IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

YouTube Comment Analyser

[1]Aditya Neve, [2]Kalpesh Pachpute, [3]Bhimashankar Mathapati, [4]Prerana Thorve

Guided by- Miss Dipika Paranjape

[1]Computer Science(Artificial Intelligent) ,
[1]Nutan College of Engineering and Research, Pune.

Abstract: The YouTube Comment Analyzer is a Python-based tool that helps content creators and marketers analyse and organize user-provided comments on YouTube videos. It uses Python's robust capabilities to efficiently process and categorize comments, overcoming challenges in the unstructured nature of comments. We categorize the user comments posted on YouTube video sharing website based on their relevance to the video content given by the description associated with the video posted. Comments are analysed for polarity and are further segregated as positive or negative. A comparative analysis of classifier using the Bag of Words and Association List approaches is presented.

Index Terms - YouTube comment, Sentiment Analysis, Classification, Machine Learning

I. INTRODUCTION

In the digital age, where online platforms serve as a hub for diverse content, YouTube stands out as a prominent space for sharing videos and fostering community engagement. With billions of users worldwide, YouTube has become a vast land space of opinions, reactions, and conversations encapsulated within its comment sections. Understanding the sentiments, trends, and dynamics embedded in these comments can provide invaluable insights into user behaviour and content reception. In an era where digital communication has become synonymous with everyday life,

YouTube serves as a digital agora, bringing together a multitude of voices, opinions, and perspectives. The comment section, often referred to as the "heartbeat" of the platform, is a dynamic space where users engage in conversations, share their thoughts, and express reactions to the vast array of videos spanning genres from educational content to entertainment and beyond.

This report endeavours to be a compass in navigating the vast sea of YouTube comments, utilizing sophisticated analytical tools to chart the depths of user interactions. Beyond the surface-level expressions of likes and dislikes, the aim is to uncover the nuanced layers of sentiment, identify the prevailing themes that captivate audiences, and ultimately unravel the intricate social fabric woven within the comment sections. YouTube comments are more than mere textual annotations; they are a collective voice that echoes the sentiments of the platform's diverse user base. This analysis seeks to answer questions that transcend the superficial, delving into the motivations, preferences, and patterns that define user engagement. By understanding the dynamics at play within YouTube comments, stakeholders can make informed decisions to improve content creation, optimize user experience, and foster a sense of community among viewers. As we embark on this journey into the heart of YouTube comments, our intent is not only to unravel the intricacies of user behaviour but also to contribute to the broader conversation on digital communication. By decoding the language of YouTube comments, we aim to shed light on the rich tapestry of human expression in the digital age and, in doing so, provide a valuable resource for those seeking to navigate the ever-evolving landscape of online interaction. Explore the prevailing sentiments expressed in YouTube comments to identify patterns of positivity, negativity, or neutrality. Understanding the emotional tone of comments can shed light on how users

perceive and react to content. Group comments into clusters based on common themes or topics. Unveiling the prevalent subjects of discussion can aid content creators, marketers, and platform administrators in adapting strategies to meet audience expectations.

1.1 HISTORY

The YouTube Comment Analyser project originated from the need to understand and analyse user engagement on the popular video-sharing platform, YouTube. The project's inception was marked by a recognition of the significance of user comments as a valuable source of feedback, sentiment, and interaction within the online community. During the initial phase, extensive research was conducted to identify existing tools, technologies, and methodologies for analysing YouTube comments. The team engaged in a comprehensive requirement analysis to define the scope, objectives, and features of the YouTube Comment Analyser. With a clear understanding of the project requirements, the development phase commenced. Prototyping played a crucial role in visualizing the user interface and functionality. The team experimented with different algorithms for sentiment analysis, natural

language processing, and data visualization. The first major milestone was achieved with the finalization of the user interface design. User friendly and intuitive interfaces were prioritized to ensure accessibility for a diverse user base. Feedback from potential users was incorporated to enhance the overall user experience. Simultaneously, the back end development was underway. This involved the implementation of algorithms for comment parsing, sentiment analysis, and data storage. Integration with YouTube's API was a critical aspect, allowing real-time data retrieval and analysis. An alpha version of the YouTube Comment Analyser was released for internal testing. This phase involved rigorous testing to identify and address any bugs, usability issues, or performance concerns. Feedback from the testing phase was invaluable in refining the tool and ensuring its reliability. The beta version marked a significant step towards a broader audience. The YouTube Comment Analyser was made available to the public for testing and feedback. User input during this phase helped fine-tune features, improve accuracy, and address specific user needs. After thorough testing and refinement, the YouTube Comment Analyser reached its final release. Comprehensive user documentation was provided to assist users in effectively utilizing the tool. The project was officially launched, and user feedback continued to be monitored for future updates. The YouTube Comment Analyser project remains an ongoing initiative with a commitment to continuous improvement. Regular updates are released to address emerging trends, user requirements, and platform changes, ensuring the tool's relevance and effectiveness in the ever-evolving landscape of online content.

1.2 LITERATURE SURVE

TABLE OF LITERATURE SURVEY

Sr. no	Author	Title of paper	Year	Technology	Outcome	Limitation
1	Alexander Lighter & Bedor Tekinerdogan	Systematic reviews in sentiment analysis: a tertiary study	(2021)	SLR: These studies are performed to aggregate results related to specific researchquestions.	This tertiary study provides a comprehensive overview of key topics and approaches for various tasks in sentiment analysis.	The selection process has limitations as the inclusion criteria are restricted to SLR and SMS papers.
2	RitikaSingh	Youtube comment sentiment analysis	(2021)	We used python based machine learning library named Scikit-Learn for implementing the system. Scikit-Learn is a well-known machine learning library tightly integrated with Python language and provides easy-to-interact interface.	Machine learning algorithms were utilized for classification, with evaluation metrics indicating the best F-score in macro, RF in micro average, LR in micro average, and Uni-gram in superior performance. .	SVM outperforms other classifiers, with Naïve Bayes following suit. SVM excels in macro average Fscore and accuracy measures, while random forest excels in micro average.

3	Muhammad Zubair Asghar1, Shakeel Ahmad2, Afsana Marwat1, Fazal Masud Kundi1	Sentiment Analysis on YouTube: A Brief Survey	(2020)	The text provides a list of techniques for analyzing sentiment, including an unsupervised Lexicon-based Sentiment Approach, a Breadth-first Search, a SentimentNet Thesaurus, a Setiment Analysis, and a Knowledgebase System.	The process involves identifying a user, clustering videos, filtering unrated comments, and accurately analyzing settings for emotions and valence annotation. .	Detecting sentiment polarity on social media, especially YouTube, is challenging due to limitations in current sentiment dictionaries, which lack proper sentiments for community-created terms.
4	Brian Alafwan1 ,Manahan Siallagan1 , Utomo Sarjono Putro1	Comments Analysis on Social Media: A Reviw	(2023)	The research in this field is limited, with a need for a comprehensive literature review to identify strengths and limitations, identify methodologies, regions, tools, and platforms used.	Research on social media comments is divided into content analysis and sentiment analysis. Content analysis objectively identifies messages' properties, while sentiment analysis assesses opinion, polarity, and author's assessment of a topic's features.	The research in this field is currently under-explored, with limited understanding of its strengths and limitations, including methodologies, regions, tools, research objects, and platforms used.
5	Franziska Oehmer-Pedrazzi ·Sabrina Heike Kessler · Edda Humprecht · Katharina Sommer · Laia Castro	Standardisierte Inhaltsanalyse in der Kommunikationswissenschaft – Standardized Content Analysis in Communication Research		The focus should be on providing transparency and reproducibility rather than a map to measure meaning in big data.	The study explores the future of coding by comparing hand-coding with three different computer-assisted text analysis methods. .	The deduction is limited to satirical TV shows from a single country, with intentional bias in corporate communication targeting specific stakeholder groups.

GAP FINDINGS

1. Unstructured and Informal Language

YouTube comments often contain informal language, slang, and expressions that challenge traditional sentiment analysis algorithms. Understanding the nuances of user sentiments in this unstructured environment remains a significant challenge.

2. Scale and Volume

Popular videos on YouTube can accumulate thousands or even millions of comments. Analyzing such vast amounts of data in real-time poses scalability challenges. Efficient algorithms are required to process and extract insights from large datasets without compromising performance.

3. Lack of categories

The system retrieves YouTube video comments, preprocesses, tokenizes, extracts keywords, categorizes, sorts, presents, and refines algorithms, optimizes for speed and scalability, handles errors gracefully, and continuously improves performance.

System architecture and Methodology:

The purpose of the methodology is defined in this section. Our methodology is depicted in Fig. 1. First of all, we used the annotated dataset. We used python based machine learning library named Scikit-Learn for implementing the system. Scikit-Learn is a well-known machine learning library tightly integrated with Python language and provides easy-to-interact interface. First of all our system reads the data stored in the file having (Tab Separated Values) format. After reading, pre-processing phase is applied to clean and prepare the data for the use of machine learning algorithms. Directly text data cannot be given to machine learning algorithms, it should be converted into a suitable type. Using Scikit-Learn module named "countvectorizer ", the text data firstly convert into numeric format and prepare the matrix of tokens count. Now the data is ready for machine learning algorithms. Then 60% of data is splitted randomly to train the classifier and 40% for testing the classifier' accuracy. We perform our experiments in two phases, firstly we just apply N-grams (Length 1-3) features on data

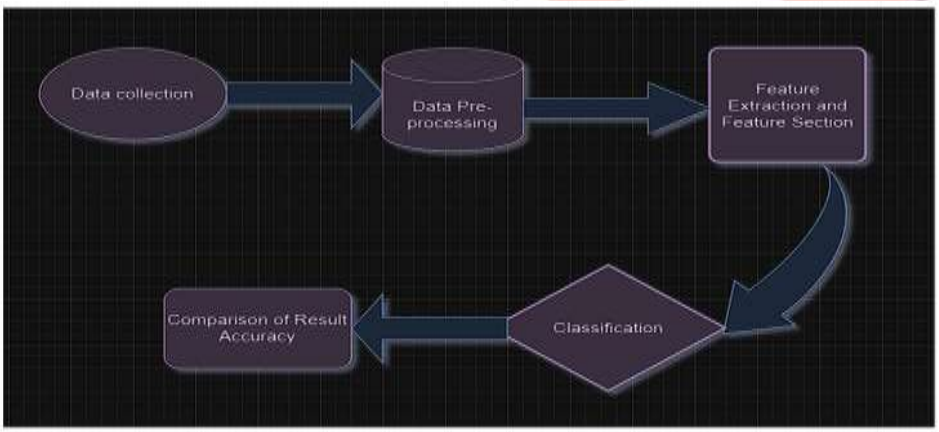


Fig:- Feature on Data

This work attempted to utilize six machine learning techniques for the task of sentiment analysis. The modeling of all techniques is briefly discussed below.

After pre-processing and features selection the very next step is to apply classification algorithms. Many text classifiers have been purposed in literature. We have used 6 algorithms of machine learning including Naïve-Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), and Random Forest (RF). Naïve Bayes: Naïve- Bayes is the most popular classification algorithm due to its simplicity and effectiveness. This classifier works according to the concept of Bayes theorem. It's a kind of module classifier that follows the idea of probabilities for the purpose of classification.The benefit of using Naïve Bayes on text classification is that it needs less dataset for training. removal of numeric, foreign words, html tags and special symbols yielding the set of words. This pre-processing produces word-category pairs for training set. Consider a word 'y' from test set (unlabeled word set) and a window of n-words (x1, x2, …… xn) from a document.

The conditional probability of given data point 'y' to be in the category of n-words from training set is given by:

$$P(y/x1,x2,\ldots\ldots xn)=P(y)\times\prod i=1nP(xi/y)P(x1,x2,\ldots\ldots xn)$$

Types of Classification Algorithms:

1. Support Vector Machines (SVM)
2. Decision Trees
3. Random Forest

Support Vector Machines (SVM) SVMs, or Support Vector Machines, are a type of supervised learning models that arecommonly used for classification and regression tasks. They gained popularity in theearly1990s due to their ability to perform well in high-dimensional spaces andtheireffectiveness across various types of data.

The main goal of an SVM is to find the best hyperplane that can separatedatapoints of different classes with the maximum margin. It works by identifying a hyperplanein a feature space that divides the data points. The optimal hyperplane is the onethat maximizes the distance or margin between the classes. The support vectors, whicharethedata points closest to the hyperplane, influence its position and orientation. Onlythesepoints are used to determine the optimal hyperplane.

To put it simply, the hyperplane is like a decision boundary that separates different classes. In a two-dimensional space, it's like a line, while in a three-dimensional space, it'slike a plane. The margin, which is the distance between the hyperplane and the nearest support vectors from each class, is crucial. A larger margin generally leads tobetterperformance when the model encounters new, unseen data.

In mathematical terms, the SVM optimization problem can be expressed as follows. Given a training dataset with feature vectors (xi) and class labels (yi), where theclasslabels can be either -1 or 1, the goal is to find the weight vector (W) and bias term(b) that define the hyperplane. This optimization problem is mathematically formulated as:

$$\min_{w,b} \frac{1}{2}||w||^2$$

By understanding these concepts and utilizing the SVM optimization problem, wecaneffectively use SVMs for classification and regression tasks.

Subject to constraint:-

$$y_i(w.x_i + b) \geq 1 \forall_i$$

Where:
W is the weight vector prependicular to the hyperplane.
b is the bias term.
|| w || denotes the norm of the weight vector

Decision Trees :-

A decision tree is a really powerful and intuitive algorithm that's used for bothclassification and regression tasks in supervised learning. It's like a tree with nodes andbranches that represent decisions and their potential outcomes. This structure makes it supereasy to visualize and understand. The topmost node in the tree is called the Root Node, and it represents the wholedataset. It's split into child nodes based on a feature that separates the data the best, accordingto a certain criterion. These child nodes are called Internal Nodes, and each one correspondsto a test on an attribute.

Depending on the outcome of that test, the node branches out. Ontheother hand, we have Leaf Nodes, also known as Terminal Nodes. These nodes represent thefinal decision or classification. In classification, each leaf node corresponds to a class label, while in regression, it represents a continuous value. The connections between nodes are called Branches, and they showthe outcomeof atest at an internal node. Now, let's talk about the process of building a decisiontree. It involves splitting the data at each node based on the feature that gives the best separationofthe target variable. We usually use measures of impurity to determine this, like Gini Impurityor Information Gain.

$$Gini(D) = 1 - \sum_{1}^{n} p_i^2$$

Where pi is the probability of class i.

Gini Impurity measures the frequency of different classes at a node. Lower values meanpurer nodes.

$$InformationGain = Entropy(D) - \sum_{1}^{k} \frac{|D_i|}{D} Entropy(D_i)$$

Information Gain, on the other hand, is based on the concept of entropy and measures howmuch the entropy decreases from the parent node to its children

$$Entropy(D) = -\sum_{1}^{n} p_i \log_2 p_i$$

Random Forest :-

Random Forest is a really cool method that can improve the performance of classificationand regression tasks. It's basically a team of decision trees working together to make predictions. This method was developed by Leo Breiman and Adele Cutler to address thelimitations of individual decision trees, like overfitting and accuracy issues Random Forest builds multiple decision trees and combines their results to makeafinal prediction. It's like a team huddle where everyone's opinion counts. For classification, they use majority voting, and for regression, it's all about averaging. RandomForest uses atechnique called bagging. Each tree is trained on a random subset of the training data, andthey do this by sampling with replacement. This helps create a diverse group of trees, whichis pretty cool. 21 On top of bagging, Random Forest introduces even more randomness. At eachsplit inthe tree, they select a random subset of features to consider. This ensures that the trees arediverse and not too similar

1. Accuracy: This measures the percentage of instances that are correctly classified

$$Accuracy = \frac{Number\ of\ Corrct\ Predictions}{Total\ Number\ of\ Predictions}$$

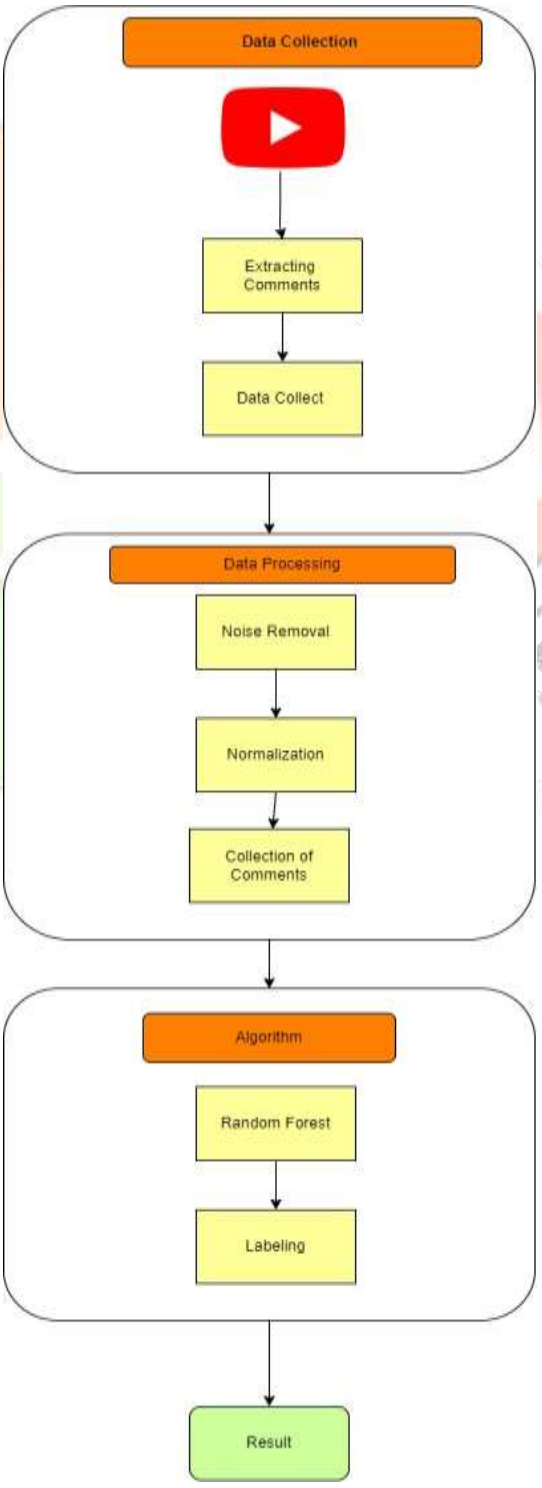
2. Precision, Recall, and F1 Score: These metrics are useful for evaluating classificationmodels, especially when working with imbalanced datasets.

3. Mean Squared Error (MSE): This is used in regression tasks to calculate the averageof the squared errors. It compares the predicted values to the actual values (yi).

$$MSE = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2$$

Where yi are the actual values and are the predicated values. Random Forest is a pretty cool and versatile way of learning that jazzes up decisiontrees. Basically, it takes a bunch of trees and adds some randomness to them, which helps prevent overfitting and makes predictions more accurate. It's a bit of a computer hog, but it's great for handling big sets of data and figuring out which features are most important. That's whyit'ssuch a handy tool in all sorts of fields

FLOWCHART



EXPERIMENTAL RESULTS :

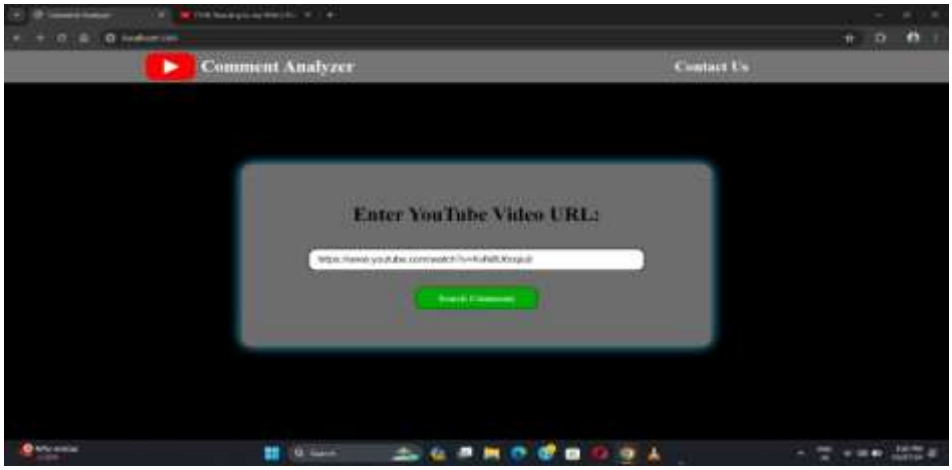


Fig: comment analyzer

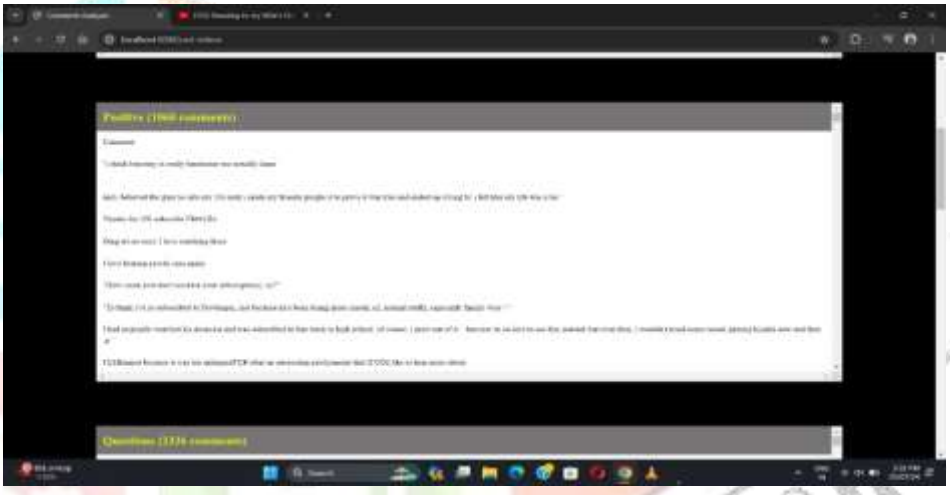


Fig: Positive Comment

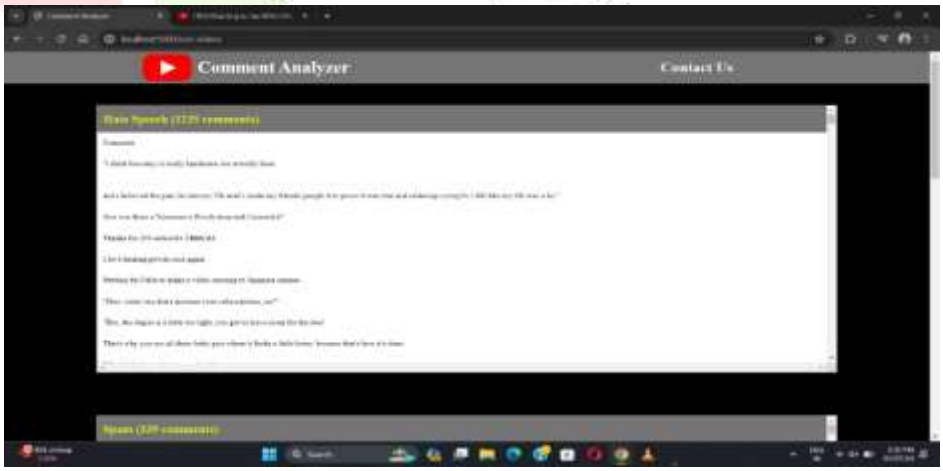


Fig: Hate Speech Comment

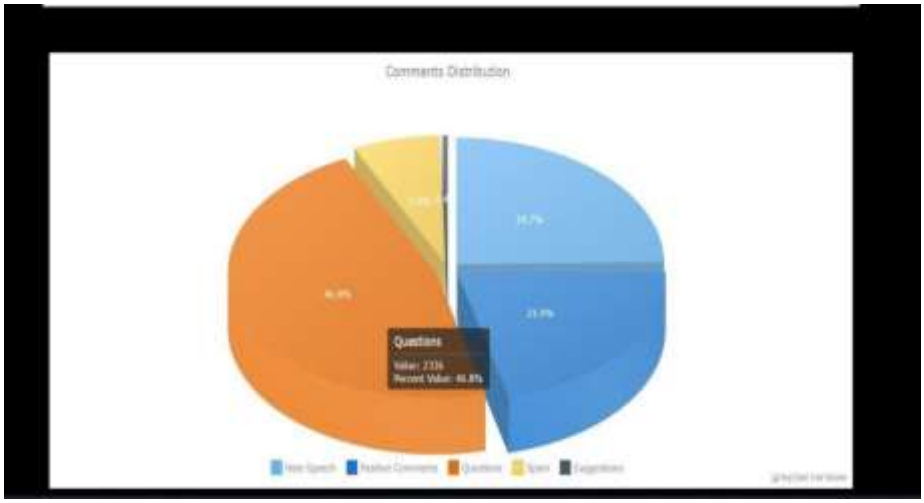


Fig: Pie Distribution of Comments

CONCLUSIONS

This paper presents a classifier-based tool for automatically classifying YouTube comments into relevant, irrelevant, positive, and negative categories. It compares bag of words and association list-based feature extraction methods. We have used supervised learning in the project.

Future work aims to classify multilingual phrases and non-contiguous phrases. In future we can use Natural Language Processing in the project using different methodology.

REFERANCES:

[1] Alexander Lighter & Bedor Tekinerdo gan," Systematic reviews in sentiment analysis: a tertiary study", (2021) .

[2] A. Kantchelian, J. Ma, L. Huang, S. Afroz, A. Joseph, J. Tygar, Robust detection of comment spam using entropy rate, in: Proceedings of the ACM Conference on Computer and Communications Security, 2012, pp. 59–70.

[3] A. Madden, I. Ruthven, D. McMenemy, A classification scheme for content analyses of youtube video comments, Journal of documentation (2013).

[4] A. Severyn, O. Uryupina, B. Plank, A. Moschitti, K. Filippova, Opinion mining on youtube (2014).

[5] M. Z. Asghar, S. Ahmad, A. Marwat, F. M. Kundi, Sentiment analysis on youtube: A brief survey, arXiv preprint arXiv:1511.09142 (2015).

[6] T. C. Alberto, J. V. Lochter, T. A. Almeida, Tubespam: Comment spam filtering on youtube, 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) (2015) 138–143.

[7] A. U. R. Khan, M. Khan, M. B. Khan, Naïve multi-label classification of youtube comments using comparative opinion mining, Procedia Computer Science 82 (2016) 57–64.

[8] J. Savigny, A. Purwarianti, Emotion classification on youtube comments using word embedding, in: 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017, pp. 1–5.

[9] S. Sharmin, Z. Zaman, Spam detection in social media employing machine learning tool for text mining, 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (2017) 137–142.

[10] A. O. Abdullah, M. A. Ali, M. Karabatak, A. Sengur, A comparative analysis of common youtube comment spam filtering techniques, 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (2018) 1–5.

[11] E. Poché, N. Jha, G. Williams, J. Staten, M. Vesper, A. Mahmoud, Analyzing user comments on youtube coding tutorial videos, in: 2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC), 2017, pp. 196–206.

[12] A. Aziz, C. F. M. Foozy, P. Shamala, Z. Suradi, Youtube spam comment detection using support vector machine and k–nearest neighbor, 2018.

[13] Burhanudin, Y. Musa'adah, Y. Wihardi, Klasifikasi komentar spam pada youtube menggunakan metode naïve bayes, support vector machine, dan k-nearest neighbors, 2018.

[14] N. M. Samsudin, C. F. M. Foozy, N. Alias, P. Shamala, N. F. Othman, W. Din, Youtube spam detection framework using naïve bayes and logistic regression, Indonesian Journal of Electrical Engineering and Computer Science 14 (2019) 1508.