IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPARATIVE STUDY ON YOUTUBE SPAM COMMENT DETECTION USING VARIOUS MACHINE LEARNING ALGORITHMS

[1]P. Vimala Manohara Ruth, [2]Mohammed Sohail Khan, [3]Y Siva Prasad Reddy

[1]Assistant Professor, [2]Bachelor of Engineering Student, [3]Bachelor of Engineering Student

[1]Department of Computer Science and Engineering,

[1]Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India

Abstract: People now feel more comfortable socializing over the internet through popular social networking and media websites than face to face interaction. Thus, the social media websites are thriving more nowadays. YouTube is a vastly popular social media site which is expanding at very fast pace. YouTube depends mostly on user created contents sharing and spreading videos. However, due to this popularity, YouTube has become more susceptible to unwanted and malicious spammers. The increase in quality of online social networks, it has become easy for the spammers to implement malicious activities into the websites by adding spam messages in the comment section of the videos. This paper classifies whether the youtube comments are legitimate or spam comments. Few sample datasets are taken and trained on different models. Each model classifies spam and non-spam comments. An optimal model is being chosen (high accuracy model) and is deployed into web application. This web application takes a YouTube video link as input and detects spam comments in it. The accuracy of model is been observed as 95%.

Index Terms **– spam comments, classification, logistic regression, random forest**

I. INTRODUCTION

Social networking has been occupying major time of daily schedule. It has become an integral part of human lives. People look up to social media for other human interaction, sharing ideas, obtaining knowledge, entertainment and being informed about the events happening around the world. Among these websites, YouTube is one of the most popular website for sharing and viewing video content. This popularity of YouTube has its own side effects because it attracted spammers, who upload videos with the sole purpose of polluting the system content and causing dissatisfaction among other viewers. The spam videos can have lot of videos which may be unrelated to their title or may contain pornographic content. Therefore, it is very important to find a way to identify such videos and report them before they are viewed by users. YouTube themselves have been blocking the comments like URLs in the comment section. Such methods have proven to be extremely ineffective as spammers have found ways to bypass such heuristics. Standard machine learning classification algorithms have proven to be somewhat effective but there is still room for better accuracy with new approaches. To design a model which detects YouTube spam comments in real time using machine learning classification algorithms by taking YouTube video URL as input. This technology can be leveraged for various purposes and in fields. As every process these days is being automated here are a few applications of this project idea. Among different kind of undesired content, YouTube is facing problems to manage the huge volume of undesired text comments posted by users that aim to self-promote their videos, or to disseminate malicious links to steal private data. To stop such kind of activities this spam detection can be supportive.

II. LITERATURE SURVEY

Simran Kanodia et.al. [1] in the paper titled "A Novel Approach for Youtube Video Spam Detection using Markov Decision Process" used markov decision process to detect the YouTube spam comments. This model gave approximately 78.8 % accuracy. A Markov Decision Process (MDP) is a mathematical model that is used to make decisions in a stochastic way. A stochastic environment has outcomes which are partially random. They are under the decision maker's control. The environment in MDP is modeled as a discrete-time state-transition system with a set of actions and states. The state is represented as the outcomes of the decision-making process, an action is represented as a decision which can be taken from a particular state. The agent receives a reward as a consequence of the decision it chooses from a particular state. The aim of the agent is to maximize the total reward

of the model calculated over all the states in the environment. For generation of data, this paper used publicly available APIs as YouTube crawlers and extracted the required attributes of a given video.

Abdullah O. Abdullah et.al. [2] in the paper titled "A Comparative Analysis of Common YouTube Comment Spam Filtering Techniques" used Super Vector Machine model for testing and training the datasets. Support Vector Machine (SVM) is one of the algorithm used for Supervised Learning, that is used for Classification as well as Regression problems. However, in machine learning, it is majorly used for classification problems. The major goal of the SVM algorithm is to identify best line or decision boundary that can differentiate n-dimensional space into classes so that the new data points can be easily put in the correct category for future classification. This is one of the best decision boundaries which is called as hyperplane. SVM creates a hyperplane by choosing the extreme points/vectors. Extreme cases are formed which are termed as support vectors, and therefore the algorithm is termed as Support Vector Machine. The datasets used are extracted from YouTube using YouTube Data API. The advantages of this method is easy to implement and better accuracy. The drawback is it cannot be used in real time analysis by taking youtube links.

Shreyas Aiyar et.al. [3] in the paper titled "N-Gram Assisted Youtube Spam Comment Detection" the probability of a given N-gram is predicted within any given word sequence of English language. If a good N-gram model is developed, the prediction p(w | h) – the probability of identifying the word when a previous word h is given in history – where the history contains n-1 words. Naive bayes, random forest and SVM classifiers were used for analysis. Random forest yielded better results. This project manually extracted around 13000 comments from different channels across Youtube using the public Youtube API. The advantages are it used hybrid model and works on real-time data.

Hayoung Oh [4] in the paper titled "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model" proposed a method based on comparative research which is a representative study on YouTube spam comment detection. This method applied six machine learning techniques (i.e., CART (Decision Tree), LR (Logistic Regression), NB-B (Bernoulli Naïve Bayes), RF (Random Forest), SVM-L (Support vector machine with linear kernel), and SVM-R (Support vector machine with Gaussian kernel)) to improve the performance of the Cascaded Ensemble Machine Learning Model aware YouTube Spam Comments Detection Scheme. An ensembled model was proposed combining them and the performance was evaluated. The experimental environment used version 3.7.1 of Python and version 0.20.1 of the Cicely Library on Jupiter notebooks.

Oviya Selvaraj [8] in the paper "Youtube Spam Comments Detection" predicted the spam comments present in the comments section of Youtube videos using machine learning. Supervised learning approach depends on a very large number of labelleddatasets. . The proposed classification algorithm (LogisticRegression) is used in order to predict the spam comment.The purpose of project is to introduce briefly the techniquesof machine learning and to outline the prediction technique.Being much more superior to the conventional data analysistechniques, machine learning can open a new opportunityto explore and increase the prediction accuracy.Spam remarks are regularly completely immaterial to thegiven video and are normally created via mechanized botscamouflaged as a client. The comments section is target byspammers to post completely irrelevant messages,comments, links and ideas. AI is the strategy for extraction,changing, stacking and anticipating the significant data fromenormous information to remove a few examples andfurthermore change it into justifiable structure for additionalutilization. Grouping and expectation are two sorts ofdissecting information which portray principal classes ofinformation and forecast of patterns in future information.The noxious spam remarks will ruin the positive perspectiveof the contents present in the videos posted. Thecontingency for anticipating the spam remarks has startedbut has yet not been concluded and built up for an exactforecast of spam remarks.

III. METHODOLOGY

To implement the model, the standard datasets are extracted from Kaggle. The extracted datasets are based on two main characteristics: the popularity of the channel and the availability of up-to-date comments. There was no other consideration apart from these two characteristics. Therefore, the datasets were randomly selected (not based on celebrities or whatsoever). The total number of used YouTube channels is 100 and the overall samples are 10,000. Nevertheless, not all channels have presented the exact same sample share.

Before starting with the model preparation, preprocessing of the comments is done like checking whether all the characters must be in lowercase. The word which is both in uppercase and lowercase must be considered as same words and not as two different words. Then tokenization is done for each comment in the data set.

The main advantage of using the words present in the dataset is that it is capable of reducing uncertainty in the prediction of the final results as those phrases have a remarkable effect of frequency count in spam and ham comments in YouTube.

Attribute significance is a supervised characteristic that ranks attributes in a step-by-step manner with their significance in predicting an aim. Here Count Vectorizer is used which convert a "collection of text documents to a matrix of token counts.

A. SYSTEM ARCHITECTURE

Fig. 1: Sequential Flow of operations

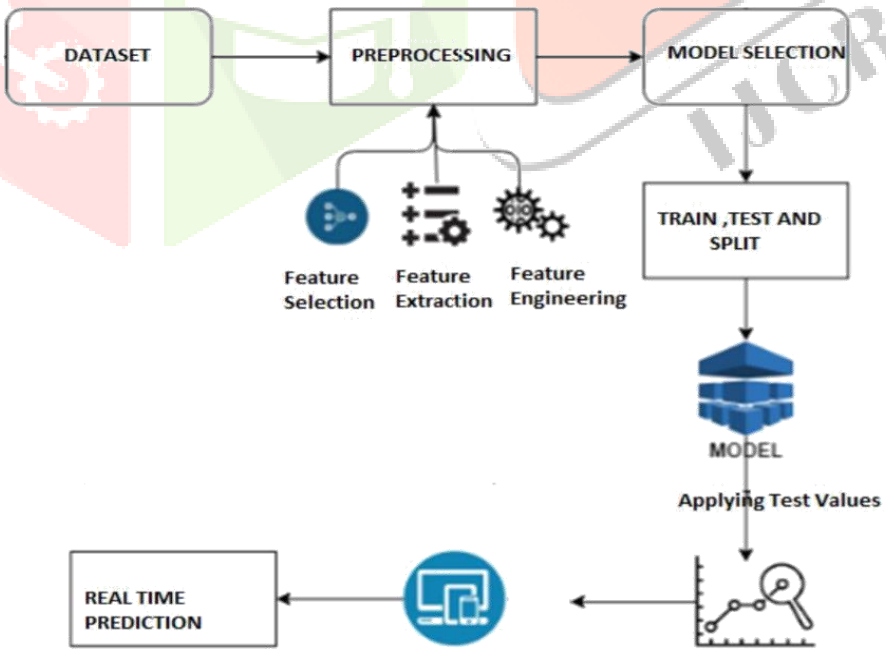
The datasets are trained on various models. Each model accuracy is calculated and compare the results. A model selection will be done based the performance and high accuracy. The bestmodel is being implemented and deployed into web application.

B. MODEL SELECTION

The datasets are trained on the following models and model with the high accuracy is selected. The datasets which are considered are Youtube-01 Psy.csv, Youtube-02 KatyPerry.csv, Youtube-03 LMFAO.csv, Youtube-04 Eminem.csv, Youtube-Shakira.csv and Spam.csv

Table 1: Datasets used

Dataset Name	Spam	Ham	Total
Youtube-01 Psy.csv	175	175	350
Youtube-02 KatyPerry.csv	175	175	350
Youtube-03 LMFAO.csv	236	202	438
Youtube-04 Eminem.csv	245	203	448
Youtube-05 Shakira.csv	174	196	370
Spam.csv	4497	672	5169



Random forest

Random forest is an ensemble of decision tree algorithms. It is an extension of bootstrap aggregation (bagging) of decision trees and can be used for classification and regression problems. In bagging, a number of decision trees are created where each tree is created from a differentbootstrap sample of the training dataset. A bootstrap sample is a sample of the training datasetwhere a sample may appear more than once in the sample, referred to as sampling with replacement.

Bagging is an effective ensemble algorithm as each decision tree is fit on a slightly different training dataset, and in turn, has a slightly different performance. Unlike normal decision treemodels, such as classification and regression trees (CART), trees used

in the ensemble are unpruned, making them slightly overfit to the training dataset. This is desirable as it helps to make each tree more different and have less correlated predictions or prediction errors. Predictions from the trees are averaged across all decision trees resulting in better performancethan any single tree in the model.

Logistic regression

Logistic regression is a technique that is used for predicting binomial or multinomial values of a variable. A statistical approach is used to find the outcome of the variable which is binary in nature. It uses a logit function for the prediction of probability of occurrence of binary outcome, it follows bernoulis distribution, so the outcome generated will be accurate either true or false. The dataset is given as input and it works on dataset and predicts x or y that is spam or ham. Logistic regression is named for the logistic function that is used as the core method.

The logistic function, which is also called the sigmoid function, developed by statisticians was to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It can take any values and map it into a any value between 0 and 1 into an S-shaped curve, but never exactly at those limits.

1 / (1 + e^-value)

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function [3].

Representation Used for Logistic Regression

Logistic regression uses an equation as the representation, very much like linear regression.Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.
Below is an example logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.
The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's) [3].

Naïve Bayes

Naive Bayes is used for classification of data for binary class and multi-class classifications. This is one of the easiest technique to understand when described using binary or categorical input values.
It is called Naive Bayes because the probabilities are calculated for each hypothesis and are simplified to make their calculation tractable. Instead of re-calculating the values of each attribute value P(d1, d2, d3|h), already existing values are assumed to be conditionally independent based on the target value which is calculated as P(d1|h) * P(d2|H) and so on.
There is an assumption that the attributes interact in real data and this assumption is not proved well for many cases. Nevertheless, the approach performs surprisingly well on data where this analysis does not hold good.

Representation Used By Naive Bayes Models [1]

The representation for naive Bayes is probabilities.

A list of probabilities are stored to file for a learned naive Bayes model. This includes:

- Class Probabilities: The probabilities of each class in the training dataset.
- Conditional Probabilities: The conditional probabilities of each input value given each class value.

Multinomial Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.
Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature.The presence or absence of a feature does not affect the presence or absence of the other feature. Naive Bayes is a powerful algorithm that is used for text data analysis and with problems with

multiple classes. To understand Naive Bayes theorem's working, it is importantto understand the Bayes theorem concept first as it is based on the latter. [5]

Bayes theorem, formulated by Thomas Bayes, calculates the probability of an event occurringbased on the prior knowledge of conditions related to an event. It is based on the following formula:

$P(A|B) = P(A) * P(B|A)/P(B)$

Where we are calculating the probability of class A when predictor B is already provided.P(B) = prior probability of B

P(A) = prior probability of class A

P(B|A) = occurrence of predictor B given class A probability.

SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine (SVM) is a supervised Machine Learning Algorithm that is used for either classification or regression of data. It is popularly used for data classification and sometimes used for regression. It is more preferred for classification over regression. The major task of SVM is to find a hyper-plane which identifies a boundary between the various types of data. In 2-dimensional space, this hyper-plane is predicated to be a line or boundary that separated two sets of data.

In SVM, we plot each data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data. Next, find the optimal hyperplane to separate the data. So by this, you must have understood that inherently, SVM can only perform binary classification (i.e., choose between two classes). However, there are various techniques to usefor multi-class problems. Support Vectors. Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.

Hyperplane: A hyperplane is a decision plane which separates between a set of objects havingdifferent class memberships.

Margin: A margin is a gap between the two lines on the closest class points. This is calculatedas the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a badmargin.

IV. IMPLEMENTATION

PyCharm is used as a cross-platform editor which was developed by JetBrains. Pycharm is an application that provides a package of tools which are required for productive Python development. pycharm has to be installed and flask application is run on local host. The packages required are selenium package. Selenium is an open-source automation testing tool which is used for automating tests carriedout on different web-browsers. Selenium is a software which is basically used for automating the testing in various web browsers. It supports browsers like google chrome, mozilla firefox, internet explorer and safari, and the browser is easily automated for testing across these browsers using Selenium WebDriver. Now, the clients can use live automated tests which are being performed on their computer screen. There is another package called Pandas, which is an open-source package, which is most widely used for data science/data analysis and machine learning analysis. Python is built on top of Numpy, which is another package that is used to provide support for multi-dimensional arrays.

Web scrapping is performed on the comments. Prediction of the comments is done using predict() function. Preprocessing of data set is being performed then dataset is divided into 33% testing set and 67% training set. Random forest classifier model is chosen because it gave high results among the four models. The results of classification is noticed in the following figures where four machine learning algorithms were tested on the dataset.

```
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.88	0.97	0.92	181
1	0.97	0.89	0.93	211
accuracy			0.93	392
macro avg	0.93	0.93	0.93	392
weighted avg	0.93	0.93	0.93	392

Fig. 2: Classification report for logistic regression

Fig. 2 depicts classification report for logistic regression and the accuracy calculated for 0.2 test size is 93%.

```
[ ] print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.91	0.85	0.88	181
1	0.88	0.92	0.90	211
accuracy			0.89	392
macro avg	0.89	0.89	0.89	392
weighted avg	0.89	0.89	0.89	392

Fig. 3: Classification report for Naïve Bayes

Fig. 3 depicts classification report for Naïve Bayes and the accuracy calculated for 0.2 test size is 89%

```
[ ] print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.90	0.97	0.93	181
1	0.97	0.91	0.94	211
accuracy			0.94	392
macro avg	0.94	0.94	0.94	392
weighted avg	0.94	0.94	0.94	392

Fig. 4: Classification report for Random forest classifier

Fig. 4 depicts classification report of Random Forest and the accuracy calculated for 0.2 test size is 94%

```
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.89	0.97	0.93	181
1	0.97	0.90	0.93	211
accuracy			0.93	392
macro avg	0.93	0.93	0.93	392
weighted avg	0.93	0.93	0.93	392

Fig. 5: Classification report for SVM

Fig. 5 depicts classification report of SVM and the accuracy calculated for 0.2 test size is 93%.

The test data size proportions were varied and f1-score is calculated for the algorithms logistic regression, random forest, multinomial and support vector machine.

Table 2: Comparison of Accuracy for various models with different proportions of test data

Test data size Proportion	Logistic Regression	Random Forest	Multinomial	SVM
0.2	0.93	0.93	0.89	0.93
0.33	0.93	0.95	0.89	0.93
0.43	0.94	0.95	0.87	0.94

These are the various accuracies for different models at different test size proportions. It is observed that as the test data size is increased, random forest yields higher accuracy that is 95%.

V. Conclusion and Future Work

YouTube a social networking feature website providing one of the largest video content publication. This project used four machine learning models and tested accuracy with different test size proportions. Among all the models, the Random classifier has given good accuracy that is 95% for the standard datasets. Unlike other existing projects, this project has the advantage of taking youtube video url and able to classify the spam and ham comments in real time. When the comments are very high for a youtube video then it takes more time and sometimes the machine may not yield results. In future, the model can be modified so that more accurate results can be obtained in low processing time and the size of datasets can be increased for better results.

References

[1] A Novel Approach for Youtube Video Spam Detection using Markov Decision Process, Simran Kanodia, Rachna Sasheendran,Vinod Pathari-IEEE, 2018.
[2] N-Gram Assisted Youtube Spam Comment Detection, Shreyas Aiyar, Nisha P Shetty- IEEE, 2018.
[3] A Comparative Analysis of Common YouTube Comment Spam Filtering Techniques , Abdullah O. ,Abdulkadir Sengur,Murat Karabatak, Mashhood A. Ali- IEEE, 2018.
[4] A Study of Spam Filter Comments on Youtube videos, Rafaqat Alam Khan, Research Gates, 2019.
[5] A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model,Hayoung Oh,IEEE,2016
[6] An Approach for Spam Detection in YouTube Comments Based on Supervised Learning,Amir Ali,Muhammed Zain Amin, Research Gates,2017
[7] A Framework for Detection of Videos Spam on Youtube ,Niyanta Ashar,Hitarthi Bhatt,Shraddha Mehta,prof.(mrs.) Chetashri Bhadane, Research Gates,2019
[8] Youtube Spam Comment Detection,Oviya Selvaraj,Anuradha Konatham,Dr. Paavai Anand, Academia,2020

[9] P. Vimala Manohar Ruth, Dr. Y.Rama Devi, E.Haritha, N.Shiva Kumar, "Prediction of Phishing Website For Data Security Using Various Machine Learning Algorithms", International Journal of Creative Research Thoughts(IJCRT), Volume:9, Issue:6, ISSN:2320-2882, June-2021.

[10] Agrawal, K., Bhargav, G., Spandana, E. (2021). Diabetes Diagnosis Prediction Using Ensemble Approach. In: Nath, V., Mandal, J.K. (eds) Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems. Lecture Notes in Electrical Engineering, vol 673. Springer, Singapore. https://doi.org/10.1007/978-981-15-5546-6_66

[11] T. Prathima, B Anjana, V Apoorva, BR Sreedhar Ensemble Based Hybrid Recommender Systems International Journal of Innovative Technology and Exploring Engineering (IJITEE) 826-833 Jan-2020 10.35940/ijitee.C8460.019320