ISSN 2063-5346

RETRAINING A SPAM DETECTION MODEL TO HANDLE NEW EDGE CASES

Mahalakshmi .D[1], Dr P.J Sathish[2]

Article History: Received: 01.02.2023 Revised: 07.03.2023 Accepted: 10.04.2023

Abstract

As the quality of online social networks has improved, spammers have discovered that they can easily exploit them to lure users to do undesirable things by posting spam comments on videos. In this project, spam detection has been done as well as consideration of YouTube comments. YouTube Bookmaker and Google Safe Browsing technologies both identify and filter spam. measures made by YouTube to fight spam. These programmes will block hazardous links, but they will not be able to protect the user as soon as possible in real time. As a result, companies and academics have developed a wide range of spam-free social networking systems. The survey to determine the most effective method of identifying spam comments has been completed. What we want is achieved by the bag-of-words paradigm, which translates the sentences or phrases and counts the occurrences of a comparable term. In the world of computer science, a "big" is a data structure that organises information similarly to a "array" or "list," except in this case, the order is meaningless and we simply keep track of the count rather than repeating it if an object occurs more than once.

[1]PG scholar, Department of computer science and Engineering, Panimalar Engineering College, Chennai, India.

[2]Professor, Department of computer science and Engineering, Panimalar Engineering College, Chennai, India.

Mahalakshmi1607cs@gmail.com[1], sathishjraman@gmail.com[2]

DOI:10.31838/ecb/2023.12.s1-B.322

INTRODUCTION:

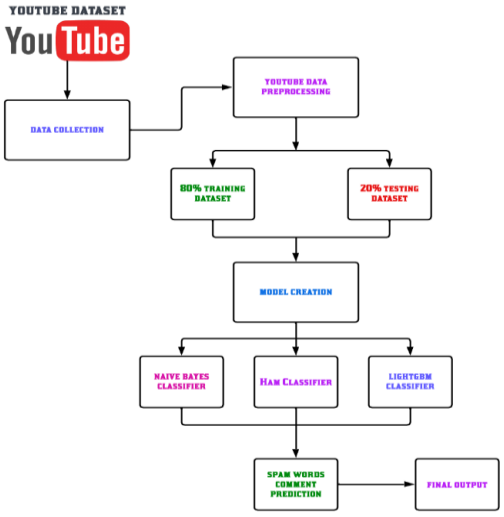
Informal online communities such as Facebook and YouTube have become increasingly common in people's daily lives in recent years. People use social media to communicate with friends and family, as well as to blog about their opinions and beliefs. Due to this new trend, these platforms draw a sizable user base and are well-liked by spammers. Currently, YouTube is the most prominent informal youth assembly. For example, bloggers known as "beauty gurus" or "beauty influencers" have launched a slew of makeup tutorials, with the vast majority of viewers being young women. Currently, 400 million users generate YouTube adds 200 million new videos every day**.** YouTube provides this extensive setting. enables spammers to direct readers to irrelevant content they have produced. These Users are the targets of an attack from irrelevant or unwelcome communications that entice them to click on links to dangerous websites. See phishing, scams, and risky websites with viruses. For instance, many makeup manuals One of YouTube's most visible features is the comments area, that show below each video which a user uploads. Users are able to express about their ideas and thoughts with this tool. The project's study of spam comments using the notion of machine learning is already available in the comments section of YouTube videos and is an accomplished subset of artificial intelligence. The application of supervised learning requires a large collection of labelled datasets. The proposed classification(Logistic Regression) methodology. The spam comment is predicted using regression analysis. The project's simple description of machine learning and prediction algorithms is its aim. Machine learning, which is significantly more efficient than conventional data processing techniques, can open up new research avenues and increase the accuracy of predictions Spam comments are frequently completely unrelated to the video in issue and are typically created by computer bots that imitate real users. Spammers often publish messages, comments, links, and thoughts in the comments section that are completely irrelevant. Artificial intelligence (AI) is a technique for extracting, changing, stacking, and anticipating the crucial data from enormous volumes of data in order to weed out a few samples and then further turn it into a legal structure for additional usage. The primary information classes are displayed and future information patterns are predicted using the information analysis techniques of grouping and anticipation. The inflammatory spam comments will destroy any negative perception of the material in the submitted videos. Despite having started, the preparation for anticipating spam comments has not yet been done and is not yet ready for a precise forecast of spam remarks.

RELATED WORK:

Thi-kim-hien Le, Yi-zhen Li1, and Sheng-tun Li proposed a system in 2022 [1], Hierarchical Logistic Regression Model (HLR) algorithm employed IHLR can categorise review spammers and false reviews with more accuracy than traditional machine-learning techniques. The length of prior reviews may have an impact on the success of detection as the duration analysis generated a greater level of detection accuracy than the recency analysis. A method employing Naive Bayes, Random Forest, and Support Vector Machine was proposed by HAYOUNG OH in 2021 [2]. We increase accuracy by employing an ensemble machine learning model. The data set with 5,000 comments fared worse than the data set with 1,000 genuine comments and 1,000 spam comments.Shreyas Aiyara , Nisha P Shetty proposed a system in 2018 [3] , Random Forest (RF), SupporVector Machine (SVC), Multinomial Naive Bayes (MNB), andCombining proprietary

algorithms like N-Grams, which have been demonstrated to be highly effective in identifying and removing spam comments, with more established machine learning techniques like Random Forest, Support Vector Machine, and Naive Bayes. Neural networks can be used to improve word representation. Amar Singh, NidhiChahal, Simranjit Singh proposed a system in 2021 [4], Naive Bayes and Support Vector Machine Algorithm used, with the help of the suggested method, we evaluate better accuracy in separating spam from non-spam data. Although classifiers are employed to address the issue of spam detection and enhance the results, spammers continue to create new ways to send spam messages. Thus that the work might be extended in the future to avoid various types of attacks such as Denial of Service attacks. Maximum entropy, logistic regression, support vector machines, naive bayes classifier, decision trees, rule-based classifiers, are some methods suggested by Maria Novo-Louries in 2022[5].to find Compression characteristics and/or knowledge about the likelihood of accurately predicting the language of target texts are frequently utilised to enhance categorization. The authors have found traits that are detrimental to spam filtering, like the inclusion of terms written in capital letters that are influenced by certain trends and usage patterns of users of various Internet services. Low-accuracy classifiers using synset-based representations for short texts (YouTube Comments Dataset).These issues are most likely a result of text disambiguation. In huge texts (emails), employing NER to extract pertinent data (amounts of money, addresses, dates, etc.) did not result in appreciable improvements in the datasets under analysis. SAQIB ALI proposed a system in 2020 [6] K-means clustering, KNN (k-nearest neighbors) algorithm used, to gather client feedback for Kansei engineering's Uber service in the India-Pakistan region, perform aspect-based sentiment analysis. Online reviews are gathered for this reason through the Uber service's Facebook page, where users share their opinions in an unfiltered and transparent manner. It can help ridesharing companies grow their businesses by enhancing their offerings to better meet customer needs. Roman Urdu lacks a standardized dictionary and is a poor resource language. Depending on the context, a phrase may be spelt very differently by the same person or even different people. ASIF KARIM proposed a system in 2021 [7]Utilizing Hierarchal clustering, K-means clustering, KNN (k-nearest neighbours), and an anomaly detection technique, The best clustering was obtained by OPTICS, which outperformed DBSCAN by 0.26% in terms of average efficacy. OPTICS and DBSCAN's combined average accuracy was discovered to be 75.76%. Clustering quality still has to be enhanced. Xin Tong, Jingya Wang proposed a system in 2020 [8]. Naive Bayes and Support Vector Machine Algorithm used. The dynamic routing method is optimised, and the topology of the traditional capsule network is improved for feature mining and classification. This creates a model with outstanding accuracy without sacrificing running performance. Sensitive word splitting and character order alterations, which are common tactics used by spammers to create adversarial samples in practice, might cause the spam detection system to make the wrong call without sacrificing readability. TIAN XIA proposed a system in 2020 [9]. Naive Bayes, Support Vector Machine Algorithm, KNN Algorithm used. The processing speed of the algorithm is constant and irrespective of the number of rules or the amount of its vocabulary. Constant time complexity is made achievable by the development of a new unique data structure called a hash forest and a technique for encoding rules. The suggested algorithm detects spam terms by inspecting a relatively small subset of all phrases rather than iterating through each

spam term in the rules. The algorithm lacks flexibility because there aren't many logical operators init. In 2020, DONGJIE LIU presented a system [10]. There are two algorithms used: Naive Bayes and Support Vector Machine. He presented a revolutionary detection technique that uses images of bogus websites and user perspectives to disprove Web spam. It works better and is suitable for use in actual Web environments. This study only takes into account the picture feature, but it may integrate the picture and text features to provide a more accurate detection model. ZHIJIE ZHANG proposed a system in 2020 [11] Naive Bayes algorithm used, the suggested I2FELM effectively distinguishes between balanced and unbalanced datasets. Additionally, the I2FELM can more accurately detect spam with less variables taken into account, demonstrating the efficiency of the algorithm. More factors are not taken into account in order to precisely detect spam (e.g., semantic analysis and emotion analysis).RhitabratPokharel, Dixit Bhatta proposed a system in 2021 [12] Naive Bayes, Decision Tree, and Support Vector Machine algorithm used. YouTuber to clearly see the queries raised regarding the video and the solutions offered to enhance the information. This can save the YouTuber from having to scroll through hundreds of comments and manually filter them for each video. The remarks utilising two feature selection methods and five distinct model combinations. A system was suggested in 2017 by Aqlima Aziz [13]. Using the K-Nearest Neighbor technique and the Support Vector Machine The study devised a method for identifying spam comments on YouTube, which has recently experienced incredible growth, using a Cascaded Ensemble Machine Learning Model. Due to the increase in outliers and missing values, the data set with 5,000 comments less effective than the data set with 1,000 spam comments and 1,000 genuine comments. MUHAMMAD BILAL proposed a system in 2020 [14] Features Mapping algorithm used Changes in the company's reputation are investigated, along with patterns in user rating and business choice. Additionally, new features for businesses and reviewers are introduced, and their influence on the usefulness of the reviews is examined together with that of already-existing features. Only the reviews on Yelp.com are the subject of this investigation. It only concentrates on the retail sector. The findings of this study should be confirmed against datasets of short-term evaluations because it uses a dataset of long-term reviews (2005-2018).JINGDONG WANG, HAITAO KAN proposed a system in 2020 [15] Dictionary-Based Sentiment Intensity Extraction Algorithm ,SVM,RF algorithm used, The proposed method is 3.5% more accurate than the baseline procedures with an accuracy of 84.45%. When compared to the most recent deep learning model, its baseline precision has increased by 5.3%. The most effective classifiers according to statistics are the random forest (RF) and support vector machine (SVM). Professional fake authors' writing methods will alter when the e-commerce platform's detecting mechanism is updated. It is harder to locate using the detecting technique.



PROPOSED SYSTEM:

Spam or typical labeled datasets are discussed in this work. In the 6,407 videos that had the most views in the United States between October 31, 2011 and January 17, 2012, a total of 6,431,471 crawling comments were found, of which 481,334 were spam. As there were both English and non-English comments in this dataset, we exclusively retrieved English comments for the experiment. Additionally, we gathered 1,000 spam comments and 1,000 genuine comments, compared them to 5,000 samples, and adjusted the data size to be similar to the data size used in the experiment of 3. ANN (Artificial Neural Network) technology and the methods from 3 were combined in the experiment. Following that, we multiplied each curve by 1,000 to obtain the Precision, Recall, F1-score, and ROC curves. 1,000 to 5,000 data points are needed.

METHODOLOGY:

YouTube Dataset

The advantage of selecting these terms based on their entropy score in the characteristic-set is that we were able to reduce uncertainty in the prediction findings since such phrases have a considerable impact on frequency count in spam and non-spam. YouTube.

PREPROCESSING

Before beginning preparation, it is required to pre-process the messages. First and foremost, all of the characters must be in lowercase. It is required to treat a phrase that appears in both capital and lowercase letters as a single entity rather than as two independent entities. After that, each message in the data collection needs to be tokenized.

FEATURE SELECTION

The main advantage of using the terms in the dataset is that it can reduce uncertainty in final result prediction because those phrases have a substantial impact on the frequency count in spam and ham comments on YouTube.

FEATURE EXTRACTION AND FEATURE ENGINEERING

The supervised feature called "attribute significance" ranks characteristics according on how effectively they sequentially predict an objective. In this case, Count Vectorizer is used to convert a "collection of text documents into a matrix of token counts". This undergoes the following technique:

• N-grams: N-grams can be used to improve precision. One word is dealt with, but if there are two terms that are mutually exclusive, the meaning changes completely. As a result, the variation in accuracy is better when text is broken into tokens of two or more words rather of being a single word. Analyzer: “Whether the element should be composed of n-grams of words or characters. Option 'char_wb' generates character n-grams solely from text within word borders; n-grams at word edges are padded with space."

NAIVE BAYES CLASSIFIER

The Naive Bayes classifier is named after the Bayes theorem, from which it derives. It is a straightforward probabilistic model that provides incredibly speedy predictions. Naive Bayes deals with dependent events and the likelihood that an event will occur in the future based on the likelihood that the same event will occur in the past. On YouTube, this algorithm can be used to categorize spam videos; the crucial aspect is probability. If certain terms are regularly found in spam but not in ham, this incoming YouTube is most

likelyspammed.

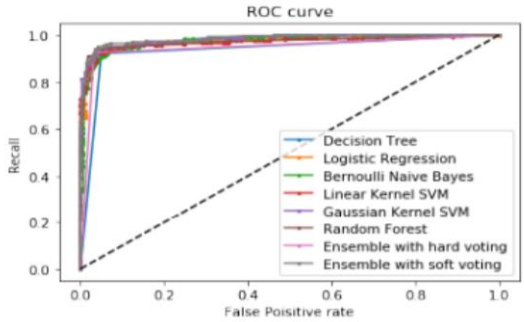


Fig: ROC curve of the proposed classifiers scheme

Spam words filters now frequently use the Naive Bayes classifier algorithm. Every phrase has a predetermined chance of turning up in spam or ham in its database. At the training stage, Naive Bayes create a lookup table in which they store all possible probabilities that will be utilised by the algorithm to forecast the outcome. The filter will classify the dataset into one of two categories if the sum of the word probabilities exceeds a predefined threshold. And let's say you give the algorithm a test point to forecast the outcome. The algorithm will retrieve the values from the lookup table, which contains all of the probability possibilities, and use that value to predict the outcome.

CONCLUSION:

In this study, we presented a strategy for recognizing spam observations on YouTube, which have dramatically increased recently. It examined prior studies on YouTube spam comment filtering and conducted classification tests using the comment data and There are six different machine learning algorithms (Decision Tree, Logistic Regression, Bernoulli Naive Bayes, Random Forest, Support Vector Machine with Linear Kernel, Support Vector Machine with Gaussian Kernel), each with a different purpose. In order to incorporate these tactics, two ensemble models—Ensemble with Hard Voting and Ensemble with Gentle Voting—were used. The experimental findings supported the ESM-S model's in this study outperformed the others across four of the five assessment metrics. Unlike previous studies that employed a single method, we created a new model that included various techniques to increase the detection performance of one model.

REFERENCE:

[1] Li, Yi-Zhen, and Sheng-Tun Li. "Do Reviewers' Words and Behaviors Help Detect Fake Online Reviews and Spammers? Evidence From a Hierarchical Model." *IEEE Access* 10 (2022): 42181-42197.

[2] Oh, Hayoung. "A YouTube spam comments detection scheme using cascaded ensemble machine learning model." *IEEE Access* 9 (2021): 144121-144128.

[3] Aiyar, Shreyas, and Nisha P. Shetty. "N-gram assisted youtube spam comment detection." *Procedia computer science* 132 (2018): 174-182.

[4] Singh, Amar, et al. "Spam detection using ANN and ABC Algorithm." *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*.IEEE, 2021.

[5]Novo-Lourés, María, et al. "Enhancing representation in the context of multiple-channel spam filtering." *Information Processing & Management* 59.2 (2022): 102812.

[6] Ali, Saqib, Guojun Wang, and ShaziaRiaz. "Aspect based sentiment analysis of ridesharing platform reviews for kansei engineering." *IEEE Access* 8 (2020): 173186-173196.

[7] SHANMUGAM, BHARANIDHARAN, and KRISHNAN KANNOORPATTI. "An Unsupervised Approach for

Content-Based Clustering of Emails into Spam and Ham through Multiangular Feature Formulation."

[8] Tong, Xin, et al. "A content-based Chinese spam detection method using a capsule network with long-short attention." *IEEE Sensors Journal* 21.22 (2021): 25409-25420.

[9] Xia, Tian. "A constant time complexity spam detection algorithm for boosting throughput on rule-based filtering systems." *IEEE Access* 8 (2020): 82653-82661.

[10] Liu, Dongjie, and Jong-Hyouk Lee. "CNN based malicious website detection by invalidating multiple web spams." *IEEE access* 8 (2020): 97258-97266.

[11] Zhang, Zhijie, RuiHou, and Jin Yang. "Detection of social network spam based on improved extreme learning machine." *IEEE Access* 8 (2020): 112003-112014.

[12] Pokharel, Rhitabrat, and Dixit Bhatta. "Classifying YouTube Comments Based on Sentiment and Type of Sentence." *arXiv preprint arXiv: 2111.01908* (2021).

[13] Aziz, Aqliima, et al. "YouTube Spam Comment Detection Using Support Vector Machine and K–Nearest Neighbor." *Indonesian Journal of Electrical Engineering and Computer Science* 5.3 (2017): 401-408.

[14]Bilal, Muhammad, et al. "Profiling users' behavior, and identifying important features of review "helpfulness"." *IEEE Access* 8 (2020): 77227-77244.

[15]Wang, Jingdong, et al. "Fake review detection based on multiple feature fusion and rolling collaborative training." *IEEE Access* 8 (2020): 182625-182639.