PAPER SENT TO CONFERENCE, WAITING FOR ACCEPTANCE

YouTube Spam Comment Detection

[1]Anu Antony, [2]Anusha Rajendran, [3] Deepa G
[1,2] PG Scholars, [3] Assistant Professor
[1,2,3] Department of Computer Science and IT
[1,2,3] Amrita Vishwa Vidyapeedam, Kochi, India
[1] antony.anu289@gmail.com, [2] anusha47646@gmail.com
[3] deepa@asas.kh.amrita.edu

Abstract. YouTube is an open platform where anyone can create content and share it with the world via their own YouTube Channel. YouTube offers a wide range of content for its audience of all ages, including comedy shows, vlogging, cooking recipes, hacks, unboxing videos, and other content are available.

If you're a frequent YouTube viewer and have spent time reading YouTube comments, you might have noticed a repetitive pattern, along with abuse or childish insults and dubious viewpoints, which is commonly referred to as spam. The presence of such spam comments has a negative impact on a channel's reputation as well as the experience of normal users.

YouTube has addressed the problem with very limited methods, such as YouTube Studio, which offers a "Report" or "Report as spam or abuse" option for comments, allowing the community to control the number of spam comments left on videos. While reviewing comments in YouTube Studio, creators can also report spam. However, these methods have attested to be not fruitful, as spam creators discovered methods to circumvent such heuristics. As a result, in this paper, we use six traditional machine learning algorithms to detect spam comments: Gaussian Naive Bayes, Support Vector Machine, linear classifier, AdaBoost, Random Forest, and decision tree classifier. The data for the model was obtained from the Kaggle Repository, and the experiments were carried out using Google Collaboratory.

Thus, we can achieve an accuracy of 91.39 percent with the AdaBoost Classifier, beating the current course of action by around 20%, and have shown to be extremely successful at detecting and removing spam comments.

Keywords: Spam; comments; Machine learning; YouTube; Random Forest; Support Vector Machine; linear classifier; AdaBoost Algorithm; Gaussian Naive-Bayes; decision tree classifier

Introduction

YouTube; an online video sharing and social media platform owned by Google that allows users to watch, like, share, comment on, and upload videos. On February 14, 2005 Chad Hurley, Steve Chen and Jawed Karim founded it and it is the 2nd most visited site in the world. It has around one billion users monthly who watch almost one billion hours of video per day. On YouTube, you'll find everything from adorable pets to funny science lessons, quick fashion tips, quirky cooking demos and much more. It attracts two types of users: those who create channels and upload videos to them, and those who watch videos, interact with them, and subscribe to them. As YouTube is owned by Google, all you need to create a YouTube account and start using it is a Google account. The sheer number of user-generated videos available on YouTube is one of the reasons it is so popular. Every minute, about 100 hours of video are uploaded to YouTube, so there's always something new to watch!

One of YouTube's most popular features is its commenting system, which allows users to leave comments on videos uploaded to different channels allowing the users to connect with each other while watching a video and share their thoughts, feelings, and so on. However, malicious users have taken advantage of this to spread content for promotion, also known as spam. These pieces of content that are overly shared, repetitive, or untargeted, and which do one or more of the following: Makes a promise to viewers that they will see something, but instead directs them to another website.

Obtains YouTube views, clicks, or traffic by promising viewers that they will make money quickly. Since the beginning of 2018, YouTube has been chastised for its inability to moderate user-generated content. Children make up a large portion of the YouTube user base, and they are frequently exposed to harmful and malicious comments. YouTube has tried to overcome this by removing all links from comments. Spammers have resorted to better creative methods, like inserting whitespace between links. There has been increasing pressure to find innovative solutions to the problem. We believe that recent advances in computing power have paved the way for traditional Machine Learning algorithms to be applied to such problems. Thus, in this paper, we try to figure out which algorithms can effectively filter spam comments and which heuristics can figure out spam with a larger F1 Score precisely.

Literature Review

[1] In this paper, they use ML methods like Random Forest, Naive Bayes, and Support Vector Machine, as well as proprietary heuristics like N-Grams, to detect spam comments. These algorithms have been shown to be quite fruitful in detecting and removing spam comments. They carefully extracted roughly 13000 comments from a variety of YouTube channels. They utilized a simple spam filter to extract Spam comments to speed up the search and increase the classifier's performance. They provided a strategy for automatic machine assisted identification of spammy comments in the YouTube platform in this study, emphasizing the superiority of character grams, substrings of n characters, over word-grams in terms of classification accuracy.

[2] In order to detect spam comments on YouTube, a fully connected feed forward neural network was utilized as a classification strategy in this study. RapidMiner studio was used to develop the ANN Model because it is a machine learning software that has various beneficial capabilities. Parameter optimization, regularization, data pre-treatment, data sampling, and model performance measurement are among these aspects. The optimal values of the ANN Model learning rate and the ANN Model L2 regularization were found using a grid parameter optimization operator. A split validation operator was utilized within the optimization operator, which was responsible for dividing the data into training and testing datasets. The linear sampling method was utilized with a split ratio of 0.7.To divide the data, the split ratio was adjusted to 0.7 and linear sampling was utilized. An ANN operator was employed inside the split validation operator, which was set to have two hidden layers.

The model's accuracy and classification performance were measured using a performance operator. The ANN Model outperformed Alberto's Models in terms of accuracy, Fl measure, and MCC in the majority of datasets, while having a lower or equivale not BH rate in the majority of datasets.

[3] The study examines the performance of five well-known univariate filter-based feature selection approach- es using text for comment spam filtering on YouTube, namely information gain, Gini index, distinguishing feature selector, discriminative features selection, and relative discriminative criterion, with classifiers like Naive Bayes-NB and decision tree DT. In the trials, 5 datasets containing spam comments from various subjects were used. The Macro-F1 success metric was used to assess the project. In addition, for a fair performance re view, 3-fold cross- validation is preferred. The TF-IDF approach is used to weight terms. In addition to weighing terms with TFIDF, lowercase conversion and Porter stemming were utilized. In this experiment, three-fold cross validation is employed for a smooth assessment of the performance. As a result of the Experiments, it was discovered that the DFS (Distinguishing feature selector) and GI (Gini Index) feature selection methods produced the best classification results. In most circumstances, however, the DT classifier outperforms the NB classifier when it comes to spam filtering on YouTube.

[4] This study describes a method for detecting spam comments on YouTube videos. The goal of this study is to identify spam users, those who provide comments solely for the purpose of self-promotion, and to identify individuals who leave comments that are irrelevant to the video. YouTube's new monetization policy for user channels, as well as the placement of various advertising on YouTube videos, has attracted a great number of people. As the number of
users on YouTube has grown, so has the number of unscrupulous individuals whose job it is to construct automated bots for commenting and subscribing to various YouTube channels. These harmful user comments harm the channel's reputation as well as the experience of regular users. Different strategies for separating spam comments from normal user comments in order to improve classification and a current trend in this field are briefly discussed in order to address this key issue. Support Vector Machine SVM and K-Nearest Neighbour KNN are made use in this study to create a YouTube detection framework (k-NN). This study has five (5) phases: data collecting, pre-processing data, feature selecting, classification, and spam detection. The trials are carried out with the help of Weka and RapidMiner. SVM and KNN accuracy results utilizing both machine learning algorithms indicate good accuracy. Another way to avoid spam attacks is to avoid clicking on links in comments.

[5] Text mining has exploded in popularity in recent years, because of the widespread availability of user- generated content. Comment mining is a common application, with opinion mining and sentiment analysis receiving a lot of attention. Text pre-processing is a crucial phase in the comment mining process, in which each of the linguistic terms is given a corresponding weight that increases along with its appearance on the analysed data but is neutralized by the term's frequency in the domain of interest. To calculate these weights, many people use the tf-idf formula. The bias contributed between participants' lecture to the research on comments in the social media is revealed in this paper, and an adjustment is proposed. We discovered that content retrieved from speech is frequently strongly associated resulting in dependence structuring between the data, causing statistical bias. Neglecting this bias results in a weak analysis, and a completely incorrect conclusion at worst. We suggest a TFIDF adjustment for this bias. They used data from seven Facebook fan sites to demonstrate the impacts of bias and correction in a variety of categories, like politics, sport, shopping, news, finance, and entertainment.

[6] Nowadays, social media allows users to connect and communicate by uploading comments or videos. In truth, comments are a website content that can be used by spammers to distribute phishing, malware, or advertisements. Because malicious individuals might transmit malware or phishing through comments, this research presents a method for detecting video sharing spam remarks. The dataset gathering phase is the initial step in the methodology utilized in this study. Datasets from the Machine Learning repository named UCI are used in this experiment. The building up of a framework and experimentation are done in the next step. Tokenization and lemmatization will be used to pre-process the dataset.Then the features to detect spamminess were chosen, and classification trials were carried out using six classifiers: KStar, Decision Table, Naive Bayes, Decision Stump, Random Forest and Random Tree.

Methodology

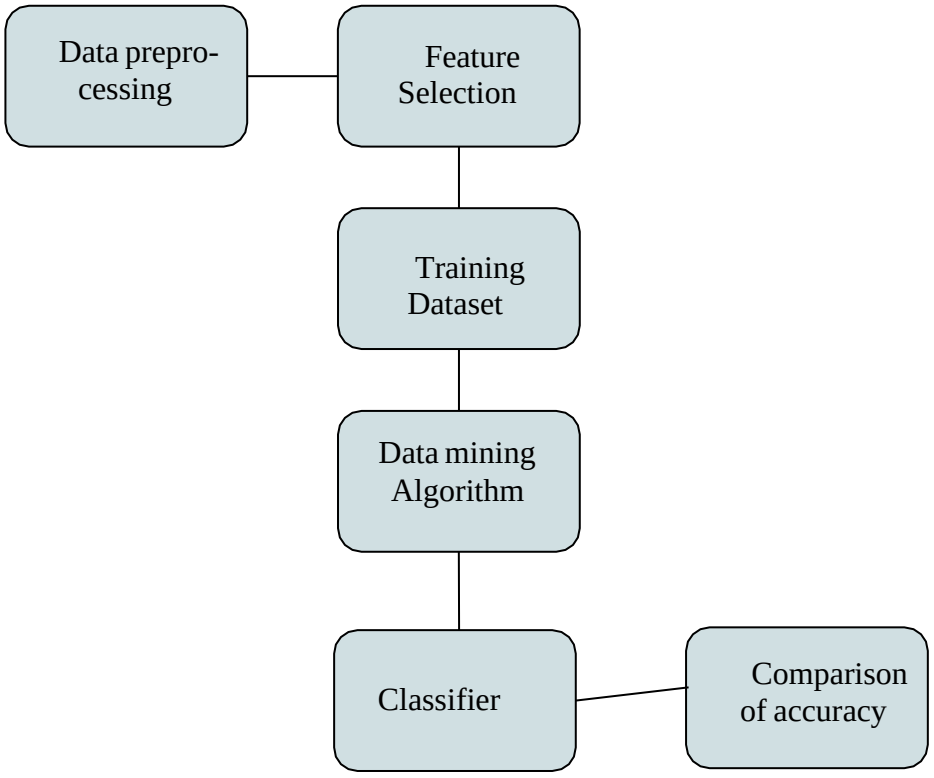

Fig 1. System Diagram

- Data Collection: The dataset for our research has been taken from UCI repository and Kaggle. The data are comments of 5 different YouTube videos in which 4 of them are used for training which includes 1400 data values and the last dataset is used for testing which has 371 comment values.

- Data pre-processing: As we cannot work with raw data it is important to clean the dataset by transforming raw data into an understandable format using suitable features like removal of punctuation marks, spaces followed by stemming, stop word removal etc.

- Stop word removal: here we have eliminated the words that are so commonly used that they carry very little useful information such as the, we, you, have, has, was, me, mine, his, her etc.

  Stemming: Stemming is reduction of a word into its base word that affixes to suffixes and prefixes. It reduces inflected words to their word stem, base/root word.

  Feature Selection: Feature selection is where suitable features such as keywords, hashtag etc. are identified to be tested by various classification algorithms

- Training Dataset: the features of the algorithm that is used to extract dataset is trained

- Data mining algorithms

- Classifier: Finally, the result is Classified based on the accuracy result shown by different classifiers through different data mining tools

- Comparison of result accuracy

B. Technical Module 1. Supervised Learning

Supervised learning comes under machine learning where machines get schooled using training data followed by testing them to predict the result.

2. Gaussian Naive Bayes

A Gaussian Naive Bayes algorithm is a NB algorithm that uses a Gaussian or normal distribution as its basis. When the features have continuous values, it's used specifically.

Because you only need to assess the mean and standard deviation from your training data's, this model is also the simplest to use.

3. Random Forest

Random Forest is an assemblage learning model that fuses many decision trees on various dataset subsets followed by averaging the output to upgrade the dataset's predictive accuracy.

4. SVM Classifier

SVM is a supervised ML algorithm which is used both for classification and regression. SVM is primarily used in Machine Learning to solve classification problems. SVM is a simple concept: It takes historical data as input and outputs a line or hyperplane that divides mixed data into classes.

5. Decision Tree

The internal nodes of a decision tree represent the dataset features, the branches define the decision constraints, and each leaf represents the results. It is represented graphically to obtain all the possible solutions to a decision based on certain parameters.

6. Linear Classifier

Linear classifiers are widely used in classification and serve as the foundation for more sophisticated classification methods. It's a statistical classification method for determining an object's class based on its characteristics.

7. AdaBoost

Adaptive boosting, or AdaBoost, is a process that select the feature that need to be improved in order to improve the model's prediction power. By combining multiple "weak classifiers" into a single "strong classifier," to improve performance, it can be combined with a variety of other learning algorithms.

Results

In the proposed paper we have obtained a result by comparing 6 ML Algorithms namely Support Vector Machine, Random Forest, Gaussian Naive Bayes, AdaBoost, linear classifier, and decision tree classifier that AdaBoost provides the highest Accuracy for determining the spam comments in the YouTube Comment section.

ALGORITHMS	ACCURACY	F1-Score
Gaussian Naïve Bayes	0.8756	0.8524
Decision Tree	0.8924	0.8717
Linear Classifier	0.9032	0.8860
Random Forest	0.8924	0.8684
SVM	0.9032	0.8860
AdaBoost	0.9139	0.9024

Fig 2. Performanceand Analysis graph

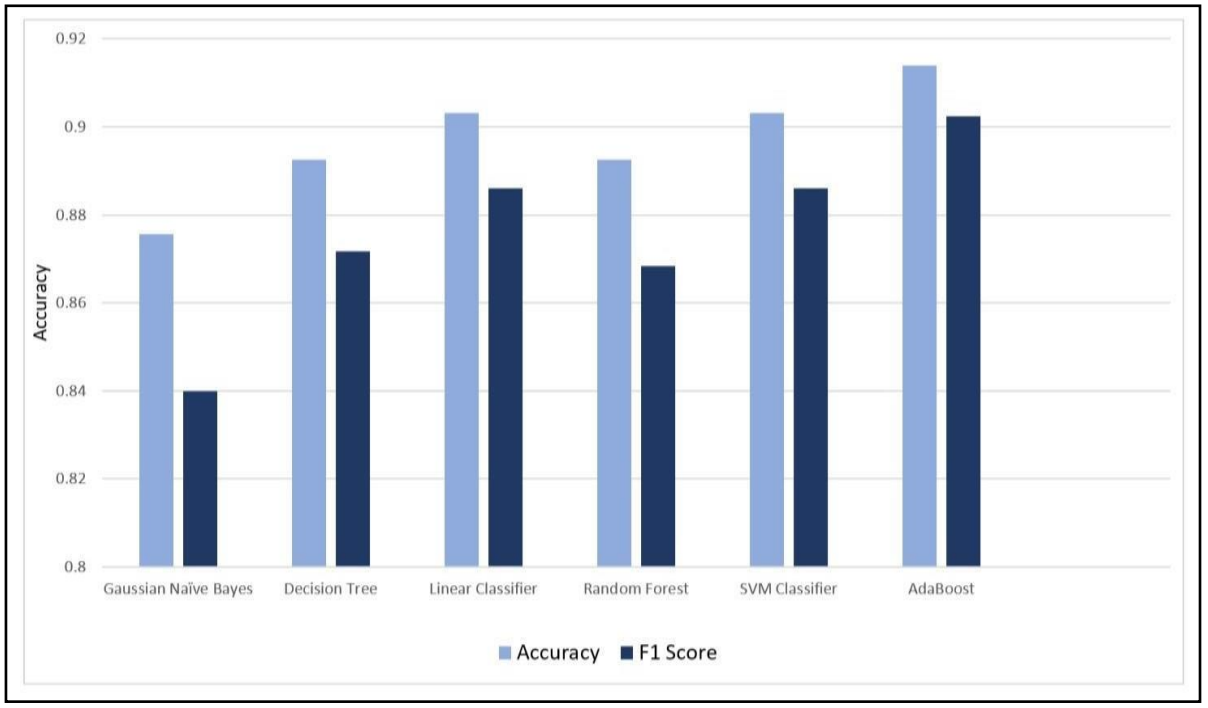

Fig 3. Analysis graph

Conclusion and Future Scope

Various techniques are used to classify YouTube comments as spam and not spam. Here, in this work as a result of comparing 6 algorithms we have come to a conclusion that AdaBoost Classifier gives the most precise Spam Comment prediction and provides the most effective spam comment detection. For the Future work, using other machine learning algorithms we can again classify the spam comments into different categories on the basis of what kind of spam Comment it is.

References

1. N-gram assisted YouTube spam comment detection, Shreyas Aiyar, Nisha P Shetty 2018

2. YouTube Spam Comments Detection Using Artificial Neural Network,
Thulfiqar Abd, Hussein Altabrawee and Samir Qaisar Ajmi Al Muthanna University, 2018

3. Feature Selection for Comment Spam Filtering on YouTube, Alper Kursat Uysal Eskisehir Technical University,2018

4. Spammer Detection: A Study of Spam Filter Comments on YouTube Videos, Rafaqat Alam Khan Lahore Garrison University 2019

5. Comments Mining With TF-IDF: The Inherent Bias and Its Removal, IEEE, Inbal Yahav; Onn Shehory ; David Schwartz

6. Video spam comment features selection using machine learning techniques,Nabilah Alias, Cik Feresa Mohd Foozy,
Sofia Najwa Ramli, N Zainuddin Indones. J. Electr. Eng. Comput. Sci 15 (2), 1046-1053, 2019

PLAGIARISM REPORT

ORIGINALITY REPORT

4%	2%	1%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Inbal Yahav, Onn Shehory, David Schwartz. "Comments Mining With TF-IDF: The Inherent Bias and Its Removal", IEEE Transactions on Knowledge and Data Engineering, 2019 Publication	1%
2	1tapdollarup.com Internet Source	1%
3	sciencepubco.com Internet Source	1%
4	Shreyas Aiyar, Nisha P Shetty. "N-Gram Assisted Youtube Spam Comment Detection", Procedia Computer Science, 2018 Publication	1%
5	editora.iap.org.br Internet Source	<1%
6	1library.net Internet Source	<1%
7	www.hindawi.com Internet Source	<1%