IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12671

Hate crime detection on social media (YouTube) using ML Techniques

Prathamesh Gade[1], Prof. Yogeshchandra Puranik[*2]

SYMCA, MCA Department, PES Modern College of Engineering, Pune-411005, Maharashtra, India[1]

Assistant Professor, PES Modern College of Engineering, Pune-411005, Maharashtra, India[2]

Abstract: With the increasing popularity of social media platforms like YouTube, monitoring and regulating user-generated content has become a significant challenge. In particular, detecting hate speech and hate crimes within YouTube comments is a critical task to ensure user safety, foster inclusive online communities, and comply with legal regulations. This abstract presents an overview of a research study focused on hate crime detection in YouTube comments using machine learning (ML) techniques. The objective of the study is to develop an automated system capable of identifying and flagging hate speech and potential hate crimes within user comments.

The research involves collecting a large dataset of YouTube comments labeled for hate speech, hate crimes, and non-offensive content. Our work encompasses two main contributions. Firstly, we have developed a detailed taxonomy for classifying hateful online comments, considering both the types of hate speech and the targets of such comments. This taxonomy enables a more granular understanding and analysis of hate speech occurrences in online social media.

Secondly, we have conducted an extensive machine learning experiment using various algorithms, including Logistic Regression, Decision Tree, Random Forest, Adaboost, and Linear SVM. The goal was to create a multiclass, multilabel classification model capable of automatically detecting and categorizing hateful comments within the realm of online social media.

To ensure the reliability of our model, we performed validation tests to assess its predictive capability. Additionally, this research has provided valuable insights into the distinct types of hate speech prevalent on social media platforms.

Keywords: Hate, YouTube, social media, offensive, Muslim, jihad, fool.

I. LITERATURE SURVEY

The field of computer science's scientific exploration of hate speech is relatively new, and the number of studies in this area remains limited. During our literature review, we encountered only one survey article [1] that provided a concise yet comprehensive overview of automatic hate speech detection in the domain of natural language processing. The authors of this survey article presented a structured and critical examination of the field, covering various aspects.

To begin, they introduced the necessary terminology for studying hate speech and delved into an in-depth analysis of the features employed in addressing this problem. Additionally, they focused on research concerning bullying, highlighting its relevance to hate speech. The article also discussed the applications of hate speech detection, particularly in anticipating alarming societal changes. Classification methods, challenges, and data-related considerations were dedicated sections within the survey.

As online platforms increasingly serve as a popular means of delivering news [2][3], the comment sections of these platforms have become significant spaces where users engage with the content, content providers, and each other to express their opinions. However, the ease of expressing opinions on online social platforms has unfortunately led to the misuse of this medium through the posting of toxic comments [4].

Consequently, numerous researchers have investigated various inappropriate behaviors within comment sections on different websites. While the majority of prior research has focused on designing mechanisms to classify or detect inappropriate comments, only a few studies have considered user experience and engagement.

Machine learning approaches have been widely adopted in this domain. For instance, Davidson et al. [5] proposed a multi-class classifier that distinguishes between hate speech, offensive language, and non-offensive content.

DOI: 10.17148/IJARCCE.2023.12671

II. INTRODUCTION

Hate speech, which refers to derogatory comments targeting specific groups or individuals (Walker, 1994), is prevalent and pervasive in online environments. Firstly, they contribute to a destructive cycle of insults and hostility, often referred to as "online firestorms," creating toxic and unproductive comment threads (Pfeffer et al., 2014).

We intend to utilize this taxonomy to train a machine learning model capable of automatically detecting hateful comments and identifying the specific targeted groups. Automation is necessary due to the labor-intensive and often neglected nature of manual comment moderation, particularly for media organizations producing a high volume of content on platforms like YouTube. Therefore, a fully automatic or computer-assisted moderation system is crucial to maintain the well-being of online communities. In pursuit of this goal, our research seeks to answer the following research questions:

To automatically identify and categorize hateful comments on social media, we will first create a classification system by carefully analyzing and coding different types of hateful comments. Using this system, we will then train machine learning models using annotated data to detect and classify hateful comments accurately.

Additionally, we will explore the main targets of hate speech online. To achieve this, we will apply our trained model to a dataset obtained from a well-known online media company.

By addressing these research questions, our aim is to contribute to the development of effective techniques for automatically detecting and categorizing online hate speech. Simultaneously, we seek to gain valuable insights into the primary targets of such hate speech.

III. METHODOLOGY

We obtained our dataset from a prominent online news and media company that has a wide international audience. This company maintains an active presence on social media platforms such as YouTube and Facebook, regularly publishing numerous videos each day. As a result, their videos often attract a substantial number of comments, with thousands of comments per video being quite common. During our exploration of this media company's social media presence, we noticed the prevalence of hateful language within the comments, motivating us to find automated methods for detecting and categorizing such comments. Considering the scale and real-world relevance of this issue for online news publishers, it was logical to leverage the data collected through YouTube APIs to create a valuable dataset for studying online hate speech.

By utilizing official APIs, we were able to extract a total of 137,098 comments from videos posted on YouTube. For our analysis, we focused specifically on English comments. To explore hate within this dataset, we constructed a simple dictionary based on two sources: a) public repositories of hateful words and b) a qualitative analysis. By examining the data and identifying commonly used terms associated with hateful comments, we made adjustments to the list of hateful words obtained from the public sources.

In summary, our dataset was collected from a major online news and media company through YouTube APIs, where we pulled a substantial number of English comments. We took measures to identify and address hateful language by creating a dictionary of hateful words based on public sources and qualitative analysis, refining it to reflect common terms found in the dataset. Figure 1 illustrates the most commonly used words.

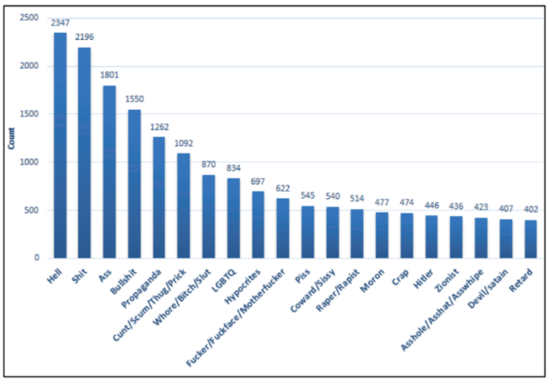
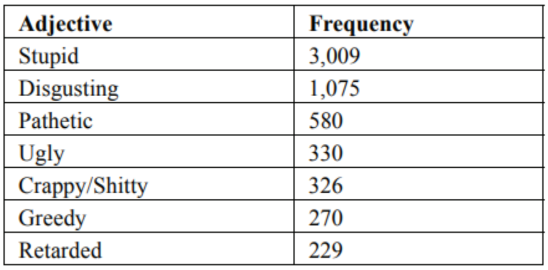


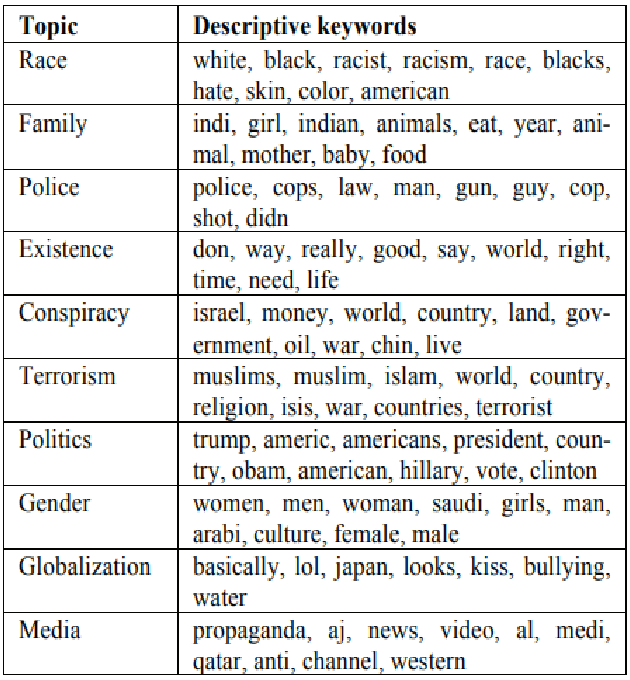
Figure 1: Distribution of Offensive words

Table 1 common offensive words.



Adjective	Frequency
Stupid	3,009
Disgusting	1,075
Pathetic	580
Ugly	330
Crappy/Shitty	326
Greedy	270
Retarded	229

Table 1: Distribution of Offensive Adjectives in the Dataset.



Topic	Descriptive keywords
Race	white, black, racist, racism, race, blacks, hate, skin, color, american
Family	indi, girl, indian, animals, eat, year, animal, mother, baby, food
Police	police, cops, law, man, gun, guy, cop, shot, didn
Existence	don, way, really, good, say, world, right, time, need, life
Conspiracy	israel, money, world, country, land, government, oil, war, chin, live
Terrorism	muslims, muslim, islam, world, country, religion, isis, war, countries, terrorist
Politics	trump, americ, americans, president, country, obam, american, hillary, vote, clinton
Gender	women, men, woman, saudi, girls, man, arabi, culture, female, male
Globalization	basically, lol, japan, looks, kiss, bullying, water
Media	propaganda, aj, news, video, al, medi, qatar, anti, channel, western

Table 2: Topics from LDA Analysis, Named by Researchers.

Architecture Model:

Our model operates by gathering diverse data from YouTube comments across various videos and classifying them as either hateful comments, specifically those related to cyberbullying, or non-hateful comments. The implementation of the cyberbullying identification and detection system on YouTube relies on utilizing features extracted from the YouTube dataset obtained through the YouTube API. This involves applying machine learning algorithms and utilizing Python libraries and code to facilitate the process.

Data Collection:

At first, we generated a YouTube API key. By using this API key, we implemented it in a python code to collect YouTube data from various YouTube videos. The raw data collected was stored in a csv file using python code and the data contained YouTube comments, and their respective usernames. This data was further sent for the splitting and filtration process.

Data Filtration:

To extract features from the YouTube comments dataset CSV file, we employed algorithms and libraries that enabled us to split and extract the data effectively. Our approach involved considering each unique word present in the document and tallying its frequency within the CSV file, where each row corresponds to a piece of data. The extracted features were represented in a vectorized format, ensuring a well-organized arrangement. This process involved selecting unique words from the entire document and keeping track of the number of occurrences for each word.

CONCLUSION

In conclusion, hate crime detection on YouTube using machine learning (ML) techniques is a crucial area of research and development in order to mitigate the spread and impact of hateful content on the platform. By leveraging ML algorithms and tools, it is possible to automate the process of identifying and flagging hate crimes in videos, leading to faster and more efficient detection.

ML techniques offer several advantages in hate crime detection on YouTube. They can analyze large volumes of data, including video and textual content, to identify patterns, keywords, and behavioral indicators associated with hate crimes. ML models can be trained on labeled datasets, consisting of examples of hate crimes, to learn the distinguishing characteristics and features of such content. This enables the models to generalize and identify potential hate crimes in real-time, even in cases where new and previously unseen patterns emerge.

Implementing ML-based hate crime detection on YouTube can have significant societal benefits. It can help protect users from exposure to harmful content and foster a safer and more inclusive online environment. By swiftly detecting hate crimes, platform administrators can take appropriate action, such as removing or flagging offensive content, suspending accounts, or even reporting instances to law enforcement agencies. This proactive approach can deter potential perpetrators and reduce the overall prevalence of hate crimes on the platform.

However, it is important to acknowledge the challenges and limitations of ML-based hate crime detection on YouTube. ML models are not infallible and can be susceptible to biases and false positives/negatives. Proper training and validation of models using diverse and representative datasets are crucial to minimize these issues. Additionally, hate crimes can be complex and context-dependent, requiring nuanced understanding and human judgment to accurately identify them. Therefore, ML techniques should be seen as a tool to assist human moderators and not as a substitute for human involvement and oversight.

In conclusion, hate crime detection on YouTube using ML techniques has the potential to be an effective and efficient approach in combating the spread of hate crimes. However, it should be accompanied by continuous research, improvement, and collaboration between ML experts, platform administrators, and relevant stakeholders to address the challenges and ensure the development of robust and fair detection systems.

Future Enhancement:

Improve the explain ability and transparency of ML models for hate crime detection on YouTube. This would involve developing techniques to provide clear explanations for the detection decisions, enabling content creators and users to understand why certain content is flagged as hate speech. Transparent models can help build trust and facilitate informed discussions about hate speech detection policies.

Implement ML models that can detect hate speech in real-time, enabling prompt intervention and mitigation. By analyzing live streaming content and providing immediate feedback or moderation alerts, harmful content can be addressed more efficiently, preventing its spread and potential impact.

Enhance ML models to be more resilient against adversarial attacks aimed at evading hate speech detection. Adversaries may try to modify or obfuscate hate speech content to bypass the system. By developing robust ML models that can withstand such attacks, the detection system can maintain its effectiveness over time.

Implement mechanisms to actively involve users in hate crime detection. ML models can be designed to proactively solicit user feedback on potentially problematic content, allowing users to flag hate speech and contribute to the training and improvement of the detection system. This approach can create a collaborative environment for hate speech monitoring.

REFERENCES

[1] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In Proceedings of the Workshop on Natural Language Processing for SocialMedia (SocialNLP'17). 1.

[2] Locklear, M. More people get their news from social media than newspapers. https://tinyurl.com/y8ht3ubr, 2018. Accessed: 2020-16-04.

[3] GEIGER, A. Key findings about the online news landscape in America. tinyurl.com/y44m63xu, 2019. Accessed: 2020-16-04

DOI: 10.17148/IJARCCE.2023.12671

[4] Massaro, T. M. Equality and freedom of expression: The hate speech dilemma.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” arXiv Prepr. arXiv1703.04009, 2017

[6] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, Bernard J. Jansen Qatar Computing Research Institute, Hamad Bin Khalifa University Turku School of Economics at the University of Turku Independent Researcher.

[7] Antonios Anagnostou, Ioannis Mollas, Grigorios Tsoumakas School of Informatics, Aristotle University of Thessaloniki {anagnoad,iamollas,greg}@csd.auth.gr.