FLAGGING INAPPROPRIATE CONTENT IN VIDEOS WITH NATURAL LANGUAGE PROCESSING AND AUDIO RECOGNITION

Javin Djapri, Putu Stephanus Gerald, Michael Guston

Bina Nusantara University, Indonesia

Abstract

Profanity or Swear filters are commonly used tools in censoring unwanted words in a text or video. However, racial slurs, derogatory content, and hateful content which are harder to detect has become a problem in current society. We devised a method to recognize and flag these inappropriate contents by combining both Natural Language Processing and Audio Recognition techniques.

© 2020 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Global Science and Technology Forum Pte Ltd

*Keywords:* YouTube, Natural Language Processing, Audio Recognition, Inappropriate content

1. Introduction

Content filtering is essential for internet users. There are various misusage and unethical acts by irresponsible people which takes advantage of WWW users. The massive volume of internet content is followed by massive amount of internet users. Nowadays, one can easily access improper contents at will. Without proper countermeasures, internet users are at risk of being negatively influenced (Thangaraj & Karthikeyan, 2014).

YouTube is the largest video-sharing platform in the world which boasts more than one billion active users worldwide. It is positive that the platform has become a source of entertainment, education, and information to large amount of people. This is the reason why it is important that video contents inappropriate for the public, be it racial slurs, hateful comments, or unreasonable violence, need to be filtered out and removed from the platform.

In 2017, more than 20 million videos have been removed and 85% of these were identified by machine learning algorithms. It shows that Artificial Intelligence techniques are very potent and effective. However, it is a fact that many other videos showing inappropriate content has gone undetected under the radar.

In this paper inappropriate contents are defined as contents that contain profane words, racial slurs, and malicious/hate comments in the English language. One way to filter out the malicious content is by using information filtering techniques. Information filtering techniques are used to extract useful information and block unwanted information according to certain rules (Imran & Sukhur, 2015). However, common tools such as profanity filters are not powerful and diverse enough to remove the unwanted contents.

We employed audio-recognition technology and natural language processing to increase the accuracy and effectivity of filtering the contents. Audio recognition is mainly used to detect speech patterns as well as identify the video contents. Then, various NLP methods are used in the actual flagging of those malicious contents.

The main objective of this study is to propose a new content filtering method using String Matching. It is desired that by combining both audio recognition and natural language processing, flagging inappropriate video contents will be much more effective.

1. Literature Review

In studies relating to text censorship, various researchers were seen using several techniques such as Bayesian Filtering, k-Nearest Neighbor, Bayesian Network, and R\*Tree based technique.

In the paper by (Imran & Sukhur, 2015), they used Bayesian Filtering and Approximate String Matching to create a new text censoring system. Bayesian Filtering is used to detect the profanity, while Approximate String Matching enhances the effectiveness of detecting the words itself.

(Qingshan et al, 2011) employed Bayesian Network to increase the filtering accuracy. The paper discussed the use of an adaptive function which can learn new malicious content and is able to censor the latest malicious content.

‘A New KNN Categorization Algorithm for Harmful Information Filtering’ by (Du & Yi, 2012) used k-NN algorithm to identify profanity. They used the method of constructing virtual samples based on the actual samples. By having more samples, it subsequently increased the classification accuracy of the k-NN classifier.

(Yoon, Park & Cho, 2010) proposed using R\*Tree based searching algorithm to align a query word against all vulgar words registered in a database. This method overcame the otherwise expensive use of resources in computation. With a database of more than 9300 words, the system had a computation speed of 0.006 second with 1000 words.

1. Methodology

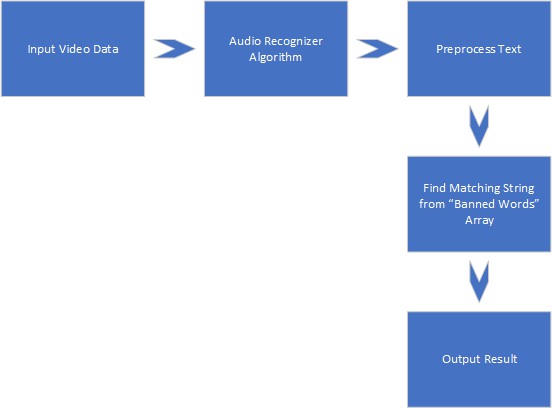


Fig. 1: Word Flow Diagram

The methodology of this project divided into several phases:

1. Video Downloading and Formatting

YouTube DL library is used for video downloading and function utilization. The link of the video from YouTube will be extracted by the download function. The program will access the video page on YouTube and immediately download the video. There are several parameters that being used to convert the downloaded video:

* + 1. format: the format parameter of the video is best audio.
    2. outtmpl: change the file name.
    3. postprocessor:
       1. key: using FFmpeg library for extracting the video audio
       2. preferredcodec: codec format is mp3
       3. preferredquality: the quality is 192

1. Audio Formatting

The mp3 audio file will be converted into Waveform Audio Format (WAV). WAV is used because python3 is unable to detect the audio file in another format. AudioSegment function from pydub library was used for converting the mp3 into WAV format.

1. Audio Recognition

For the audio recognition process, speech recognition library was used for recording the audio file and store the record data into a variable. The variable which contains the audio data will be recognized by the google recognition function. The function will convert the data from the audio into text. The converted text data will be stored into a different variable.

1. Text Filtering

Text Filtering begins will begin by creating the sample data in a labeled array, the sample data contains forbidden words which being used for text filtering. Then, the converted text data will be filtered by the forbidden words array using find function in the looping process. For every matching word, the program will print a warning and the word.

1. Result

The prototype of application that flagging any inappropriate words in a video, there are several steps for this application to get to work. The video’s link is the key so we can download and extract the video. The functions from Youtube.FL library will be used as the downloader and extractor for the youtube video.

FFmpeg is used to extract the audio of the youtube video. Converting the extracted audio will be the next step of the process, after converting mp3 to Waveform Audio Format (WAV) file. The mp3 and WAV file will be added to the folder.

We used speech recognition so we can get a machine to single out words or phrases in a spoken language from output.wav. But in this case, it is difficult to get a perfect speech recognition result using only a standard audio recognition technology. From the wav file, the audio will be converted into text data.

The next phase is to filter the text data, before the filter phase begin, a sample data that contains forbidden words were made. Then the text data will be filtered and matched with the sample data. If the system found any matches between the text data and the sample data, the program will notice user and give warning.

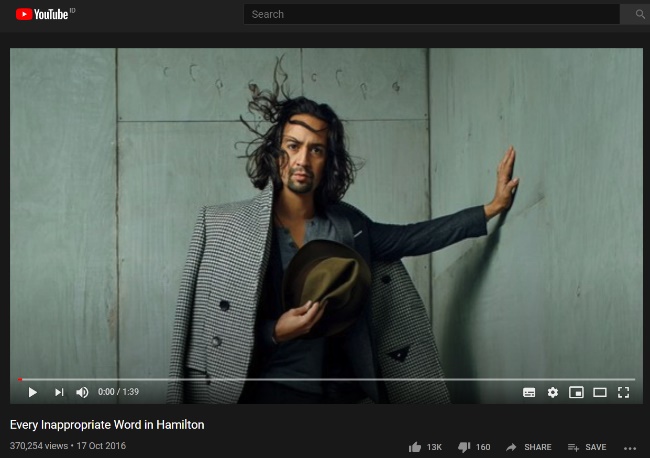


Fig.2: Youtube Video

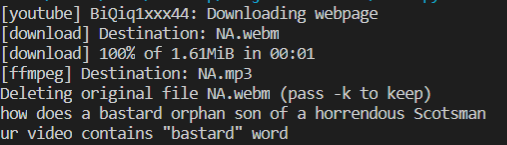


Fig.3: Result

1. Conclusion

Natural language processing and audio recognition can be used as a parameter that can define which the video friendly enough or maybe containing some bad words or appropriate words. By extracting the video and get the audio of the video, we can filter the text from the converted audio based on our filter word data. However, the recognition build in model seems cannot recognize the audio perfectly. Which can lead to mistranslation of the words.

For future work, we will try to build our own speech recognition model and implement it to our application.

Acknowledgements

These and the Reference headings are in bold but have no numbers. Text below continues as normal.

References

Imran, Khairil & Sukhur, Muhammad. (2015). Text Censoring System for Filtering Malicious Content Using Approximate String Matching and Bayesian Filtering. 331. 149-158. 10.1007/978-3-319-13153-5\_15.

Qingshan, Zhao & Shaobing, Wu & Ying, Cui & Xueming, Jia. (2011). The Research of Information Filtering Technology Based on Bayesian Network. Procedia Environmental Sciences. 11. 545–551. 10.1016/j.proenv.2011.12.086.

Du, Juan & Yi, Zhi. (2012). A New Knn Categorization Algorithm for Harmful Information Filtering. Proceedings - 2012 5th International Symposium on Computational Intelligence and Design, ISCID 2012. 1. 489-492. 10.1109/ISCID.2012.128.

Yoon, Taijin & Park, Sun-Young & Cho, Hwan-Gue. (2010). A Smart Filtering System for Newly Coined Profanities by Using Approximate String Alignment. 643-650. 10.1109/CIT.2010.