iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51973

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Analysing Sentiments for YouTube Comments using Machine Learning

Sainath Pichad[1], Sunit Kamble[2], Rohan Kalamb[3], Sumit Chavan[4]
Information Technology, Pune Institute of Computer Technology, Pune, India

Abstract: Sentiment analysis is a method for learning what users think and feel about a service or a product. YouTube is one of the most popular platforms for sharing videos. Millions of views are attained. These get a lot of comments, many of which offer helpful information that raises the posted content's rating levels. Natural language processing and machine learning techniques are used to make use of these remarks. There have been several academic attempts with two classes (positive or negative), three classes (two with neutral), or multiple classes (happy, sad, fear, surprise, and rage). Consequently, there had been efforts to utilise study of comments on YouTube to determine the polarity. This study examines the perception of strategies and methods for analysis that may be used to the material on YouTube.

Keywords: Sentiment Analysis, Opinion Mining and YouTube, YouTube comments, User comment, Text mining and YouTube, Classification and YouTube, Machine learning and YouTube.

I. INTRODUCTION

Sentiment analysis on YouTube is becoming increasingly important for businesses and content creators. It allows them to understand how their audience is responding to their content and make informed decisions about future content creation or marketing strategies. Sentiment analysis can also help businesses and content creators identify areas for improvement and address negative feedback to enhance their brand reputation. With help of Metadata(comments) we have the potential to correctly find popularity of the video. Since the invention of computer linguistics, text mining, and sentiment analysis, determining the polarity of words in a certain context has been possible. Machine learning-based methods use a supervised learning mechanism for sentiment. This indicates the polarity information (i.e., positive, negative, and neutral) and an assigned numeric value to score how positive or negative a given word is [1]. The lexicon-based methods, NLP methods are improved method for sentiment analysis. One of the biggest challenges in sentiment analysis on YouTube is the large volume of comments that need to be analyzed. Moreover, the informal nature of YouTube comments and the use of slang and jargon can make it difficult for machine learning algorithms to accurately interpret the sentiment of comments. Therefore, natural language processing techniques and sentiment lexicons need to be carefully selected to achieve accurate sentiment analysis on YouTube comments.

A. Applications of Sentiment Analysis on YouTube

Sentiment analysis on YouTube can have various applications, such as:

1) *Identifying Popular Videos:* By analyzing the sentiment of comments on a video, it is possible to determine the level of popularity of the video.
2) *Identifying Key Themes:* By analyzing the sentiment of comments on a video, it is possible to identify key themes that are being discussed by the audience. This can help content creators and businesses understand what their audience is interested in and create content around those themes.
3) *Understanding Audience Sentiment:* By analyzing the sentiment of comments on a video, it is possible to understand how the audience is feeling about the content. This can help content creators and businesses identify areas where they can improve the content or address negative feedback to enhance their brand reputation.
4) *Techniques for Sentiment Analysis on YouTube:* There are Various Techniques that can be used for sentiment analysis on YouTube, including:
5) *Lexicon-based Methods:* These methods use a predefined set of words and their associated sentiment scores to analyze the sentiment of text.
6) *Machine learning-based Methods:* These methods use supervised learning algorithms to train a model on a labeled dataset of comments to accurately classify the sentiment of new comments.
7) *Deep learning-based Methods:* These methods use neural networks to learn complex patterns in the data and achieve more accurate sentiment analysis on YouTube comments.

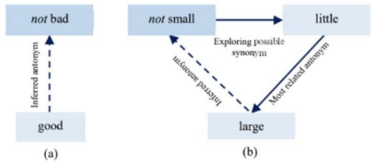
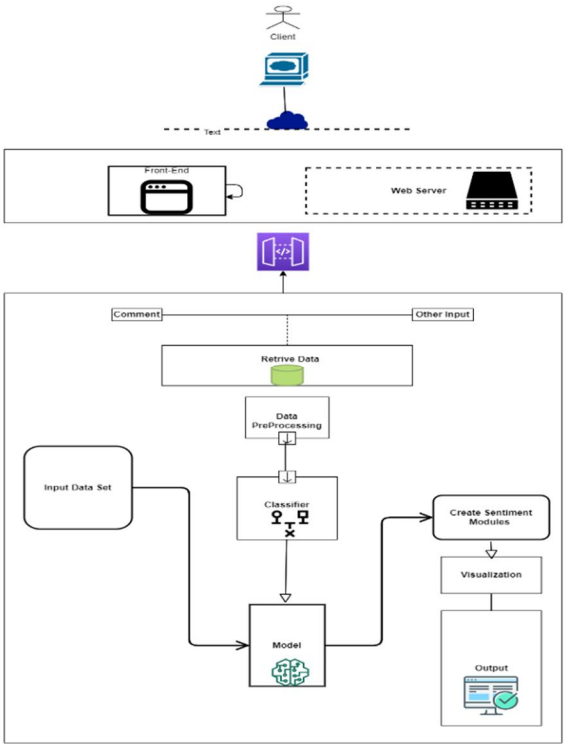


FIG. 1 ANTONYM WORD REPLACEMENT. [2]

Sentiment analysis is required for majority of text mining applications. For the classification of deep learning-based methods are been used. For this PURPOSE HIGH, quality training set containing highly accurate labels are required. Building an accurate training set is difficult task. For this PURPOSE, A new labelling strategy called two level long short-term memory network is constructed as sentiment classifier. A new encoding strategy is also used which is p-hot encoding. The sentiment polarity may vary from sentence to sentence. For this PURPOSE, a flipping model is used for polar flipping of words [5].

More people are beginning to publicly share their ideas on websites as a result of the Internet's and social media platform recent explosive growth. Therefore, the big data of user comments is generated on the Internet. For example, the product comments are generated on E-commerce websites such as Jindong and Taobao, and hotel comments are generated on travel websites such as Ctrip and ELong. With the explosive increasing of comments, it is difficult to analyse them manually. Text categorization in the form of sentiment analysis incorporates NLP, machine learning, data mining, information retrieval, and other study areas. The sentiment orientation analysis of comment corpus, which shows that people express positive, negative, or neutral attitudes toward items or events, is the primary emphasis of sentiment analysis of comments.



II. SCOPE

The project will focus on assessing the effectiveness of out-of-the-box machine learning classifiers in classifying trending YouTube videos using a limited testing dataset. The classifiers will be applied with default settings from Scikit-learn, a popular Python library. The study will only use publicly available data from the YouTube trending top lists. By evaluating the performance of different classifiers, the project aims to identify the best approach for this task. To ensure originality, proper citation and referencing will be employed throughout the project.

III. LITERATURE SURVEY

The link between the opinion targets of a document and the polarity values that correspond to them is defined by aspect-based sentiment analysis. Since aspects are frequently implicit, identifying them and determining their appropriate polarity is a very difficult process. Numerous techniques, tactics, and enhancements, such as word frequency and reverse document frequency approaches, corpus or lexicon-based approaches, have been proposed to solve these issues at various levels. Heuristic techniques are more effective than frequency- and lexicon-based approaches, but they take longer because of the various ways characteristics might be combined. The CNN's hyperparameters used in aspect-based sentiment analysis are tweaked using genetic algorithms (GA). The suggested approach outperformed state-of-the-art procedures, according to experimental findings, with accuracy rates of 95.5%, precision rates of 94.3%, recall rates of 91.1%, and f-measure rates of 96.0% [3].

Data collection, pre-processing, semantic feature extraction, word2vec representation, and CNN implementation were the phases of the proposed study.

The majority of currently used opinion mining techniques rely on text-level analysis and can only find clearly articulated opinions. (Phases of aspect-based sentiment analysis [4])

The aim of ABSA was to pinpoint a subject's attributes and the opinions expressed about each attribute. Following were the Phases of ABSA [4].

IV. PROPOSED METHOLODOGY

A. Recurrent Neural Networks

The traditional neural network model is ineffective in dealing with the sequence learning because it is impossible to describe the correlation between the front and back of the sequence. RNN (Recurrent Neural Networks) is a sequence learning model that connects nodes between hidden layers and can learn sequence feature dynamically. RNN which is applied to Chinese text sentiment analysis. The input text is "**酒店的**环**境不**错

"(The environment of the hotel is good). After word segmentation, it becomes "**酒店**/**的**/环**境**/**不**错". Each word is converted into the corresponding word vector (w1 , w2 , w3 , w4), and then the corresponding word vector (w1 , w2 ,w3, w4) is sequentially input into the RNN.

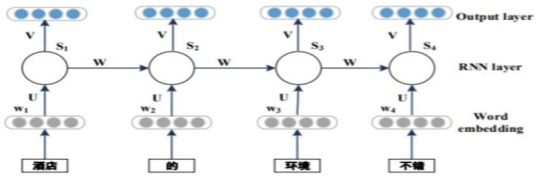


Fig 3. The sentiment analysis model of RNN [3]

The calculation process of RNN is as follows:
1) At the time t, wt is input to the hidden layer.
2) t s is the hidden layer's output of the step t. t s is based on wt and -t 1 s .)**(t □ t □ t□1 sWwUfs , where f is usually the non-linear function, such as tanh or ReLU.
3) Finally, the output D is calculated according to)*max(t t □ sVsofto [3].

B. VADER

Web scraping is used in the proposed methodology to collect three different types of data sets (hotel, cars, and movies), which are then pre-processed to extract useful information. The supplied reviews were utilised as labels throughout the deployment of VADER after being carefully analysed [4]. VADER is a text analysis tool designed primarily to extract emotions from text. In accordance with their semantic orientation, VADER provides a collection of lexical characteristics (such as words) that are often categorised as positive or negative. When it comes to social media messages, movies, and product evaluations, VADER is surprisingly successful [4]. The dataset in the suggested strategy was trained on vectors. Three million words and three million phrases total in the model. It is feasible to find precise relationships by running such a huge corpus. In the network's first layer, words are transformed into feature vectors. GA chooses the best option from the available options for the given problem [4]. The genetic algorithm takes five steps into account. first population, Selection of the fitness function, crossover, and mutation. Experimental findings revealed that the extraction of semantic characteristics yields incredibly refined data. The accuracy is improved by the information gained by looking at the specific domain re- views, which lowers the false negative and false positive rates. The performance of the integrated technique outperformed SVM, decision trees, LDA, RF, and LG by 4%, 23%, 17%, 10%, and 12%, respectively. The suggested approach offers the best precision, recall, accuracy, and f-metric when compared to prior work. The performance of the integrated technique outperformed SVM, decision trees, LDA, RF, and LG by a combined 6%, 21%, 15%, 11%, RF, and 10%. The experimental findings demonstrate that, with 95.5%, 94.3%, 91.1%, and 96.6% for accuracy, precision, recall, and f-measurement, respectively, the proposed approach beats all other current methods.

C. P-Hot Encoding

When different categories are independent, one-hot encoding is the most popular encoding strategy for categorical data[c]. For instance, if three categories, positive, neutral, and negative, are represented using one-hot encoding, the encoding vectors for the three categories are [1, 0, 0]T, [0, 1, 0]T, and [0, 0, 1]T, respectively. Numerous lexical signals in the work are categorical data, and various categories are independent. One-hot encoding can directly reflect these lexical cues. Then, additional vectors, including character/word embedding, are concatenated with the encoded lexical cue vectors. Our empirical analyses have led us to suggest a more efficient encoding than the traditional one-hot encoding.

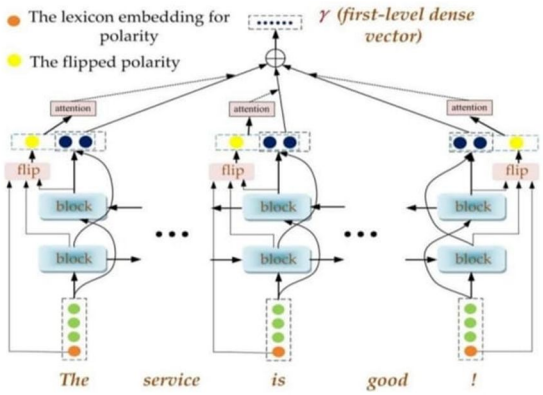


Fig. 4 First-level LSTM with lexicon embedding in both the input and attention layers. [5]

The experimental comparisons mentioned above show that our suggested technique provides the following benefits.
More supervised information may be provided by the two-stage labelling technique, which is helpful for model training, The suggested "two-hot" encoding is more adaptable than "one-hot," making it more suited for cue embedding and additional helpful information about sentiment orientations is included thanks to the embedding of more linguistic signals. The flipping module is useful for enhancing performance as well.

V. CONCLUSIONS

This project aims to summarize, detect, and classify different types of user-generated text, including tweets, comments, and other forms of text. Machine learning techniques will be used to classify the text into various types. The insights obtained from the analysis will be presented in an interactive dashboard for easy interpretation. Proper citation and referencing will be used throughout the project to ensure originality.

VI. FUTURE SCOPE

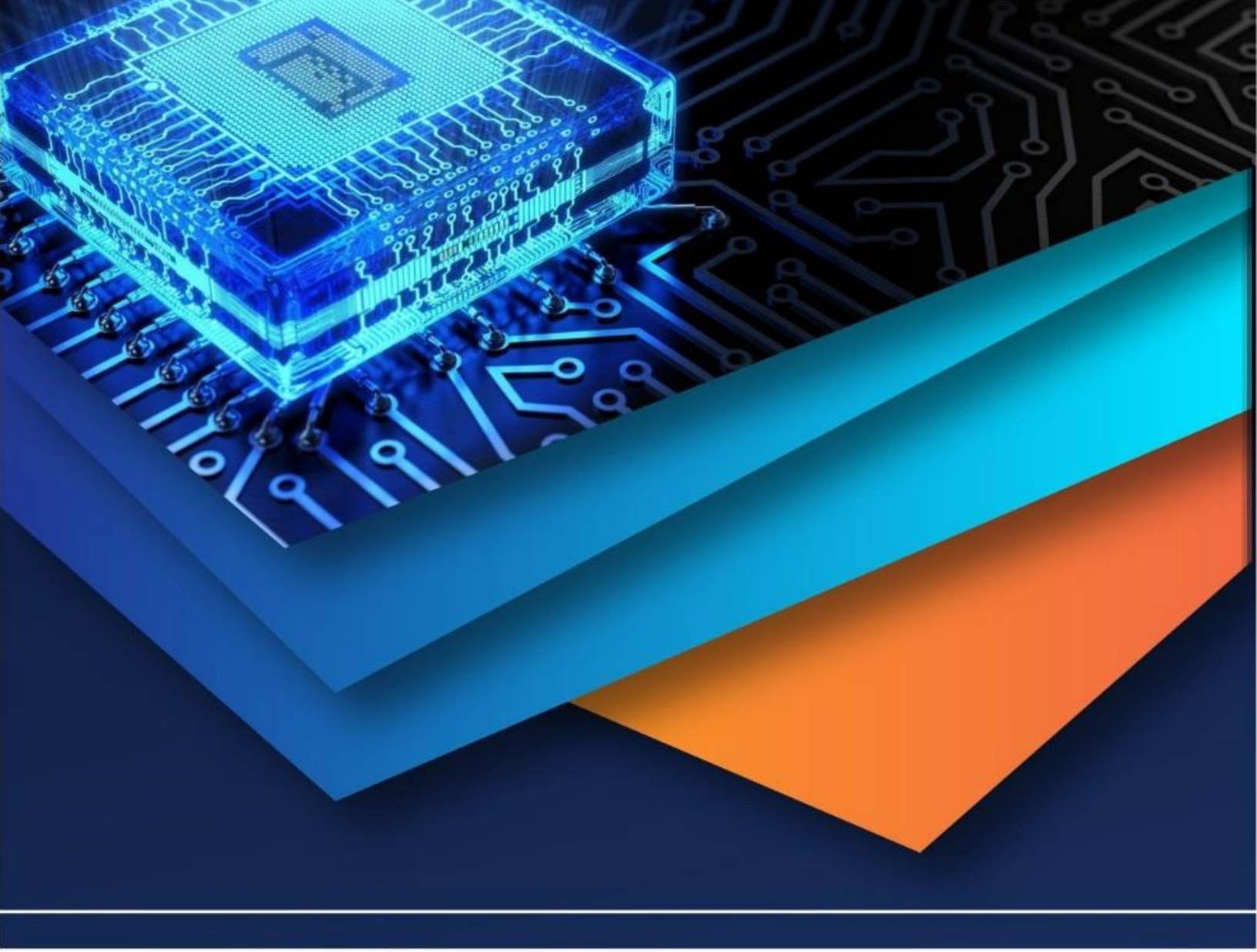
As YouTube continues to grow and more users share their opinions on the platform, sentiment analysis will become increasingly important for businesses and content creators. Moreover, the advancements in natural language processing and machine learning techniques will make it easier to accurately analyze the sentiment of YouTube comments. Therefore, it is important for businesses and content creators to stay up-to-date with the latest trends and techniques in sentiment analysis to remain competitive on the platform.

Platform will focus on developing more accurate and efficient methods for detecting and analysing sentiment in textual data. This may involve exploring new approaches to word polarity detection and sentiment classification using a variety of lexical resources and languages.

Future direction of research lies in applying word polarity identification in the context of moreall-encompassing sentiment analysis applications using different lexicon resources and languages, such as Arabic. The accuracy of NLP based sentiment method is around 75% which can be further improved. Future approaches should study elements linked to metaheuristics and combine parallel computing to accelerate computation. To have favoured employment. An automated web-based ontology system will be used. High-quality labels are crucial for learning systems. Nevertheless, texts with mixed sentiments are difficult for humans to label in text sentiment classification. In this study, a new labelling strategy was introduced to partition texts into those with pure and mixed sentiment orientations. These two categories of texts were labelled using different processes. A two-level network was accordingly proposed to utilize the two labelled data in our two-stage labelling strategy. Lexical cues (e.g., polar words, POS, and conjunction words) are particularly useful in sentiment analysis. These lexical cues were used in our two-level network, and a new encoding strategy, that is, ρ-hot encoding, was introduced. ρ-hot encoding was motivated by one-hot encoding. However, the former alleviates the drawbacks of the latter. Due to labelling noise or context, the polarity of a word varied in different texts. A flipping model was proposed to model the polarity flipping process. Three Chinese sentiment text data corpora were compiled to verify the effectiveness of the proposed methodology. Our proposed method achieved the highest accuracies on these three data corpora. On English data corpora, the proposed method outperformed state-of-the-art algorithms. The proposed two-level network and lexicon embedding can also be applied to other types of deep neural networks. In our future work, we will extend our main idea into several networks and text mining application

REFERENCES

[1] Rawan Fahad Alhujaili; Wael M.S. Yafooz “Sentiment Analysis for Youtube Videos with User Comments” in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), March 2021.
[2] Khaled Abdalgader; Aysha Al Shibli “Experimental Results on Customer Reviews Using Lexicon-Based Word Polarity Identification Method” in IEEE Access (Volume: 8), October 2020.
[3] Guixian Xu; Yueting Meng; Xiaoyu Qiu; Ziheng Yu; Xu Wu “Sentiment Analysis of Comment Texts based on BiLSTM” in IEEE (Volume 7), April 2019 .
[4] Adnan Ishaq; Sohail Asghar; Saira Andleeb Gillani “Aspect-based sentiment analysis using a hybridized approach based on CNN and GA” in IEEE Access (Volume: 8), July 2020.
[5] Ou Wu; Tao Yang; Menyang Li; Ming Li “Two-Level LSTM for Sentiment Analysis With Lexicon Embedding and Polar Flipping” in IEEE Transactions on Cybernetics (Volume: 52, Issue: 5, May 2022).
[6] Hanif Bhuiyan “Retrieving YouTube Video by Sentiment Analysis on User Comment” in Conference: ICSIPA 2017 : IEEE International Conference on Signal and Image Processing Applications At: Kuching, Malaysia, May 2018.
[7] Abhilasha Sancheti, Amit Dixit, and Sudeep D. Thepade. "A review on sentiment analysis techniques and their applications in social media platforms." International Journal of Computer Applications, 184(27), 2018.
[8] Amna Hafiz, Uzma Raja, and Muhammad Farhan. "Sentiment analysis using deep learning: A review." In 2020 IEEE 12th International Conference on Quality, Reliability, Infocom Technology and Industrial Application (ICQRITIA), pp. 320-324. IEEE, 2020.
[9] Jiashen Liu, Jia Liu, Xinyi Wang, and Jian Zhang. "Sentiment analysis on social media: A survey." IEEE Transactions on Computational Social Systems, 7(3), pp. 682-705, 2020.
[10] Sonam Gupta, R. S. Anand, and Jitendra Kumar. "Sentiment analysis in social media using machine learning techniques: A survey." In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 342-347. IEEE, 2019.
[11] Xiaoxu Liu, Jianquan Liu, Xianping Tao, and Yuexiang Yang. "A review of sentiment analysis research based on deep learning." Journal of Ambient Intelligence and Humanized Computing, 12(2), pp. 1475-1489, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY