**YOUTUBE SPAM DETECTION: LEVERAGING ENSEMBLE ALGORITHMS FOR ROBUST FILTERING**

**ABSTRACT:**

This paper proposes a technique to detect spam comments on YouTube, which have recently seen tremendous growth. YouTube is running its own spam blocking system but continues to fail to block them properly. Therefore, we examined related studies on YouTube spam comment screening and conducted classification experiments with six different machine learning techniques (Decision tree, Logistic regression, Bernoulli Naïve Bayes, Random Forest, Support vector machine with linear kernel, Support vector machine with Gaussian kernel) and two ensemble models (Ensemble with hard voting, Ensemble with soft voting) combining these techniques in the comment data from popular music videos Psy, Katy Perry, LMFAO, Eminem and Shakira.

**SCOPE OF PROJECT:**

Research on detecting spam content and users focus on various fields. Many studies focused on spam on websites (e.g., portal sites and blogs).As YouTube gains popularity as a video sharing platform, spammers target it with low quality content or promotions. Since spammers that harm the YouTube community are increasing, detecting them becomes an interesting source to research. So, we divide the literature of detecting spam into two sections, spam on websites and spam on YouTube.

**OBJECTIVE:**

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. At present, more than 400 hours of video are uploaded and 4.5 million videos are watched 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.

**INTRODUCTION:**

In the previous years, informal online communities like Face book and YouTube have become progressively common platform in an individual person’s day to day life. People use social media as a virtual community platform to stay in touch with friends and family and to also share thoughts and ideas in blogs. Due to this developing pattern, these platforms pull in an enormous number of clients and are easy targets for spammers. YouTube has become the most well-known informal community among youngsters. For example, many makeup tutorials have been started by bloggers who are referred to as “beauty guru” or “beauty influencers” in which majority of the audiences are teenage girls. These days, 200 million clients produce 400 million new YouTube content (videos) every day. This extensive environment provided by YouTube also creates an opportunity for spammers to create irrelevant content directed to users. These irrelevant or unsolicited messages are aimed to attack users by luring them into clicking links to view malicious sites containing malware, phising and scams. One of the most highlighted features of YouTube is the comments section below every video posted by a user. This feature allows users to share opinions and ideas. In this project, the prediction of the spam comments present in the comments section of Youtube videos using the concept called machine learning, it is also known as subset of artificial intelligence, is done. Supervised learning approach depends on a very large number of labeled datasets.The proposed classification algorithm (Logistic Regression) is used in order to predict the spam comment. The purpose of project is to introduce briefly the techniques of machine learning and to outline the prediction technique. Being much more superior to the conventional data analysis techniques, machine learning can open a new opportunity to explore and increase the prediction accuracy. Spam remarks are regularly completely immaterial to the given video and are normally created via mechanized bots camouflaged as a client. The comments section is target by spammers to post completely irrelevant messages, comments, links and ideas. AI is the strategy for extraction, changing, stacking and anticipating the significant data from enormous information to remove a few examples and furthermore change it into justifiable structure for additional utilization. Grouping and expectation are two sorts of dissecting information which portray principal classes of information and forecast of patterns in future information. The noxious spam remarks will ruin the positive perspective of the contents present in the videos posted. The contingency for anticipating the spam remarks has started but has yet not been concluded and built up for an exact forecast of spam remarks.

**MOTIVATION OF PROJECT:**

Spam, poses a huge challenge for researchers aiming to find ways to detect or exclude these unwanted messages from nearly every online media. Many studies focused on traditional methods, such as content-based analysis or extraction of features from the content or their information. Analysis of users’ behavior is often, used to improve the accuracy and performance of spam detection. In an experiment based on content analysis, Rathod and Pattewar analyzed a body of the Gmail dataset to classify legitimate and spam email using the Bayesian classifier model. The proposed bayesian classifier was able to achieve as high as 96.46% accuracy. Moreover, similarity and relevance are important features used to distinguish spams from regular messages. Liu et al. developed two algorithms to identify false reviews on Amazon.com based on the similarity of the reviews and how much the review content is related to the product and to describe some common behavior features of spammers in the spam review. According to the observation, it was found that if the similarity of the two reviews is greater than 70%, then the second review was identified as a copied review. The results showed that 54% of mobile phone reviews on Amazon.com are copied from existing reviews. Jindal and Liu categorized customer reviews of Amazon.com into three types: false opinion, brand review, and non-reviews.

**LITERATURE SURVEY**

**PAPER 1**

**TITLE NAME:**CASCARO: Cascade Of Classifiers For Minimizing The Cost Of Prediction

**AUTHOR:** Blaise Hanczar, Avner Bar-Hen

**YEAR:** 2021

**ABSTRACT:**

Although the prediction performance is crucial for a classifier, its cost of use is also an essential issue forpracticalapplication, however, this question is rarely addressed in the literature. The aim of this article is topropose a prediction method that controls not only the error rate but also the cost of the construction of theclassifier. The main idea is that some examples are easier to predict than others and can be predicted using fewervariables i.e. with a lower prediction cost. Our method, called CASCARO, is based on a cascade of rejectclassifiers of increasing cost. The first classifier of the cascade required only one variable, if the prediction is notreliable the second classifier requiring one more variable is used. The principle is repeated until the last classifierusing all variables. We solve the two main problems for the construction of this type of cascade: its architecture(the order of the classifier) and the simultaneous computation of the rejection regions of the classifiers. Theexperiments show that CASCARO produces significant improvements in the use cost without decreasingprediction performance

**PAPER 2**

**TITLE NAME:** Twitter Spam Detection Based on Deep Learning

**AUTHOR:** Tingmin Wu, Shigang Liu, Jun Zhang, Yang Xiang

**YEAR:** 2021

**ABSTRACT:**

Twitter spam has long been a critical but difficult problem to be addressed. So far, researchers havedeveloped a series of machine learning-based methods and blacklisting techniques to detect spamming activities onTwitter. According to our investigation, current methods and techniques have achieved the accuracy of around 80%.However, due to the problems of spam drift and information fabrication, these machine-learning based methodscannot efficiently detect spam activities in real-life scenarios. Moreover, the blacklisting method cannot catch upwith the variations of spamming activities as manually inspecting suspicious URLs is extremely time-consuming. Inthis paper, we proposed a novel technique based on deep learning techniques to address the above challenges. Thesyntax of each tweet will be learned through WordVector Training Mode. We then constructed a binary classifierbased on the preceding representation dataset. In experiments, we collected and implemented a 10-day real Tweetdatasets in order to evaluate our proposed method. We first studied the performance of different classifiers, and thencompared our method to other existing text-based methods.

**PAPER 3**

**TITLE NAME:** Two-Stage Unsupervised Approach For Combating Social Spammers

**AUTHOR:** Darshika Koggalahewa, Yue Xu, Emest Foo

**YEAR: 2**0**20**

**ABSTRACT:**

Spammers use Online Social Networks (OSNs) as a popular platform for spreading malicious content andlinks. The nature of OSNs allows the spammers to bypass the combating techniques by changing their behaviours.Classification based approaches are the most common technique for spam detection. “Data labelling”, “spam drift”,“imbalanced datasets” and “data fabrication” are the most common limitations of classification techniques thathinder the accuracy of spam detection. The paper presents a two-stage fully unsupervised approach using a user’speer acceptance within OSN to distinguish spammers from genuine users. User’s common shared interest overmultiple topics and the mentioning behaviour are used to derive the peer acceptance. The contribution of the paperis a pure unsupervised method to detect spammers based on users’ peer acceptance without labelled datasets. Ourunsupervised approach is able to achieve 95.9% accuracy without the need for labelling

**PAPER 4**

**TITLE NAME:** BotNet Detection on Social media

**AUTHOR:** Aniket Chandrakant Devle, Julia Ann Jose, Abhay Shrinivas Saraswathula, Shubham Mehta,

Siddhant Srivastava, Sirisha Kona, Sudheera Daggumalli

**YEAR:** 2020

**ABSTRACT:**

As our reliance on social media platforms and web services increase day by day, exploiters view theseplatforms as an opportunity to manipulate our thoughts and actions. These platforms have become an openplayground for social bot accounts. Social bots not only learn human conversations, manners, and presence butalso manipulate public opinion, act as scammers, manipulate stock markets, and so on. There has been evidenceof bots manipulating people’s opinions and thoughts which can be a great threat to democracy. Identification andprevention of such campaigns that release or create these bots have become critical. Our goal in this paper is toleverage web mining techniques to help detect fake bots on social media platforms such asTwitter, therebymitigating the spread of fake news.

**PAPER 5**

**TITLE NAME:** A Novel Recruitment Policy to Defend against Sybils in Vehicular Crowdsourcing

**AUTHOR:** Federico Concone, Fabrizio De Vita, Ajay Pratap, Dario Bruneo, Giuseppe Lo Re, Sajal K.Da

**YEAR:** 2021

**ABSTRACT:**

Vehicular Social Networks (VSNs) is an emerging communication paradigm, derived by merging the concepts of Online Social Networks (OSNs) and Vehicular Ad-hoc Networks (VANETs). Due to the lack ofrobust authentication mechanisms, social-based vehicular applications are vulnerable to numerous attacksincluding the generation of sybil entities in the networks. We address this important issue in vehicularcrowdsourcing campaigns where sybils are usually employed to increase their influence and worsen thefunctioning of the system. In particular, we propose a novel User Recruitment Policy (URP) that, after extractingthe participants within the event radius of a crowdsourcing campaign, detects and filters out the sybil vehicles byusing a novel sybil detection approach, called SybilDriver. This technique combines the advantages of VANETsand OSNs by means of an innovative concept of proximity graph obtained from the physical vehicular network, inconjunction with a community detection and Random Forest techniques adopted in the OSN domain.

**EXISTING SYSTEM:**

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 comments and propose the Cascaded Ensemble Machine Learning Model aware YouTube Spam Comments Detection Scheme to improve the performance of the model. 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.

**DISADVANTAGE:**

* The user evaluation can collect desired information such as usage problems or fitness of recommendation to users who have been invited for evaluation.
* This is a method for determining whether the prediction result is reliable in the case of the data scarcity problem

**PROPOSED SYSTEM:**

In this paper on spam or normal labeled datasets. The dataset consists of 6,431,471 crawled comments of which 481,334 comments were spam in the 6,407 videos that were most viewed between October 31, 2011 and January 17, 2012 in the United States. This dataset was mixed with English and non-English comments, so we extracted only English comments for the experiment. In addition, to make it similar to the data size used in the experiment of 3, we extracted 1,000 spam comments and normal comments, and compared them with 5,000 samples. In the experiment, we used an Extra-Tree Classifier with the techniques used in 3. Finally, we plotted the Precision, Recall, F1-score, and ROC curves by adding 1,000 data points from 1,000 to 5,000.

**ADVANTAGE:**

* The advantages are that it is very efficient in terms of storage space and computation time and handles noise and missing data well.
* Offline evaluation has the advantage of standardizing evaluation methods and evaluation items, and there are various evaluation items such as accuracy, coverage, confidence, and novelty.

**SYSTEM CONFIGURATION:**

# H/W SYSTEM CONFIGURATION:

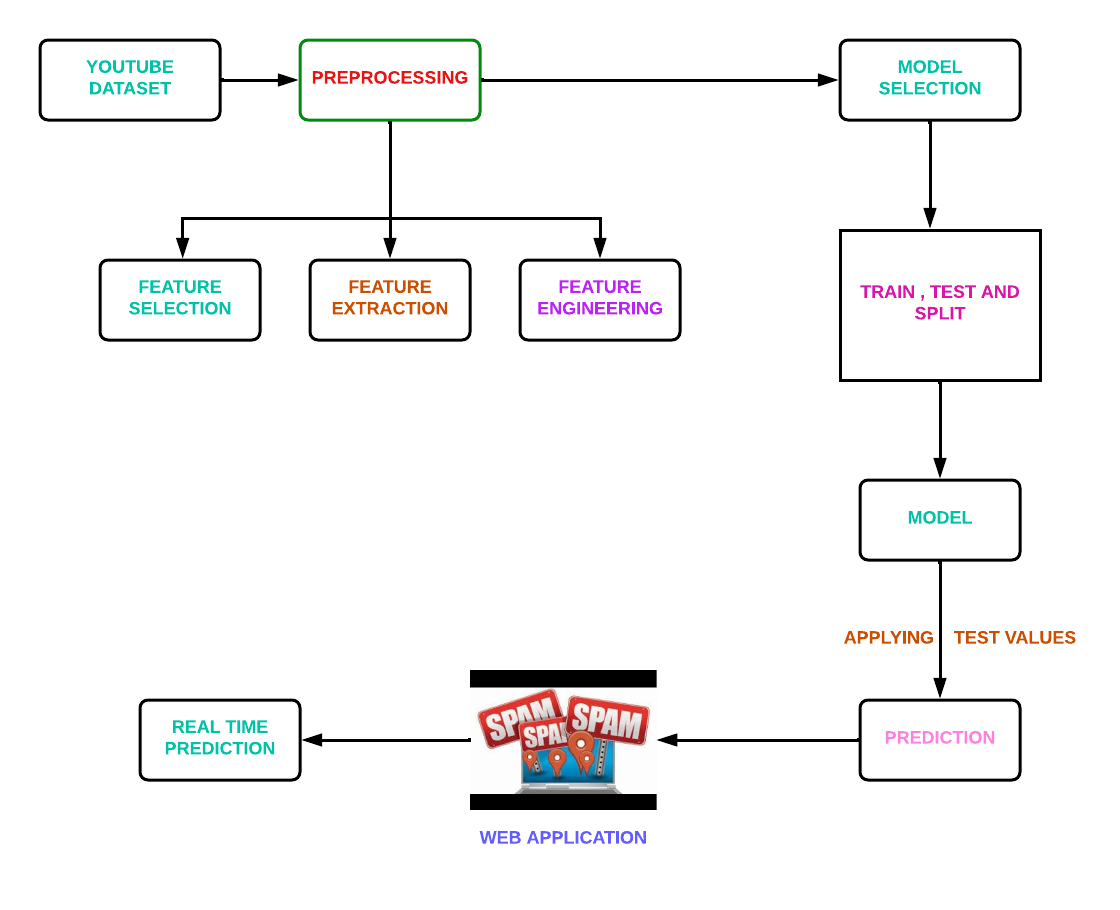
# Processor - I3, I5,I7

* RAM - 8 Gb
* Hard Disk - 500 GB

# S/W SYSTEM CONFIGURATION:

* Operating System - Windows 7/8/10
* Front End - Html,Css
* Scripts - python language

**ARCHITECTURE DIAGRAM:**



**MODULES DESCRIPTION**

**MODULES LIST:**

* Youtube dataset.
* Preprocessing
* Feature selection
* Feature extraction and feature engineering
* Extra-Tree Classifier classifier

**YOUTUBE DATASET:**

The benefit of using these words based on their entropy score in the characteristic-set is that we have been capable of lessen uncertainty in the prediction final results as those phrases have a exceptional effect of frequency count in spam and non-spam YouTube.

**PREPROCESSING:**

Before starting with preparation preprocessing of the messages must be done. First 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 must be done for each message in the data set.

**FEATURE SELECTION:**

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.

**FEATURE EXTRACTION AND FEATURE ENGINEERING:**

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 . This undergoes the following technique:

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

Analyzer: “Whether the feature should be made of word or character n-grams. Option ‘char\_wb’ creates character n-grams only from text inside word boundaries; n-grams at the edges of words are padded with space.”

**EXTRA TREE CLASSIFIER:**

The Extra Trees classifier algorithm, a member of the ensemble learning family, shares similarities with Random Forests but introduces distinctive features in its tree-building process. Notably, Extra Trees employs a higher degree of randomization during the creation of individual decision trees. This involves selecting random subsets of features and utilizing random thresholds for node splitting, enhancing the model's robustness to noisy data. This algorithm excels in parallelization, allowing for efficient construction of each tree independently. By embracing randomness, Extra Trees mitigates overfitting concerns, making it less sensitive to the intricacies of the training data. As a result, Extra Trees stands as a powerful and efficient tool for predictive modeling tasks, offering reliable predictions across diverse datasets.

**CONCLUSION:**

In this paper, we proposed a technique to detect spam comments on YouTube, which have recently seen tremendous growth using a Cascaded Ensemble Machine Learning Model. It examined related studies on YouTube spam comment screening and conducted classification experiments with six different machine learning techniques (Decision tree, Logistic regression, Bernoulli Naïve Bayes, Random Forest, Support vector machine with linear kernel, Support vector machine with Gaussian kernel) and two ensemble models (Ensemble with hard voting, Ensemble with soft voting) combining these techniques in the comment data. The experimental results showed that the ESM-S model proposed in this paper had the best performance in four of five evaluation measures. We proposed a new model, combining various techniques that improved the performance results unlike previous studies that used one model for detection

**REFERENCE:**

[1]. H. Shaban. (Sep. 19, 2019). Nearly Half of Cellphone Calls Will be Scams by 2019, Report Says. The Washington Post. Accessed: Feb. 21, 2020.

[2]. O. Abayomi-Alli, S. Misra, A. Abayomi-Alli, and M. Odusami, ‘‘A review of soft techniques for SMS spam classification: Methods, approaches and applications,’’ Eng. Appl. Artif. Intell., vol. 86, pp. 197–212, Nov. 2019.

[3]. O. M. E. Ebadati and F. Ahmadzadeh, ‘‘Classification spam email with elimination of unsuitable features with hybrid of GA-naive Bayes,’’ J. Inf. Knowl. Manage., vol. 18, no. 1, Mar. 2019.

[4]. M. S. Hanif and M. Bilal, ‘‘Competitive residual neural network for image classification,’’ ICT Exp., vol. 6, no. 1, pp. 28–37, Mar. 2020.

[5]. H. Liu, M. Zhou, and Q. Liu, ‘‘An embedded feature selection method for imbalanced data classification,’’ IEEE/CAA J. Automatica Sinica, vol. 6, no. 3, pp. 703–715, May 2019.

[6]. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” Feb. 2020.

[7]. S. Venkatraman, B. Surendiran, and P. A. R. Kumar, “Spam e-mail classification for the internet of things environment using semantic similarity approach,” J Supercomput, vol. 76, no. 2, pp. 756-776, 2020.

[8]. K. Sheridan. (2020). FBI: Business Email Compromise Cost Businesses 1.7B in 2019, Dark Reading. Accessed: Mar. 21, 2021.

[9]. A. Ali. (2020). Visualizing the Social Media Universe in 2020. Accessed: Jan. 14, 2021.

[10]. R. Lerman and H. Denham. (2020). 3 Charged in Massive Twitter Hack, Including Alleged Teenage ‘Mastermind’.The Washington Post. Accessed: Jan. 14, 2021.