# Spam Detection On Social Media Platforms

R Abinaya

*Department of Computer Science*
*St.Joseph's College of Engineering*
Chennai, India
abinayaramesh83@gmail.com

Bertilla Niveda E

*.Department of Computer Science*
*St.Joseph's College of Engineering*
Chennai, India
bertillaniveda@gmail.com

P Naveen

Assistant Professor
*.Department of Computer Science*
*St.Joseph's College of Engineering*
Chennai, India
navee4@gmail.com

***Abstract*— With the increased quality of online social platforms, spammers have come up with various techniques to lure users into accessing malicious links. This is done by generating spam on the comment section of various social media networks. In this paper, we've taken YouTube comments as the dataset and performed spam YouTube comment detection. The current methods to stop spammers include using tools such as Google Safe Browsing which help to detect and also block irrelevant spam on YouTube. Although these tools help in blocking harmful links, it fails to secure the users in real-time scenarios. Thus, many different approaches have been applied to form an environment that is spam free. A few of them are solely supported user-based options whereas others are based on YouTube content. We have assessed our answer with four completely different algorithms based on machine learning, namely – Logistic Regression, Decision Trees Classifier, Random Forest, Ada Boost Classifier and Support Vector Machine . With Logistic Regression, an accuracy of 95.40% is possible, surpassing the current solution by roughly 18%.** **Keywords— logistic regression, spam detection, YouTube, machine learning, support vector machine.**

## I. Introduction

Spam is the irrelevant data which needs to be filtered out in order to avoid wasting the users' time as well as to maintain the quality and credibility of the platform. Whereas ham is the content which is important and required for various purposes.

Online social networks such as Face-book and YouTube have become increasingly popular outlets over the past few years and are an integral part of the everyday life of people. People spend a lot of time posting their messages on micro blogs websites, sharing their thoughts and making friends around the world. Such sites are also attracting a large number of users and spammers to transmit their messages to the world due to this rising phenomenon. YouTube is ranked by teenagers as the most successful social network. Yet YouTube's explosive growth encourages more unsolicited activities on this site as well. Currently 200 million users create up to four hundred million new YouTube videos every day. This rapid expansion of YouTube platform encourages more spammers to create spam on YouTube platform that contains malicious links to external websites that contain malware downloads, phishing, drug sales, or scams.

One of YouTube's most well-known features is its comment system, in which different users may comment on uploaded videos. YouTube users have channels of their own. YouTube lets channels upload, share, rate, post comments on videos and subscribe to different channels. This remarkable feature allows users to communicate with each other and express their emotions, thoughts, etc. on the content. But, this has also made it possible for malicious users to post irrelevant or self-promotional content often called as spam. Spam comments are mostly entirely unrelated to the video and are usually created by automated bots posing as legitimate users. These bots have the ability to execute spam campaigns (coordinated large scale posting of spam comments.)

In our paper, we classify the spam comment of the YouTube. We use the machine learning concept which is a subset of artificial intelligence. Four types of machine learning modules are available, namely supervised learning, semi-supervised learning, unsupervised learning, and strengthening. Machine learning is the method of extraction, transforming, loading and predicting the meaningful information from huge data to extract some patterns and also transform it into understandable structure for further use. Prediction and classification and are the two kinds of data analysis techniques that are used by mine models that identify the most relevant data classes and predict future data trends.

The purpose of our paper is to briefly introduce machine learning techniques as well as to outline the prediction technique. As it is far more superior to the conventional data analytics practices used in agricultural research, machine learning has the capability to create new openings to explore and also to improve the accuracy of prediction. The main objective of this research work is to provide a prediction so that it can perform descriptive analytics on prediction of spam comments in YouTube. In this study we are using supervised learning. This method depends upon a huge number of labeled datasets. We propose the classification algorithm called logistic regression for classifying spam comments.

## II. SYSTEM ANALYSIS

### *A.Existing System.*

YouTube has developed filters to combat spam. Since releasing them, they have seen a 40 percent decrease in vital spam metrics[1]. The rules could be updated in a centralized manner by the maintainer of the spam filtering tool. One of the weak aspects of this program is that it doesn't protect a victim from new spam, i.e. it isn't an effective tool for detecting YouTube spam in real time. 90 per cent of users can visit a new spam link before the blacklist is blocked [3]. To get past the more advanced filters that rely on word frequencies, spammers append a large quantity of "usual words" to the end of a message.

978-1-7281-7223-1/20/$31.00 ©2020 IEEE

### *B.Proposed System.*

We prepare our dataset by collecting 400,000 videos of videos corresponding. To get information from YouTube text, we extract certain terms that may be strong indicators for classifying YouTube into one of the classes: spam or non-spam.

Feature selection is done to provide a collection of terms that can be used for classification. To do this, Bag-of-Word Model's Information Gain is used[2]. Such terms are combined here to form top-30words that we use in our feature set.

After the classification process, the data is split into the training set and testing set and the best machine learning algorithm that would provide optimum accuracy in classification is selected and applied.

Taking these top 30 terms into account would help us accurately identify the comments for each class[4]. Frequent pattern mining of YouTube's text could also be the crucial factor for identifying real-time YouTube spam.

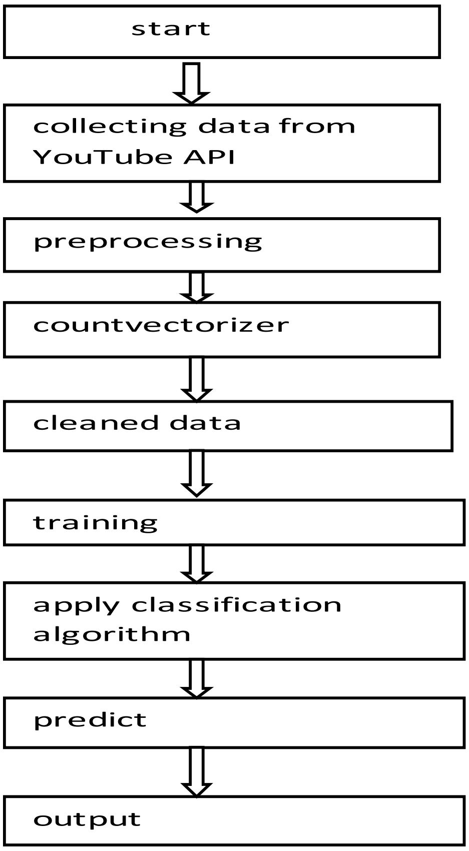


Fig 2.1 Block Diagram for the Proposed Work

## III. Module description

### *A.Module 1:Feature Selection*

The dataset is cleaned and processed. Cleaning the dataset involves removing the null values, meaningless words and extracting only the required set of meaningful words. These words are given scores based on their occurrence. The advantage of using these terms in the feature-set based on their entropy score is that we have been able to minimize uncertainty in the outcome of the prediction as these terms have a different impact on spam and non-spam YouTube frequency counting[5].

### *B.Module 2: Feature Extraction and Feature Engineering*

Feature selections in the attribute value mining feature are assisted by Oracle Data Mining. Attribute importance is a supervised function that ranks attributes in accordance with their significance when predicting a target. Count Vectorizer is applied to transform a set of texts into a token count matrix . This undergoes the following process

- Preprocessing
- N-grams
- Analyzer
- Mixed
- Max_features
- Vocabulary
- Binary

### *C.Module 3:Model Building*

Supervised learning is the process of interpreting a function from training data(labeled). We can find the most efficient model parameters to predict labels that are unknown on the test set consisting of other objects, by fitting to the labeled training set. If the label is a real number, regression is applied. If the label originates from a small list of unordered values, then classification is used[6].Classification with different learning methodology:

- SVM(Support Vector Machine)
- K-NN(K Nearest- Neighbour)
- DT(Decision Tree)
- NB(Naïve Bayes)

### *D.Module 4: Prediction*

The above modules collectively yield a prediction as to which category(spam or ham) the comment belongs to.

## IV. METHODOLOGY

### LOGISTIC REGRESSION:

Using logistic regression it is possible to determine the likelihood of a result that has only two values. [9]. This prediction is based on the use of one or more (categorical and numeric) predictors. For two reasons, a linear regression alone is insufficient to predict the value of a binary variable:

• A linear regression predicts values that lie beyond the permissible range (e.g. probability prediction for values that are beyond range 0 to 1).

• Since these binary experiments can have only one out of two values for every experiment performed, the residuals will not be distributed normally along the line that is predicted.[7].

A logistic regression, on the contrary, generates a logistic curve, which is restricted to the values lying between 0 to 1.

Logistic regression is similar to linear regression in some ways, but instead of using probability, the curve is built using the natural logarithm of target variable "odds." In addition, the predictors are not expected to be normally distributed in each group or have equal variance.

The constant (b0) pushes the curve from left to right in the logistic regression, and the steepness of the curve is determined from the slope. The equation of logistic regression is written with respect to odds -ratio by a simple transformations.

$$p/(1\text{-}p) = \exp(b_0 + b_1x)$$

The equation is written based on log-odds (logit),which is a linear function of the predictors, after taking natural log on both sides. The amount by which the logit (log-odds) changes in x with a change in one unit is given by the coefficient (b1).

$$\ln(p/(1\text{-}p)) = b_0 + b_1x$$

As previously stated, logistic regression is cable of handling any number of categorical and/or numerical variables [8].

$$p=1/(1+e^{-(b0+b1x1+b2x2+\ldots.+bpxp)})$$

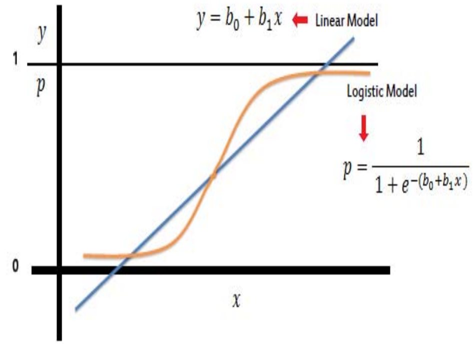
A variety of comparisons exist between logistic regression and linear regression. Like in linear regression, logistic regression makes use of maximum likelihood estimation (MLE) in order to produce the ideal coefficients which link the predictors and the target[10]. After that the initial function is then calculated and this process is continued until LL (Log Likelihood) substantially stops changing

$$\beta^1 = \beta^0 + [X^TWX]^{-1}.X^T(y\text{-}\mu)$$

β is a vector of logistic regression coefficients.

W is a square matrix of order N with elements $n_i\pi_i(1\text{-}\pi_i)$ on the diagonal zeros everywhere else.

μ is a vector od length N with elements $\mu_i = n_i\pi_i$



## VI . CONCLUSION AND RESULT

The findings indicated most of the analyzed methods of classification are for filtering spam comments on YouTube. Most of them were able to provide small or no blocked ham rates and it was aslo possible to achieve accuracy rates that were more than 90 per cent. Our solution has also been checked with real-time YouTube detection and has outperformed the current approach by 18 %.

| | |
|---|---|
| **Logistic Regression** | 0.9540 |
| **Support Vector machine** | 0.5051 |
| **Random Forest** | 0.8469 |
| **AdaBoostClassifier** | 0.9438 |
| **DecisionTreeClassifier** | 0.9438 |

## VII. REFERENCES

[1] J. Zhan, P. Chopade and M. Bikdash. Node attributes and edge structure for large-scale big data network analytics and community detection. In International Symposium on Technologies for Homeland Security (HST), pages 1–8, 2015.

[2] F. Checconi, X. Que, F. Petrini, and J. A. Gunnels. Scalable community detection with the louvain algorithm. In Parallel and Distributed Processing Symposium (IPDPS), pages 28–37, 2015.

[3] Z. Wang, P. Cui and Z. Su. What videos are similar with you?:Learning a common attributed representation for video recommendation. In ACM International Conference on Multimedia (MM),pages 597–606, 2014.

[4] M. Halappanavar, H. Lu, A. Kalyanaraman, and S. Choudhury. Parallel heuristics for scalable community detection. In International Parallel & Distributed Processing Symposium Workshops (IPDPSW), pages 1374–1385, 2014.

[5] N. Sverdlik and S. Oreg. Source personality and persuasiveness:Big five predispositions to being persuasive and the role of message involvement. Journal of Personality, 82(3):250–264, 2014.

[6] A. Kanavos, E. Kafeza, C. Makris, and D. Chiu. Identifying personality-based communities in social networks. In Legal and Social Aspects in Web Modeling (Keynote Speech) (LSAWM), 2013.

[7] D. Gao, C. Chen, W. Li, and Y. Hou. Inferring topic-dependent influence roles of twitter users. In SIGIR, pages 1203–1206, 2014.

[8] E. A. Stepanov, F. Alam and G. Riccardi. Personality traits recognition on social network - facebook. Computational Personality Recognition, 2013.

[9] L. Rossi and F. Celli. The role of emotional stability in twitter conversations. In Semantic Analysis in Social Media, pages 10–17, 2012.

[10] J.Scott.. Social Network Analysis. Sage, 2012

978-1-7281-7223-1/20/$31.00 ©2020 IEEE