodologies in social media analytics. We expect that continued research and innovation in the field of sentiment analysis will provide a deeper understanding of public sentiment and opinions in the digital age.

REFERENCES

1) Arwa Alshamsi, Reem Bayari, Said Salloum,3, “Sentiment Analysis in English Texts” Advances in Science, Technology and Engineering Systems Journal Vol. 5, No. 6, 1683-1689 (2020)
2) Aliza Sarlan, Chayanit Nadam, Shuib Basri “Twitter Sentiment Analysis” 2014 International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia
3) Munir Ahmad, Shabib Aftab, Iftikhar Ali “Sentiment Analysis of Tweets using SVM” International Journal of Computer Applications (0975 – 8887) Volume 177 – No.5, November 2017
4) Mohammad Rezwanul , Ahmad Ali, Anika Rahman “Sentiment Analysis on Twitter Data using KNN and SVM”
5) Milan Tripathi “Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM” Journal of Artificial Intelligence and Capsule Networks (2021)
6) https://www.kaggle.com/datasets/kazanova/sentiment140?select=training.1600000.processed.noemoticon.csv
7) https://github.com/rv-2003/tweet-sentiment-analysis.git