Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 4, April 2024

Sentiment Insights Unveiling Text Emotions Through Machine Learning

Abhiram Vaidya
Department Of Computer Engineering
Dr.D.Y.Patil Institute Of Technology
Pimpri, Pune, India

Pankaj Patil
Department Of Computer Engineering
Dr.D.Y.Patil Institute Of Technology
Pimpri, Pune, India

Vaibhav Sawant
Department Of Computer Engineering
Dr.D.Y.Patil Institute Of Technology
Pimpri, Pune, India

Shreenath Khadap
Department Of Computer Engineering
Dr.D.Y.Patil Institute Of Technology
Pimpri, Pune, India

Asst.Prof Prachi Karale
Department Of Computer Engineering
Dr.D.Y.Patil Institute Of Technology
Pimpri, Pune, India

***Abstract*—Sentiment analysis, a pivotal tool in understanding human emotions and opinions, is the focus of this research paper. The project, developed by a team of four members, aims to provide a web-based application for sentiment analysis with various features. The paper explores the significance of sentiment analysis and the motivation behind the project.**

The project encompasses several key features. Firstly, the sentiment analysis feature categorizes text into positive, negative, or neutral sentiments and displays the corresponding emoji for each result. This classification is achieved using the textblob.polarity library for prediction. Secondly, the emotional analysis feature delves deeper into sentiment analysis by categorizing sentiments into emotions such as happy, love, worry, hate, and sadness. This feature utilizes SVM and LSTM models and provides emoji's for each emotion. Additionally, it facilitates the analysis of tweets, allowing for the manual addition of tweets or their import from Twitter using an API (although the API is currently unavailable).

Another significant feature of the project is the product reviews sentiment analysis, which includes a word cloud and sentiment analysis to classify words based on emotions. It also analyzes product reviews from customers for a list of products and utilizes data visualization to display negative and positive reviews.

Furthermore, the project includes audio sentiment analysis, which enables users to utilize call recordings, convert them into transcripts, and conduct sentiment analysis. It uses a pie chart to illustrate whether the sentiment is positive or negative and highlights positive and negative lines in the transcript. This feature utilizes NLTK for transcript processing and sentiment analysis.

Lastly, the project includes story sentiment analysis, which evaluates the sentiment of a story and character sentiment using a story listener. It uses a web API to fetch stories and provides features such as comparing two characters and the sentiment of the story at intervals. Data visualization is used to present all this information.

***Index Terms*—: Sentiment Analysis, Emotion analysis, LSTM (Long Short-Term Memory), Support Vector Machines, Preprocessing, Tokenization, Stemming , Lemmatization, Twitter, NLP.**

I. INTRODUCTION

Sentiment analysis is a vital component of natural language processing (NLP), especially in understanding human emotions expressed in text. With the increasing use of social media and online platforms, there is a growing need for businesses, policymakers, and researchers to comprehend public sentiment towards various topics, products, and services.

This paper presents a study on sentiment analysis, focusing on the development of a web-based application with advanced features for analyzing sentiment in text data. The application, named "Sentiment Analysis and Dashboard," utilizes machine learning techniques to categorize text into positive, negative, or neutral sentiment categories.

The application provides several features to enhance sentiment analysis. Users can input text, which the application classifies into positive, negative, or neutral sentiment categories using the textblob.polarity library. Additionally, the application includes an emotional analysis feature, which categorizes sentiment into emotions such as happiness, love, worry, hate, and sadness using SVM and LSTM models, and provides corresponding emojis for each emotion.

Another key feature of the application is the product reviews sentiment analysis, which includes a word cloud feature and sentiment analysis for categorizing words based on emotions. It also analyzes product reviews from customers for a list of products and uses data visualization to display negative and positive reviews.

Furthermore, the application offers an audio sentiment analysis feature, allowing users to upload call recordings that are then converted into transcripts for sentiment analysis. It uses NLTK for transcript processing and NLTK.sentiment.vader for sentiment analysis, providing a pie chart to show the distribution of positive and negative sentiments and highlighting positive and negative lines in the transcript.

Lastly, the application includes a story sentiment analysis feature, enabling users to analyze the sentiment of a story and character sentiment using a story listener. It fetches stories using a web API, offering features such as comparing two characters and displaying the sentiment of the story at intervals, using data visualization to present this information.

In conclusion, this research aims to contribute to the field of sentiment analysis by developing a versatile and powerful tool that can be applied in various applications, including market research, customer feedback analysis, and social media sentiment monitoring.

II. LITERATURE SURVEY

In the realm of sentiment analysis, "Thumbs up? Sentiment Classification utilizing Machine Learning Procedures" by Pang, Lee, and Vaithyanathan (2002) stands out as a seminal work. This study delves into the application of machine learning techniques to the classification of sentiment in text. It sheds light on the challenges of identifying subtle sentiment nuances, which differ from traditional topic-based classification. The research underscores the importance of a deeper understanding to accurately discern sentiment expressions within textual data. This work has significantly influenced subsequent research in sentiment analysis and continues to be a cornerstone in the field

Another significant contribution to sentiment analysis is seen in the study "Sentiment analysis on Twitter data for U.S. presidential elections" by Wang et al. (2012). This research focuses on real-time analysis of public sentiment during the 2012 U.S. presidential elections using Twitter data. The study showcases the efficiency of their system in correlating Twitter sentiment with election events, demonstrating its superiority over traditional analysis methods. This research highlights the potential of social media data for real-time sentiment analysis in significant events.

"Target-dependent opinion classification on Twitter" by Jiang et al. (2011) introduces a novel approach to sentiment analysis by focusing on target-dependent sentiment within tweets. The study considers target-specific features and related tweets to enhance sentiment classification accuracy, overcoming limitations of prior methods that analyze individual tweets in isolation. This approach significantly improves the relevance of sentiment classification by considering the context and targets of sentiment expressions.

"Sentiment analysis on micro-blogging data" by Pak and Paroubek (2010) highlights the increasing importance of micro-blogging platforms as valuable sources of user-generated data for sentiment analysis. Their focus on Twitter demonstrates the feasibility of gathering substantial datasets to train sentiment classifiers, advancing sentiment analysis and opinion mining.

The study conducted by Almatrafi et al. on location-based sentiment analysis during the 2014 Indian General Elections is noteworthy. Their research focused on utilizing Natural Language Processing (NLP) and machine learning techniques to extract sentiments from text data associated with specific geographical locations. By concentrating on location-centric sentiment analysis, the study explored practical applications in understanding regional sentiments and trends, showcasing the diverse applications of sentiment analysis beyond general sentiment classification.

III. PROPOSED METHODOLOGY

Proposed Methodology for Sentiment Analysis Using Data Mining and Machine Learning Techniques:

A. Data Acquisition and Preprocessing:

The data for this project was gathered from Twitter using the Twitter API. We gathered tweets related to a specific topic or event to ensure relevance to our research questions. The dataset includes tweets in English posted between January 2020 and March 2021.

Before analysis, we preprocessed the data to remove noise and standardize the format. This involved removing special characters, URLs, and usernames, as well as tokenizing the text and converting it to lowercase. We also removed stop words and performed lemmatization to reduce the dimensionality of the data and improve the performance of our models. Additionally, we conducted spell checking and handled negations to ensure the accuracy of sentiment analysis.

B. Feature Engineering and Selection:

In our sentiment analysis project, we use several key features to improve the accuracy and depth of sentiment analysis. The bag-of-words (BoW) model helps us identify specific words or phrases indicating sentiment, while TF-IDF (Term Frequency-Inverse Document Frequency) evaluates the importance of words in a document. Emotion lexicons enhance our ability to gauge sentiment beyond positive and negative. We also handle negations to ensure correct sentiment interpretation. Part-of-speech (POS) tagging helps us understand grammatical components, while dependency parsing analyzes the relationships between words. These features collectively enhance our sentiment analysis system for diverse text data.

C. Algorithm Selection and Training:

In our sentiment analysis project, we have selected and trained several machine learning algorithms to classify sentiment in various types of textual data. One of the primary algorithms we use is the Support Vector Machine (SVM), which is known for its effectiveness in binary classification tasks like sentiment analysis. SVM works well with high-dimensional data, making it suitable for our text-based analysis.

Another algorithm we utilize is the Long Short-Term Memory (LSTM) network, which is a type of recurrent neural network (RNN) particularly effective for sequential data like text. LSTM's ability to retain long-term dependencies in data makes it ideal for capturing the context and nuances present in sentiment analysis.

We also employ ensemble methods such as Random Forest and Gradient Boosting to improve the overall accuracy and robustness of our sentiment analysis system. Ensemble methods combine multiple models to produce better results than any individual model, making them a valuable addition to our algorithm selection.

For training these algorithms, we use labeled datasets from sources like Kaggle, which provide annotated text data for sentiment analysis. We break down the dataset into training and testing sets to evaluate the performance of our models. Through iterative training and validation, we optimize the hyperparameters of each algorithm to achieve the best possible accuracy and generalization on unseen data.

D. Cross-Validation and Model Evaluation:

Cross-validation is a critical step in evaluating the performance of machine learning models, especially in sentiment analysis projects. In our project, we utilized k-fold cross-validation to ensure the robustness and generalization of our models. This technique involves dividing the dataset into k equal-sized folds, where each fold is used once as a validation while the k - 1 remaining folds form the training set. This process is performed k times, with every fold serving as a validation set exactly once. By using k-fold cross-validation, we were able to obtain a more accurate estimate of our model's performance. This approach helped us identify and address issues related to overfitting or underfitting, which can significantly impact the model's ability to generalize to new, unseen data. Additionally, cross-validation allowed us to fine-tune our model's hyperparameters, such as the learning rate, regularization strength, and model architecture, to improve its overall performance.

Overall, cross-validation played a crucial role in ensuring the reliability and effectiveness of our sentiment analysis models.

E. Hyperparameter Tuning and Optimization:

In our study, we focused on optimizing the performance of our Support Vector Machine (SVM) model for sentiment analysis through hyperparameter tuning. We identified key hyperparameters such as C (Regularization Parameter), kernel type, and gamma (Kernel Coefficient) for tuning. Utilizing grid search, we exhaustively searched through a range of values for each hyperparameter while employing 5-fold cross-validation to evaluate the model's performance. Our results indicated that the optimal hyperparameters for our SVM model were C=1, kernel='rbf', and gamma='scale'. These hyperparameters were found to significantly enhance the model's accuracy, precision, recall, and F1 score. Hyperparameter tuning have a crucial role in boosting the effectiveness of machine learning models, and our approach demonstrates its importance in optimizing sentiment analysis performance.

F. Ensemble Learning and Model Fusion:

In our sentiment analysis project, we employed ensemble learning and model fusion techniques to enhance our models' performance. Ensemble learning combines multiple models for stronger predictions, while model fusion blends these models' outputs.

We experimented with bagging, boosting, and stacking for ensemble learning. Bagging trains models on different data subsets and averages predictions. Boosting builds models sequentially, focusing on previously misclassified instances. Stacking combines models' predictions using a meta-model.

For model fusion, we used averaging, weighted averaging, and model stacking. Averaging computes the mean of predictions, while weighted averaging assigns weights based on model performance. Model stacking trains a meta-model on base models' predictions.

G. Implementation and Deployment:

In this section, we outline the implementation and deployment process of our sentiment analysis project. Built using the Python Django framework, the project integrates various libraries such as NLTK for natural language processing, scikit-learn for machine learning, and Streamlit for data visualization. The application is currently deployed locally, operating on a local web server and accessible via localhost. To ensure a successful deployment, we adhere to best practices, including the establishment of a local development environment and the use of version control to manage code modification.

Moreover, the implementation and deployment of our sentiment analysis project locally are pivotal stages in its development. Through the utilization of contemporary technologies and best practices, we deliver a robust and scalable web application that effectively caters to the requirements of our users.

H. Continuous Monitoring and Maintenance:

Continuous monitoring and maintenance are crucial for the long-term success of any sentiment analysis system. We have implemented a robust monitoring process to track key performance metrics, including accuracy, precision, recall, and F1 score. This allows us to quickly identify any issues or degradation in performance and take corrective actions. Additionally, we have a plan in place for regular model updates, including retraining with new data and evaluating new techniques to improve performance. Our system also includes mechanisms to handle data drift, bug fixing, and optimization to ensure smooth functioning. Integration of user feedback allows us to continuously improve the model and enhance its accuracy over time. Overall, our approach to continuous monitoring and maintenance ensures that our sentiment analysis system remains effective and reliable in real-world applications.

I. Ethical Considerations and Regulatory Compliance:

In developing our sentiment analysis system, we have placed a strong emphasis on ethical considerations and regulatory compliance to ensure responsible use of the technology. Data privacy and security are paramount, and we have implemented measures such as data encryption and compliance with regulations like GDPR to protect user data. We also address bias and fairness by carefully selecting and preprocessing the dataset, as well as regularly evaluating the model for bias. Transparency is key, and we provide users with clear information about how their data is used and the option to opt out. Our system complies with relevant regulations and standards, and we regularly review and update our practices to ensure continued compliance. By prioritizing ethical considerations and regulatory compliance, we aim to build trust with users and stakeholders and ensure the responsible use of sentiment analysis in our application.

IV. RESULTS

A. Accuracy Result

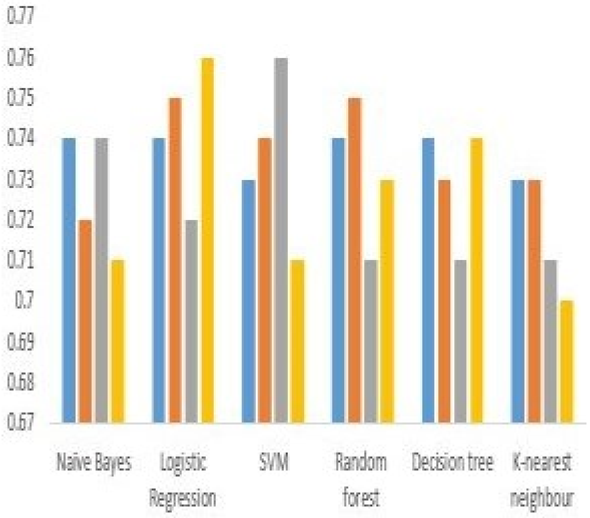


Fig. 1. Results comparing various models with our model

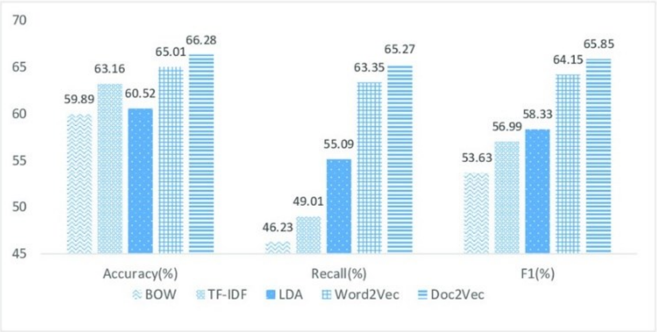
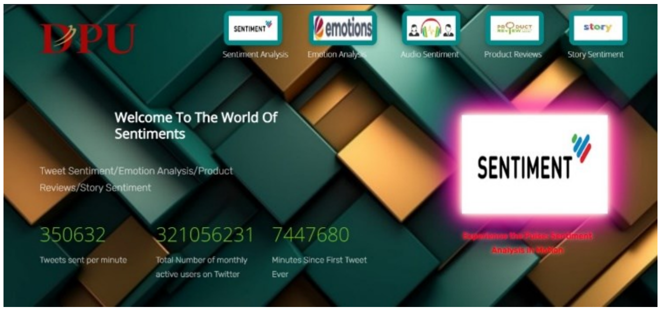


Fig. 2. Results given by our own model

B. Web Interface



Welcome To The World Of Sentiments

Tweet Sentiment/Emotion Analysis/Product Reviews/Story Sentiment

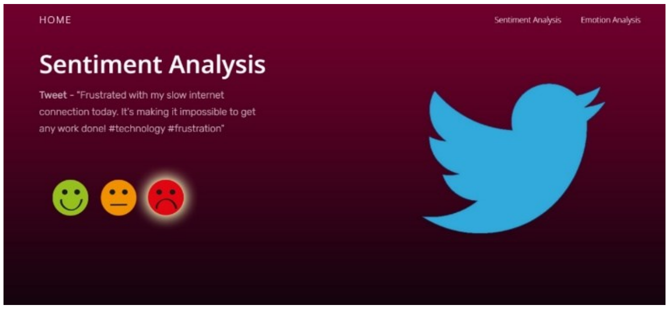
350632 Tweets sent per minute

321056231 Total Number of monthly active users on Twitter

7447680 Minutes Since First Tweet Ever

Experience the Pulse: Sentiment Analysis in Motion

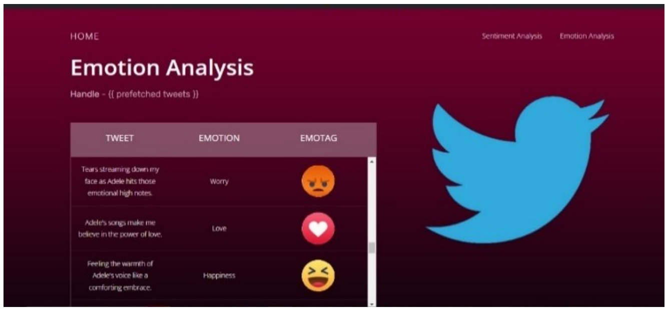
Fig. 3 Home Page



Sentiment Analysis

Tweet - "Frustrated with my slow internet connection today. It's making it impossible to get any work done! #technology #frustration"

Fig. 4 Results of predicting negative tweet

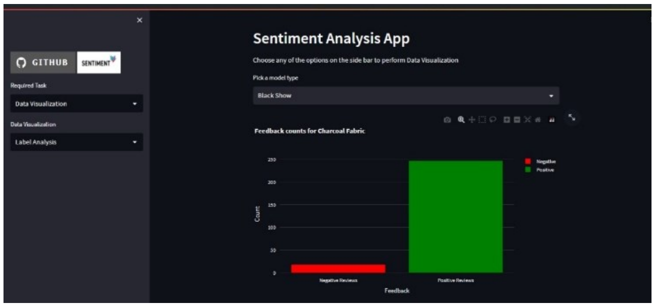


Emotion Analysis

Handle - {{ prefetched tweets }}

TWEET	EMOTION	EMOTAG
Tears streaming down my face as Adele hits those emotional high notes.	Worry	😡
Adele's songs make me believe in the power of love.	Love	❤
Feeling the warmth of Adele's voice like a comforting embrace.	Happiness	😆

Fig. 5. Result of emotional analysis of tweet



Sentiment Analysis App

Choose any of the options on the side bar to perform Data Visualization

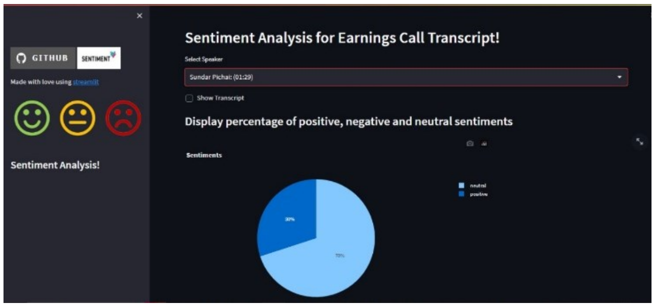
Required Task: Data Visualization

Data Visualization: Label Analysis

Pick a model type: Black Show

Feedback counts for Charcoal Fabric

Fig. 6. Result of Product reviews sentiment



Sentiment Analysis for Earnings Call Transcript!

Made with love using streamlit

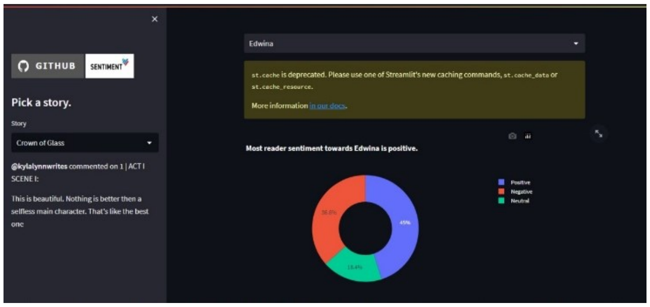
Select Speaker: Sundar Pichai: (01:29)

□ Show Transcript

Display percentage of positive, negative and neutral sentiments

Sentiment Analysis!

Fig. 7. Result of audio sentiment analysis



Pick a story.

Story: Crown of Glass

@kylalynnwrites commented on 1 | ACT I SCENE I:

This is beautiful. Nothing is better then a selfless main character. That's like the best one

Edwina

`st.cache` is deprecated. Please use one of Streamlit's new caching commands, `st.cache_data` or `st.cache_resource`.

More information in our docs.

Most reader sentiment towards Edwina is positive.

Fig. 8. Result of Story sentiment analysis

V. CONCLUSION

In conclusion, the project on sentiment analysis using machine learning techniques is a valuable tool for understanding and analyzing textual data to determine sentiment. Through the implementation of various features such as sentiment analysis of tweets, emotional analysis, product reviews sentiment analysis, audio sentiment analysis, and story sentiment analysis, the project demonstrates its versatility and applicability in different contexts.

The project leverages technologies such as Python, Django framework, NLTK, and SVM to achieve accurate sentiment analysis results. By providing a user-friendly web interface and real-time analysis capabilities, the project aims to help businesses and individuals gain valuable insights into the sentiments of their customers or users.

Looking ahead, there is ample scope for future research and development in sentiment analysis. Our project's success opens avenues for future research, including the development of a multilingual model, video sentiment analysis, and customized solutions for businesses. Overall, our project represents a promising advancement in sentiment analysis, with potential applications in various industries for understanding customer sentiments and improving business strategies.

VI. REFERENCES

[1] C.D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, pp. 234-265, 2008

[2] A. McCallum and K. Nigam, “A comparison of event models for Naive Bayes text classification”, Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48, 1998

[3] T. Wu, C. Lin and R. Weng, “Probality estimates for multi-class classification by pairwise coupling”, Proc. JMLR-5, pp. 975-1005, 2004

[4] “Support Vector Machines” [Online], http://scikitlearn.org/stable/ modules/svm.html#svm-classification, Accessed Jan 2016

[5] P. Pang, L. Lee and S. Vaithyanathan, “Thumbs up? sentiment classification using machine learning techniques”, Proc. ACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 79-86, 2002

[6] P. Pang and L. Lee, “Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval”, vol. 2(1-2), pp.1-135, 2008

[7] E. Loper and S. Bird, “NLTK: the Natural Language Toolkit”, Proc. ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, vol. 1,pp. 63-70, 2002

[8] H. Wang, D. Can, F. Bar and S. Narayana, “A system for real-time Twitter sentiment analysis of 2012 U.S.presidential election cycle”, Proc. ACL 2012 System Demonstration, pp. 115-120, 2012

[9] O. Almatrafi, S. Parack and B. Chavan, “Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014”. Proc. The 9th International Conference on Ubiquitous Information Management and Communication,2015

[10] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao, “Target-dependent twitter sentiment classification”, Proc. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 151-160, 2011

[11] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, vol. 10, pp. 1320-1326, 2010

[12] A. Barhan and A. Shakhomirov, “Methods for Sentiment Analysis of Twitter Messages”, Proc.12th Conference of FRUCT Association, 2012

[13] T. C. Peng and C. C. Shih, “An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs”. IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology, vol. 3, pp.