2021 International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 25-27, 2021

Utilizing Topic Modelling To Identify Abusive Comments On YouTube

Shubhanshu Shekhar
Department of Computer Engineering
Delhi Technological University
New Delhi, India
shubhanshushekhar_2k17co338@dtu.ac.in

Akanksha
Department of Computer Engineering
Delhi Technological University
New Delhi, India
akanksha_2k17co34@dtu.ac.in

Aman Saini
Department of Computer Engineering
Delhi Technological University
New Delhi, India
amansaini_2k17co46@dtu.ac.in

2021 International Conference on Intelligent Technologies (CONIT) | 978-1-7281-8583-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CONIT51480.2021.9498368

***Abstract*—Online video platforms such as YouTube was once regarded as a haven for entertainment, educational, and promotional purposes. Now they have become a breeding ground for spreading toxic behavior, radicalizing content, and political propaganda. We have used YouTube data API v3 to scrape several data related to youtube videos like videos URL, description, view count, and comment information like commenters, comments, and replies. We investigated some hot topics which are prone to abusive comments. For example, bullies are highly existent in racial, teenage lifestyle and appearance, LGBTQ topics, or targeted mainly towards women and girls. We randomly selected a couple of videos from these YouTube channels and used them to research and analyze this project. We have used LDA (latent Dirichlet allocation) to identify dominant topics in a particular uploader's comment section. We have also used the TextBlob library of python for determining the polarity and subjectivity of comments for each uploader.**

***Index Terms*—YouTube, LDA (Latent Dirichlet Allocation), TextBlob, Abusive Comments , Polarity and Subjectivity**

I. Introduction

The last decade is known as Internet Bloom as in previous ten years, many people had joined the internet and caused massive participation on a platform like Youtube. YouTube is the world's second-largest search engine after Google having over 1.9BN users and over 1,000,000,000 hours of YouTube videos are watching in a day, which is more than Netflix and Facebook videos combined. Hate speech, offensive language, sexual harassment, racism, and other forms of abusive behavior have become very common. That is why we have used youtube for our analysis. The part of our concern is the comment section. The comment section is the section that is accessible to everyone in the world to write anything as some people use it to communicate with other peoples or connect with content creators. Some people misuse this platform and use abusive languages against the audience and the content creator, which is non-appreciated. These may reflect a destructive impact on other people and influence them to use more inappropriate contexts online.

Anyone with a user account can indulge in inappropriate behavior by commenting on the videos, which can often lead to other negative comments and spreads like a disease. Moreover, this toxic behavior is unfortunate and hard to deal with it. Such negativity embarrasses and humiliates subjects. For teenagers and young adults, it can be hard to deal with and can be psychologically jarring. Some abusive cross the line into unlawful or criminal behavior. There are several incidents where young adults and teenagers have committed suicide because of online harassment. Due to the overwhelming audience and large user-base, there must be a check on the video's content and the comments because they are visible to everyone. There is a considerable number of people who make abusive comments. Despite social media efforts to combat abusive behavior online, the problem is still emerging[12]. Now they have joined the arms race with criminals, who are constantly changing tactics to escape these platforms' detection algorithms.

Moreover, only one out of ten people will inform their parents or trusted adults about their abuse. It increases the child's risk of anxiety, low self-esteem, depression and can even lead to drugs and alcohol, and insomnia.

The major challenge is that the cyberbullying attack can come at any point of the day. As most teens have their smartphone and 24 hrs access to the internet so they can bully or are being bullied at any time virtual means plus it is very challenging to defend oneself from cyberbullying as we live in a digital world.

One survey has also suggested that 19% of the teens have reported that someone has written or posted embarrassing things on their social network, and 71% of the survey participant do not feel like social platforms are doing enough to fight the problem. Also, 42% of LGBT youth have experienced cyberbullying, while 35% have received online threats. In contrast, 58% have been a victim of hate speech at least once. Hence, we have chosen YouTube mainly because it is prevalent among people. It provides a fertile environment for these abusive behaviors, as most of the videos have a comment section open in them[13]. Some of the questions we will answer in this project are the topics that people are most abusive about and how people are exploiting, that is, what is the worst form of abuse. We present the below contribution in this paper.

1. Exploratory Data Analysis

1.1. Top words that are highly correlated with the uploader and are not appropriate 1.2. Formation of the word cloud to visualise the data and getting insights from it 1.3. Graphically visualising the profanity of comments

2. Topic Modelling using LDA (Latent Dirichlet allocation) , i.e., looking at words in a text and determining what topic is the text all about. Here, we will try to find topics using LDA (Latent Dirichlet Allocation), and from the topics that people talk to each other in the comment we will Figure out which uploader is more prone to abuse.

3. Subjectivity and polarity i.e, we will use inbuilt TextBloblibrary of python and determine the subjectivity and polarity of the comments. By plotting the same, we can get the insights of a positivity - negativity, and fact-options of each uploader's comments. We can then compare the abusiveness to each YouTuber's comment.

II. RELATED WORK

There is no dataset available online for abusive youtube content. Therefore, to do any analysis, we need to extract our data, create our dataset, and research our own accord. Thus there could be analysis on the dataset, which is done through manually labeling the data. Researchers do the work on working on their dataset and doing separate studies.

Previous works on this topic include Inoka Amarasekara and Will J Grant showed the gender gap in Youtube channels[1]. Chen, H., Mckeever, S., Delany, S. J. worked on detecting the Abusive Content of social media by harnessing text mining power [2]. Fox, J., Tang, W. Y. and Whittaker E. Kowalski R.-M who showed the statistics about sexual harassment to women on the social platform[3][5]. Obadimu A., Mead E., Hussain M.N., Agarwal N. identified different toxicity types in the Youtube Video Comments[4].

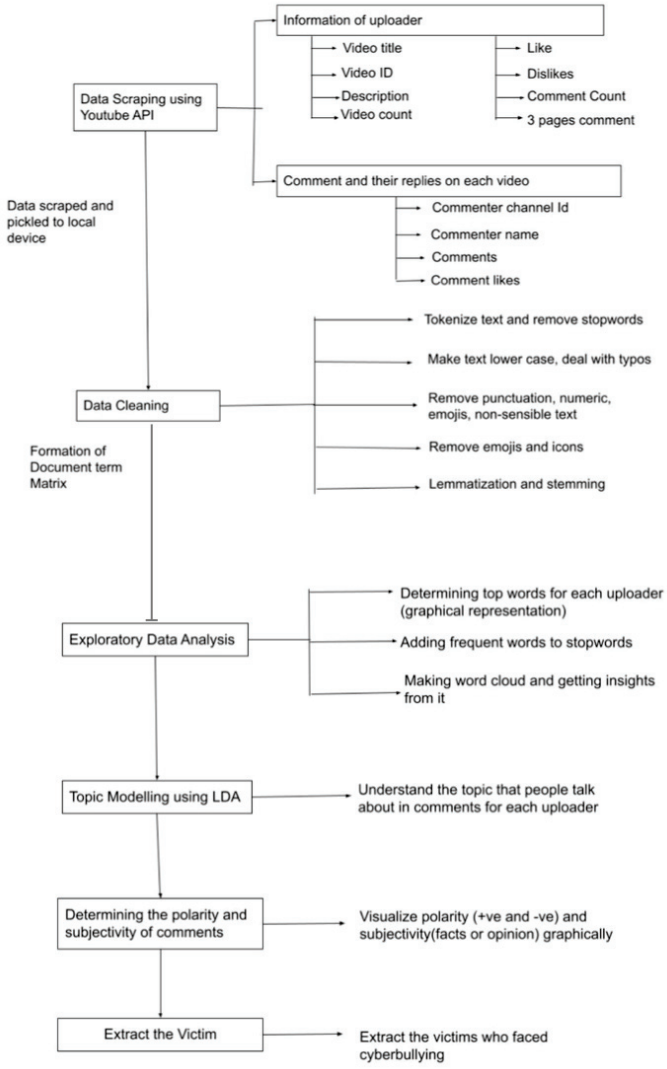
By referencing the above research and their methods, we developed an analysis to Find which Youtuber had the most abusive comments. Thus it needs to report this information to youtube and turn off the comment section for the good well-being of Youtuber and the people who use Youtube.

III. METHODOLOGY

A. Data collection

We have investigated some hot topics which are prone to bullying, for example, teenage lifestyle and appearance, LGBTQ topics, or cyberbullying is targeted mainly towards women and girls. We have also chosen some of the YouTube celebrities who were bullied or cyberbullied at some point in their life and randomly selected a couple of videos from their YouTube channel to research and analyze this project.

We have used YouTube Data API v3 to scrape the metadata of these handpicked videos that is video URLs, description, views and likes statistics, comment information like commenter's name, comments, comment likes, And their replies. We have also analyzed the commentator's and their comments by scrapping all three-page comments per video as a single text. For analyzing what kind of comment each uploader is getting and saving comments and their replies in a separate CSV File for each video.



1. Flow Chart

B. Data pre-processing

The phrase garbage in garbage out is also applicable to the machine learning projects out there. Data preprocessing is an essential step before any NLP procedure. The data scraping part is very loosely controlled and further results in larges chunks of the data out of which only some amount of data is of our need. Hence we need to clean that. Here we combine multiple videos into one by concatenating all the comments into one for a particular user. We are interested in finding what kind of comments a particular uploader gets. After that, we

applied traditional text cleaning methods, i.e., making all the text lower case, removing punctuation, removing numeric values, removing non -sensical text, tokenize text, and removing stop words and emoticons, and also dealing with typos. Here we are skipping lemmatization and stemming. It is because they change the spelling and map word to their base word, moreover. In this project, we are looking for abuse words too. We do not want such words to be changed. Our goal is to make a Document term matrix (DOM), i.e., uploader name along the row side and number of words along the column side. Finally, we have eight rows and 7143 columns as an output of this step.

C. Exploratory Data Analysis

EDA is used to understand better the main features of data, variable, and relationship that holds them, identifying the crucial variable for us, which is not directly visible.Exploratory data analysis majorly uses statistical graphs and many other data visualizing methods. It tells us beyond what the actual model can tell us, i.e., it gives us the fundamental insight into the data and gives us an intuition of what model to use for machine learning. Here we AIM to explore our data set and find the hard facts about the comments, such as top words for all the videos, top 30 words for each person.

Fig 2. Word Cloud

We will add frequent words to the stop words as if the words are coming in all the unloaders, so it does not make any sense to have it now as it will not help us to differentiate between the videos, hence include that words also to the stop words and remove then while preprocessing. We also here view the graphical relationship between several mild and swear words. Moreover, we have also made the word cloud to analyze the most common words for each uploader.Graphical analysis is done to identify a number of unique words in the comments section for each uploader, i.e., a frequency graph of unique words

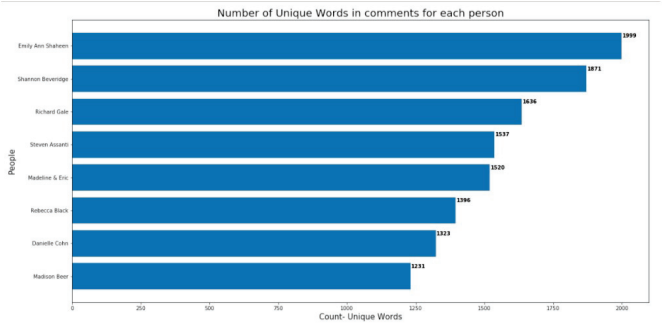


Fig 3. Unique word graph

D. Topic Modelling

Topic Modelling is a statical modelling technique use for discovering the abstract topic from a document. This is an unsupervised classiZication between documents. We can also say that it is analogous to clustering in numerical data. LDA (latent Dirichlet allocation) is the most popular topic modelling technique among all the topic modeling ways. Hence we are using it. Inbuilt python modules are used for this purpose with the parameters of num_topics=4 and passes=20. We have applied LDA on a large chunk of comments proposed before. Here we will understand the topics, i.e., topics for each uploader's comments, and after that, analyze which uploader is more prone to abuse or cyberbullying.

E. Sentiment Analysis

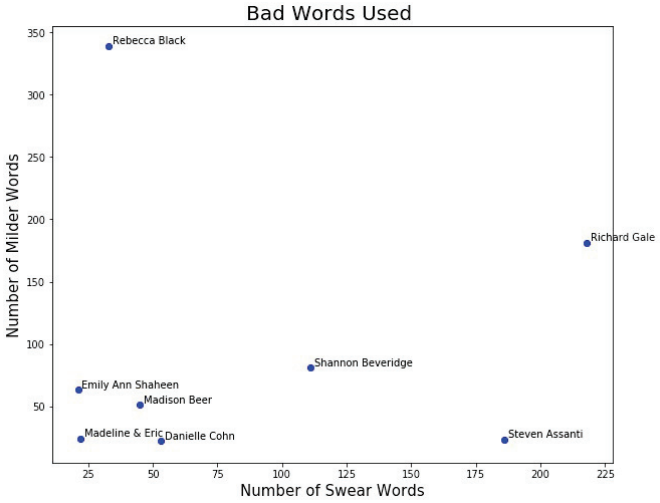
Sentiment analysis is a process of detecting positive, negative, or neutral sentiments in a text. It is mainly used on textual data. Hence we are using it in our project for analytics purposes. Since everyone expresses their feelings and thoughts more openly than ever, analyzing the sentiments had become a significant part of any NLP pipeline.
Using sentiment analysis, we not only focus on polarity, i.e., positivity or negativity, but also on urgency, emotions (happy, sad, low), and even intentions. We used the TextBlob library of python for the sentiment analysis and applied the methodologies in each of these subjects' comments, and analyze what they could mean. We will find polarity and Subjectivity of comments for all the uploaders to determine each uploader's sentiments. Polarity tells us how much positive or negative connotation a sentence or group of the sentence has. It ranges between -1 to 1, where 1 is a super positive sentence while -1 otherwise, and Subjectivity tells us if a sentence states a fact or opinion. It ranges from 0 to 1. 0 is where a sentence or group of a sentence is factual while one is highly opinionated.

IV. RESULTS AND DISCUSSION

1) In the section of Exploratory data analytics, a graph is made between the number of mild words and the number of abusive words for the understanding purpose about the

frequency of mild and abusive words for each uploader . We can get the following insights from the graph

- Highest Number of Swear Words: Richard Gale
- Second Highest Number of Swear Words: Steven Assanti
- Highest Number of Milder Abuse Words: Rebecca Black
- Least swear words: Emily Ann Shaheen , Madeline Eric
- Least milder abuse words: Madeline Eri



From the chart above, we can confirm that videos on Steven Assanti and Richard Gale has more abuse words. Even from the word cloud we saw that they have words that hurts someone psychologically.

We have used LDA for the topic modelling and the results of topic modelling are as follows

SN.	Topics	Uploader name
1	song/music/voice, Friday/ party/weekend	Rebecca Black
2	beautiful/perfect/amazing song/music/voice	Madison Beer
3	gay/lesbian, sex, kid, fat, victim, b**ch	Shannon Beveridge, Steven Assanti, Richard Gale
4	arab, gold, girls, family, beautiful, pregnant	Emily Ann Shaheen, Madeline & Eric

Table 1. Topic corresponding to each uploader comments

From topics people talk about in the comments for each of the uploaders, we can see that yet Shannon Beverage, Steven Assanti, Richard Gale are prone to harsh comments than any other people on the list. As the number of swear words is much more in the comments of these YouTuber.

Sentiment analyses are done on each uploader's comments to determine the subjectivity and polarity of the comments. Sentimental analysis to get the insights from each uploader's comments, i.e., to determine how positive or negative a comment or how much opinionated each uploader's comments are? i.e., Subjective in a sentence generally refer to feeling, emotions or opinions and judgment, whereas objective refers to hard facts. Polarity is how positive or negative a comment is, i.e., it ranges from negative 1 to positive one, and comments can be positive, negative, or natural depending on the sign of value.

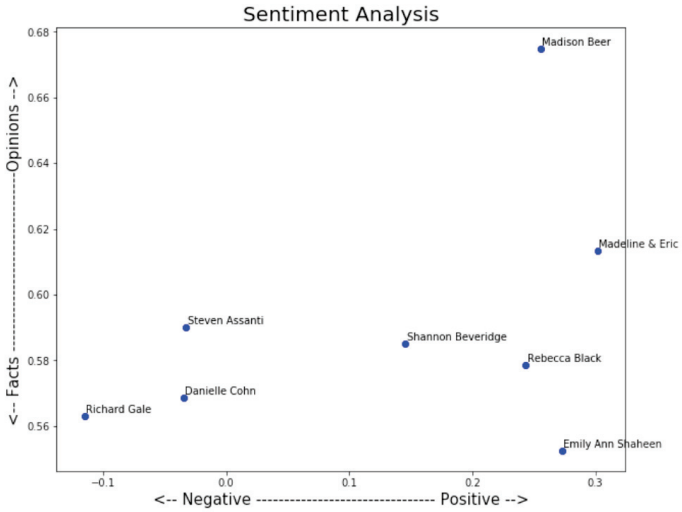


Fig 4. Sentimental Analysis

From the chart we can infer that:

- Madison Beer gets more opinionated positive comments. She is a singer, so it is obvious for her to getopinions at her singing.
- Emily Ann gets positive but factual comments. She gets comments mostly about her Arabic heritage. It is good to know that she does not get much xenophobic comments.
- Madeline and Eric get mostly positive and fairly opinionated comments. They are independent teen parents blogging their livelihood. Most people find it commendable.
- Rebecca Black who used to be heavily bullied for her Friday song now receives factual and positive comments. Though, most of the words in comments of her new video still refer to the 'Friday' song, she is admired for her persistence in singing career.
- Steven Assanti and Danielle Cohn gets more negative comments than above individuals. Steven gets more opinionated comments than Dani. The reason could be he is obese and bit rebellious towards comments on his obesity which generates worse comments. Danielle on other hand is a teenager who displays her odd choices of life in social media.

- Richard Gale videos get the worst comments and also contain factual comments. He was a bully who started a fight with his classmate. However, it is disheartening to see a minor being commented on harshly on social media. There should be some moral policing around this.

V. Conclusion

Looking at all previous comparisons, we find that Richard Gale and Steven Assanti are worse victims of cyber- bullying. Since Richard was/is a minor, we do not want him to take the wrong steps because of such a negative impact. Hence we would contact YouTube administrators or cyber police to ban comments on his videos as the first measure. Moreover, some of the hard-hitting question we can get from here are how to prevent this type of abusive behavior, can mocking will also be considered in the bullying, and do cyberbullying victims, and perpetrators Zit any stereotypical profile or not .

References

[1] Amarasekara I, Grant WJ. Exploring the YouTube science communication gender gap: A sentiment analysis. Public Understanding of Science. 2019;28(1):68-84. doi:10.1177/0963662518786654

[2] Chen, H., Mckeever, S., Delany, S. J. (2016). Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media. Advances in Computational Intelligence Systems, 187-205. doi:10.1007/978-3-319-46562-3_12

[3] Fox, J., Tang, W. Y. (2015). Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsive- ness, withdrawal, and coping strategies. New Media Society, 19(8), 1290– 1307. https://doi.org/10.1177/1461444816635778

[4] Obadimu A., Mead E., Hussain M.N., Agarwal N. (2019) Identifying Toxicity Within YouTube Video Comment. In: Thomson R., Bisgin H., Dancy C., Hyder A. (eds) Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2019. Lecture Notes in Computer Science, vol 11549. Springer, Cham. https://doi.org/ 10.1007/978-3-030-21741-9_22

[5] Whittaker E. Kowalski R.-M. (2015). "Cyber-bullying via Social Media." Journal of School Violence, Vol. 14, No.1. DOI: http://dx.doi.org/ 10.1080/15388220.2014.949377 Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] Thomas Davidson, Debasmita Bhattacharya, Ingmar Weber. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets.

[7] Lange, P. G. (2014). Commenting on YouTube rants: Perceptions of inap- propriateness or civic engagement? Journal of Pragmatics, 73, 53–65. doi:10.1016/j.pragma.2014.07.004

[8] R. Kaushal, S. Saha, P. Bajaj and P. Kumaraguru, "KidsTube: Detection, characterization and analysis of child unsafe content promoters on YouTube," 2016 14th Annual Conference on Privacy, Security and Trust (PST), Auckland, 2016, pp. 157-164, doi: 10.1109/PST.2016.7906950.

[9] Kapadia, S. (2020, December 29). Topic Modeling in Python: Latent Dirichlet Allocation (LDA). Medium. https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

[10] Obadimu, A., Mead, E., Hussain, M. N., Agarwal, N. (2019). Identifying Toxicity Within YouTube Video Comment. Lecture Notes in Computer Science, 214–223. doi:10.1007/978-3-030-21741-9_22

[11] Obadimu, A., Mead, E., Hussain, M. N., Agarwal, N. (2019, July). Identifying Toxicity Within YouTube Video Comment. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (pp. 214-223). Springer, Cham.

[12] Hosseini, H., Kannan, S., Zhang, B., Poovendran, R. (2017). Deceiving google's perspective api built for detecting toxic comments. arXiv preprint arXiv:1702.08138.

[13] Gr¨ondahl, T., Pajola, L., Juuti, M., Conti, M., Asokan, N. (2018, January). All You Need is" Love" Evading Hate Speech Detection. In Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (pp. 2-12).