**“YOUTUBE COMMENT ANALYZER”**

***Major-Project report submitted***

***in***

***Partial fulfillment of requirement for the award of the***

***degree of***

**Bachelor of Technology**

**in**

**Data Science**

***by***

**Mr. Tushar Nandurkar Ms. Pranita Ambulkar**

**Mr. Shubham Mowade Ms. Sakshi Dhage**

***Under the guidance of***

**Dr. Bhupesh Lonkar**

(Assistant Professor,

Data Science)

****

**Department of Data Science**

G H Raisoni Institute of Engineering and Technology, Nagpur

(An Autonomous Institute Affiliated to Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur)

Accredited by NAAC with A+ Grade

**2023-2024**

**“YOUTUBE COMMENT ANALYZER”**

***Major-Project report submitted***

***in***

***Partial fulfillment of requirement for the award of the***

***degree of***

**Bachelor of Technology**

***in***

**Data Science**

***by***

**Mr. Tushar Nandurkar Ms. Pranita Ambulkar**

**Mr. Shubham Mowade Ms. Sakshi Dhage**

***Under the guidance of***

**Dr. Bhupesh Lonkar**

(Assistant Professor,

Data Science)

****

**Department of Data Science**

G H Raisoni Institute of Engineering and Technology, Nagpur

(An Autonomous Institute Affiliated to Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur)

Accredited by NAAC with A+ Grade

**2023-2024**

© G H Raisoni Institute of Engineering and Technology, Nagpur.

**G H Raisoni Institute of Engineering and Technology, Nagpur**

**(An Autonomous Institute)**

**Department of Data Science**

****

**Certificate**

The report of project titled “**Youtube Comment Analyzer”** submitted by **Shubham Mowade, Pranita Ambulkar, Sakshi Dhage & Tushar Nandurkar** in the partial fulfillment of the degree of Bachelor of Technology in **Data Science** during academic year 2023-24, has been carried out under my/our supervision at the Department of Data Science of G H Raisoni Institute of Engineering and Technology, Nagpur. The work is comprehensive, complete and fit for evaluation.

**Name of Guide Prof. Dinesh Banabakode**

**Dr. Bhupesh Lonkar Associate Professor, DS Dept.**

**Dr. Madhuri Tayal Dr. Vivek Kapur**

**HOD, DS Dept. Director, GHRIETN**

**G H Raisoni Institute of Engineering and Technology, Nagpur**

**(An Autonomous Institute)**

**Department of Data Science**



### Declaration

We certify that

1. The work contained in this project has been done by us under the guidance of my supervisor(s).
2. The work has not been submitted to any other Institute for any degree or diploma.
3. We have followed the guidelines provided by the Institute in preparing the project report.
4. We have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
5. Whenever we have used materials (data, theoretical analysis, figures, and text) from other sources, we have given due credit to them by citing them in the text of the report and giving their details in the references. Further, we have taken permission from the copyright owners of the sources, whenever necessary.

**Name & Signatures of the Projectees**

|  |  |
| --- | --- |
| Name of the candidate | Signature |
| Mr. Shubham Mowade |  |
| Ms. Pranita Ambulkar |  |
| Ms. Sakshi Dhage |  |
| Mr. Tushar Nandurkar |  |
|  |  |

**G H Raisoni Institute of Engineering and Technology, Nagpur**

**(An Autonomous Institute)**

**Department of Data Science**

### Acknowledgement

We are extremely thankful to our guide **Dr. Bhupesh Lonkar** under whom our project took the shape of reality from mere idea. We are thankful to our guide for enlightening us with his/her precious guidance and constant encouragement. We thank our guide for providing us with ample support and valuable time. We are indebted to our guide who constantly provided a stimulus to reach our goals.

We are grateful to **Dr. Madhuri Tayal** , HOD, Data Science, GHRIET, for kind co-operation and timely help.

We express our gratitude towards **Dr. Swati Dixit**, Dean Academics, GHRIET, for her never-ending support, planning and motivation.

We are blessed to have **Dr. Vivek Kapur**, our beloved Director Sir as our mentor who is an immense source of motivation and encouragement.

Lastly, we would like to thank all those who were directly or indirectly related to our project and extended their support to make the project successful.

**Name & Signatures of the Projectees**

|  |  |  |
| --- | --- | --- |
| Name of the candidate | | Signature |
| Mr. Shubham Mowade | |  |
| Ms. Pranita Ambulkar | |  |
| Ms. Sakshi Dhage | |  |
| Mr. Tushar Nandurkar | |  |
|  |

**INDEX**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE NAME** | **PAGE NO.** |
|  | **List of Figures** | I |
|  | **List of Tables** | II |
|  | **Abstract** | III |
| **1** | **INTRODUCTION** |  |
|  | 1.1 Overview | 1-2 |
|  | 1.2 Problem Statement | 3 |
|  | 1.3 Thesis Objectives | 4 |
| **2** | **LITERATURE SURVEY & REVIEW** |  |
|  | 2.1 Review Of Literature | 5-7 |
| **3** | **WORKDONE** |  |
|  | 3.1 Methodology | 8-15 |
|  | 3.2 Flow Chart Of Project Process | 16 |
|  | 3.3 Block Diagram | 17-19 |
| **4** | **RESULTS AND DISCUSSION** |  |
|  | 4.1 Result | 20-22 |
|  | 4.2 Limitation | 23 |
| **5** | **SUMMARY AND CONCLUSIONS** |  |
|  | 5.1 Conclusion | 24 |
| **6** | **REFERENCES** |  |
|  | 6.1 References | 25-26 |
| **7** | **APPENDICES** |  |
|  | Authors Note | 27 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **FIGURE**  **NO.** | **FIGURE NAME** | **PAGE**  **NO** |
| 3.1 | The process of scraping | 9 |
| 3.2 | Number of comments in each class | 10 |
| 3.3 | Data Extraction (Phase 1) | 17 |
| 3.4 | V Data Preprocessing (Phase 2) | 18 |
| 3.5 | Classification Model (Phase 3) | 19 |
| 4.1 | Bar graph representation of spam comment. | 21 |
| 4.2 | Line graph representation of spam comment. | 21 |
| 4.3 | Scattered Graph Representation of Spam Comment using QQ Plot | 22 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **TABLE**  **NO.** | **TABLE NAME** | **PAGE**  **NO.** |
| 3.1 | Stopwords ignored in each class | 12 |
| 4.1 | Showing sample predict the result of spam comment | 20 |

**Abstract**

As a YouTube channel grows, each video can potentially collect enormous amounts of comments that provide direct feedback from the viewers. These comments are a major means of understanding viewer expectations and improving channel engagement. However, the comments only represent a general collection of user opinions about the channel and the content. Many comments are poorly constructed, trivial, and have improper spellings and grammatical errors. As a result, it is a tedious job to identify the comments that best interest the content creators. In this paper, we extract and classify the raw comments into different categories based on both sentiment and sentence types that will help YouTubers find relevant comments for growing their viewership. Existing studies have focused either on sentiment analysis (positive and negative) or classification of sub-types within the same sentence types (e.g., types of questions) on a text corpus. These have limited application on non-traditional text corpus like YouTube comments. We address this challenge of text extraction and classification from YouTube comments using well-known statistical measures and machine learning models. We evaluate each combination of statistical measure and the machine learning model using cross validation and 𝐹1 scores. The results show that our approach that incorporates conventional methods performs well on the classification task, validating its potential in assisting content creators increase viewer engagement on their channel.

**CHAPTER 1**

**INTRODUCTION**

* 1. **Overview**

In a digital landscape dominated by video content, particularly on platforms like YouTube, understanding audience engagement has become paramount. The interaction between creators and viewers, as reflected in comments, offers valuable insights into sentiments and opinions. However, manually analyzing a substantial volume of comments is a daunting task prone to subjectivity and bias. This is where sentiment analysis emerges as a game-changer. By automating the process of extracting emotional nuances from text data, sentiment analysis enables a systematic approach to deciphering audience sentiments. we will delve into the intricacies of sentiment analysis, categorizing sentiments into positive, negative, relevant(opinion giving) ,irrelevant(spam comments) and explore its far-reaching applications.

In 2018 YouTube users reached 1.5 billion from around the world and YouTube has more than 2.70 billion monthly active users as of 2023. YouTube is the largest video platform in the world by displaying a variety of media content created by companies or individuals that include music videos, product promotion videos, blog videos, review videos, educational videos [2]. The increase in users is directly proportional to the increasing number of video content uploaded on YouTube. This has become a place for everyone to compete in creating video content and earning income from uploaded videos.

Therefore an approach can be made to find out the perception of YouTube user on video content by using sentiment analysis data obtained from the textual content. Text mining approach becomes the best alternative to interpret the meaning of each comment. The classification of positive and negative content becomes very important for the YouTube user to assess how meaningful the content that has been published is based on user opinion comments.

There have been multiple studies in the field of sentiment analysis such as Twitter sentiment analysis [1], YouTube polarity trend analysis [12], user comment sentiment analysis on YouTube [4], and so on. However, not enough research has been carried out on sentiment analysis through classification of a sentence based on its type. We have approached this issue from the perspective of YouTube comments. Consequently, it is a challenging task

to categorize the comments into different sentence types because of various factors such as non-standard language, spelling errors, unformatted texts, and trivial comments. Apart from these, sometimes there are multiple sentences of different classes on a single comment. The combination of these issues poses a unique challenge in sentiment analysis based on sentence types. One of the simplest ways to address the problem is to categorize the comments purely based on lexicon [3] e.g., the interrogative comments can be identified from keywords such as what, how, and why. Similarly, positive sentences can be identified from keywords like good, best, and wonderful. However, this approach is naive and does not address unique challenges presented by informal texts. Moreover, this method performs poorly if a single comment comprises of multiple categories. Such comments can be categorized more efficiently by appropriately extracting features from the text corpus and using supervised machine learning techniques [9]. Neural networks [21] can be used as a potential solution, however, they are difficult to tune and are not readily explainable. Explain ability is especially important for comments that fall under multiple categories to clearly understand why a resulting category was selected. Our approach is to extract features from preprocessed data and use those features to train well-known supervised learning algorithms. The supervised learning models can then be fine tuned to get the best results. Since the performance of the model is dependent on the text corpus, we select multiple popular fine tuned algorithms for this task and observe their performance. We experiment our YouTube comments dataset with five different fined tuned classification models using two different feature extraction methods to obtain the best results for each model. The accuracy of the models are calculated in terms of cross validation score and 𝐹1 score. Although our approach is simple, the results are effective enabling content creators to view their feedback of interest easily.

There have been multiple studies in the eld of sentiment analy-

sis such as Twitter sentiment analysis [

1

], YouTube polarity trend

analysis [

12

], user comment sentiment analysis on YouTube [

4

],

and so on. However, not enough research has been carried out on

sentiment analysis through classication of a sentence based on

its type. We have approached this issue from the perspective of

YouTube comments. Consequently, it is a challenging task to catego-

rize the comments into dierent sentence types because of various

factors such as non-standard language, spelling errors, unformatted

texts, and trivial comments. Apart from these, sometimes there are

multiple sentences of dierent classes on a single comment. The

combination of these issues poses a unique challenge in sentiment

analysis based on sentence types.

One of the simplest ways to address the problem is to categorize

the comments purely based on lexicon [

3

] e.g., the interrogative

comments can be identied from keywords such as what, how, and

why. Similarly, positive sentences can be identied from keywords

like good, best, and wonderful. However, this approach is naive and

does not address unique challenges presented by informal texts.

Moreover, this method performs poorly if a single comment com-

prises of multiple categories. Such comments can be categorized

more eciently by appropriately extracting features from the text

corpus and using supervised machine learning techniques [

9

]. Neu-

ral networks [

21

] can be used as a potential solution, however, they

are dicult to tune and are not readily explainable. Explainabilit

**1.2 Problem Statement**

YouTube is one of the world's largest video-sharing platforms, hosting a vast amount of content across various genres. Video creators and businesses rely on audience feedback in the form of comments to gauge the reception of their content and gain insights into their viewers' opinions. Analyzing YouTube comments provides valuable data that can be used for content improvement, trend identification, and audience engagement.

These comments can be a valuable source of feedback engagement and insights for content creators. However, it is very challenging to manage and extract meaningful information, from the enormous amount of comments.

The YouTube Comment Analyzer project aims to develop a comprehensive system for collecting, processing, and analyzing comments on YouTube videos. This system will empower content creators, businesses, and researchers to gain insights from user comments, monitor sentiment, and identify key topics and trends.

**1.3 Objectives**

**The objectives of proposed work are as follows:**

### To improve the efficiency and content of the channel by integrating the YouTube comments.

### To categorize the comments based on their form, such as questions, feedback, criticism or suggestions.

### To implement the existing techniques to get the better results with more accuracy.

### To save the time of an content creator by categorizing the comments.

### To study the comments which are affecting to the performance of the channel.

### 

**CHAPTER 2**

**LITERATURE SURVEY & REVIEW**

**2.1 Review of Literature**

* The comment sections of YouTube videos, one of the most prominent video-sharing platforms, serve as a digital forum where users express their opinions, emotions, and feedback. These comments hold a wealth of information, making them an invaluable resource for content creators, researchers, and platform administrators alike. An essential aspect of analyzing YouTube comments is the ability to categorize them into meaningful groups, such as positive, negative, neutral sentiments, and distinguishing spam from genuine user contributions.
* This literature survey embarks on a comprehensive exploration of the methodologies and tools employed to categorize YouTube comments into positive, negative, neutral sentiments, and to identify spam comments. Our goal is to provide a thorough understanding of the evolving landscape of research and applications in this field.The research paper, authored by Rhitabrat Pokharel and Dixit Bhatta, focuses on classifying comments on YouTube videos by employing a combination of sentiment analysis and content-based categorization techniques. Within the scope of various models considered, the paper underscores that Logistic Regression yields the most favorable outcomes in terms of cross-validation and F1 score. This study proposes potential future directions, including the expansion of comment categories and the enhancement of models to accommodate multi-category comments.
* This involves determining the sentiment or emotional tone of a comment, such as positive, negative, or neutral. Sentiment analysis can be performed using various techniques, including machine learning models or rule-based approaches. This is the process of categorizing comments based on their content or topics. This can be done using natural language processing (NLP) techniques like text classification. The paper highlights that among various machine learning models used in the study, Logistic Regression yielded the best results in terms of cross-validation and F1 score. This means that Logistic Regression was the most effective model in classifying comments based on both content and sentiment. Cross-validation is a technique used to assess the performance of a machine learning model. It involves splitting the dataset into multiple subsets, training the model on some of these subsets, and testing it on the remaining subsets. The F1 score is a common metric used to measure the model's accuracy and balance between precision and recall.
* In their study, the authors present a sentiment analysis system that harnesses the capabilities of multiple machine learning classifiers, including Naive Bayes (NB), Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression (LR), k-Nearest Neighbors (KNN), and Random Forest (RF). These classifiers are accompanied by the incorporation of diverse features, which are strategically selected to enhance the accuracy and efficiency of sentiment classification. By utilizing a range of classifiers and features, the research paper aims to optimize the outcomes of sentiment analysis, providing valuable insights into the realm of understanding user sentiment in the context of YouTube comments.
* The research paper titled "N-Gram Assisted Youtube Spam Comment Detection" authored by Shreyas Aiyara and Nisha P Shetty, explores a method for automated machine-assisted detection of spam comments on YouTube. The study investigates the effectiveness of various machine learning algorithms, including Support Vector Machines and Random Forests, in classifying spam comments. The research findings indicate that these algorithms outperform other traditional machine learning methods, particularly when dealing with high-dimensional datasets. This suggests that Support Vector Machines and Random Forests are well-suited for the task of identifying and flagging spam comments on the platform, potentially enhancing the overall user experience by reducing the prevalence of such undesirable content.
* In summary, this research is significant because it addresses a common issue faced by online platforms like YouTube: the presence of spam comments. By leveraging machine learning algorithms like Support Vector Machines and Random Forests, the study demonstrates an effective way to automatically detect and filter spam comments. This research has practical implications for content moderation on YouTube, potentially leading to a cleaner, safer, and more user-friendly environment for its vast user base.
* The literature survey examines the correlation between YouTube video likes and positive comments through sentiment analysis, with a focus on enhancing accuracy. Several key areas for improvement have been identified, including the utilization of more robust training data, implementing effective comment filtering mechanisms to eliminate irrelevant data, and expanding the depth of video analysis. These proposed enhancements aim to advance the accuracy and reliability of sentiment analysis as a means of understanding the relationship between user engagement metrics and the sentiment expressed in YouTube comments.
* Abbi Nizar Muhammad, Saiful Bukhori ,” Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier” Proc. ICOMITEE 2019, October 16th-17th 2019, Jember, Indonesia The document discusses a research study on sentiment analysis of YouTube comments using the Naïve Bayes - Support Vector Machine (NBSVM) classifier. The study aims to understand and extract sentiment information from YouTube video comments to assess the positive and negative content. The research found that the combination of Naïve Bayes and Support Vector Machine produced better accuracy and performance. The testing results showed a precision of 91%, recall of 83%, and f1 score of 87%.
* HAYOUNG OH , A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model Received September 26, 2021, accepted October 17, 2021, date of publication October 19, 2021, date of current version October 28, 2021. This paper proposes a technique to detect spam comments on YouTube using a Cascaded Ensemble Machine Learning Model. The experimental results show that the ensemble model with soft voting (ESM-S) performs the best in terms of accuracy, F1-score, and MCC. The proposed model is tested on datasets from various categories, and it consistently demonstrates good performance in detecting spam comments.

**CHAPTER 3**

**WORKDONE**

**3.1 Methodology**

The interaction utilized in Figure 1 for spam remark discovery has been expounded in this part. The previously mentioned informational collection establishes both spam and ham remarks. The all-out dataset has been combined and embedded for preparation. The preparation dataset has been pre-handled in the following stage and the equivalent has been handled through the Naïve Bayes calculation.

A model has been anticipated and the testing information is embedded in the anticipated model to check the forecast aftereffect of the model. In this model the information dataset is ordered into two sections, first and foremost, utilized for preparing and also, for testing. The dataset utilized for preparing has been expounded in the next three stages. In the first phase, the unorganized data gets converted into a more organized form which in term helps to filter the data. The subsequent stage runs after the highlight assortment. If the said information is moved to Naïve Bayes calculation, a few copy information can be acquired which is then changed over or limited to a bunch of capacity known as vector. In the third stage, Count-Vectorizer is utilized to make a lattice where each particular word is addressed by the segment of the framework and the cell worth of the network addresses the coordinating with word check.

After the utilization of Count Vectorizer, information fitting has been refined. the information fitting ought to be executed so that it fabricates a model with the innocent Bayes calculation .

The information which is utilized as Counter-Vectorizer is being carried out by a capacity known as Multinomial NB which is reasonable for arrangement through Naive Bayes classifier and has been utilized in this exploration to acquire the normal outcome.

**Data Collection:**

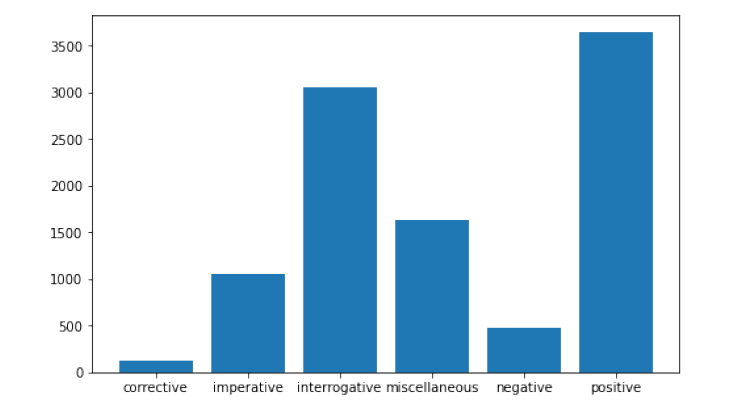
We build our dataset by scraping YouTube comments. We use the YouTube Data API for authenticating and accessing the comments on videos . First, the API is used for authentication and then the credentials obtained are fed into the comment extractor. The comment extractor then scrapes comments from the comment section by scrolling through all the comments and loading them dynamically.

The dataset consists of 10,000 comments picked from different tutorial videos 1 . We chose tutorial videos for our experiments because they contain a wide variety of comments. We then manually label the comments into 4 different classes: Positive, Negative, Neutral & spam. These classes are defined based on general needs of the YouTubers. Note that further categories can be established if or as needed. These classes belong to broader class: Sentiment Analysis (positive and negative). The classes are explained in more detail next. Positive tells that the viewers perceived the content as worthy and that the content created a positive impact on them. Negative provides information on what is wrong with the content and why the viewers are not attracted to it.

****

**Figure 3.1:** The process of scraping

All other trivial comments. We manually labelled the miscellaneous category from the creators’ point of view. As mentioned previously, some of the comments can belong to more than one class. For instance, the first sentence in "Your solution is not practical. Can you suggest another one?", suggests the negative class while the second sentence suggests the interrogative class. In such situations, we classified the comment based on the importance to the content creator. In this example, we assumed it as an interrogative sentence because it is more important to answer the question to increase odds of the viewer to return and stay engaged in the content of the channel. Fig. 3.2 shows the visual presentation of the quantity of comments in each class.

****

**Figure 3.2:** Number of comments in each class

**Data Pre-processing:**

It is important to clean the data and have them in appropriate format to improve classification. The data pre-processing step handles the following factors that make the classification process difficult:

(1) Non-standard language: The texts used in the comments section do not always employ standard English. Comments often contain slangs and improper form of words, making it

difficult to extract features from them.

(2) Spelling errors: Commenters often do not pay attention to spellings due to the informal setting. Such spelling errors need to be corrected; otherwise, the words would add unnecessary features to the classification model, decreasing the overall accuracy of the classifier and impacting efficiency of the approach. For example, the words "plz" and "please" convey the same meaning in informal writing, but if the incorrect spelling is ignored, they would be treated as two different words by the classifier.

(3) Unformatted texts: These refer to comments containing computer codes. These do not contribute to the feature extraction

accuracy; rather, they add unnecessary load to the feature matrix.

(4) Trivial comments: Not all the comments posted were about the video or related to the channel.Alarge number of viewers comment in order to market their products or just to show their presence. These comments are not useful to the content

creators and only add unnecessary overhead.

Above issues are common in platforms like YouTube because of the informal nature of communication. We addressed these issues using the following pre-processing steps:

• lowercasing

• removing URLs

• removing new line character ("\n")

• removing punctuations

• removing integers

• removing emojis

• correcting spelling errors

• lemmatizing

• removing stopwords

Given the nature of this study, lowercasing was relevant because the same word would have been identified as a different feature had some letters been capitalized (for example, "Love" and "love" are two different words for a computer). The removal of URLs, new line characters, punctuation, integers, and emojis was performed because they did not provide useful information for feature extraction; rather adding unnecessary complexity to the model.

Since there are a lot of spelling errors in informal writing, we corrected the typos using the autocorrect-library from Python. We used Lemmatization to analyze the words morphologically and group the similar words together. Furthermore, there are certain frequent words that do not add significant meaning to the sentence such as "is", "are", and "it". They were removed from the corpus. However, stopwords were not removed from all the classes of comments because stopwords for one category might be important for another category. For example, "not" and "no" are important for the negative class whereas they are not important for other classes. Stopwords were used from nltk English corpus, which consists of 179 stopwords. Table. 3 shows the stopwords that were not considered in each category.

**Converting text into features :**

Once pre-processing is completed, the pre-processed text is filtered so that only the necessary features remain in the corpus. This helps in reducing the load to the classifiers and also increases the accuracy of selected models [19]. The well-known techniques for vectorizing a corpus of text include document frequency, tf-idf vectorizer, hashing vectorizer, and Word2Vec. We selected document frequency vectorizer and tf-idf vectorizer for this paper. Using these two methods we can study the behaviour of different classification models under two different Converting text into features Once pre-processing is completed, the pre-processed text is filtered so that only the necessary features remain in the corpus. This helps in reducing the load to the classifiers and also increases the accuracy of selected models .

**Class Ignored stopwords**

**Positive** NA

**Negative** no, not

**Interrogative** how, what, which, who, whom, why, do, is

does, are, was, were, will, am, are, could, would

should, can, did, does, do, had, have

**Imperative** could, would, should, can

**Corrective**  NA

**Miscellaneous** NA

**Table 3.1:** Stopwords ignored in each class

The well-known techniques for vectorizing a corpus of text include document frequency, tf-idf vectorizer, hashing vectorizer, and Word2Vec. We selected document frequency vectorizer and tf-idf vectorizer for this paper. Using these two methods we can study the behaviour of different classification models under two different conditions. Document frequency (df ) vectorizer gives importance to the term that has higher frequency in the document; whereas, tf-idf can incorporate the terms that are rarely present in the document. Unlike hashing vectorizer, we can examine the text features which are important to the model using the vectors generated by df and tf-idf. For rare or out of vocabulary terms (which might be important to a model), Word2Vec can not create an ideal vector for them and it is difficult to interpret those vectors because of hidden layers.

When calculating document frequency (Eqn. 1), if the same term is present multiple times in a comment, then its additional counts are not considered. Also, the terms that appear in less than or equal to 5 comments are ignored because they do not add value to the features. In the same way, if any term appears in the majority of the comments, it does not add value to the feature because it is not the distinguishable feature for a class. These terms are likely already filtered by the stopwords removal process. However, we ensure that only terms that significantly add value to their comment’s class are considered.

𝑑 𝑓 = 𝑛/𝑁 (1)

where, 𝑑 𝑓 denotes document frequency, 𝑛 denotes number of documents in which the term appears, and 𝑁 denotes total number of documents where document means comment. After above steps, 2210 terms (features) were derived and scaled from 0 to 1 using a min-max scalar (normalization). We performed this because some of the machine learning models cannot handle large ranges of data. Doing so also helps in speeding up some of the calculations.

𝑀𝑖𝑛𝑀𝑎𝑥𝑆𝑐𝑎𝑙𝑒𝑟 (𝑥) = 𝑥 – 𝑥𝑚𝑖𝑛 / 𝑥𝑚𝑎𝑥 – 𝑥𝑚 (2)

where, 𝑥 is observed value, 𝑥𝑚𝑖𝑛 is the minimum value of that class and 𝑥𝑚𝑎𝑥 is the maximum value of that class. The second feature extraction technique used in this paper is tfidf (term frequency - inverse document frequency). It not only considers the frequent terms, but also the rare terms.

𝑡 𝑓 -𝑖𝑑 𝑓 = 𝑡 𝑓 ∗ log 1/ 𝑑 (3)

where, 𝑡 𝑓 is the term frequency and 𝑑 𝑓 is the document frequency For td-idf, we got 4304 features when both unigram and bigram were taken into account.

**Model Selection and Hyperparameter Tuning :**

Each machine learning model has its own strengths and weaknesses when dealing with a text corpus. The fitness of a model for a specific dataset depends on the characteristics of the model as well as the features of the dataset. Since our text corpus is dense and has numerous features which significantly affects the classification task, the machine learning models selected are based on the density of the features and the number of classification categories (binary or multiple). Linear Support Vector Classification (Linear SVC), Logistic Regression, Multinomial Naive Bayes (Multinomial NB), Random Forest Classifier, and Decision Tree Classifier are selected as we work with a dense dataset and multiple classes. Especially,

Random Forest Classifier helps prevent the overfitting problem and mitigates the impact of outliers, whereas Decision Tree Classifier can easily deal with the irrelevant features on the dataset.

For each model selected, the outcome can be enhanced by optimizing the hyperparameters. To understand how different parameters can significantly affect the performance of each model, we experimented all the models with the minimum subsets of those

parameters (grid search strategy). The values of the parameters that were tested for both 𝑑 𝑓 and 𝑡 𝑓 −𝑖𝑑 𝑓 measures. The columns "Best Value for" present the value with the best result, which were used for the final experiment.

**3.2 Flow Chart of Project Process**

Data Collection from Open Source

Data processed Manually

Input Training Dataset

Data Pre-processing

Machine Learning Algorithm (Navie Bayes)

Input Testing Dataset

Predict Model

Prediction of not spam

Prediction of spam

**3.3. Block Diagram**





Positive Class

Extracting Comments

Labeling

Google API

(YouTube Data API v3)

Neutral Class

Negative Class

**Figure 3.3: Data Extraction (Phase 1)**

We select a set of YouTube videos or channels to analyze. We use the YouTube Data API to fetch video comments.We obtain API access credentials to fetch the data.

**Data Preprocessing**



Noise Removal

Normalization

Tokenization

Positive Class

Negative Class

Neutral Class

Collected

Datasets

**Figure3.4: Data Preprocessing (Phase 2)**

We clean the data by removing timestamps, emojis, and URLs. We tokenize the comments and perform basic text processing like lowercasing.

Classification Model

Classification Model

Classification Model

Classification Model

**Classification Model**

Classification Model



Neutral Class

Negative Class

Positive Class

Naïve Bayes

Result

**Figure 3.5: Classification Model (Phase 3)**

In the classification phase a model is trained over a labeled data the Naïve Bayes Algorithm is used for it . The preprocessed comments are passed through the model and provides result based on those classificaiton

**CHAPTER 4**

**RESULT AND DISCUSSION**

* 1. **Result and Observation**

|  |  |  |
| --- | --- | --- |
| PREDICTIONS SPAM | ORIGINALS SPAM | COMMENTS |
| Spam | spam | http://glearn.io/2z8qp.com |
| ham | ham | hey there you are |
| spam | ham | you are a nice guy |
| ham | ham | ðŸ‘• |
| Spam | spam | get some nuts baby |
| ham | ham | look son there who is standing |
| spam | spam | I love you.com |

**Table 4.1** Showing sample predict the result of spam comment

Table 4.1, showing test foresee the consequence of spam remark of YouTube channel utilizing Naïve Bayes classifier. Rejecting has been performed with next to no distinction in both as far as the precision level. Concerning the above table, one can say that a spam sifting strategy utilizing the Naive Bayes classifier forecognizable proof of spam comment on YouTube. gives a good and adequate outcome.

The predicted dataset by machine and manually scrapping has been performed with very little

difference in both in terms of the accuracy level. With the assistance of this dataset, Bar diagram, Line Graph and QQ plot are made for accuracy checking displayed in underneath:



**Figure 4.1** Bar graph representation of spam comment.

In the Figure 4.1 shown above represent a bar graph where x axis represents number of YouTube video result and y axis represent the no of spam comment, and the blue line shows the prediction spam comment andred line show the original spam comment and the graph is shown that we have gathered comment from 15 different YouTube channel and the data has been plotted which is shown that no of prediction comment an no of original comment.

****

**Figure 4.2** Line graph representation of spam comment.

In Figure 4.2: the x axis represent number of YouTube videos and y axis represents the number of spam comments. In this bar graph the blue line shows the number of prediction spam comment and the red line represents the original spam comments. Mean and SD has been calculated 4.5 and SD respectively.

Standard Deviation is defined as how the calculations for a group are scattered out from the average (mean or expected value). A low standard deviation implies that the vast majority of the numbers are near the average, while a high standard deviation implies that the numbers are more fanned out. Since the SD of this data is 1.68[38]. So, it can be concluded that the model has produced good result and the accuracy level of the model is approximately 98%.

****

**Figure 4.3** Scattered Graph Representation of Spam Comment using QQ Plot

In Figure 4.3 the x axis represents the prediction distribution and y axis represents the original distribution. The middle part of the chart forms a linear plot which means that the middle range of the prediction distribution Correctly map the middle range of the original distribution. The two ends of the graph represent the data in the tail is not linear

**4.2 Limitations**

**[1]Access to Comments:** YouTube's API and data policies may limit the number of comments you can access and analyze, so you might not be able to analyze all comments on a popular video.

**[2]Privacy and Data Protection:** Ensure that you are not violating YouTube's terms of service or any privacy regulations when collecting and analyzing comments. Protect the privacy of users and respect their consent.

**[3]Comment Quality:** The quality of comments can vary widely, from insightful and relevant to spam or offensive. Your analyzer may need to filter out irrelevant or inappropriate comments**.**

**[4]Sentiment Analysis:** Sentiment analysis can be challenging as comments may contain sarcasm, irony, or cultural nuances that affect the accuracy of sentiment classification.

**[5]Multilingual Comments:** If your project involves comments in multiple languages, you'll need a robust language detection and translation system to handle diverse content.

**[6]Frequency** **Limits:** YouTube's API may impose rate limits on the number of requests you can make, so you may need to manage your requests accordingly.

**[7]Real-time Data:** If your project requires real-time comment analysis, be aware that there might be delays in retrieving and processing comments through the API.

**[8]YouTube API Changes:** Keep up to date with changes to the YouTube API, as YouTube can modify its policies and access methods.

**[9]Copyright and Fair Use:** Respect copyright and fair use policies when using YouTube comments and data in your project.

**[10]Ethical Considerations:** Ensure your project adheres to ethical guidelines, especially if it involves sensitive topics or user data.

**CHAPTER 5**

**CONCLUSION**

**­5.1 Conclusion**

The YouTube Comment Analyzer, integrated with the Google YouTube API, proves to be a valuable asset for content creators and platform administrators. It streamlines the process of analyzing comments, offering insights into audience sentiment, engagement, and demographics. This data-driven approach enhances content creation and audience engagement, ultimately contributing to the success and growth of YouTube channels. As digital content continues to thrive, this tool remains instrumental in understanding and responding to the YouTube community effectively.

Moreover, the analyzer generates a summary of common feedback and recurring themes within comments. It acts as a compass, guiding content creators toward areas that require attention or expansion. By addressing frequently raised issues and integrating viewer suggestions, creators can fine-tune their content and enhance audience satisfaction. It's a crucial tool for continuous improvement and maintaining a harmonious relationship with the audience.

**CHAPTER 6**

**REFERENCE**

**6.1 References:**

[1]Abinaya, R., E, B. N., & Naveen, P. (2020). Spam Detection On Social Media Platforms. *2020 7th International Conference on Smart Structures and Systems (ICSSS)*. <https://doi.org/10.1109/icsss49621.2020.9201948>

[2]S. Aiyar and N. P. Shetty, ``N-gram assisted Youtube spam comment detection,'' *Proc. Comput. Sci.*, vol. 132, pp. 174\_182, Jan. 2018, doi: 10.1016/j.procs.2018.05.181.

[3]Kantchelian, J. Ma, L. Huang, S. Afroz, A. Joseph, and J. D. Tygar,``Robust detection of comment spam using entropy rate,'' in *Proc.* *5th ACM Workshop Secur. Artif. Intell. (AISec)*, 2012, pp. 59\_70, doi:10.1145/2381896.2381907.

[4]Madden, I. Ruthven, and D. Mcmenemy, ``A classi\_cation scheme for content analyses of Youtube video comments,'' *J. Documentation*, vol. 69, no. 5, pp. 693\_714, Sep. 2013, doi: 10.1108/JD-06-2012-0078.

[5]Severyn, A. Moschitti, O. Uryupina, B. Plank, and K. Filippova, ``Opinionmining on Youtube,'' in *Proc. 52nd Annu. Meeting Assoc. Comput.* *Linguistics (Long Papers)*, vol. 1, 2014, pp. 1\_10, doi: 10.3115/v1/P14-1118.

[6]T. C. Alberto, J. V. Lochter, and T. A. Almeida, ``TubeSpam: Comment spam \_ltering on Youtube,'' in *Proc. IEEE 14th Int. Conf. Mach. Learn.* *Appl. (ICMLA)*, Dec. 2015, pp. 138\_143, doi: 10.1109/ICMLA.2015.37.

[7]S. Sharmin and Z. Zaman, ``Spam detection in social media employing machine learning tool for text mining,'' in *Proc. 13th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Dec. 2017, pp. 137\_142, doi:10.1109/SITIS.2017.32.

[8]Alhujaili, R. F., & Yafooz, W. M. S. (2021). Sentiment Analysis for Youtube Videos with User Comments: Review. *IEEE*. <https://doi.org/10.1109/icais50930.2021.9396049>

[9]Muhammad, A. N., Bukhori, S., & Pandunata, P. (2019). Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier. *IEEE*. <https://doi.org/10.1109/icomitee.2019.8920923>

[10]Rong, Y., Singh, S., Cao, P., Chi, E., & Fu, B. (2016). Video Watch Time and Comment Sentiment: Experiences from YouTube. *IEEE*. https://doi.org/10.1109/hotweb.2016.13

[11]Douiji, Y., Mousannif, H., & Hassan, A. M. (2016b). Using YouTube comments for text-based emotion recognition. *Procedia Computer Science*, *83*, 292–299. <https://doi.org/10.1016/j.procs.2016.04.128>

[12]Siersdorfer, S., Chelaru, S., Nejdl, W., & Pedro, J. S. (2010b). How useful are your comments? *Ieee*. <https://doi.org/10.1145/1772690.1772781>

[13]Severyn, A., & Moschitti, A. (2015b). Twitter Sentiment Analysis with Deep Convolutional Neural Networks. *Ieee*. <https://doi.org/10.1145/2766462.2767830>

[14]Cunha, A. a. L., Costa, M. C., & Pacheco, M. a. C. (2019b). Sentiment analysis of YouTube video comments using deep neural networks. In *Lecture Notes in Computer Science* (pp. 561–570). <https://doi.org/10.1007/978-3-030-20912-4_51>

[15]Kavitha, K., Shetty, A. B., Abreo, B., D’Souza, A., & Kondana, A. (2020). Analysis and

classification of user comments on YouTube videos. *Procedia Computer Science*, *177*, 593–

598. <https://doi.org/10.1016/j.procs.2020.10.084>

**CHAPTER 7**

**APPENDICES**

**7.1 AUTHOR’S NOTE**

|  |  |  |
| --- | --- | --- |
|  | | |
|  |  | Name: Shubham Mowade  [Email Id: shubham.mowade.ds@ghrietn.raisoni.net](mailto:Email%20Id:%20shubham.mowade.ds@ghrietn.raisoni.net)  Address: Ward No.2, Nandagomukh,Saoner, Nagpur Contact No: 9096060646 |

|  |  |  |
| --- | --- | --- |
|  | | |
|  |  | Name: Tushar Nandurkar  [Email Id: deepak.tushar.ds@ghrietn.raisoni.net](mailto:Email%20Id:%20deepak.tushar.ds@ghrietn.raisoni.net)  Address: Plot No.174 Shastri Nagar, Bhandara  Contact No: 9403816548 |

|  |  |  |
| --- | --- | --- |
|  | | |
|  |  | Name: Pranita Ambulkar  [Email Id: pranita.ambulkar.ds@ghrietn.raisoni.net](mailto:Email%20Id:%20pranita.ambulkar.ds@ghrietn.raisoni.net)  Address: Plot No. 130 Chandramani Nagar, Nagpur  Contact No: 9172627246 |

|  |  |  |
| --- | --- | --- |
|  | | |
|  |  | Name: Sakshi Dhage  [Email Id: sakshi.dhage.ds@ghrietn.raisoni.net](mailto:Email%20Id:%20sakshi.dhage.ds@ghrietn.raisoni.net)  Address: Plot No.4 Geeta Nagar, Nagpur  Contact No: 9322077301 |

**Photo of Projectees along with the Guide and Project**

