Indian Journal of Data Mining (IJDM)
ISSN: 2582-9246 (Online), Volume-4 Issue-1, May 2024

Youtube Comment Sentimental Analysis

Aiswarya A S, Haritha Rajeev

Abstract:*The amount of textual data has grown dramatically over time, opening up new avenues for machine learning (ML) and natural language processing (NLP) study. These days, sentiment analysis of comments on YouTube is a really fascinating subject.Although there are a lot of user reviews and comments on many of these films, the low consistency and quality of the material in these comments has prevented much work from being done in terms of identifying trends from them thus far. In this research, we use machine learning techniques and algorithms to perform sentiment analysis on YouTube comments pertaining to popular themes. We show that a clear picture of how real-world events affect public sentiment can be obtained by analyzing the attitudes to identify trends, seasonality, and projections. The findings indicate a strong correlation between the sentiment trends of users and the actual occurrences linked to the corresponding keywords. This study uses a YouTube extractor to perform sentiment analysis on comments on YouTube using citation sentences.To remove the noise from the corpus of comments, various data normalization algorithms were applied to the data. We created a system using six distinct machine learning techniques, including Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF), to perform classifying on this data set.*

Keywords:*Youtube Comments, NLP, Youtube Extractor, Machine Learning Algorithms.*

I. INTRODUCTION

The project aims to gather data from public YouTube comments and analyze the users' attitudes towards various parts of a video as expressed in their written words. Sentiment analysis is a valuable tool for efficiently comprehending the overall concept by analyzing a vast amount of textual data, and it can aid in understanding the user's perspective. Sentiment analysis, also known as opinion mining, aims to discover and classify the positive, negative, or neutral opinions views, attitudes, perceptions, emotions, and sentiments expressed in a given text.

YouTube offers several social tools to assess user sentiment and engagement, such as voting, rating, favoriting, sharing, and leaving negative comments. It is crucial to acknowledge that YouTube offers more than just the ability to share videos.

Manuscript received on 15 April 2024 | Revised Manuscript received on 02 May 2024 | Manuscript Accepted on 15 May 2024 | Manuscript published on 30 May 2024.

*****Correspondence Author(s)*

Aiswarya A S*, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. Email: aiswaryasandhya1234@gmail.com, ORCID ID: 0009-0007-4416-6781

Haritha Rajeev, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. E-mail: haritharajeev19@gmail.com

In addition to publishing and watching videos, users have the option to subscribe to video channels and engage with other users by leaving comments. YouTube is primarily a platform that facilitates both implicit and explicit interaction between users. The user-to-user social aspect of YouTube, sometimes known as the YouTube social network, has been recognized as a significant distinguishing feature when compared to other conventional content providers. Text analytics refers to the process of analyzing unstructured data found in natural language text using a variety of machine learning tools and methodologies. Text analysis provides a cost-effective approach to measure public sentiment. In this research work, we are taking pre-labeled comments of 5 popular youtube videos for training & testing which had over 2 columns namely – label and Comments. The file consists of over 6500 entries or rows.We've developed a system based on six different machine learning algorithms including Multinomial Naïve-Bayes, Support Vector Classifier, Logistic Regression, Decision Tree and Random Forest.Accuracy of the classification algorithms has been evaluated using different evaluations measures e.g., Accuracy score to evaluate the classification system' correctness. To improve our system' performance, we've used different features selection techniques like lemmatization, tokenization, stop words and punctuation removal.

II. LITERATURE REVIEW

[1][7][8] P. Durga and D. Godavarthi , This paper describes the new sentiment analysis model to predict sentiments effectively that can be used to improve product quality and sales. The proposed approach is an integrated model combining several techniques, such as the pre-trained model BERT-large-cased (BLC) for training the dataset. BLC model contains 24-layer, 1024-hidden, 16-heads, 340M parameters. Optimization algorithms can fine-tune pre-trained models, such as BERT, for sentiment analysis tasks. Fine-tuning involves training the pre-trained model on a specific sentiment analysis task to improve performance. Stochastic Gradient Descent (SGD) is the optimized algorithm that helps to analyze the sentiments effectively from the given datasets. The next step is the combination of pre-processing techniques such as Tokenization, Stop Word Removal, etc. The next step focused on Bag-of-Words (BoW) and word embedding techniques like Word2Vec used to extract the features from the datasets. The deep sentiment analysis (DSA) based classification is designed to classify the sentiments based on aspect and priority model to achieve better results. The proposed model combines Aspect and Priority-based Sentiment analysis with a Decision-based Recurrent Neural Network (D-RNN).

The experiments are conducted using Twitter, Restaurant, and Laptop datasets available publicly on Kaggle—the proposed model's performance is analyzed using a confusion matrix. The proposed approach addresses various challenges in analyzing the sentiment analysis. Python programming language with several libraries such as Keras, Pandas, and others extracts the sentiments from given datasets. The comparison between the existing and proposed models shows the effectiveness of the sentiment outputs.

[2] A.Nazir, Y. Rao, L. Wu and L. Sun , this survey emphasized on the issues and challenges that are related to extraction of different aspects and their relevant sentiments, relational mapping between aspects, interactions, dependencies, and contextual-semantic relationships between different data objects for improved sentiment accuracy, and prediction of sentiment evolution dynamicity. A rigorous overview of the recent progress is summarized based on whether they contributed towards highlighting and mitigating the issue of Aspect Extraction, Aspect Sentiment Analysis or Sentiment Evolution. The reported performance for each scrutinized study of Aspect Extraction and Aspect Sentiment Analysis is also given, showing the quantitative evaluation of the proposed approach. Future research directions are proposed and discussed, by critically analysing the presented recent solutions, that will be helpful for researchers and beneficial for improving sentiment classification at aspect-level[3]. D. Prabha and R. Rathipriya ,This work aims to propose an algorithm which is combination of Capsule Network (CN) with Gravitational Search Algorithm (GSA) to analyze people's sentiments from twitter data. In text data mining, CN works to an excessive extent for sentiment analysis compared with other models. The performance of the proposed approach is studied using existing benchmark datasets and COVID-19 twitter posts. The results showed that the proposed approach could automatically classify the sentiments with high performance. It works better compared to other algorithms and results also encourage further research.

[4][9][10][11] K. Cheng, Y. Yue and Z. Song , Currently, various attention-based neural networks have achieved successes in sentiment classification tasks, as attention mechanism is capable of focusing on those words contributing more to the sentiment polarity prediction than others. However, the major drawback of these approaches is that they only pay attention to the words, the sentimental information contained in the part-of-speech(POS) is ignored. To address this problem, in this paper, we propose Part-of-Speech based Transformer Attention Network(pos-TAN). This model not only uses the Self-Attention mechanism to learn the feature expression of the text but also incorporates the POS-Attention, which uses to capture sentimental information contained in part-of-speech. In addition, our innovative introduction of the Focal Loss effectively alleviates the impact of sample imbalance on model performance. We conduct substantial experiments on various datasets, and the encouraging results indicate the efficacy of our proposed approach [5].

D. Prabha and R. Rathipriya, COVID-19 is an extremely contagious virus that has rapidly spread around the world. This disease has infected people of all ages in India, from children to the elderly. Vaccination, on the other hand, is the only way to preserve human lives. In the midst of a pandemic, it's critical to know what people think of COVID-19 immunizations. The primary goal of this article is to examine corona vaccination tweets from India's Twitter social media. This study introduces CompCapNets, a unique deep learning approach for Twitter sentiment classification. The results suggest that the proposed method outperforms other strategies when compared to existing traditional methods.

[6] N. Zhao, H. Gao, X. Wen and H. Li, Aspect-based sentiment analysis (ABSA) aims to identify views and sentiment polarities towards a given aspect in reviews. Compared with general sentiment analysis, ABSA can provide more detailed and complete information. Recently, ABSA has become an important task for natural language understanding and has attracted considerable attention from both academic and industry fields. The sentiment polarity of a sentence is not only decided by its content but also has a relatively significant correlation with the targeted aspect. For this reason, we propose a model for aspect-based sentiment analysis which is a combination of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU), utilizing the local features generated by CNN and the long-term dependency learned by GRU. Extensive experiments have been conducted on datasets of hotels and cars, and results show that the proposed model achieves excellent performance in terms of aspect extraction and sentiment classification. Experiments also demonstrate the great domain expansion capability of the model.

III. METHODS

Data collection is a critical aspect of the Youtube Comment Sentimental Analysis project. we are taking pre-labeled comments of 5 popular youtube videos for training & testing which had over 2 columns namely – label and Comments. The file consists of over 6500 entries or rows.

Extract features from the preprocessed text data to represent each comment.Technique used in this project is TF-IDF (Term Frequency-Inverse Document Frequency). These techniques help convert textual data into numerical vectors that machine learning algorithms can process.

Assigned sentiment labels to the comments in the dataset. Typically, sentiments are categorized into classes like positive, negative, or neutral.

Predictive modeling is the process of using statistical and machine learning techniques to create a model that can make predictions about future outcomes based on historical data. The aim of predictive modeling is to identify patterns in the data and use them to make accurate predictions about future events. The main objective of predictive modeling is to select an appropriate algorithm or machine learning technique that will be used to build the predictive model. Common techniques include linear regression, decision trees, and neural networks. In this project we have used Logistic Regression, Naive Bayes, Random Forest. Logistic Regression, Naive Bayes, Random Forest were machine learning algorithms used in the existing system, while SVC algorithm has been included in this to obtain a better result than previously used algorithms and compare them based on performance.

The test-train split is an important step in developing and evaluating machine learning models and helps ensure that the models are accurate and reliable. In this project we have split the dataset into 75-25, 75% being the training set and 25% being the testing set. Evaluate the trained model's performance on the testing dataset using appropriate evaluation metrics such as accuracy. Once satisfied with the model's performance, deployed it to perform sentiment analysis on new YouTube comments. Created a user-friendly interface where users can input a YouTube video URL, fetch comments, and view sentiment analysis results. Deployed the model on a web server.

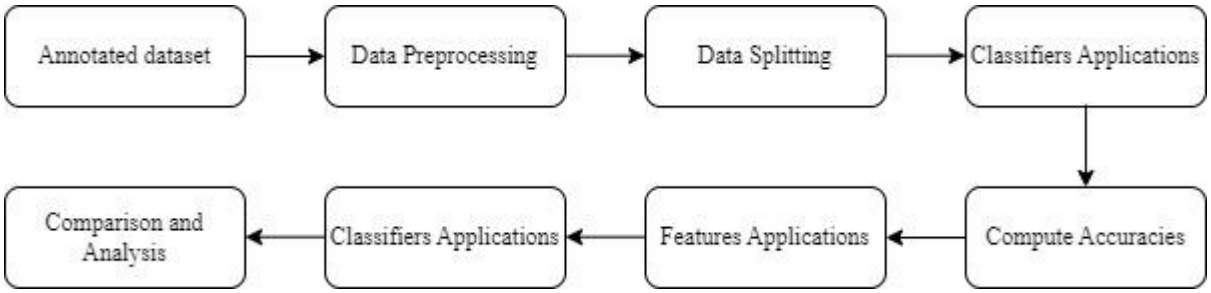

Fig.1. Block Diagram

IV. RESULTS AND DISCUSSION

A. Results

The implementation of the Support Vector Classifier (SVC) algorithm for sentiment analysis of YouTube comments yielded highly encouraging results, showcasing its superior performance compared to alternative models. With an accuracy of 79.23% on the testing dataset, the SVC model demonstrated its efficacy in accurately categorizing sentiments expressed within YouTube comments. Analysis of the confusion matrix revealed low rates of misclassification, particularly in distinguishing between positive and negative sentiments, indicating the model's effectiveness in discerning subtle nuances in language and context. These promising results underscore the potential of the SVC algorithm in practical applications such as content moderation, audience engagement analysis, and community management within the YouTube platform. By leveraging the insights provided by the SVC model, content creators and platform administrators can make informed decisions to enhance user experiences and foster a positive online community environment.

Table 1. Accuracy Comparison for Each Model

Model Name	Accuracy (%)
Logistic Regression	78.31
Random Forest	79.16
Naïve Bayes	78.31
SVC	79.23
Decision Tree	79.18

B. Discussion

1. Effectiveness of the Machine Learning Algorithm: Discuss the effectiveness of the chosen machine learning algorithm in accurately classifying sentiments in YouTube comments. Evaluate its performance metrics using accuracy.
2. Data Quality and Bias:the quality of the labeled data used for training the sentiment analysis model. Discuss potential biases in the dataset, such as imbalanced class distributions, skewed representations of certain topics, or cultural biases.
   strategies for addressing biases in the data, such as augmenting the dataset with more diverse comments, applying data preprocessing techniques, or using techniques like adversarial training to mitigate biases.
3. Interpretability vs. Performance Trade-offs:Debate the trade-off between model interpretability and performance. Simple models like Naive Bayes or logistic regression offer interpretability but may sacrifice accuracy, whereas complex models like deep learning neural networks may achieve higher accuracy but are less interpretable.
   Consider the implications of model interpretability in real-world applications, especially in contexts where stakeholders require explanations for decision-making processes.
4. Future Directions and Opportunities: Identify areas for future research and development, such as enhancing sentiment analysis models with multimodal features (e.g., analyzing text alongside video and audio content), incorporating temporal dynamics, or exploring sentiment evolution over time.
   Discuss potential applications of sentiment analysis beyond YouTube, such as social media monitoring, brand reputation management, market research, and customer feedback analysis.

V. CONCLUSION

The sentiment analysis project focusing on YouTube comments using a machine learning algorithm presents a comprehensive understanding of audience sentiments and engagement within the YouTube ecosystem. Through meticulous data collection, preprocessing, and model training, valuable insights have been gained regarding the sentiments expressed by users across various videos and topics.The Support Vector Classifier algorithm has demonstrated its efficacy in accurately classifying sentiments within YouTube comments. By evaluating performance metrics and conducting thorough analyses, we have gained confidence in the model's ability to interpret and categorize sentiments with a reasonable degree of accuracy.

Looking ahead, there are promising opportunities for future research and development. Enhancing the robustness and generalizability of sentiment analysis models, addressing biases in the dataset, and integrating multimodal features for a richer understanding of user sentiments are areas ripe for exploration. Furthermore, leveraging sentiment analysis insights to inform content creation strategies, enhance user experiences, and foster community engagement represents a compelling direction for future endeavors. In essence, the sentiment analysis project serves as a valuable tool for creators, marketers, and platform administrators to gain actionable insights into audience sentiments, preferences, and behavior on YouTube. By continuing to refine methodologies, address ethical considerations, and explore new avenues of research, we can unlock even greater value from sentiment analysis in the dynamic landscape of online content creation and consumption.

DECLARATION STATEMENT

Funding	No, I did not receive
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Yes, It is relavant. The dataset used for this project was Grid Corpus Dataset GRID is an openly available corpus containing an audio-visual database from 34 speakers with 1000 utterances per speaker.
Authors Contributions	Each author has made an independent contribution to the article. The individual contributions of each author are presented below for clarity and transparency. Aiswarya A S is the main contributor and Haritha Rajeev is the project guide.

REFERENCES

1. P. Durga and D. Godavarthi, "Deep-Sentiment: An Effective Deep Sentiment Analysis Using a Decision-Based Recurrent Neural Network (D-RNN)," in IEEE Access, vol. 11, pp. 108433-108447, 2023, doi: 10.1109/ACCESS.2023.3320738.
2. A.Nazir, Y. Rao, L. Wu and L. Sun, "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," in IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 845-863, 1 April-June 2022, doi: 10.1109/TAFFC.2020.2970399.
3. D. Prabha and R. Rathipriya, "Sentimental Analysis Using Capsule Network with Gravitational Search Algorithm," in Journal of Web Engineering, vol. 19, no. 5-6, pp. 775-794, September 2020, doi: 10.13052/jwe1540-9589.19569.
4. K. Cheng, Y. Yue and Z. Song, "Sentiment Classification Based on Part-of-Speech and Self-Attention Mechanism," in IEEE Access, vol. 8, pp. 16387-16396, 2020, doi: 10.1109/ACCESS.2020.2967103.
5. D. Prabha and R. Rathipriya, "Competitive Capsule Network Based Sentiment Analysis on Twitter COVID'19 Vaccines," in Journal of Web Engineering, vol. 21, no. 5, pp. 1583-1601, July 2022, doi: 10.13052/jwe1540-9589.2159.
6. N. Zhao, H. Gao, X. Wen and H. Li, "Combination of Convolutional Neural Network and Gated Recurrent Unit for Aspect-Based Sentiment Analysis," in IEEE Access, vol. 9, pp. 15561-15569, 2021, doi: 10.1109/ACCESS.2021.3052937.
7. Das, S., S, S., M, A., & Jayaram, S. (2021). Deep Learning Convolutional Neural Network for Defect Identification and Classification in Woven Fabric. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 1, Issue 2, pp. 9–13). https://doi.org/10.54105/ijainn.b1011.041221
8. R, A. (2019). Logistics Network Optimization in Distributing Critical Medical Supplies for a Pharmaceutical Company. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 3, pp. 7767–7770). https://doi.org/10.35940/ijrte.c6320.098319
9. Thakur, T. B., & Mittal, A. K. (2020). Real Time IoT Application for Classification of Crop Diseases using Machine Learning in Cloud Environment. In International Journal of Innovative Science and Modern Engineering (Vol. 6, Issue 4, pp. 1–4). https://doi.org/10.35940/ijisme.d1186.016420
10. Sistla, S. (2022). Predicting Diabetes u sing SVM Implemented by Machine Learning. In International Journal of Soft Computing and Engineering (Vol. 12, Issue 2, pp. 16–18). https://doi.org/10.35940/ijsce.b3557.0512222
11. Tripathi, K., Gupta, A. K., & Vyas, R. G. (2020). Deep Residual Learning for Image Classification using Cross Validation. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 6, pp. 1525–1530). https://doi.org/10.35940/ijitee.f4131.049620

AUTHORS PROFILE

Aiswarya A S, currently pursuing Master of Science in Computer Science from the prestigious St. Albert's College (Autonomous), Ernakulam. Prior to this she had completed his Bachelor of Science degree in Computer Science from National College, Thiruvananthapuram. She has a wide range of interests ranging from IoT, python and Machine learning. she is given attention to details as well as she is able to think outside the box, she loves to solve problems and has been keenly observing the latest technology. She is an active member of the Computer Science community and coordinates in various events conducted.

Haritha Rajeev she joined Department of Computer Science of the prestigious college, St. Albert's College (Autonomous), Ernakulam as Assistant Professor in 2022. She has a teaching experience of 3 years and has an industry experience of 1 year. She completed her undergraduate studies from Amrita School of Arts and Science, Kochi and went to do her Master's in Computer Science (MCA) from FISAT. She has specialized in Software Engineering and Machine Learning and Computer Security. She completed her M.Phil in Computer Science and IT from Amrita Viswa Vidyapeetham. She is doing her PHD in IT at Lincoln University College (Malaysia). She has published Six papers in professional journals. She has successfully published a book Entitled Introduction to Software Engineering.