Analysis and detection of spam comments on social networking platforms like youtube using machine learning

Sonali Patil [1], Reshma Jadhav[2], Ishwari Mote[3] ,Dakshata Ramteke[4]

Department of Computer Engineering, JSPM Bhivarabai Sawant Institute of Technology and Research, Wagholi ,Pune,

Savitribai Phule Pune University, Pune, India

Abstract— **The profitability promoted by Google in its brand-new video distribution platform YouTube has attracted an increasing number of users. However, such success has also attracted malicious users, which aim to self-promote their videos or disseminate viruses and malware. As we know that YouTube offers limited tools for comment moderation, so spam increases very rapidly and that's why comment section of owners disabled. It is very difficult to established classification methods for automatic spam filtering since the messages are very short and often rife with slangs, symbols, and abbreviations. In this work, we have evaluated several top-performance classification techniques for such purpose. The statistical analysis of results indicates that, with 99.9% of confidence level, decision trees, logistic regression, Bernoulli Naive Bayes, random forests, linear and Gaussian SVMs are statistically equivalent. Therefore, it is very important to find a way to detect these videos and report them before they are viewed by innocent users.**

Keywords— **Spam Filtering, SVM , Naïve Bayes ,Machine Learning, YouTube**

I. INTRODUCTION

The popularization of wideband around the world has boosted the number of Internet users. With faster connections, video host and sharing services became popular among users. According to a press release of Sandvine1, a company focused on standards-compliant network policy control, around 55% of downstream traffic from the United States is due to video platforms like Netflix and YouTube. The availability of resources through Internet and the wideband connections allowed the appearance of sophisticated new platforms. In along these lines, YouTube is a renowned video content distribution stage with informal community highlights, for example, support for presenting content remarks on giving cooperation between channel proprietors and watchers or endorsers. The success of YouTube is expressed through recent statistics rumored by Google2: the platform has quite one billion users, 300 hours of video are uploaded every minute and it generates billions of views every day or every minute. Around an hour of a creator's views come back from outside their home country and 1/2 YouTube views are on mobile devices.

Recently, YouTube has adopted a monetization system to reward producers, stimulating them to make high-quality original content and increasing the amount of visualizations. After the deployment of this system, the platform was flooded by undesired content, usually of low-quality information known as spam. Among completely different reasonably unsought 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. The spam found on YouTube is directly associated with the engaging

profit offered by the substantiation system. According to a press release by Google, more than a million advertisers are using Google ad platforms, the mobile revenue on YouTube is up 100% year over year and the number of hours folks are look on YouTube every month is up five hundredth year over year. At the same time, according to Negate, a computer security company, just in the first half of 2013, the volume of social spam increased by 355%. For each spam found on any social network, different two hundred spams square measure found on Facebook and YouTube. The problem became therefore vital that it actuated users to make a petition in 2012, in which they ask YouTube to provide tools to deal with undesired content.

In 2013, the YouTube official blog reported efforts to deal with undesired comments through recognition of malicious links, ASCII art detection and display changes to long comments. However, many users are still not satisfied with such solutions. In fact, in 2014, the user "PewDiePie", owner of the most subscribed channel on YouTube (nearly 40 million subscribers), disabled comments on his videos, claiming most of the comments are mainly spam and there's no tool to wear down them.The problem caused by social spam began to be seriously mentioned in 2010, but an earlier work is dated from 2005.

However, unsought comments on YouTube still hurt the platform's community, evidencing such drawback needs attention and analysis. Established techniques for automatic spam filtering have their performance degraded when dealing with YouTube's comments. It is mainly due to the fact that such messages are usually very short and rife with idioms, slangs, symbols, emoticons, and abbreviations which make even tokenization a challenging task

II. LITERATURE SURVEY

Spam is nothing but undesired content with low-quality information. They are commonly found as images, texts or videos, hindering visualization of interesting things. There are many kinds of research related to spam in literature, such as web spam, blog spam, e-mail spam, and SMS spam. In social networking sites, undesired content is known as social spam. Blog comment spam is the most similar scenario.

However, the commonly known strategy to detect a blog spam comment usually is to find the best representation of language model in post-publication, using that representation to filter less related content to its original subject. Such a strategy cannot be applied on YouTube since comments are related to video content with small or no textual description, therefore language models cannot be properly mapped from the original publication. YouTube also faces malicious users that publish low-quality content videos, which is known as video spam. There are some studies in the literature to find efficient ways to handle this activity through classification methods and feature extraction from metadata, such as title, description and popularity numbers. The next common alternative is automatic blocking spammers – users that disseminate spam. However, unlike spam disseminated in other social networks and email, the spam posted on YouTube is not usually created by bots, but posted by real users aiming self-promotion on popular videos.

Therefore, such messages are more difficult to recognize due to its similarity to legitimate messages. Automatic spam filtering is useful in other tasks as well. Reported notable improvement of performance in the opinion detection task, when spam was removed before training a classifier. As noted by Bratko et al., the spam filtering task slightly differs from similar text categorization problems. They claim undesired messages have chronological order and their characteristics may change according to that. It also explains that cross-validation is not recommended, because earlier samples should be used to train the methods, while newer ones should be used to test them. Furthermore, in spam filtering, errors associated with each class should be considered differently, because a blocked legitimate message is worse than an unblocked spam.

III. SYSTEM OVERVIEW

We conceive to observe spam comments by applying standard machine learning algorithms Naive mathematician together with sure custom heuristics like N-Grams that have tested to be effective in police work and later on combating spam comments. we've got collected and created 5 databases composed by real, public and non-encoded knowledge directly extracted from YouTube through its API7. we've got designated 5 of the 10 most viewed YouTube videos. every sample represents a text comment denote within the comments section of every designated video. No preprocessing technique was performed. later on, every sample was manually labeled as spam or legitimate (ham), employing a cooperative tagging tool developed for this purpose, known as Labeling. The samples have associated information info, such as the author's name and publication date, that are preserved

IV. CLASSIFICATION MODELS

A. *Naïve Bayes*

In machine learning, naïve mathematician classifiers **are** a family **of easy** "probabilistic classifiers**"** supported applying theorem **with** study (naïve) independence assumptions between the **options**.They are among the simplest Bayesian network models. These classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.

B. Decision Trees

C4.5 is one of the decision tree family algorithm that generates trees from training data using information entropy the decision tree C 4.5 is an extension of the ID3 algorithm. At each node, the algorithm selects the attribute of the data bt splitting the samples into a set of subsets using the information gain criteria, the highest attribute value will make the decision. Finally, the processes are repeated on the smallest sub lists.

C.SVM

The objective of the support vector machine algorithm is to find a hyper plane in an N-dimensional space that distinctly classifies the data points. It is **extremely most well-liked** by **several because it** produces **vital** accuracy with less computation power. Support Vector Machine abbreviated as SVM **may be** used for **each** regression and classification tasks. But, it is widely used in classification objectives.

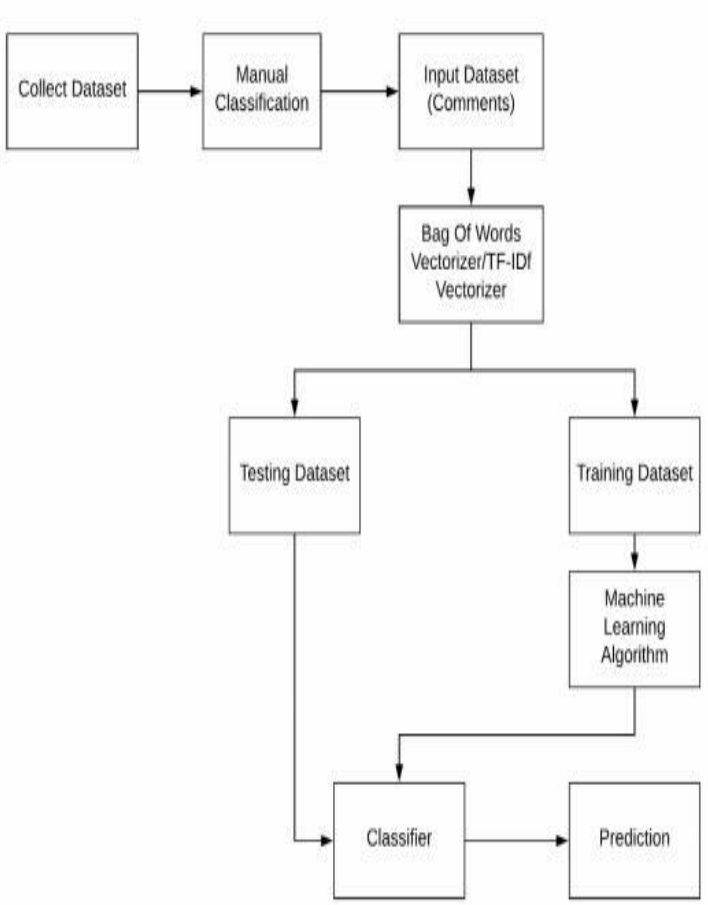

Fig. Proposed System

VI. CONCLUSIONS

Social media networks have become extremely popular and this creates the opportunity for the malicious user to publish unwanted comments. This study has introduced the feature set to be used in detecting video spammers that exist in the YouTube media. The features will be constructed based on the features obtained from the user profile and the content that they shared. Based on the undertaken experiments, it is expected that existing classifiers that were widely used in the data mining community could utilize the features in detecting comment spammers.

VII. REFERENCES

[1] Sreekanth Madisetty, Maunendra Sankar Desarkar"A Neural Network-Based Ensemble Approach for Spam Detection in Twitter",IEEE,2018

[2] Shivangi Gheewala, Rakesh Patel" Machine Learning Based Twitter Spam Account Detection: A Review", IEEE,2018.

[3] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou and G. Min, "Statistical Features- Based Real-Time Detection of Drifted Twitter Spam", IEEE Transactions, April 2017, pp.914-925.

[4] Ala' M. Al-Zoubi∗, Ja'far Alqatawna, Hossam Faris"Spam Profile Detection in Social Networks Based on Public Features",IEEE,2017

[5] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Surveyof new approaches and comparative study," Comput. Secur., vol. 76,pp. 265–284, Jul. 2017.

[6] Chen Liu, Genying Wang"Analysis and Detection of Spam Accounts in Social Networks",IEEE,2016

[7] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan and S. A. Razak, "Malicious accounts: Dark of the social networks", Elsevier, 2017, pp. 41-67

[8] Miss.Shukla Twinkle Kailas,Prof.D.B.K shirsagar,"Design of machine learning approach for spam tweet detection",IEEE,2016

[9] Chao chen, Jun Zhang ,Yi Xie and Yang Xiang ."A performance evaluation of machine learning based streaming spam tweets detection",in IEEE transaction on computational social system,2015,Vol-2 No-3.

[10] A. Gupta and R. Kaushal, "Improving Spam Detection in Online Social Networks", IEEE, 2015.

[11] M. Verma, Divya, S. Sofat, "Techniques to Detect Spammers in Twitter – A Survey", International Journal of Computer Applications, January 2014, Vol. 85, No. 10, pp. 27-32.

[12] Ziyan Zhou ,Lei Sun,"Network based spam filter on tweeter",2014