International Journal of Scientific Research in Mathematical and Statistical Sciences
Volume-8, Issue-3, pp.24-32, June (2021)
Research Paper
E-ISSN: 2348-4519

Using Jaccard Similarity Measure for Detection of Abusive Comments on Video by Indian YouTuber

Karina Bohora[1*], Sanjay Patil[2]

[1,2]Department of Computer Engineering and Information Technology, College of Engineering, Pune, India

[*]*Corresponding Author: bohoraka16.it@coep.ac.in*

Available online at: www.isroset.org

Received: 06/May/2021, Accepted: 13/Jun/2021, Online: 30/Jun/2021

***Abstract*—**In this paper, a Jaccard index based technique for detection and analysis of abusive YouTube comments is proposed. The effectiveness of the designed approach is evaluated using a popular video by one of the most subscribed YouTubers in India having over 24 million subscribers. This entertainment YouTube video titled 'The Art of Bad Words' posted by CarryMinati garnered over 25 million views and more than 0.3 million comments within 15 days of being uploaded. Offensive language used in YouTube comments is often culture-specific and hence can be challenging to identify and keep a check on. So, the focus of this study is on comments containing derogatory language prevalent in the Indian subcontinent and thereby violating YouTube's community guidelines and policies. The approach's performance is compared for 4 different threshold values of Jaccard coefficient and the impact of each value on the results obtained is illustrated.

Keywords*—*Text Mining, YouTube, Data Mining, Online Video, Content Analysis, Jaccard Distance, Popular Video, India

I. INTRODUCTION

Text mining finds application in a broad array of fields such as customer care service, spam filtering, social media analysis, fraud detection, risk management among others. [1] demonstrates the use of mining to distinguish content containing valuable information from irrelevant and spam content. [2] makes use of mining techniques to extract different features of products by analyzing pros and cons mentioned in reviews.

As a developing nation, more and more Indians are increasingly gaining easier access to the massive amount of data and content present on the Internet and various different applications. One such popular online video sharing application is YouTube. The YouTube video chosen for this analysis belongs to the entertainment category. Entertainment videos typically enjoy a higher viewership and comment count compared to political videos [3]. Taking into account that Hindi is the 3rd most spoken language in the world, the emphasis of our study and analysis is on comments in the Hindi language written using English alphabets.

The impact of YouTube comments cannot be denied. Comments on videos give insights into the reactions of audience to general important issues as well as particular videos [4]. [5] acknowledges that the options of rating and commenting while viewing videos brings new social aspects into the picture. [6] researches YouTube comments to comprehend the value that they add and the role that they play in portraying a particular image of a video and how these comments affect users' overall rating behavior. In fact, it is found that as many as 53% of users read the initial few (particularly, around two or three) comments on finishing viewing a video on YouTube.

However, not all videos and comments that appear on YouTube have a positive or even neutral effect on the viewer or reader. As a matter of fact, hateful comments on YouTube have the ability to discourage self-expression and constructive communication among participants on the site [7]. And that is likely why [8] describes YouTube videos and comments that include topics like race, sexuality, culture and other such aspects that people cannot change about themselves as sensitive.

It is not unusual to find abusive and hateful content on YouTube and specifically in users' comments under videos. [9] describes abusive comments as those that contain extremely racist remarks. YouTube comments containing common and obvious abusive terms like 'suck', 'nigger', 'dick', 'fuckin', 'asshol', 'faggot', 'cunt' among others often lead to negative voting and unacceptance of comments on the grounds of possessing offensive content [10]. Further, it is pointed out in [7] that 'haters' and hence, hateful comments, don't belong to a single category but rather are of several different types owing to the complex nature of hate-related behavior. Antagonism can be categorized into different types such as racism, homophobia and misogyny, as shown in [11].

However, it is not always easy and straightforward to detect derogatory and abusive language in the text present in comments. [9] finds the informal language styles used by YouTube commenters to be a major challenge and

deterrent to any analysis. Similarly, [8] observes that while it's easier to model and detect direct verbal abuses and profanity, expressions that are indirect and involve some degree of euphemism and sarcasm often get missed. [12] finds it to be a challenging task to identify sentiment that is expressed in a more subtle manner and without the use of keywords that can be easily detected. [13] observes a challenge in terms of it being difficult to grasp the overall theme of user-contributed comments.

A noteworthy finding in [11] is that racism on online platforms like YouTube and its comment space is not an exceptional phenomenon but rather a clear indication of the racially charged prevalent everyday culture. In fact, their study reveals that random insults and other expressions of 'hate' are correlated with networked interactions and are passed through several parts of the comment network of YouTube.

Nevertheless, YouTube does take measures to ensure community wellbeing and welfare on its platform. Under YouTube's community guidelines, there are sections of 'Sensitive content' and 'Violent or dangerous content'. Within the framework of 'Violent or dangerous content', the subsection of 'Hate speech' indicates YouTube's hate speech policy which includes but is not limited to hatred against an individual or a group based on sex/gender, ethnicity, race, sexual orientation, caste. YouTube also specifies a 'Harassment and cyberbullying policy' which includes and calls out 'name-calling and malicious insults'. YouTube specifies that content that 'encourages others to bully or harass' or 'sexually harass' violates its policy on 'Child Safety' and applies to YouTube comments as well.
The organization of the remainder of this paper is as follows. Section II includes a review of the literature and some background of previous similar and relevant works. Section III illustrates the details of the proposed methodology with respect to the steps of data collection using YouTube API, data preparation and data processing for analysis. Section IV summarizes the obtained results and discusses the performance based on four metrics namely precision, recall, accuracy and specificity. Section V concludes the study and presents potential scope for subsequent work in the future.

II. RELATED WORK

The emergence of India as the second-biggest market for YouTube in recent times and the fact that Indian video watchers favor YouTube as their preferred video sharing and watching application necessities delving into related research gone into analysis of the comment space of YouTube. This section focuses on the presence of hateful language and behavior in YouTube comments and in particular, the usage of abusive language and profanity.

Different research work and analyses make use of comments on YouTube videos that belong to several different categories. For instance, [14] focuses on the comments on coding tutorial videos and in particular those comments that include concerns and questions that the viewers feel need to be addressed. Keeping in mind YouTube's massive scale, [15] limits data collection to two categories of 'Entertainment' and 'Science & Technology'. The study in [16] analyses the comments that have comparative content and focuses efforts of research work on comparative opinion mining. [3] makes use of comments on entertainment and political videos on YouTube for presenting a comparative content analysis. [17] chooses YouTube videos that belong to the category of job interview videos. [11] directs the focus of analysis on the comments on YouTube videos of Das Racist (a provocative musical group). An observation made by [10] with regards to the different categories of YouTube videos available is that videos belonging to the domain of politics have a substantially greater number of comments that are rated negatively in comparison to all of the other categories. On the other extreme, the videos pertaining to the music category have the greatest number of comments that are positively rated.

Several performed analyses mention their steps related to pre-processing. Pre-processing of data performed in [8] involves stop-word removal, stemming and removal of unimportant sequence of characters. [18] executes a number of pre-processing steps such as converting a comment to a set of tokens, stop-word removal, removal of non-Latin-based words, removal of punctuation characters and conversion of all letters to lowercase. [19] performs pre-processing of words by removing stop-words, stemming as well as fuzzy matching which is used to deal with incorrectly spelled words and several different variants of a word. [17] performs pre-processing of YouTube comments by removal of stop-words, stemming of words and by using tfidf scores to rank words.

The solutions proposed in the studied research papers also understandably differ in terms of the methods employed, techniques used and approaches taken. The research in [20] proposes a rule-based method to detect comment spamming by mining the comment activity log of YouTube users. The solution formulated makes use of the feature that marks comments with a hasSpamHint tag. The recent commenting activity of users is retrieved using YouTube API. Four indicators are used to score a user and establish whether the user is a potential content spammer or not. These indicators include: Comment Repeatability Across Videos (CRAV), Average Time Difference between Comments (ATDC), Comment Repetition and Redundancy (CRR), and Percentage of Comments with hasSpamHint Flag (PCHF). [6] uses nine features to operationalize comment types; one of these features is the 'Offensive Hint' feature which includes comments that have a greater negative sentiment, more capitalization and/or words that are categorized as "aggressive" or "angry". [14] uses Support Vector Machines to detect viewers' comments that are useful and achieves an average accuracy of 77% in doing so. The thesis presented conducts a quantitative as well as a qualitative analysis wherein the manual qualitative analysis is used to figure

out the information value of the comments that are sampled. An extractive frequency-based summarization technique is used to capture and identify the primary concerns in the YouTube comments posted by users. The three tasks performed as part of the project work are: data collection, comments' classification and comments' summarization. The study in [10] utilizes negative votes to identify comments with inappropriate as well as offensive content. Their analysis work identifies top 50 terms that lead to unacceptance of comments and ranks them using the Mutual Information (MI) measure.

YouTube comments can be classified, annotated or categorized into different types in various ways. [6] classifies the comments into three different classes namely substantial comments, discussion posts, and inferior comments. The study in [16] classifies comments posted by users to be either relevant or irrelevant and the relevant comments are further categorized into four types namely declarative comments, comparative comments, direct opinion comments and comments containing more information. [8] annotates comments using three labels namely sexuality, race as well as culture and intelligence. The 'sexuality' label is attached for comments towards sexual minorities that are of an attacking nature as well as comments about women that are sexist in nature. The 'race and culture' label is utilized to annotate comments that are attacks on racial minorities or stereotypical mocking of cultural traditions. [9] makes use of the four divisions or classes which are self-promotion, propaganda, comments that are abusive as well as miscellaneous comments in order to flag comments in an attempt to study the different techniques used to analyze comments shared by users on YouTube videos. [17] categorizes and labels some comments as relevant while some as noisy and the noisy comments are those that are found to be spam and irrelevant.

On some occasions, certain hypotheses and assumptions are used by the studies conducted. [20] assumes abusive and hateful comments to be a subpart of a larger phenomenon named comment spamming and hypothesizes certain behavioral characteristics of comment spammers. [3] makes the hypothesis that the more positive a comment is, the more replies and likes it will receive. But this hypothesis holds only in case of entertainment videos and not in the scenario of videos that are political. Similarly, the classifier of YouTube comments of users presented in the study in [16] is based on a naïve assumption that terms (or words) which are closely associated with certain keywords are sufficient to recognize the sentiment of the commenting user with regards to his or her opinion and preference. And so, although the performance of the developed classifier is slightly less, the naïve assumption that is made offers benefit in the form of less requirement of power for computation. [13] hypothesizes that factors such as comment visibility, reputation of commenting user and the comments' content affect the community ratings of comments. Besides, they also make a hypothesis that communities within different categories of videos use their own specific set of jargon.

Each study and research use comments for different purposes and analyses. The research in [11] aims to examine YouTube comments to shed light upon the systematic, networked and entangled nature of online racism, hostility and antagonistic racial discourses that propagate a toxic culture of 'hate' speech, online trolling and abuse through the quasi-anonymous platform of YouTube's comment space. [3] focuses on comparing YouTube comments on entertainment videos with those on political videos by presenting a content analysis. [16] analyzes users' YouTube comments to gain insights regarding their preferences and opinions when they compare different products or options. [14] analyses users' comments to understand how content makers can increase their engagement with their target audience and be able to decide how to make more effective video content in the future helping them grow in popularity. By utilizing a machine learning approach, [17] analyzes user-generated content and comments in order to derive key user characteristics and social user profiles for them; with the eventual aim being augmentation of existing prevalent user models for users that are found to be similar. The purpose of the research work of [10] is to analyze the dependence of comments, views, comment ratings and topic categories on each other and primarily deals with community feedback received through YouTube comments. On the other hand, [7] aims to prove that there is little to no dependence and correlation between user anonymity and the posting of antagonistic comments on the YouTube platform. [8] intends to detect textual cyberbullying by making use of topic-sensitive binary and multiclass classifiers.

The performed research work includes several noteworthy conclusions, observations and inferences. [14] finds that the performance and effectiveness of several text summarization techniques depends upon their ability to correctly capture the primary key topics and issues raised in the comments section of the concerned coding tutorial video. [3] records that the comments under videos which can be said to belong to the entertainment category are neutral (in nature) to a greater degree as compared to the comments made on political videos. Another observed phenomenon is that comments having a considerably and comparatively stronger valence (either negative or even positive) typically get to enjoy a greater count of replies and likes. Besides, the comments on political videos are found to be more extreme in nature and often cause polarization owing to the fact that content which is of political type is often controversial in nature. [21] finds that the value for recall of sentiments which are negative is not as good as or is poorer in comparison with that of positive sentiments which can likely be attributed to the greater degree of variation present in linguistics that are used to express feelings of dissatisfaction and frustration. [7] puts forth the need to understand the extent to which phenomena such as 'online hating' is viewed as a

problematic threat and by whom and cites the productive way forward to be the exploration of the negative impact that offline bullying practices have on online hating behaviors.

III. METHODOLOGY

The method employed predominantly focuses on 3 main aspects:

- Data collection using YouTube API
- Data pre-processing, preparation and cleaning
- Processing and analyzing the comments' dataset using Jaccard measure

A. Data Collection Using YouTube API

Broadly, the two steps involved in the data collection process are:

- Gathering the YouTube comments' dataset in the JSON file format
- Converting JSON file to CSV file for processing

First of all, in order to generate and avail an API key, a project is created on the Google Console Developer platform. After creation and naming of the project, the 'Credentials' section within 'APIs & Services' is accessed wherein the API key generated is noted for future reference.

Thereafter, the API key is used to retrieve data from YouTube. The 2 inputs given for retrieval of YouTube comments data are: Video_id and API_KEY. To fetch the video's Video_id, YouTube link of concerned video i.e. the "THE ART OF BAD WORDS" video on the channel of "CarryMinati" is examined. The Video_id in the YouTube URL is the id after the '?v=' part in the URL. So, after fetching both, the Video_id and the API_KEY, the comments' data in JSON format can be availed.

Each comment in any comment thread has the following fields:

- videoId: The id of the video on which the concerned comment was posted
- textDisplay: The part of the YouTube comment posted that is actually displayed
- textOriginal: The entire text posted by a user as a YouTube comment
- authorDisplayName: The commenting YouTube user's displayed name
- authorProfileImageUrl: Link to the profile image of the YouTube user who posted the comment
- authorChannelUrl: Link of channel of the commenting YouTube user
- authorChannelId: The channel id of the channel of the YouTube user who posted the comment
- canRate: Whether one can rate (i.e., like or dislike) the comment or not (a "true" Boolean value in this field indicates that one can like or dislike the comment posted)
- viewerRating: The rating given by the viewer to the concerned comment (valid values can be either "like" or "none")
- likeCount: The total count of likes received on the comment under consideration
- publishedAt: Time and date when comment was originally written/posted
- updatedAt: Date and time when the comment was last updated

Of the several different fields present for every comment, the ones found to be of most importance and relevant for analysis were the 'textOriginal' and 'likeCount' fields. And so, the generated CSV file from the JSON file included only these 2 fields. A partial snapshot of the comments' dataset in the csv format before pre-processing and cleaning is shown below in Figure 1.

```
Itni chutiya panti Kyo is video me ,0
Chutiya sala sahi se roaster hota to bollywood zaroor roast
karta .. unsubscribed  u ,0
Carry bhai # hatsoff,0
nepotism of bollywood ko roast kro carry,0
Tare ko your father gala bolna data ha ba,0
Kay gali deteho bhai,0
Karan ko Raost kro carry bhai,0
accha.hua ..Copy right nikal gaya iss video see..luvvv u
bhaiiiii,0
4.2 m likes done what u said,0
Nice.bro,1
Karan Johar ko roast kro,2
Bhai abli video kab aye ge,2
Teri video dekhne ke liye earphone jarur chahie,1
Bhai thoda aur behtr kr skta tha bki act accha h,0
```

Figure 1. Partial snapshot of the comments' dataset in the CSV format

B. Data Pre-Processing, Preparation and Cleaning

In order to aid the subsequent analysis, the dataset thus obtained goes through 5 steps as part of the cleaning and pre-processing process. The details of each of these steps are explained below.

1) Removal of Emoticons in Comments

In this step, all emoticons present in the YouTube comments' dataset are eliminated. Although emoticons can be good indicators of the feelings of the user posting any comment, (such as feelings of anger or frustration can be conveyed through appropriate emoticons), it is not common practice to explicitly swear or abuse using emoticons. One exception to this, could however, be the 'middle finger' emoji used as an insulting gesture in some cultures.

2) Removal of Punctuation and Other Symbols

It is quite common to find comments under YouTube videos to be cluttered with too many (unnecessary and sometimes repetitive) punctuation marks. While punctuations do communicate feelings and intent (like '!" for 'strong feelings or emphasis' or '?' for 'questioning'), there is little help that they can provide in indicating the abusive, derogatory or hateful nature of a comment.

Besides, the '#' hashtag, which is popularly used to represent trending topics, is removed because it merely indicates popular trends and in no justifiable way suggests the presence of any kind of abusive content. So,

punctuations and symbols like '!', '?',',', '.', '@', '#' among others are cleaned up from the dataset.

3) Removal of Extra Spaces

This step is included keeping in mind that commenters on YouTube videos are typically not the most grammar-conscious people and often post comments without going through the typed content to check for errors. It is not unusual for comments to possess extraneous spaces merely because the commenter didn't care enough to post a perfectly error-free comment. This step, hence, takes care of these extra spaces to move towards cleaner data.

4) Removal of Non-English Characters (Devanagari Script)

Since the focus is on comments posted on a video of a famous Indian YouTuber, the commenters are primarily Indian and besides English, there is presence of comments in the Hindi language (one of the two official languages of India; the other being English) as well. Because Hindi comments written using English alphabets constitute a very important part of the analysis and study performed; they are kept. However, Hindi comments written in the Devanagari script and using Devanagari alphabets and keyboard are not included in the analysis. A sample comment which has alphabets in the Devanagari script is shown in Figure 2.

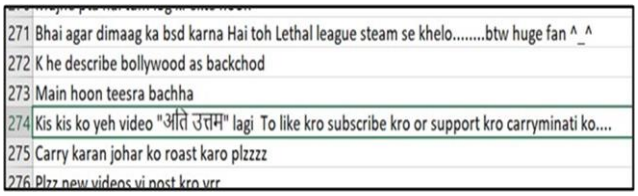


Figure 2. A sample comment having alphabets in Devanagari script

It is important to note; however; that majority of the Hindi comments observed are written using English alphabets and not the Devanagari script.

5) Removal of Numbers

As numbers do not convey significant information with regards to a comment being or not being abusive in nature, they are done away with for convenience and ease of analysis.

After the aforementioned 5 steps for preparation and preprocessing of the YouTube comments dataset are completed, a cleaner dataset is available for processing and analysis.

C. Processing and Analyzing the Comments' Dataset Using Jaccard Measure

Jaccard similarity between two sets is defined as the number of elements in the intersection of the two sets divided by the number of elements in the union of the two sets. It is a measure of similarity of the sets. Jaccard similarity = n(intersection of two sets)/n(union of two sets).

On the other hand, Jaccard distance measures difference between two sets. It is found by subtracting the Jaccard similarity from 1. Jaccard distance = 1 - Jaccard similarity. So, the greater the value of Jaccard distance, the lower the value of Jaccard similarity and vice versa.

1) Building Lexicon for Detection of Abusive Language in Comments

By taking into account the culture-specific abusive language prevalent in the Indian subcontinent and particularly in India, a lexicon of derogatory and abusive words often used is built. The words included are abusive, racial attacks, attacks directed at sexual minorities and women and/or are derogatory in nature. The lexicon that is generated is shown in Figure 3.

Figure 3. Generated lexicon to detect abusive language in comments' dataset

To ensure that the developed lexicon is concise, 1 representative word from each family of abusive words is taken. For ex: 'chutiya' can be thought as representative of other very similar abusive words like 'chutiye', chutiyo' or 'chutia'. Similarly, 'bhosdiwala' can be said to represent words such as 'bhosdiwale', bhosdiwali' or 'bhisdiwalo'. Another such case can be the use of the word 'gand' to represent words like 'gaand', 'gandi' or 'gandoo'. Table 1 shows the representative lexicon words along with their English meanings and associated families.

Table 1. Some representative lexicon words with their meanings and families

Some representative lexicon words	**Meaning in English**	**Family of similar words**
chutiya	Fucker	chutia chutiye chutiyo chut chootwali
gand	Ass	gaand gandi gandoo
chakka	derogatory term for transgenders	chakki chakko
bhosdiwala	pussy	bhosdiwali bhisdiwalo bhosdiwale bhosidi
bakchod	senseless fucker	backchod bakchodi

2) Jaccard Distance and Similarity

As an example, the Jaccard similarity between the abusive word 'chutiya' present in the lexicon of abusive words

created and another word ‘chutia’ which is a slight derivation of it can be calculated by using the formula specified. These two words selected are essentially words with different spellings but the same intended meaning and pronunciation.

The set of alphabets present in the word “chutiya is {‘c’, ‘h’, ‘u’, ‘t’, ‘i’, ‘y’, ‘a’}. The set of alphabets present in the word “chutia” is {‘c’, ‘h’, ‘u’, ‘t’, ‘i’, ‘a’}. The Jaccard similarity is, thereby calculated by using the specified formula as:

n{‘c’, ‘h’, ‘u’, ‘t’, ‘i’, ‘a’}/n{‘c’, ‘h’, ‘u’, ‘t’, ‘i’, ‘a’, ‘y’} where ‘n’ denotes ‘number of elements in the set and the ratio is found to be 6/7 i.e. 0.8571. So, the Jaccard similarity between the words ‘chutiya’ and ‘chutia’ can be said to be 85.71% and the Jaccard distance between them can be said to be 1-0.8571=0.1429 or 14.29%. Figure 4 shows the calculated Jaccard distance between a comment and a lexicon word.

```
chutia chutiya 0.14285714285714285
abusiveComment
```

Figure 4. Calculated Jaccard distance between a comment and a lexicon word

Likewise, another example of the word ‘bhosdiwala’ (having set of alphabets present in it as {‘b’, ‘h’, ‘o’, ‘s’, ‘d’, ‘i’, 'w', ‘a’, 'l'}) present in the lexicon and a word similar to it ‘bhisdiwalo’ (having set of alphabets present in it as {‘b’, ‘h’, ‘i’, ‘s’, ‘d’, 'w', ‘a’, 'l', 'o'}) can be illustrated as: Jaccard distance = n{‘b’, ‘h’, ‘i’, ‘s’, ‘d’, 'w', ‘a’, 'l', 'o'}/n{‘b’, ‘h’, ‘i’, ‘s’, ‘d’, 'w', ‘a’, 'l', 'o'} = 9/9 = 1. So, the Jaccard similarity between the words ‘bhosdiwala’ and ‘bhisdiwalo’ is found to be 100.00% and the Jaccard distance between them can be calculated to be 1-1=0.0 or 0%.

Thus, the similarity between these differently spelled words with similar abusive meaning is very well captured by the Jaccard similarity measure. Jaccard distance and similarity between several other lexicon abusive words and non-lexicon abusive words can be seen in the following Table 2.

Table 2. Jaccard measure between some lexicon and non-lexicon abusive words

Lexicon word	Comment word	Jaccard distance	Jaccard similarity
chutiya	chutia chutiye chutiyo chut chootwali	0.142857143 0.25 0.25 0.428571429 0.5	0.857142857 0.75 0.75 0.571428571 0.5
gand	gaand gandi gandoo	0 0.2 0.2	1 0.8 0.8
chakka	chakki chakko	0.2 0.2	0.8 0.8
bhosdiwala	bhosdiwali bhisdiwalo bhosdiwale bhosidi	0 0 0.1 0.333333333	1 1 0.9 0.666666667
bakchod	backchod bakchodi	0 0.125	1 0.875
sale	saale sala	0 0.25	1 0.75

3) Using Python 3.8 Script to Process Collected YouTube Comments Data

Two libraries of Python namely the Pandas and the NLTK libraries proved to particularly be useful in the analysis performed. The Pandas library is used to import data present in CSV file format in a data frame for further analysis. A function of the NLTK library is used for assistance in finding Jaccard distance.

Using the Python 3.8 script and the two libraries mentioned, the analysis is performed for four threshold values of Jaccard distance: 0.25, 0.2, 0.15 and 0.1. Partial snapshots of the outputs obtained for each of these threshold values are shown below in Figure 5, Figure 6, Figure 7, and Figure 8 respectively.

Delimiter: ,

	comment	likes
1	Itni chutiya panti Kyo is video me	0
2	Chutiya sala sahi se roaster hota to bollywood zaroor roast karta unsubscribed u	0
3	accha hua Copy right nikal gaya iss video see luvvv u bhaiiii	0
4	Bhai thoda aur behtr kr skta tha bki act accha h	0
5	My favourite part in this video Lund its thicker its stronger	1
6	Bhai tum hi to nhi ho gandu kyuki tum hi beech me aaye ho	0

Figure 5. Some of the comments classified as abusive by Jaccard distance threshold of 0.25

Delimiter: ,

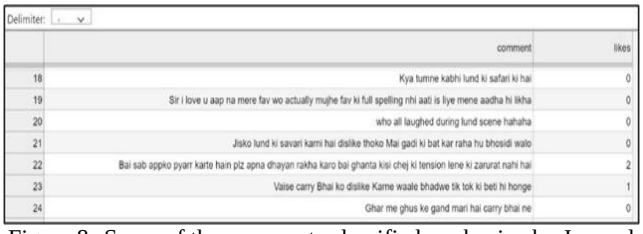
	comment	likes
14	Suprabhat bhisdiwalo	1
15	Carry vai sudim kela bolo na please	0
16	Gandu	0
17	Nagi dem jhajha saiberian bird vlog Link https youtu be vJjBTUisw Please watch the video	1
18	Ek number ka harami hai	0
19	Chap hatt bahen ki laudi	0
20	carry yu hi vo gandu hai jisna khde khde tik tok pe tala laga diya	1

Figure 6. Some of the comments classified as abusive by Jaccard distance threshold of 0.2

Delimiter: ,

	comment	likes
27	https youtu be eTYjDKm pubg hack version chahiye to comment kro	0
28	K he describe bollywood as backchod	0
29	See that yoda doing tatti in dimond play button	0
30	You suck	0
31	Hum chakko ke saath ungli mat kar saale deepakkalalgang	0
32	Aji aapne to YouTube pr gaand phaad rakhi hai sabki Kon kon iss baat se sehmat hai	1
33	Mere bhai tum jio hajaro sall or tumare se jalne vale ki Gand mera lodaa	0

Figure 7. Some of the comments classified as abusive by Jaccard distance threshold of 0.15



	comment	likes
18	Kya tumne kabhi lund ki safari ki hai	0
19	Sir i love u aap na mere fav wo actually mujhe fav ki full spelling nhi aati is liye mene aadha hi likha	0
20	who all laughed during lund scene hahaha	0
21	Jisko lund ki savari karni hai dislike thoko Mai gadi ki bat kar raha hu bhosidi walo	0
22	Bai sab appko pyarr karte hain plz apna dhayan rakha karo bai ghanta kisi chej ki tension lene ki zarurat nahi hai	2
23	Vaise carry Bhai ko dislike Karne waale bhadwe tik tok ki beti hi honge	1
24	Ghar me ghus ke gand mari hai carry bhai ne	0

Figure 8. Some of the comments classified as abusive by Jaccard distance threshold of 0.1

The number of comments that get classified as abusive differs as the threshold value changes. The obtained results or outcomes are organized as below in Table 3.

Table 3. Number of comments classified as abusive for the four threshold values

	Threshold value chosen for Jaccard Distance between words in YouTube comments and words in dictionary of abusive words			
	<= 0.25	***<= 0.2***	***<= 0.15***	***<= 0.1***
Number of comments classified as abusive	122	91	45	42

So, as can be noticed from the compiled observations, the number of comments getting classified as abusive doesn't change much when shifting from a threshold value of 0.15 to a value of 0.1. It does, however, significantly change when moving from a threshold of 0.25 to 0.2. In fact, the number of comments that get classified as abusive reduces to half when threshold value is changed from 0.2 to 0.15.
The obtained results are also plotted on a graph in Figure 9 to better understand how much change is present in the number of comments getting classified as abusive by observing the graph's slope as threshold value changes.

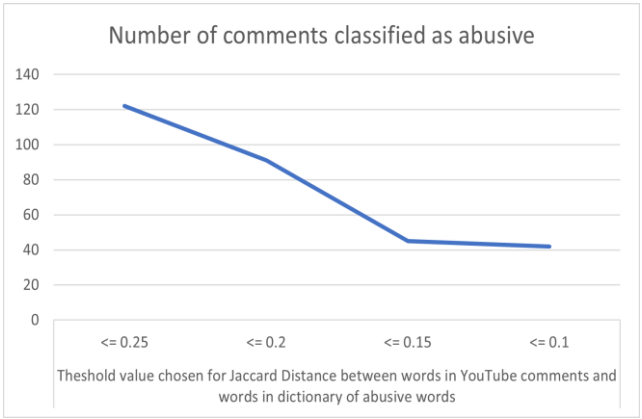


Figure 9. Plot of number of "abusive" comments for 4 Jaccard threshold values

As can be observed, the number of comments getting classified as abusive doesn't change much when the threshold value is reduced from 0.15 to 0.1 and merely decreases by 6.67% and hence, a threshold value further lower than 0.1 is not considered in the analysis.

IV. RESULTS AND DISCUSSION

A confusion matrix as displayed in Table 4 helps to better understand results obtained from the analysis performed.

Table 4. Confusion matrix

True Positive	False Positive
False Negative	True Negative

4 boxes of the matrix can be interpreted as:

- TP (True Positive): The comments that are actually abusive and indeed get classified as abusive
- FP (False Positive): The comments that actually are abusive but do not get classified as abusive
- FN (False Negative): The comments that actually are not abusive but get classified as abusive
- TN (True Negative): The comments that actually are not abusive and in fact don't get classified as abusive

In order to evaluate performance under 4 different threshold values of Jaccard distance, 4 metrics namely precision, recall, accuracy and specificity are used.

A. Precision

Precision can be formulated as (TP)/(TP+FP). So, precision for the 4 chosen threshold values can be calculated as following:

- for threshold value of <=0.25, precision = 45/(45+77) = 0.3689 = 36.89%
- for threshold value of <= 0.2, precision = 42/(42+49) = 0.4615 = 46.15%
- for threshold value of <= 0.15, precision = 38/(38+7) = 0.84444 = 84.44%
- for threshold value of <= 0.1, precision = 35/(35+7) = 0.83333 = 83.33%

The following graph in Figure 10 shows the variation in the value of precision as the Jaccard distance's threshold value changes.

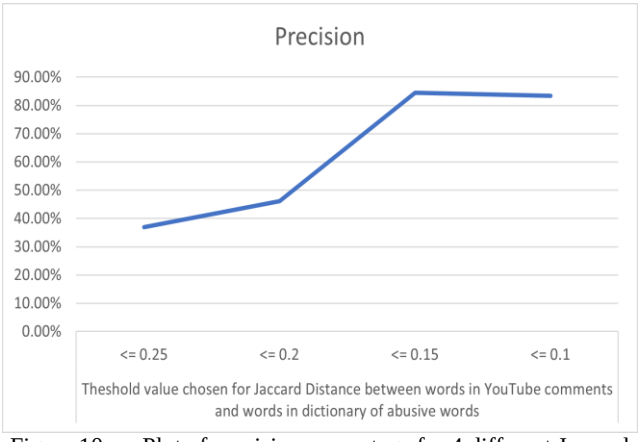


Figure 10. Plot of precision percentage for 4 different Jaccard threshold values

B. Recall (or Sensitivity)

Recall can be formulated as (TP)/(TP+FN). So, recall for the 4 chosen threshold values can be calculated as following:

- for threshold value of <=0.25, recall = 45/(45+7) = 0.8654 = 86.54%
- for threshold value of <= 0.2, recall = 42/(42+10) = 0.8077 = 80.77%
- for threshold value of <= 0.15, recall = 38/(38+14) = 0.73077 = 73.1%
- for threshold value of <= 0.1, recall = 35/(35+17) = 0.6731 = 67.31%

The following graph in Figure 11 shows the variation in the value of recall as the Jaccard distance's threshold value changes.

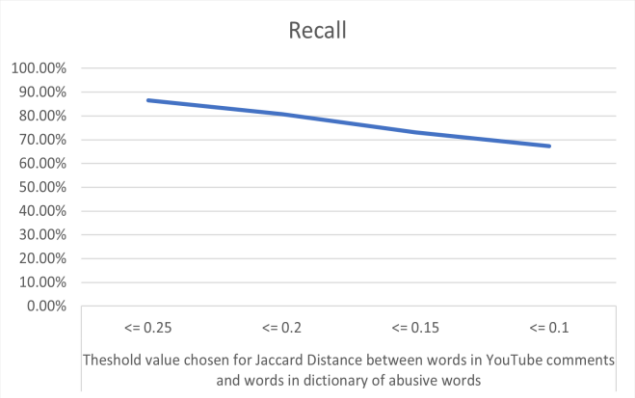


Figure 11. Plot of recall percentage for 4 different Jaccard threshold values

C. Accuracy

Accuracy can be formulated as (TP+TN)/(TP+TN+FP+FN). So, accuracy for 4 chosen threshold values can be calculated as following:

- for threshold value of <= 0.25, accuracy = 45+380/(45+380+77+7) = 425/509 = 0.835 = 83.5%
- for threshold value of <= 0.2, accuracy = (42+408)/(42+408+49+10) = 450/509 = 0.8841 = 88.41%
- for threshold value of <= 0.15, accuracy = (38+450)/(38+450+7+14) = 488/509 = 0.9587 = 95.87%
- for threshold value of <= 0.1, accuracy = (35+450)/(35+450+7+17) = 485/509 = 0.952848 = 95.29%

The following graph in Figure 12 shows the variation in the value of accuracy as the Jaccard distance's threshold value changes.

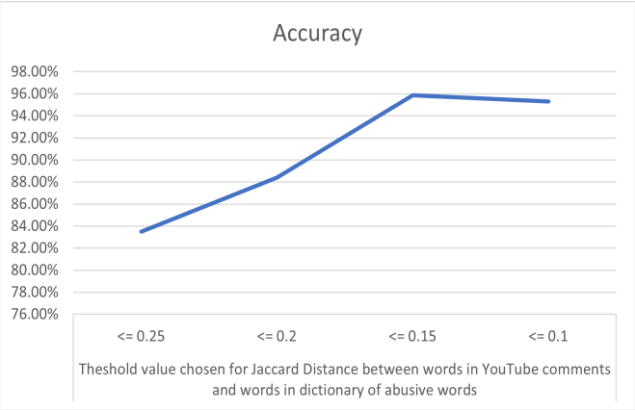


Figure 12. Plot of accuracy percentage for 4 different Jaccard threshold values

D. Specificity

Specificity can be formulated as TN/(TN+FP). So, specificity for the 4 chosen threshold values can be calculated as following:

- for threshold value of <=0.25, specificity = 380/(380+77) = 0.8315 = 83.15%
- for threshold value of <= 0.2, specificity = 408/(408+49) = 0.8928 = 89.28%
- for threshold value of <= 0.15, specificity = 450/(450+7) = 0.9847 = 98.47%
- for threshold value of <= 0.1, specificity = 450/(450+7) = 0.9847 = 98.47%

The following graph in Figure 13 shows the variation in the value of specificity as the Jaccard distance's threshold value changes.

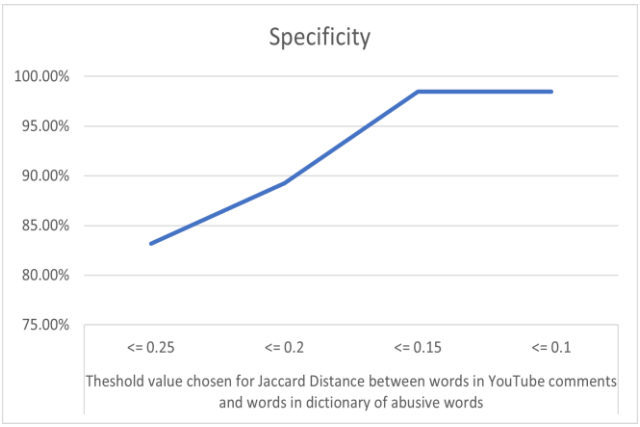


Figure 13. Plot of specificity percentage for 4 different Jaccard threshold values

These percentages of precision, recall, accuracy, and specificity for 4 threshold values of Jaccard distance are consolidated in the Table 5 that follows.

Table 5. Performance evaluation using four metrics

	Threshold value chosen for Jaccard Distance between words in YouTube comments and words in dictionary of abusive words			
	<= 0.25	***<= 0.2***	***<= 0.15***	***<= 0.1***
Precision	36.89%	46.15%	84.44%	83.33%
Recall	86.54%	80.77%	73.1%	67.31%
Accuracy	83.5%	88.41%	95.87%	95.29%
Specificity	83.15%	89.28%	98.47%	98.47%

As observed, the highest obtained percentages for precision, recall, accuracy, and specificity are 84.44%, 86.54%, 95.87% and 98.47% respectively. The threshold Jaccard distance value of 0.15 does the best in terms of obtained precision, accuracy, and specificity. As far as the recall is concerned, however, the best performance is found for the threshold Jaccard distance value of 0.25.

V. CONCLUSION AND FUTURE SCOPE

While precision, accuracy and specificity find their respective peak values for the threshold value of Jaccard distance of 0.15, its recall percentage can be improved by

making efforts to reduce the false negatives. Enhancing the lexicon used to classify comments (as abusive or non-abusive) can potentially help in reducing the number of false negatives and thereby addressing the issue of a low recall. In addition, certain provisions can be made in the designed lexicon to detect one of the quite common and prevalent swearing practices of profanity-with-asterisks. Further, the average number of likes received on abusive comments can be compared with the average number of likes received on non-abusive comments and these calculated values can be tested for statistical significance to determine whether any further noteworthy conclusions can be obtained and worthwhile inferences would be able to be possibly drawn.

REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, "Detecting Spammers on Twitter," *In Proceedings of CEAS (2010)*, Vol. **6** No. **12**.

[2] B. Liu, M. Hu, J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *In Proceedings of International Conference on World Wide Web (WWW-2005)*, **2005**.

[3] A. M. Möller, R. Kühne, S. E. Baumgartner, J. Peter, "Exploring user responses to entertainment and political videos: An automated content analysis of YouTube," *Social Science Computer Review*, Vol. **37**, Issue. **4**, pp. **510–528**, **2019**.

[4] M. Thelwall, P. Sud, F. Vis, "Commenting on YouTube Videos: From Guatemalan Rock to El Big Bang," *Journal of the American Society for Information Science and Technology*, Vol. **63**, Issue **3**, pp. **616–629**, **2012**.

[5] X. Cheng, C. Dale, J. Liu, "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study," *Technical Report arXiv:0707.3670v1 {cs.NI}, Cornell University, arXiv e-prints*, July **2007.**

[6] P. Schultes, V. Dorner, F. Lehner, "Leave a Comment! An In-Depth Analysis of User Comments on YouTube," *Wirtschaftsinformatik Proceedings, 42*, **2013**.

[7] P. G. Lange, "Commenting on Comments: Investigating Responses to Antagonism on YouTube," *Presented at the Society for Applied Anthropology Conference, Tampa, Florida*, **2007**.

[8] K. Dinakar, R. Reichart, H. Lieberman, "Modeling the Detection of Textual Cyberbullying," *In International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain*, **2011**.

[9] M. Z. Asghar, S. Ahmad, A. Marwat, F. M. Kundi, "Sentiment Analysis on YouTube: A Brief Survey," *MAGNT Research Report*, Vol. **3**, No. **1**, pp. **1250–1257**, **2015**.

[10] S. Siersdorfer, S. Chelaru, W. Nejdl, "How useful are your comments? Analyzing and predicting YouTube comments and comment ratings," *Paper presented at the 19th international conference on World wide web, Raleigh, NC*, **2010**.

[11] D. Murthy, S. Sharma, "Visualizing YouTube's comment space: online hostility as a network phenomena," *New Media & Society*, **2018**.

[12] A. J. Murali, V. S. Chooralil, "A Literature Survey on Web-Based Traffic Sentiment Analysis: Methods and Applications," *IJSER*, Vol. **6**, No. **12**, pp. **926–930**, **2015**.

[13] E. Khabiri, J. Caverlee, C. Hsu, "Summarizing User-Contributed Comments," *International AAAI Conference on Web and Social Media, North America*, July **2011**.

[14] E. H. Poche, "Analyzing User Comments On YouTube Coding Tutorial Videos," *LSU Master's Theses, 4452*, **2017**.

[15] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," *Paper presented at IMC'07: Internet Measurement Conference, San Diego, CA*, **2007**.

[16] A. U. R. Khan, M. Khan, M. B. Khan, "Naïve Multi-label Classification of YouTube Comments Using Comparative Opinion Mining," *Procedia Computer Science*, **82**, pp. **57–64**, **2016**.

[17] A. Ammari, V. Dimitrova, D. Despotakis, "Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models," *UMAP*, pp. **71–85**, **2011**.

[18] D. O'Callaghan, M. Harrigan, J. Carthy, P. Cunningham, "Identifying Discriminating Network Motifs in YouTube Spam," *arXiv preprint arXiv:1202.5216*, **2012**.

[19] M. Hu, B. Liu, "Mining and Summarizing Customer Reviews," *Proceedings of the 10th ACM SIGKDD International conference on knowledge discovery and data mining*, **2004**.

[20] A. Sureka, "Mining User Comment Activity for Detecting Forum Spammers in YouTube," *CoRR abs/1103.5044*, **2011**.

[21] S. Choudhury, J. G. Breslin, "User Sentiment Detection: A YouTube Use Case," *Proceedings of The 21st National Conference on Artificial Intelligence and Cognitive Science*, 30 August - 1 September, **2010**.

AUTHORS' PROFILES

Karina Bohora graduated from College of Engineering, Pune (COEP), India in 2020 with a major in Information Technology. She is passionate about the role of data and analytics in ensuring community wellbeing and curbing prevalent cyberbullying practices. Her research interests include text mining, data mining for detection of relation between offline bullying and online harassment as well as usage of pattern discovery to identify root causes of herd mentality of trolls, bullies.

Sanjay Patil graduated from College of Engineering, Pune (COEP), India in 2020 with a major in Information Technology. He believes that data mining techniques can help in making social media spaces safe and free from online harassment and thereby contribute towards ascertaining sound mental health. Some of his research interests are statistical tools, data analysis algorithms for detecting vulnerability to targeted attacking and hatred in online spaces, as well as machine learning.

