SPAM COMMENT FILTRATION

Agyey Arya
Stevens Institute of Technology
10459599

Mit Dani
Stevens Institute of Technology
10478091

Renuka Shilke
Stevens Institute of Technology
10478235

Veeksha Shetty
Stevens Institute of Technology
10474588

***Abstract*—Google's new video distribution network YouTube has drawn a rising number of customers due to its profitability. However, such popularity has attracted nefarious people who want to market themselves or spread viruses and malware. Because YouTube only provides limited comment moderating tools, the level of spam is shockingly increasing, leading several well-known channels to block the comments section of their videos. Because the posts are short and often replete with slangs, symbols, and acronyms, automatic comment spam filtering on YouTube is a difficulty even for proven categorization systems. We evaluated several high-performance classification techniques for this purpose in this study.**

I. Introduction

The comment sections of websites are full of comments from bots with URLs leading to malicious vectors. They often use the comment sections to post misleading information, these are referred to as spam. Spam comments have been around since 2003. These unpleasant remarks are made in order to increase the spammer's click through rate to his or her own website. The goal of this method is to improve the search engine ranking of the target site. It's worth noting that spammers rarely target specific individuals. Spammers frequently employ sophisticated software that targets websites based on a range of characteristics, such as how well your site ranks for specific keywords and the themes covered by your blog. They just wish to profit from the success of others in order to promote and expand their own websites. Even with community monitoring techniques like manual comment flagging and expert review/audit of comments on a regular basis, a lot of spam gets through. More automated and intelligent solutions are required, which will not replace manual auditing but will speed up the process. We must ensure that these systems are as exact as feasible as more people acquire access to the internet. We can boost coverage/recall by using a sufficiently flexible mechanism, such as online learning over time, to collect as many spam comments as possible.

II. LITERATURE REVIEW

Even though the first spam was delivered in 1978, it wasn't until 1982 that it was recognized as an issue in scientific literature. The article by Peter J. Denning is one of the first to address this issue. The Bayes' method, which was originally utilized by Sahami et.al in 1996 and thereafter by additional researchers, was the first mathematical instrument applied to spam filtering systems. Bayes' classifier is based on the well-known Bayes theorem, which was first published in 1960. Over the course of its more than 40-year history, the Naive Bayes Classifier (NBC) has been utilized to solve a wide range of problems, ranging from text categorization in news organizations to primary disease diagnosis in medicine.

YouTube has a number of techniques to counteract spammy comments, and it automatically deletes a large number of them. According to YouTube spokesperson Ivy Choi, the corporation eliminated ”almost 950 million comments for breaching our standards around spam, misleading, and scams” in Q4 2021 alone, using machine learning and human review. Automated flagging systems discovered ”the great majority” of the removals. You-Tuber ThioJoe's ”YouTube Spammer Purge” tool ”allows you to filter and search for spammer comments on your channel and other's channel(s) in many different ways AND delete/report them all at once,” according to the website.

III. RESEARCH QUESTION

Whether one blogs for profit or pleasure, shouldn't leave comments area unattended to allow spam to take over. If regular readers notice a high number of spam comments on the posts, it may discourage them from offering meaningful input. Additionally, spam comments with links might lead consumers to dangerous websites, potentially compromising their information or infecting their devices with a virus. Finally, if left unchecked, spam comments can divert attention away from the site and send visitors to other sites. This could result in a decrease in revenue for the company or a decrease in the number of people who follow content. Spam Comments have negative impact on both site's trustworthiness and performance. In order to maintain the credibility and speed, we need to clear the spam. Malicious comments impersonate the creators in an attempt to scam their viewers. Even though many organizations have spam filtration mechanisms in place in their comment areas, spam messages continue to get through. The question is how to improve on past outcomes and create a more effective method. To create a classifier in the past, researchers looked only at the substance of the comment. We believe that increasing the number of attributes will increase the models' performance. There are other data sets available, but only one is balanced, meaning it has the same quantity of spam and ham comments. The data has not been trained on the imbalanced data sets in recent efforts. We plan to integrate these data sets and work on the unbalanced classification and/or anomaly detection challenged

IV. METHODOLOGY

A. Data sources

- YouTube API : https://github.com/ThioJoe/YT-Spammer-Purge/wiki/Instructions:-Obtaining-an-API-Key
- YouTube API Video Comment
- http://mlg.ucd.ie/yt/

B. Data Description

Data was scraped from YouTube, and it was possible to retrieve comments for a certain video id using the YouTube data. The original content, author name, number of likes, and total number of comments were all retrieved. Two annotated data sets were used in addition to the scraped data.Our data set consists of 6433427 comments, of which 0.925 are not spam(ham) and the rest 0.075 are spam.

C. Models

We perform Tokenization where we break down text into smaller pieces. After that we remove Stop Words also known as un-informative words The dataset is divided into a Train and Test set. The training set, which includes all the conditional probabilities and language, exposes the algorithm to samples of spam and ham. The test set is the set of data that the algorithm never sees. Bag of words Technique – Count Vectorizer is used to count number of times each word appears and put them into vector. To make the model more precise, TF-IDF (Term Frequency- Inverse Document Frequency) feature is added that will tokenize the documents and its primary function is to evaluate how relevant a word is to a document in collection of documents.

a) RANDOM FOREST CLASSIFIER: Random Forest consists of a large number of individual decision trees that operate on ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. It uses bagging and feature randomness when building each individual tree to try to uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

b) DECISION TREE: A decision tree or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with an input feature are labeled with each of the possible values of the target feature or the arc leads to a subordinate decision node on a different input feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes, signifying that the data set has been classified by the tree into either a specific class, or into a particular probability distribution (which, if the decision tree is well-constructed, is skewed towards certain subsets of classes).They are are a non-parametric supervised learning method used for classification and regression.The goal is to create a model that predicts the value of a target variable based on several input variables.

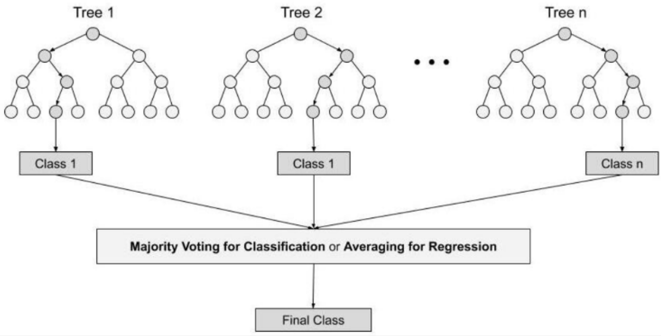

Tree 1
Tree 2
Tree n
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

Fig. 1. Model Structure

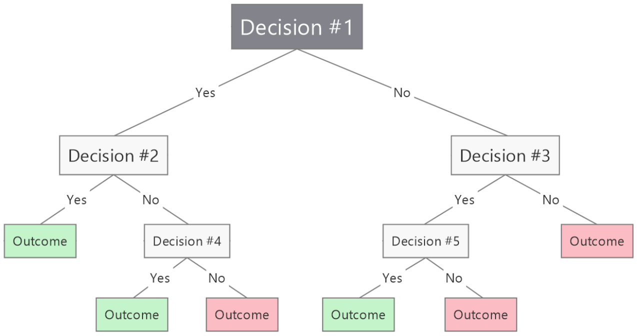


Fig. 2. Model Structure

c) MULTINOMIAL NAIVE BAYES:: The Bayes theorem is the foundation of Naive Bayes, which states that features in a dataset are mutually independent. The occurrence of one trait has no bearing on the likelihood of the occurrence of the other. Naive Bayes predicts a text's tag. They calculate each tag's probability for a given text and output the tag with the highest probability. For Example, Spam filtering in email, Diagnosis of diseases, making decisions about treatment, Classification of RNA sequences in taxonomic studies

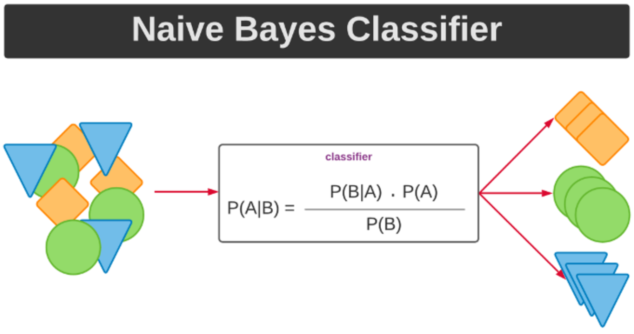


Fig. 3. Model Structure

where, PA= the prior probability of occurring A PBA= the condition probability of B given that A occurs PAB= the condition probability of A given that B occurs PB= the probability of occurring B In this project, our goal is to filter spam YouTube comments so Multinomial Naïve Bayes can be one of the models used to perform the analysis. The most significant sub-tasks in text classification are feature extraction and selection. The following are the three main criteria for good features: Salient: The features should be meaningful and important to the problem Invariant: The features are resistant to scaling, distortion, and orientation etc. Discriminatory: For training of classifiers, the features should have enough information to distinguish between text.

d) STOCHASTIC GRADIENT DESCENT - Classifier (Support Vector Classification): The Stochastic Gradient Descent - Classifier (SGD-Classifier) is an SGD-optimized linear classifier (SVM, logistic regression). SVMs are a class of supervised learning methods for classification, regression, and outlier detection. High-dimensional spaces are where support vector machines shine. Cases in which the number of dimensions exceeds the number of samples. It is memory efficient because it uses a subset of training points (called support vectors) in the decision function.

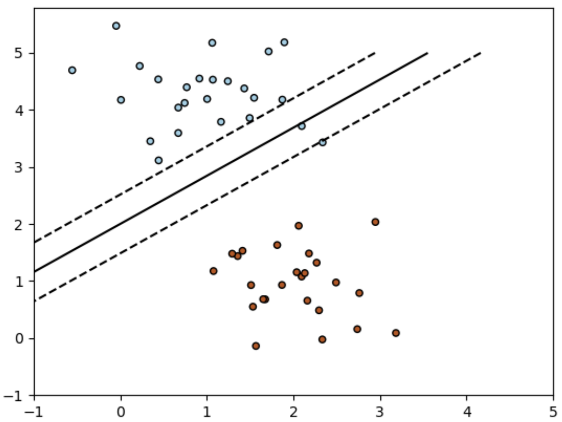


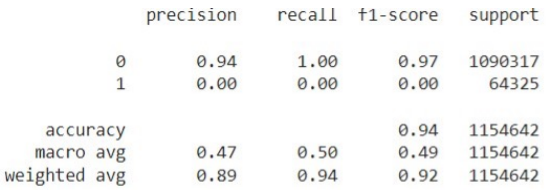
Fig. 4. Model Structure

V. Result Analysis

A. *RANDOM FOREST CLASSIFIER:*

- For Imbalanced dataset

  1) Ham F-1 Score: 0.97 , Spam F-1 Score: 0.00
  2) Spam Precision: 0.00, Spam Recall: 0.00



```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97   1090317
           1       0.00      0.00      0.00     64325

    accuracy                           0.94   1154642
   macro avg       0.47      0.50      0.49   1154642
weighted avg       0.89      0.94      0.92   1154642
```

Fig. 5. Classification Report

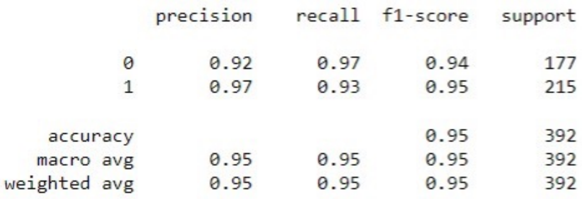
- For Random over sampling and under sampling

TABLE I
Report

Method	Oversampling	Undersampling
Spam f1 score	0.63	0.63
Spam f1 score	0.53	0.50
Ham f1 score	0.70	0.71
Spam precision	0.73	0.78
Spam recall	0.42	0.37

The imbalance in the dataset was corrected for by the above methods

- For Balanced dataset:
  1) Ham F-1 Score: 0.94 , Spam F-1 Score: 0.95
  2) Spam Precision: 0.92 , Spam Recall: 0.97

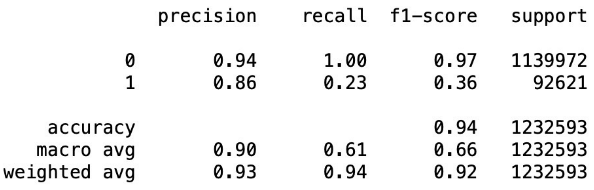


```
              precision    recall  f1-score   support

           0       0.92      0.97      0.94       177
           1       0.97      0.93      0.95       215

    accuracy                           0.95       392
   macro avg       0.95      0.95      0.95       392
weighted avg       0.95      0.95      0.95       392
```

Fig. 6. Classification Report

B. *DECISION TREE:*

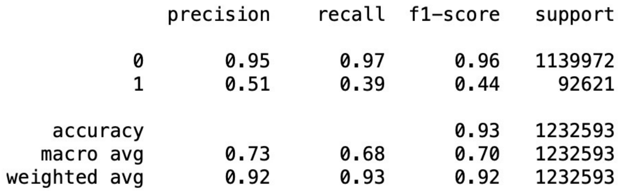
- For Imbalanced dataset:
  1) Ham F-1 Score: 0.97 , Spam F-1 Score: 0.36
  2) Spam Precision: 0.86, Spam Recall: 0.23



```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97   1139972
           1       0.86      0.23      0.36     92621

    accuracy                           0.94   1232593
   macro avg       0.90      0.61      0.66   1232593
weighted avg       0.93      0.94      0.92   1232593
```

Fig. 7. Classification Report

- For Balanced dataset
  1) Ham F-1 Score: 0.96 , Spam F-1 Score: 0.44
  2) Spam Precision: 0.51, Spam Recall: 0.39



```
              precision    recall  f1-score   support

           0       0.95      0.97      0.96   1139972
           1       0.51      0.39      0.44     92621

    accuracy                           0.93   1232593
   macro avg       0.73      0.68      0.70   1232593
weighted avg       0.92      0.93      0.92   1232593
```

Fig. 8. Classification Report

C. MULTINOMIAL NAIVE BAYES:

- For Imbalanced dataset:
  1) Ham F-1 Score: 0.97, Spam F-1 Score: 0.55
  2) Spam Precision: 0.79, Spam Recall: 0.42
- For OverSampling:
  1) Ham F-1 Score: 0.80 , Spam F-1 Score: 0.79
  2) Spam Precision: 0.82 , Spam Recall: 0.76
- For UnderSampling:
  1) Ham F-1 Score: 0.73 , Spam F-1 Score: 0.69
  2) Spam Precision: 0.74 , Spam Recall: 0.65

D. SGD-SVM:

- For Imbalanced dataset:
  1) Ham F-1 Score: 0.96, Spam F-1 Score: 0.32
  2) Spam Precision: 0.34, Spam Recall: 0.30
- For Class Balancing:
  1) Ham F-1 Score: 0.86 , Spam F-1 Score: 0.22
  2) Spam Precision: 0.14 , Spam Recall: 0.62
- For OverSampling:
  1) Ham F-1 Score: 0.78 , Spam F-1 Score: 0.80
  2) Spam Precision: 0.77 , Spam Recall: 0.82
- For UnderSampling:
  1) Ham F-1 Score: 0.67 , Spam F-1 Score: 0.68
  2) Spam Precision: 0.68 , Spam Recall: 0.68

TABLE II
FINAL RESULT

Model	Ham f1-score	Spam f1-score
Random Forest	0.94	0.95
Decision Tree	0.96	0.44
SGD-SVM	0.78	0.8
Multinomial NB	0.80	0.79

VI. CONCLUSION

The purpose of the study was to discover effective methods and settings for detecting spam comments on YouTube. Various strategies are used to categorize YouTube comments as spam and not spam (ham). This method was tested with real-time YouTube comments and yielded an overall accuracy of 0.95. Because the YouTube API is an open platform for all users, it may influence spammers' behaviour over time. In the actual world, the YouTube spam feature will not be steady; it will change rapidly.

In the future, Deep neural network-based systems such as convolutional recurrent neural networks may provide greater accuracy results for recognizing harmful Youtube comments.

REFERENCES

[1] https://vitalflux.com/class-imbalance-class-weight-python-sklearn/

[2] https://www.theverge.com/2022/4/8/23016861/youtube-comment-spam-testing-moderation

[3] https://machinelearningmastery.com/bagging-and-random-forest-for-imbalanced-classification/

[4] https://towardsdatascience.com/how-to-scrape-youtube-comments-with-python-61ff197115d4

[5] https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf