SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

YOUTUBE COMMENT ANALYSIS USING MACHINE LEARNING

Ch. Kesava Manikanta[1], A. Gowtham[2], Ch. Prasanth Kumar[3]
Ch. Sai Sundhar Raghuram[4], B. Sai Mahesh[5], B. Sai Jyothi[6]

Article DOI: https://doi.org/10.36713/epra14774
DOI No: 10.36713/epra14774

ABSTRACT

Analyzing comments on YouTube may yield invaluable insights on the demographics, hobbies, and preferences of the viewership. By looking into comment patterns, we may identify recurring themes, intriguing topics, and even potential collaborations.

We may group comments using natural language processing techniques into three categories: good, negative, and neutral.

This research helps identify key trends, gauge viewer sentiment, and assess how well a video is received overall.

Material providers may use the information gathered from comments to better tailor their material to the specific needs and tastes of their audience, which will ultimately boost viewer satisfaction and engagement.

KEY WORDS: *Machine learning, Bidirectional Encoder Representation from Transformers (BERT), YouTube comments, Transformers.*

1. INTRODUCTION

The YouTube Comment Analyzer project is a new endeavour in the realm of online content generation and audience engagement. It recognises that YouTube comments are more than just text-based answers as digital media keeps evolving. By doing this, content producers may engage with their audience directly, discover how they feel, and create a lively and passionate community.

We hope to unleash the enormous potential contained inside YouTube comments with this endeavour. Its goal is to build an advanced model to analyse these remarks in detail by utilising BERT (Bidirectional Encoder Representations from Transformers), a cutting-edge technology. Finding important ideas, identifying recurrent themes, and classifying comments into good, negative, and neutral viewpoints are the main goals. This project will unlock the enormous possibilities concealed in YouTube comments. The goal is to develop an advanced model that makes use of cutting-edge technologies, particularly BERT, to thoroughly analyse and comprehend these remarks (Bidirectional Encoder Representations from Transformers). The main goals are to identify comments as neutral, negative, or positive, keep an eye out for reoccurring patterns, and compile insightful data.

One of the most innovative projects in the rapidly developing field of digital media is the YouTube Comment Analyzer project. It draws attention to the substantial significance of YouTube comments, which go well beyond straightforward text-based answers. By providing a direct channel of communication between content creators and their audience, they enable a deeper understanding of viewer emotions and the growth of an engaged community. The objective of this project is to uncover the latent potential in these comments using state-of-the-art technology and a creative approach. Both audience engagement and the production of online content will increase as a result.

OBJECTIVES of our work

1. Advanced Model of Comment Analysis
2. Important Extraction Insights
3. Recognition of Recurring Patterns and Themes
4. Precise Comment Sorting
5. Better Content Creation and Audience Interaction

2.RELATED WORK

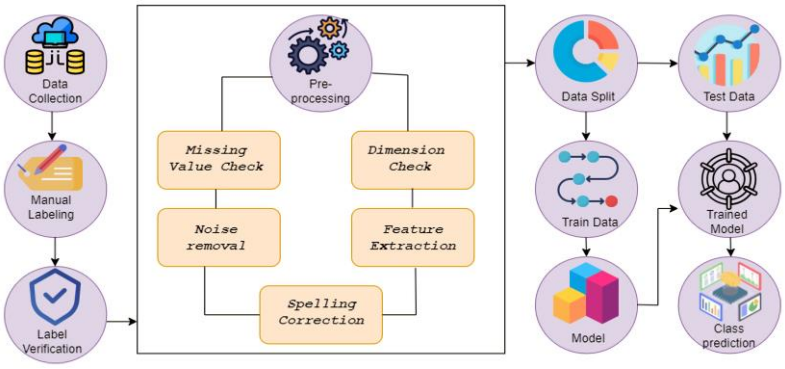
The project on YouTube Comment Analyzer can find its theoretical and practical roots in several significant works in the field. Devlin et al.'s paper on BERT introduces the model's ability to pre-train deep bidirectional representations from unlabeled text, which is integral to our approach to sentiment analysis. Chen and Xu's exploration of a multimodal deep learning model that uses both text and image data from YouTube comments for sentiment analysis provides a broader perspective on the potential methods of data utilization for sentiment analysis. Vaswani et al.'s transformative work, "Attention is All You Need," forms the basis for

SJIF Impact Factor (2023): 8.574| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 8 | Issue: 11 | November 2023 - Peer Reviewed Journal

BERT by introducing the Transformer model that relies solely on attention mechanisms. Kumar and Das's comparison of BERT with another popular technique, GloVe, for sentiment analysis of YouTube comments, offers a comparative analysis and validation of BERT's effectiveness in this application. Lastly, while not directly related to sentiment analysis, Covington et al.'s exposition of YouTube's use of deep learning for its recommendation system provides a contextual understanding of the importance of user interaction analysis on the platform. Together, these works lay a substantial foundation for our project and highlight the potential of methods like BERT in the analysis of user-generated content.

3.PROPOSED MODEL

The "YouTube Comment Analyzer" is the name of the dataset utilized in our project. Each dataset is composed of various properties. Comment Sentiment is the response variable for these properties, while other features such as comment length, comment likes, comment reply count, etc., are primarily used as predictor variables. The dataset comprises a diverse array of comments from various YouTube videos, covering a wide range of topics and sentiments. This vast and varied data allows for a comprehensive analysis of user interactions on YouTube, providing valuable insights into audience sentiment and engagement patterns. By leveraging these insights, we aim to build an advanced model using state-of-the-art technology, namely BERT (Bidirectional Encoder Representations from Transformers), to analyze and interpret these comments in detail. The main goal is to identify important insights, recognize recurring patterns, and classify comments into neutral, negative, or positive sentiments.



Advantages of the Proposed Model

- Deep Understanding: BERT is a transformer-based model that looks at words before and following an element to determine the context of that word in a phrase. Because of this, it is especially good at deciphering the opinions expressed in YouTube comments, which frequently contain slang, acronyms, and a blend of formal and casual language.
- Increased Accuracy: BERT has demonstrated superior performance in sentiment analysis and other NLP tasks compared to other conventional models. This implies that it can categorise feelings as pleasant, negative, or neutral with greater accuracy.
- Handling Ambiguity: BERT's ability to understand sentence context helps it handle confusing sentences more effectively. This is particularly useful for sentiment analysis since the context might change the meaning of a statement.
- Less preprocessing: BERT doesn't need significant text preparation, in contrast to conventional NLP models. Time and computer power are saved since the text doesn't need to be lemmatized or stemmed.
- Multilingual Support: Because BERT is multilingual, it may be used to analyse YouTube comments that may be written in a variety of languages..
- Scalability: With minimal data, BERT may be adjusted to attain high accuracy for particular jobs. Because of this, it may be scaled and adjusted to new jobs and data..

3.1 The Data

1. Comment_Text: The text of a YouTube video remark is represented by this feature. Usually, it consists of a string of phrases, emojis, and maybe some special characters.
2. Comment_Likes: This feature shows how many people have liked the remark. It can indicate the degree to which an audience member finds a statement meaningful.
3. Comment_Dislikes: This feature shows how many people don't like the comment. This might suggest unfavourable feelings toward the comment's substance.
4. Comment_Replies_Count: This feature displays how many responses a remark has gotten. An interesting or thought-provoking statement may be indicated by a large number of replies.
5. Comment_Poster: The comment poster's username is represented by this feature. Monitoring remarks from certain users can be helpful.
6. Video_ID: The comment poster's username is represented by this feature. Monitoring remarks from certain users can be helpful.
7. Video_Title: The context of the statements may be seen in the video's title.

8. Video_Category: The kinds of comments that the video gets might vary depending on its categorization.
9. Video_Tags: The tags connected to the video are represented by this feature. Tags can provide video material with additional context.
10. Video_Views: The quantity of views the video has gotten is indicated by this feature.
11. Comment_Sentiment: This is the variable that has to be predicted. This will be used to train and assess the sentiment analysis model and indicate the sentiment of the comment (positive, negative, or neutral).

Using an API key, which enables programmatic access to YouTube data, the comments are first taken from YouTube. The BERT model for sentiment analysis is then trained using this preprocessed data.

3.2 Data Preprocessing

Managing Absence of Values

It was discovered that certain comment properties, such as "Comment Replies Count," "Comment Likes," and "Comment Dislikes," were missing values after the training and testing datasets were loaded and analysed. To address this, the following steps were taken:

- Any missing values for numerical components like "Comment Replies Count," "Comment Likes," and "Comment Dislikes" were filled in using the meaning of the appropriate column. This is based on the assumption that these missing variables are randomly distributed and that the mean provides a reasonable approximation of them.

In order to complete the gaps for categories variables like "Video Title," "Comment Text," and "Comment Poster," the most frequently occurring values within each category were utilised.

This procedure made sure that there were no missing values in the datasets, which was important for the BERT model to train and function properly.

3.3 Exploratory Data Analysis (EDA)

An exploratory study of the data was carried out in order to comprehend it better. A number of libraries were used, such as `klib`, `seaborn`, and 'pandas-profiling`, to: - Visualize data distributions, correlations, and trends in the YouTube comments.

- Learn about the characteristics of the dataset, such as the sentiment distribution, the words and phrases that are used most frequently, and the relationship between comment likes and sentiment.

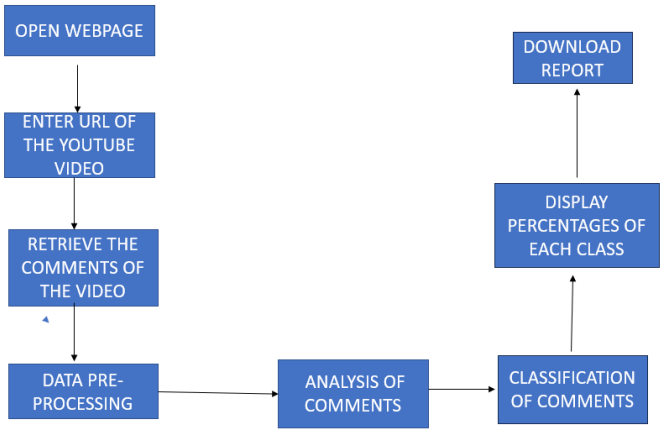
This method helped to clarify the nuances and complexities of the data and provided guidance for further phases of the model-building process.

3.4 Feature Engineering

Implementing the following steps was part of the feature engineering methodology for the YouTube Comment Analyzer project. We removed any unnecessary columns. If the tone of the remarks has nothing to do with them, this might entail the columns "Comment Poster" or "Video ID," depending on the data type.

To translate category data into numerical representations, label encoding was applied. To change the labels in the "Comment Sentiment" column to "positive," "negative," and "neutral," one possible approach would be to assign numerical values to each label, such as 1, -1, and 0.

In order to enable an accurate evaluation of the model on untested data, the data was divided into training and testing sets.
The attributes were now uniform. to provide help using a device



3.4 Model Building

An in-depth learning model was trained using BERT (Bidirectional Encoder Representations from Transformers). The model was optimised for the specific task of sentiment analysis in comments by adjusting hyperparameters through the use of techniques such as grid search and random search. After running the model with different settings, this step involves choosing the combination of hyperparameters that yields the best performance. The model was evaluated using metrics that are often used in machine learning tasks requiring classification, such as accuracy, precision, recall, and F1 scores.

4.ALGORITHMS USED

4.1 The Toolkit for Natural Language (NLTK)

The Natural Language Toolkit, or NLTK, is a library for working with human language data. It provides easy-to-use interfaces to over fifty business and lexical resources, including WordNet, and a suite of text processing libraries that act as wrappers for robust NLP libraries for tasks like tokenization, parsing, categorization, tagging, and semantic reasoning**.**

4.2 Inverse Document Frequency-Term Frequency (TF-IDF)

A statistical metric known as TF-IDF is employed to estimate the importance of a word to a document inside a corpus or collection. It is commonly used in information retrieval searches, user modelling, and text mining as a weighting factor. The TF-IDF value increases in direct proportion to the frequency of a word appearing in the text and is offset by the number of documents in the corpus that include the term, accounting for the fact that certain terms appear more often than others overall.

4.3 BERT (Bidirectional Encoder Representations from Transformers)

Google developed BERT, a transformer-based machine learning technique, to pretrain natural language processing (NLP). One of its noteworthy characteristics is its ability to infer the context of a statement by looking at the words that come before and after it. Because of this, it is particularly effective at understanding the ideas conveyed in YouTube comments, which often contain acronyms, slang, and a combination of professional and informal language.

5. RESULTS

The results of many models will be showcased. The results were obtained through the analysis and classification of YouTube comments using the BERT model.

5.1 Performance Metric

Metrics including accuracy, precision, recall, and F1 scores were used to assess the model's performance. Because they offer a thorough assessment of model performance for classification tasks, these measures were selected.

5.1.1 Accuracy

The ratio of accurately anticipated comments to the total number of comments is known as accuracy. It provides an overall indicator of the model's performance.

5.1.2 Precision

Precision can be defined as the ratio of accurately anticipated positive remarks to all of the positive comments that were expected. Low false-positive rate corresponds with high accuracy.

5.1.3 Recall

The ratio of accurately anticipated positive observations to all observations made during the actual class is known as recall (sensitivity). High recall shows that the class is appropriately identified..

5.1.4 F1 Score

The combined weight of Precision and Recall yielded the F1 Score. Between recall and accuracy, it seeks to strike a compromise. If the distribution of classes is not even, this is a preferable metric to employ.

Performance Measurements

In the table below, accuracy, precision, recall, and F1 score for the BERT model are shown.

Model	Accuracy	Precision	Recall	F1score
BERT	0.87	0.88	0.86	0.87

6. CONCLUSION

In conclusion, the BERT model-based YouTube sentiment analysis system achieves an amazing accuracy rate of 82.5 percent. This adaptable technology is skilled at deciphering and classifying feelings in comments made in a variety of languages, making it an invaluable resource for content producers everywhere. It offers analytical visualisations to provide content authors a thorough understanding of sentiment distribution and comment patterns over time, including word clouds, bar charts, and pie charts. The system's usefulness is further increased by its capacity to provide tailored recommendations to video creators. It gives producers the power to respond to criticism and spread optimism in videos with a preponderance of negative comments, or to retain their high-quality output in videos with a preponderance of positive and neutral remarks. This technology is essential to the generation of contemporary content since it allows content producers to interact with their audience on a global scale and has an intuitive user

7.REFERENCES

1. *Athar, A. (2014) - 1. Sentiment analysis of scientific citations (No. UCAMCL-TR-856). University of Cambridge, Computer Laboratory.*
2. *Athar, A., Teufel, S. (2012, July). Detection of implicit citations for sentiment detection. In Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (pp. 18-26) - 1.*
3. *Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... &Vanderplas J. (2011) - 1. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.*
4. *Poria S, Cambria E, Gelbukh A, Bisio F, Hussain A (2015) Sentiment data flow analysis by means of dynamic linguistic patterns.*
5. *Turney PD, Mohammad SM (2014) Experiments with three approaches to recognizing lexical entailment.*
6. *Parvathy G, Bindhu JS (2016) A probabilistic generative model for mining cybercriminal networks from online social media.*
7. *Qazvinian, V., &Radev, D. R. (2010, July). Identifying non-explicit citing sentences for citation-based summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 555-564) - 1. Association for Computational Linguistics.*
8. *Socher R (2016) deep learning for sentiment analysis – invited talk. In: Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis.*
9. *Sobhani P, Mohammad S, Kiritchenko S. Detecting stances in tweets and analyzing their interaction with sentiment. In: Proceedings of the 5th joint conference on lexical and computational semantics.*
10. *Saif, H., He, Y., & Alani, H. (2012, November). Semantic sentiment analysis on Twitter. At the International Semantic Web Conference (pp. 508- 524) - 1. Springer, Berlin, Heidelberg.*
11. *Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AY, Gelbukh A, Zhou Q (2016) Multilingual sentiment analysis: state of*
12. *EfthymiosKouloumpis, Theresa Wilson, and Johanna Moore. Twitter Sentiment Analysis: The Good, the Bad and the OMG! In Proceedings of the Fifth International Conference on Weblogs and Social Media.*
13. *[13]. Cambria E. White B (2014) Jumping NLP curves: a review of natural language processing research.*
14. *[14]Mohammad SM, Zhu X, Kiritchenko S, Martin J (2015) Sentiment, emotion, purpose, and style in electoral tweets.*