N-GRAMS ASSISTED YOUTUBE SPAM COMMENTDETECTION

Mrs. SRILATHA PULI,
Assistant Professor in Sreyas Institute of Engineering and Technology, JNTUH, India,
srilatha.puli@sreyas.ac.in

TEAM MEMBERS:
Yella Sowmya Reddy
CSE, Sreyas institute of engineering and Technology,Telangana, India
yellasowmya2001@gmail.com
Thirupathi Nithin
CSE, Sreyas institute of engineering and Technology, Telangana,India
thirupathinithin@gmail.com
Vemuganti Meghana
CSE, Sreyas institute of engineering and Technology, Telangana, India
vemugantimeghana2000@gmail.com
Pamulaparthi Sindhu
CSE,Sreyas institute of engineering and Technology, Telangana, India
pamulaparthisindhu@gmail.com

ABSTRACT -

This paper proposes a novel technique for detecting intrusive comments or spam on the video sharing website - YouTube. We describe spam comments as those which have persuasive intent or those who deem to be contextually irrelevant for a given video. The prospects of monetization through advertising on popular social media channels over the years have attracted an increasingly large number of users. This has in turn led to the growth of the malicious users who have begun to build up automated bots, which are proficient of large scale orchestrated distribution of spam messages across multiple channels simultaneously. These intrusive comments damage the fame of a channel and also the experience of regular users. YouTube themselves have embarked upon this issue with some finite methods which blocks unsolicited comments that consists of links. Those methods have proven to be exceptionally unproductive as spammers have found strategies to bypass such heuristics. Standard machine learning classification algorithms are operative but there is always a possibility for better accuracy with novel methods. In this work, we aim to identify such comments by implementing conventional machine learning algorithms such as Random Forest, Support Vector Machine, and Naive Bayes along with certain custom heuristics such as Count Vectorizer which have proven to be very effective in detecting and subsequently combating spam commentary.

***KEYWORDS** – Text Classification, YouTube Spam Detection, MLPs, SPM.*

1. INTRODUCTION

YouTube, the world’s largest video sharing site, was founded in 2005 and acquired by Google in 2006. YouTube has grown tremendously as a video content platform, with the recent shift in online content to video. Therefore, in the video sharing space YouTube has rised as a leading adversary. Users of Youtube are known as channels, and it allows channels to upload, rate, share, add to favorites, report, comment on videos, and subscribe to other users. One of the most utilised features of Youtube is its commenting system where users can comment on videos uploaded to other channels. Commenting on the video allows the users to interact with each other and share their feelings, ideas etc. Yet, this has also turned as a chance for malicious users to share divulgatory content also known as spam. YouTube spam comments are very important because they are the ones which bring an impression in viewers on the video because most of the viewers go through the comments even before watching the video.Spam comments are often wholly irrelevant to the given video and are usually generated by automated bots disguised as a user. The ability of such bots to accomplish spam campaigns - large scale orchestrated posting of malignant comments has been probed in .

In this way the comments add a value to the video. But there are a lot of comments which are completely irrelevant and also contain promotional content which badly effects the video. So we have a strong dedication to detect such comments and help the YouTube community to not face any problems for the things they haven't committed. At present, more than 400 hours of video are being uploaded and 4.5 million videos are viewed every minute on YouTube. It is easy for users to watch and upload videos without any restrictions. This great accessibility has increased the number of personal media, and some of them have become online influencers. YouTube creators can monetize if they have more than 1,000 subscribers and 4,000 hours of watch time for the last 12 months.

Accordingly, spam comments are being created to promote their channels or videos in popular videos. Some creators closed the comment function due to aggression such as political comments, abusive speech, or derogatory comments not related to their videos. YouTube has its own spam filtering system, though there are still spam comments that are not being caught. In this paper, we review related studies on YouTube spam. In previous studies, various machine learning techniques were applied to each dataset to detect spam comments and compare their performance. Therefore, in this paper, we propose an ensemble machine learning method that combines the results of several models to produce the final result. Using machine learning method auguring of the spam comments present in the comment section, it is also known as variant of artificial intelligence.
Supervised learning approach depends on a very large number of classified data. Logistic Regression a proposed classification algorithm is used in order to predict the spam comment. The purpose of paper is to introduce briefly the techniques of machine learning and to outline the prediction. Recently as of 2017, Youtube has faced increasing criticism about its inability to moderate uploaded content. A large digerati of Youtube has children who are often proned to malicious and deleterious material in the form of comments. Youtube has attempted to

combat this by blocking all comments containing links. However, this has led to spammers resorting to more creative techniques such as introducing whitespace characters between links. The pressure to solve the problem using creative solutions has been increasing. We believe that the increase in computing power in the recent years have paved the way for applying conventional Machine Learning algorithms to solve such problems. In this paper, we attempt to identify the algorithms and apply specific heuristics that can accurately detect spam.

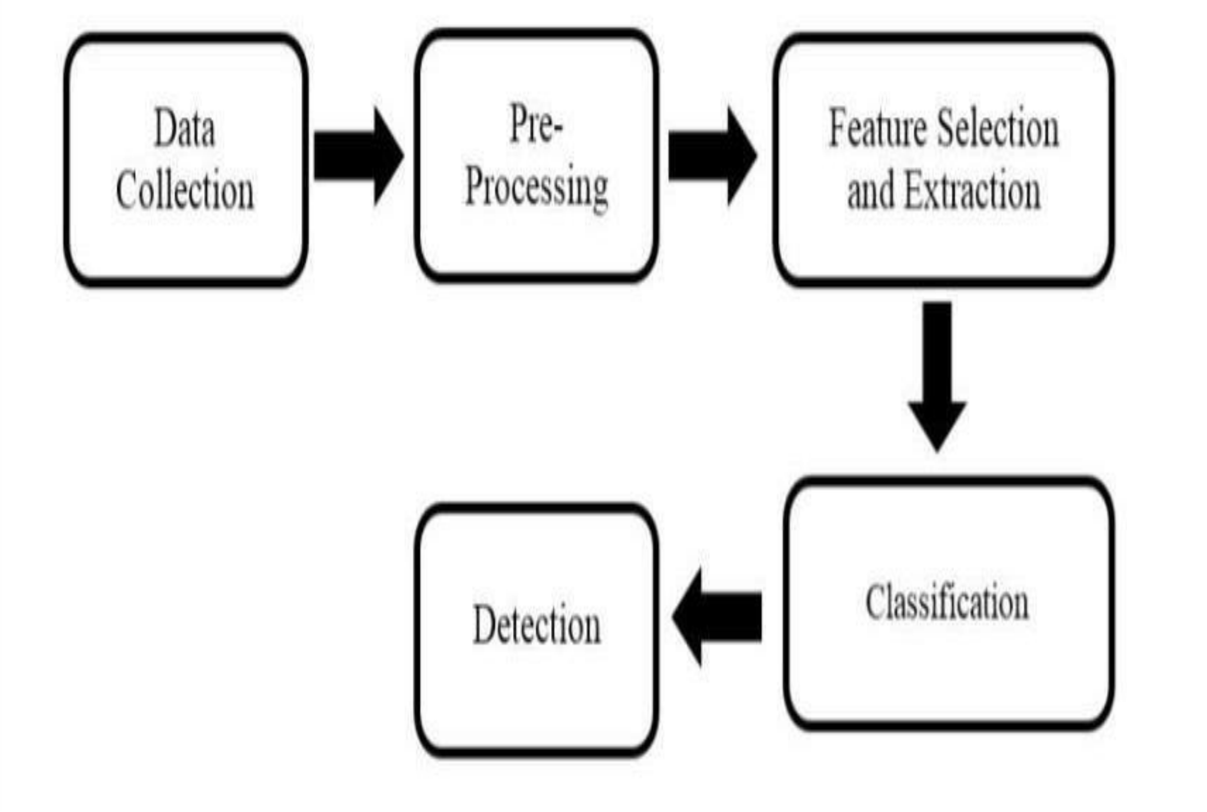


Fig 1Work flow

2. LITERATURE SURVEY

Spam attack had been world widely distributed. In addition to social media such as Facebook, Twitter, YouTube, SMS and e-mails are facing spam attacks. In case of email spam is defined as undesirable emails sent by various users daily. According to Tran et al, email spam brings the meaning of unsolicited bulk emails received by users. Unsolicited or unwanted message received on a mobile phone is known as Short Message Service (SMS) spam.

In web spamming, spam refers to the meaning of a proposed activity to mislead search engine into ranking some page higher than they deserve. Spam comments had been identified as those which has irrelevant commercial content that is inappropriate to the discussion with unwanted requests. In addition, video spam also been defined by Yusof and Sadoon asunrelated, unwanted content compared to its video’s title. Spam is usually interrelated to undesired content with futile information. They are commonly found as images, texts or videos, hindering visualization of exciting content. There are many researches related to spam in literature, such as web spam, blog spam e-mail spam and SMS spam undesired messages are known as social spam.

A runtime SD scheme called BARS: Blacklist-Assisted Runtime SD which assembled a database of Spam URLs against which URL of every new post was scrutinized to determine if the post was spam or ham which was anticipated by Enhua Tan et al. Effectiveness of detection also increased due to the clustering of User IDS based on shared URLs. Anyhow, the efficiency of this approach is a result of how successfully the boycotted URL list is promoted.

Alex Kantchelian et al. developed a spam detection technique which can compute futile and redundant features in blogs, making purposeful stories handier to the perpetual stakeholders. They advised extension of their work to broaden the annotation of spam such as URLs, short message removal, etc. moreover inclusion antagonist awareness, online deployment to facilitate prediction of futuristic comments and so on.

Seungwoo Choi et al. tested their algorithm on Ted-Talks videos to identify the comment offering broadcasting and information about the video contents. Nevertheless, the proposed method was found scanty in figuring out the feelings and opinions expressed in platforms such as YouTube.

Igor Santos et al. enforced the concept of anomaly detection where in the divergence from genuine emails was used as a metric to organize emails as spam or ham.Precision was achieved owing to the limited training sets as seen in labelling based systems.

M. McCord et al. furnished machine learning algorithms which were trained with content and user-centred features to identify spammers. They experimented their algorithms with twitter data and discovered that the Random Forest Classifier provided the best results.

Web Spammers goal is to increase their revenue by redirecting users to visit their sites, or to spread malicious content through the installation of malevolent software. Writer Shekoofeh Ghiam et al. studied the different web-spam practices and relevant exposure ways. Writer Wojciech Indyk et al. guided Map reduce Algorithm to effectively locate spam masses is their research

3.SYSTEM DESIGN AND PROPOSED METHODS

Detection framework has steps such as Data Collection, Data Pre-processing, Feature Extraction, Classification and Comparison of Results. This framework is chosen from because it can provide the result with good accuracy.This framework also provides the phase to compare the resultsof Naive- Bayes and Random Forest technique**.** In this model, count vectorizer is used for extracting features from a given dataset and design model by generating tests and training sets from given data then the Random Forsest and Naive Bayes classifiers are applied for clustering and the test and training set is given as input. Based on this data, the given comment is tested, whether it is spam or ham.

TECHNOLOGY DESCRIPTION

Jupyter Notebook

Jupyter is a free, open-source, interactive web tool known as a computational notebook, used to combine software code, computational output, explanatory text and multimedia resources in a single document by inquisitors. It provides us with an user friendly, interactive data science environment across various programming languages that doesn't only work as an IDE(Integrated Development Environment), but also as apresentation or education tool.

Visual Studio Code

Visual Studio Code is a source code editor redefined and optimized for building and debugging modern web and cloud applications. It is a streamlined code editor with support for development operations such as debugging, task running, and version control. It aims to provide just the tools a developer needs for a quick code-build-debug cycle and leave more complex workflows to fuller featured IDEs, like Visual Studio IDE. The app permits you to install and run visual studio code (next vscode) on your android device.IMPORTANT, this app is not vscode, but it permits you to INSTALL vscode in just one command. One thing that all the users have to do is download termux, insert a command from my app. After executing the code in termux, vscode can run with just a single command.

SYSTEM ARCHITECTURE

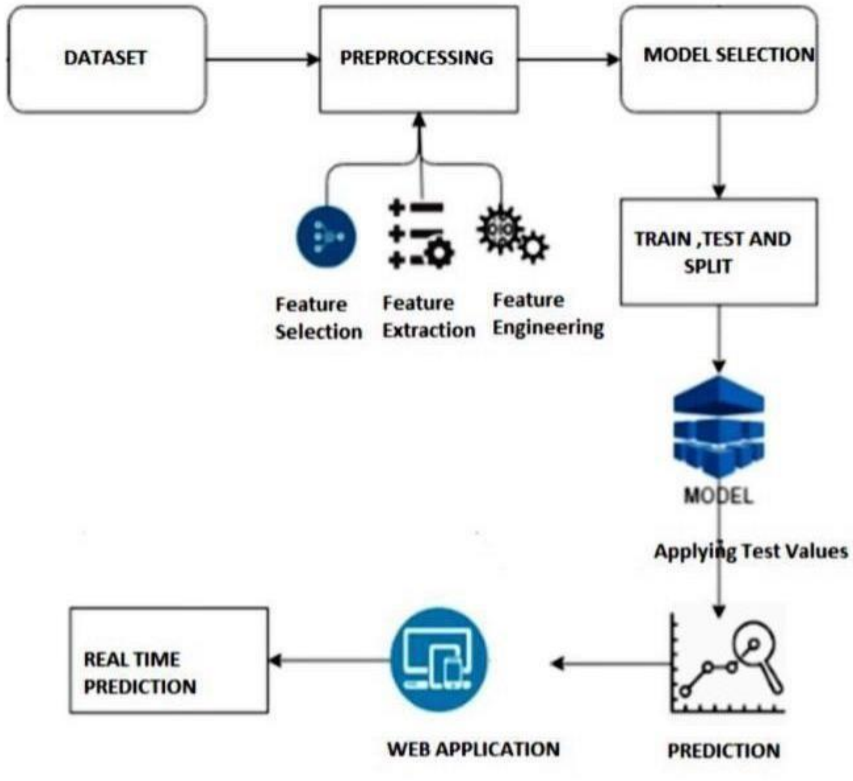


Fig 2 System architecture

4 IMPLEMENTATIONS

Data Collection

The YouTube Spam Collection Data Set Collect from UCI Repositories. It has fivedatasets composed of 1,956 real and non-encoded messages that were labelled as

legitimate (ham) or spam. Each sample constitutes a text comment posted in the comments section of each selected video. No pre-processing technique was performed. Subsequently, each sample was manually labelled as spam or legitimate (ham).The samples have associated a piece of metadata publication date, which have been preserved.

Qualitative data collection examines several factors to provide a depth of understanding to raw data. While qualitative methods comprise the collection, analysis, and management of data, instead of tallying responses or collecting numeric data, this method aims to evaluate factors like the thoughts and feelings of research members. Qualitative data collection methods surpass recording events to create context.

Data Preprocessing

The data-set used here is split into two parts: 80% for the training and the remaining 20% for the testing. In any text mining problem, the first step is text cleaning, where we remove those words from the document which do not contribute to the information that we want to extract. YouTube Comments may contain a lot of undesirable characters like punctuation marks, stop words, digits, etc which may not be helpful in SD. After cleaning the text, we fed our dataset into a Term Frequency-Inverse document frequency (TF-IDF) vectorizer which converts words into numerical features (numpy arrays) for training and testing.

Data Cleaning

Data cleaning is the process of adjusting or removing incorrect, corrupted, wrongly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many chances for data to be duplicated or mislabeled. It is also important because the data quality will be improved and doing so, increases overall productivity. When you clean your data, all outdated or incorrect information no longer exists — leaving you with the highest quality information.

Feature Selection

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

Feature Extraction and Feature Engineering

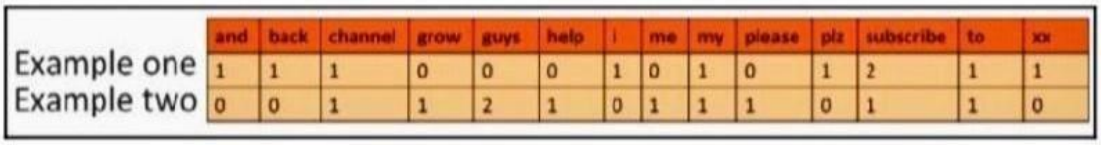
BoW

The BoW model does exactly we want, that is to convert the phrases or sentences and count the number of times the words that come out similar. In the world of computer science, a bag refers to a data structure that keeps track of objects such as an array or list does, but in such cases the order is unimportant and if an object appears inure than once, we just keep track of the count rather we keep repeating them.

Consider the following sentences and look around for what makes the first pair of phrases similar to the second pair:

As you can notice, the first phrase from the diagram, has a bag of words that has words like"channel", with one event, "plz", with one event, "subscribe", two events, and soon. Then, we would collect all these counts in a vector, where one vector per phrase or sentence or document, depending on what you are working with. Anew, the sequence in which the words appeared originally don’t bother.

In later stages, We make a larger vector with all the unique words across both phrases, we get a proper matrix representation. With each row representing a different phrase, notice the use of 0 to indicate that a phrase doesn't have a word:



	and	back	channel	grow	guys	help	i	me	my	please	plz	subscribe	to	xx
Example one	1	1	1	0	0	0	1	0	1	0	1	2	1	1
Example two	0	0	1	1	2	1	0	1	1	1	0	1	1	0

Fig 3 feature extraction

If you want to have a bag of words with lots of phrases, documents, or we would need to gather all the unique words that occur across all the examples and create a huge matrix, N x M, where N represents the number of examples and M represents the number of occurrences

Additionally, there are some points which we need to look after before preparing a bag of words model

- Lowercase every word • Drop punctuation • Drop very common words (stop words) Remove plurals (for example, bunnies => bunny) • Perform Stemming(for example, reader => read, reading
= read)

N-grams

N-grams is used to improve the accuracy. It is apportioned with single word but when there are two mutual words the complete meaning will be changed. So, the difference of accuracy is better appeared when text is split into token of two or more words rather than being a single word.

Analyzer

“Whether the feature should be built with word or character n-grams. Option ‘char_wb’ generates character n-grams only from text inside word boundaries; n-grams at the edges of words are padded with space.”

Vocabulary

“Either a Mapping where keys are phrases and values are indices in the feature matrix, or repeatable over terms. If not provided, a vocabulary is determined from the input documents.

Indices in the mapping should be unique and should not have any gap between 0 and the largest index.

Binary

If True, all non-zero counts are assigned as 1. This is advantageous for discrete probabilistic modelsthat model binary events rather than integer counts.”

Model Building

After Preprocessing there has to be a method of constructing a version to retain the abilities of the function of the project in conformity to the labeled model, which is fabricated as per the Supervised set of regulations.

Classifier Techniques

After Feature Extraction the transformed dataset is fed into classifier techniques Multilayer Perceptrons(MLPs), Support Vector Machine(SVM), Naïve Bayes(NB), Random Forest(RF), Decision Tree(DT), Logistic Regression(LG), and k-Nearest Neighbor(kNN) pipelines respectively**.**

Multilayer Perceptrons (MLPs)

An MLPs is a feed-forward network with one input layer, one output layer, and at least one hidden layer [33]. To classify data that is not linear in nature, it uses non-linear activation functions, mainly hyperbolic tangent or logistic function [34]. The network is fully connected, which means that every node in the current layer is connected to each node in the next layer. This architecture with the hidden layer forms the basis of deep learning architecture which has at least three hidden layers. Multilayer Perceptrons are used for speech recognition and translations.

Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier that can be used for both classification and regression problems. 2nd National Conference on Emerging Technologies, December 6-7, 2016, University of South Asia, Lahore, Pakistan 6 The goal of SVM is to identify an optimal separating hyperplane which maximizes the margin between different classes of the learning set. By the way, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which classifies new examples to build the largest possible distance to decrease an upper bound. Supports Vectors are just the coordinates of data points that are closest to the optimal separating hyperplane provide the most useful information for SVM categorization. In addition, an appropriate kernel function is used to transform the data into a high-dimension to use linear Discriminate functions.

K-Nearest Neighbors (kNN)

The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it holds a bunch of labeled points and utilize them to gain knowledge how to label other points. To label a new point, it looks at the labeled points that are nearer to that new point (those are its closest neighbors), and has those bystanders vote, so whatever label the maximum neighbors have is the label for the new point (the “k” is the number of neighbors it checks).

Logistic Regression

It is a statistical way of analyzing a data set in which there are one or more independent variables that decide an outcome. The outcome is computed with a dichotomous variable (in which there are only two possible outcomes). The aim of logistic regression is to find the best fitting model to define the relation between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a group of independent variables.

Naïve Bayes (NB)

It is a simple classification technique for constructing classifiers based on the Bayesian network with a belief of independence among predictors. In basic terms, a Naive Bayes classifier assumes that the presence of a specific feature in a class is not related to the presence of any other feature. Even if these features are mutually dependent or depend upon the presence of the other features, all of these properties individually contribute to the probability. NB is based on probability estimations, called posterior probability.

Random Forest

Random forests or random decision forests are as a whole learning method for classification, regression, and other works, that operate by building a multitude of decision trees at training time and manufacturing the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests right for decision trees’ habit of over fitting to their training set.

Decision Tree

A decision tree is a tree whose internal nodes can be considered as tests (on input data patterns) and whose leaf nodes can be considered as categories (of these patterns). These tests are processed down through the tree to get the correct output to the input pattern. Decision Tree algorithms can be practiced and used in various different fields. It can be used as an alternative for statistical procedures to find data, to extract text, to find missing data in a class, to improve search engines and it also finds various applications in medical fieldss. Many Decision tree algorithms have been devised. They possess different accuracy and cost-effectiveness. It is also crucial for us to know which algorithm is best to use. The ID3 algorithm is one of the oldest Decision tree algorithms. It is very useful while making simple decision trees but as the complications increase its accuracy to make good Decision trees decreases. Therefore IDA (intelligent decision tree algorithm) and C4.5 algorithms have been devised.

5 TESTING

TESTING DEFINATION:

- Software testing is the process of evaluating and verifying that a software product or application does what it is supposed to do.

- The benefits of testing constitute preventing bugs, decreasing development costs and improving performance.

TESTING AND TEST CASES:

- Software testing is the process of evaluating and verifying that a software product or application does what it is supposed to do.
- The benefits of testing constitute preventing bugs, decreasing developmentcosts and improving performance.

TYPES OF TESTING:

White Box Testing

In white-box testing, the developer will inspect code line by line before transferring it to the testing teamor the concerned test engineers.

Black Box Testing

Another kind of manual testing is black-**box** testing. In this testing, the test engineer will examine the software against requirements, identify the defects or bug, and sends it back to the development team.

Functional Testing

The test engineer will check all the components systematically against requirement specifications is known as functional testing. Functional testing is also known as Component testing.

Non-function Testing

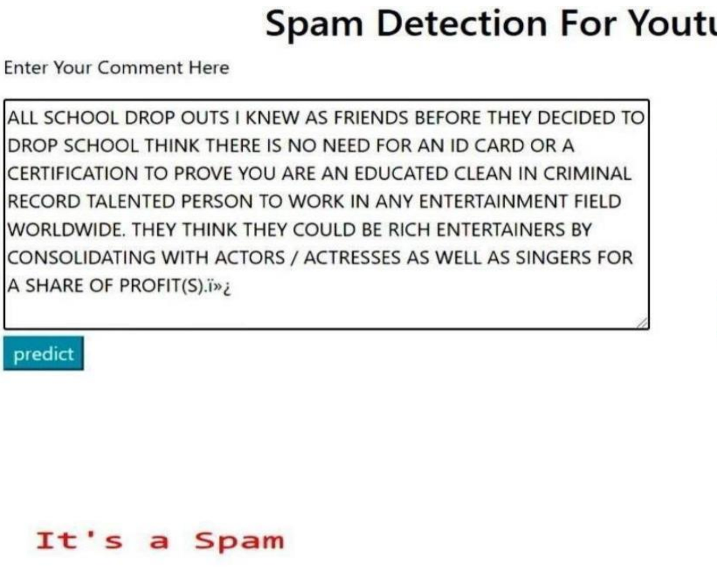
The next segment of black-box testing is non-functional testing. It supplies comprehensive information onsoftware product performance and utilised technologies.

Grey Box Testing

Another segment of manual testing is Grey box testing. It is an association of black box andwhite box testing.

Since, the grey box testing constitutes access to internal coding for designing test cases.
Grey boxtesting isdone by a person who knows coding and also testing.

6 RESULTS

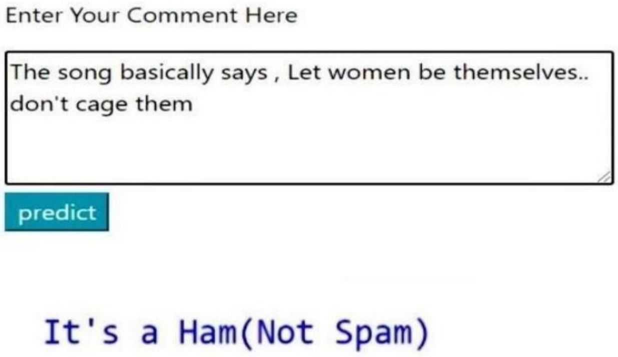


Spam Detection For Youtube

Enter Your Comment Here

ALL SCHOOL DROP OUTS I KNEW AS FRIENDS BEFORE THEY DECIDED TO DROP SCHOOL THINK THERE IS NO NEED FOR AN ID CARD OR A CERTIFICATION TO PROVE YOU ARE AN EDUCATED CLEAN IN CRIMINAL RECORD TALENTED PERSON TO WORK IN ANY ENTERTAINMENT FIELD WORLDWIDE. THEY THINK THEY COULD BE RICH ENTERTAINERS BY CONSOLIDATING WITH ACTORS / ACTRESSES AS WELL AS SINGERS FOR A SHARE OF PROFIT(S).ï»¿

It's a Spam

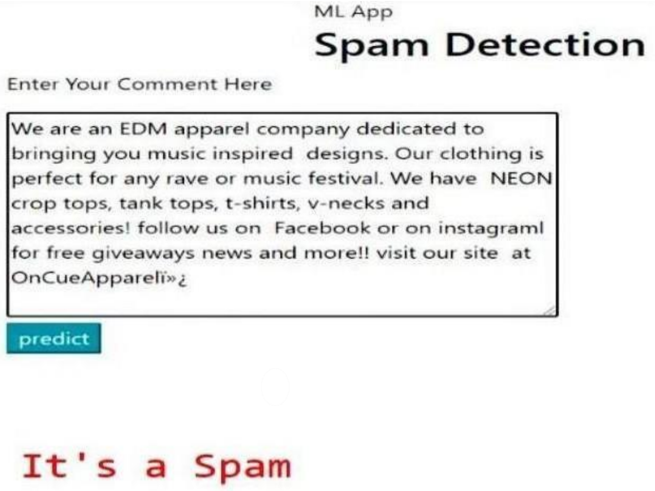


Enter Your Comment Here

The song basically says , Let women be themselves.. don't cage them

It's a Ham(Not Spam)

Spam comment **Ham comment**



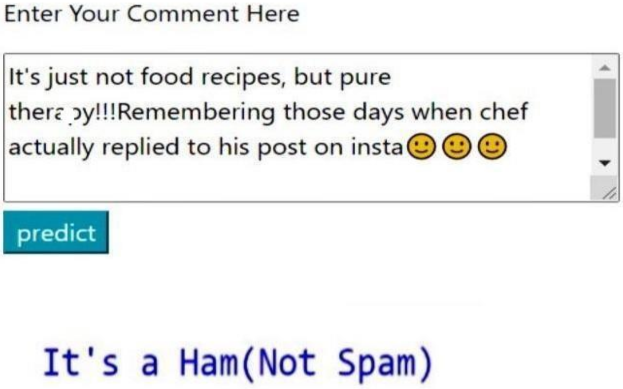
ML App

Spam Detection

Enter Your Comment Here

We are an EDM apparel company dedicated to bringing you music inspired designs. Our clothing is perfect for any rave or music festival. We have NEON crop tops, tank tops, t-shirts, v-necks and accessories! follow us on Facebook or on instagraml for free giveaways news and more!! visit our site at OnCueApparelï»¿

It's a Spam



Enter Your Comment Here

It's just not food recipes, but pure thera py!!!Remembering those days when chef actually replied to his post on insta😊😊😊

It's a Ham(Not Spam)

Spam comment **Ham comment**

We assess the performance of our spam comment detection system with the help of the cross validation and k-fold approach. The dataset is shuffled using a random number. A five-fold cross-validation procedure was used. The whole dataset was divided into five equal parts, in each fold a different segment is used as the test set and the remaining parts as training set. The final F1 Score is acquired from averaging the results of each fold. The algorithm is trained using the training subset and the accuracy of the classifier is cross validated using the testing subset.

Evaluation Metrics

We adapted Accuracy, Precision, Recall and F1 Score as metrics for evaluation. The performance of the algorithm is however, concluded using the F1 score since we would like to obtain both a high precision as well as a high recall.

Accuracy = tp + tn /tp + f p + tn + fnPrecision = tp /tp + f p
Recall = tp/ tp + fn
F1Score = 2 * (precision · recall)/precision + recall
Where tp, tn, fp, fn represent the true positive, true negative, false positive, false negative ratesrespectively.

N-Gram Performance Evaluation

We tested the performance of the algorithms using both word and character grams with different values of n. Support Vector Machine(SVM) with n value as 6 and using character-gram produces the highest F1score. Word-grams outrun character-grams at lower values of n. However, we see a notable increase in the F1 Score at higher values of n in the case of character-grams yielding improvements of nearly 1% over word-monograms.

Figure 4 shows a graphical representation of the F1 Score(YAxis) against tested values of n (X-Axis) inthe case of character-grams. The F1 Score gradually increases and at peaks at an n value of 6.

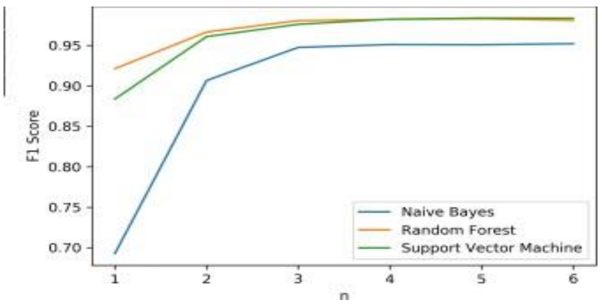
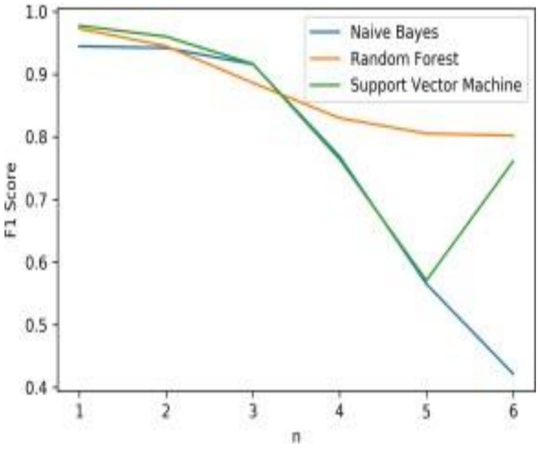


Figure 5 illustrates the same graphical representation in the case of word-grams.
Figure 6 illustrates the improvement of using character-grams over word-grams for different algorithms.



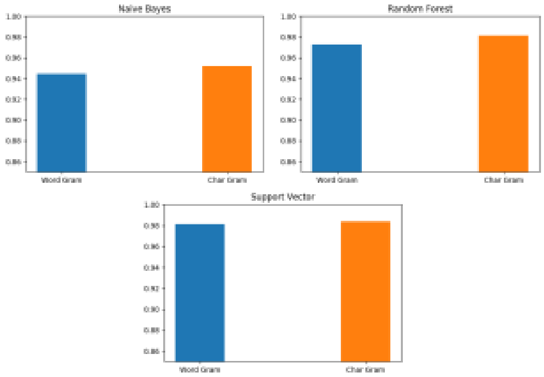


Fig5 Word gram plot

Fig6 improvement of char-gram over word gram

7 CONCLUSION AND FUTURE SCOPE

CONCLUSION

Social media networks have become extremely popular and this creates the opportunity for the malicious user to publish unwanted content such as video spam. This study has introduced the featureset to be used in detecting video spammers that exist in the YouTube media. The features were built based upon the features acquired from the user profile and the content that they shared.Based on the undertaken experiments, it is learned that existing classifiers that were widely used in the data mining community could utilize the features in detecting video phishers. The average detection accuracy was as high as 96% Such a result provides insight on the usefulness of the proposed video spammer feature set. In order to investigate more, additional experiments need to be performed, comparing the results against existing feature set for spammer detection

FUTURE SCOPE

The spam comment detection is done using the modern Machine Learning algorithms such as Random Forest, Naive Bayes classifier. But it can be further improved by using the Deep Learning model for spam detection where the model mainly focuses on : Word Embedding and Bi-directional Deep Learning model + GRU (Gated Recurrent Units). When Deep Learning is used, the model identifies the features during the training process and we can build the network using Keras where the above components are involved. Thus, it could be much advantageous and useful when these kindof methods are explored accurately.

8 REFERENCES

[1] **Y. Yusof** and **O. H. Sadoon**, “Detecting Video Spammers on Youtube Social Media,”no. 082,pp.2017.

[2] **U. K. Sah** and **N. Parmar**, “An approach for Malicious Spam Detection InEmail withcomparison of different classifiers”, IRJET,vol4.

[3] **S. Gandra**, “Implementation Of Prototype To Detect Spam In YouTube Using The ApplicationTube Kit And Naïve Bayes Algorithm”, 2014.

[4] **M. Esmaeili**, et al., “An Anti-Spam System using Naive Bayes Method and Feature SelectionMethods,” International Journal of Computer Applications, vol. 165, 2017.

[5] **Uysal, A. K**. **Gunal**, **S. Ergin**, & **Gunal, E. S.** (2013). The impact of feature extraction and selection on SMS spam filtering,67–72. https://doi.org/10.5755/j01.eee.19.5.1829

[6] **I. Santos**, **C. Laorden**, **X. Ugarte-Pedrero, B. Sanz**, **P.G. Bringas** Spam Filtering through sAnomaly Detection, Springer, Berlin Heidelberg, Berlin, Heidelberg (2012), pp. 203-216 doi:10.1007/978-3-642-35755-8 15. URL https://doi.org/10.1007/978-3-642- 35755-8_15

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

[8] R. Chowdary, N. M. Adnan, **G. A. N. Mahmud**, and **R. M. Rahman**,

“A Data Mining Based Spam Detection System for YouTube,” pp. 373–378, 2013.