Finding Spam in YouTube Comments: An Approach for Detection

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD, Audisankara college of engineering & Technology, india
Mrs A. Bharathi, Assistant Professor, Department of CSE, Audisankara college of engineering & Technology ,india
Maruboina Sushma, Department of CSE, Audisankara college of engineering & Technology, india

Abstract: Sending an unwelcome message over a messaging service is the practice known as spamming. On YouTube, user information is really abundant. Like videos, people may subscribe to channels and comment on them; YouTube has attracted more and more users because to these capabilities. This motivates spammers to flood the comments with spam, therefore luring them. Comment moderation tools on YouTube are somewhat limited, and as the spam volume is rising quickly, video owners are blocking their comment areas. Thanks to machine learning, caustic remarks are not common anymore. Learning, numpy, and pandas: the three sklearns These materials could help you to better understand the foundations of machine learning. Comment areas of spam comments on YouTube allow malware to be spread, therefore increasing user PC vulnerability.

Index terms *- —YouTube spam detection, machine learning, text classification, logistic regression, natural language processing (NLP), comment moderation, social media security, supervised learning.*

1. INTRODUCTION

The unofficial internet communities like Facebook and YouTube that have grown up recently are progressively influencing people's daily life. Social media is used by many to write about their beliefs and stay in contact with friends and relatives. These services are becoming used by a lot of individuals due to this tendency, so they are ideal targets for spams. For young people, YouTube has lately become the most used unofficial social network. Many "beauty guru" or "beauty influencer" bloggers have been creating cosmetics lessons targeted mostly at adolescent girls' audience[3]. More than 200 million users everyday utilise YouTube in the current era to upload 400 million fresh videos. YouTube offers spammers a lot of chances to upload material unrelated to its viewers. Clicking on links in these unwelcome or unsolicited communications allows one to access malicious websites including malware, phishing, and other scams[1]. Among the most noticeable elements of YouTube is its comment area. lets individuals exchange their ideas and opinions. This study predicts spam comments in the YouTube video comments area using machine learning—a subset of artificial intelligence[4]. Successful application of the supervised learning approach depends on a large number of tagged datasets. Proposed classification method (Logistic Regression) allows one to anticipate spam comments before they are published. This project aims to provide a quick review of machine learning methods together with

their relevance to prediction. Machine learning can open a fresh path for research and increase prediction accuracy when compared with conventional techniques of data analysis. Automated bots pretending as customers often publish spam comments completely unrelated to the topic of the video they are referencing[5]. Using the comments section, spammers post messages, remarks, links, and other ideas unrelated to the current conversation. AI aims to extract significant data from a lot of data and convert it into a structure that can be justified for more use by means of manipulation, modification, stacking, and prediction of data. Two forms of information dissection exist: grouping and forecasting[6]. The negative attitude of the material in the videos will be tarnished by the poisonous spam comments. Although spam comments are already expected, the backup strategy to get ready for it has not been adequately carried out[7].

2. LITERATURE SURVEY

a) **An efficient modularity based algorithm for community detection in social network*:***
https://ieeexplore.ieee.org/document/7562715

Community identification procedure aims to identify clusters in a social network (SN), in which nodes inside the cluster are densely linked as opposed to nodes outside the cluster. Especially in the field of social networking, this technique is among the difficult problems in the era of big data analytics. SN is commonly represented using graph data structures, in which case actors are represented by nodes and interactions among the actors by edges. Although there are various techniques for community recognition in an SN, each one has some shortcomings in identifying community over a large size network. This work presents a feasible modularity based community identification method. Using some of the most widely used social network datasets, the suggested method has been evaluated against other current community detection systems. The algorithm's performance has been evaluated under several criteria like modularity, clustering coefficient, execution time etc.

b) **A Scalable Distributed Louvain Algorithm for Large-Scale Graph Community Detection**
https://ieeexplore.ieee.org/document/8514887

Based on the Louvain approach, we propose a fresh distributed community discovery strategy for big graphs. We guarantee the workload and communication balance among processors by means of a distributed delegate division. Furthermore, we provide a novel heuristic approach to guarantee the convergence of the distributed clustering method and precisely coordinate the community construction in a distributed environment. Using up to 32,768 CPUs, our thorough experimental work has shown the scalability and accuracy of our technique with several large-scale real-world and synthetic graph datasets.

c) **An Approach for Spam Detection in YouTube Comments Based on Supervised Learning*:***

https://www.researchgate.net/publication/337826806_An_Approach_for_Spam_Detection_in_YouTube_Comments_Based_on_Supervised_Learning

Online social media platforms including YouTube, Twitter, Facebook, LinkedIn, etc. are very common in the lately developed civilisation. People use social media to connect with others, learn new things, exchange ideas, have fun, and keep updated on happenings all across the planet. Among these websites, YouTube is now the most often used one for

viewing and sharing video material. But such popularity has also drawn hostile users who want to spread malware and virus or self-promote their movies. These spam films could include obscene material or have no connection whatsoever to their title. Finding a means of spotting these movies and reporting them is thus rather crucial. In this article, for this aim we have assessed numerous top-performance classification methods. Good accuracy results are shown by the Multilayer Perceptron and Support Vector Machine according to statistical analysis of outcomes.

d) **SOAP: A Social network Aided Personalized and effective spam filter to clean your e-mail box**

https://www.semanticscholar.org/paper/SOAP%3A-A-Social-network-Aided-Personalized-and-spam-Li-Shen/f989941bda92faa2fa353a4c16459d1b03b2637d

The enormous rise in unwanted emails has spurred the creation of many spam filtering methods. Because it can continually adapt to target fresh spam by learning terms in fresh spam emails, a Bayesian spam filter is better than a static keyword based spam filter. Still, avoiding spam keywords and adding many benign terms to the emails will readily poison Bayesian spam filters. Besides, they require a lot of time to adjust to a new spam depending on user comments. Moreover, few present spam filters use social networks to help spam identification. In this study, we present a SOcial network Aided Personalised and effective spam filter (SOAP), thereby developing an accurate and user-friendly spam filter. Unlike other filters emphasising analysing keywords (e.g., Bayesian filter) or establishing blacklists, SOAP uses the social interaction among email correspondents to detect the spam automatically and dynamically. Three elements—social closeness-based spam filtering, social interest-based spam filtering, and adaptive trust management—are included into the fundamental Bayesian filter via SOAP. We assess SOAP performance in relation to Facebook's trace data. Experimental results indicate that SOAP may significantly raise the accuracy, attack-resilience, and efficiency of spam detection of Bayesian spam filters. Furthermore observed is that Bayesian spam filters perform at the lowest bound of SOAP.

e) **Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers*:***

https://jisajournal.springeropen.com/articles/10.1007/s13174-010-0014-7

With a significant economic influence on society, email spam has grown to be a really major issue. Luckily, there are several methods that let most of such communications be automatically identified and deleted; the most well-known methods are derived from Bayesian decision theory. One well-known challenge of such probabilistic methods, however, is the enormous dimensionality of the feature space. Many term-selection techniques have been suggested to help us escape the dimensionality curse. Still unclear, though, how the strategy used to lower the dimensionality of the feature space influences the Naive Bayes spam filter performance. We investigate in this work the performance of numerous term-selection methods using several Naive Bayes spam filter models. Our studies were carefully planned to

provide statistically strong outcomes. Furthermore, we investigate the often used measures to assess the spam filter quality. At last, we also look at the advantages of performance measurement using the Matthews correlation coefficient.

3. METHODOLOGY

i) Proposed Work:

The proposed system focuses on detecting spam in YouTube comment sections using machine learning techniques. Since YouTube's native comment moderation features are limited, this system aims to automatically classify spam comments before they appear publicly. The process begins with collecting a large dataset of YouTube comments labeled as spam or not spam. Preprocessing steps such as removing stop words, punctuation, and converting text to lowercase are applied to clean the data. Then, machine learning libraries like Scikit-learn (sklearn), Numpy, and Pandas are used to extract features and build the classification model.

Logistic Regression is employed as the core classification algorithm due to its efficiency in binary classification tasks. The model is trained on the processed dataset to differentiate between spam and non-spam comments based on patterns and keyword usage. By identifying malicious links and unrelated promotional content, the system can help prevent malware spread and phishing attacks through comment sections. This solution not only improves the quality of user interaction on YouTube but also reduces the workload of manual moderation through automated, real-time spam detection.

ii) System Architecture:

The architecture of the proposed system comprises four main components: data collection, preprocessing, feature extraction, and classification. Initially, YouTube comment data is gathered and labeled for training. The preprocessing module cleans and normalizes the text data by removing stop words, punctuation, and applying tokenization. Next, the feature extraction module converts the text into numerical vectors using techniques like TF-IDF. These features are then passed to the classification module, where a Logistic Regression model is trained to distinguish between spam and non-spam comments. Finally, the system evaluates new comments in real-time, marking spam comments and filtering them out automatically. All components work together to ensure seamless and efficient spam detection on YouTube.

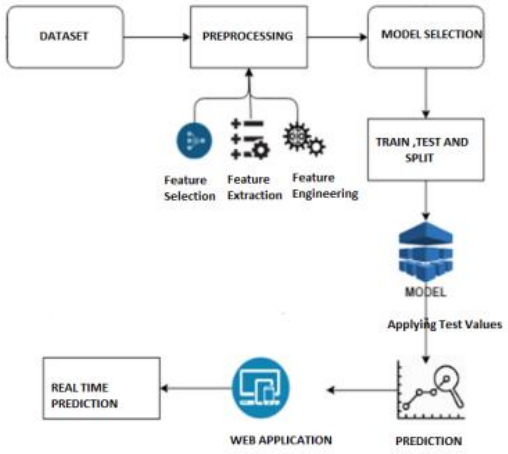

Fig: proposed architecture

iii) Modules:

a. **Data Collection Module**

- Collects YouTube comments dataset (labeled spam/non-spam).
- Supports CSV, JSON, or API-based data inputs.

b. **Data Preprocessing Module**
- Cleans text by removing stop words, punctuation, and symbols.
- Applies tokenization, stemming, and case normalization.

c. **Feature Extraction Module**
- Converts text into numerical format using TF-IDF or CountVectorizer.
- Prepares input for machine learning algorithms.

d. **Classification Module**
- Trains the Logistic Regression model using sklearn.
- Classifies new comments as spam or non-spam.

e. **Spam Detection & Filtering Module**
- Automatically detects and flags spam comments.
- Filters or hides spam comments before publishing.

f. **Evaluation Module**
- Measures accuracy, precision, recall, and F1-score.
- Validates performance of the classification model.

iv) Algorithms:

i. Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for binary classification problems such as identifying whether a YouTube comment is spam or not. It works by calculating the probability that a comment belongs to the spam class using a sigmoid function. Based on the features extracted from the comment text, it assigns weights to each feature and uses them to predict a probability score. If the score crosses a defined threshold, the comment is marked as spam. Its simplicity and effectiveness make it suitable for real-time spam detection.

j. TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF is a statistical feature extraction technique used to convert text into numerical values. It measures how important a word is in a particular document (comment) relative to all other documents in the dataset. “Term Frequency” represents how often a word appears in a comment, and “Inverse Document Frequency” reduces the weight of words that appear frequently across all comments. This technique helps the classifier focus on unique and relevant terms that are more likely to indicate spam.

k. Text Preprocessing Techniques

Before applying any machine learning model, the comment text must be cleaned and prepared using text preprocessing techniques. These include **tokenization** (splitting sentences into individual words), **stop words removal** (eliminating commonly used but irrelevant words like “the”, “and”), and **stemming or lemmatization** (reducing words to their root forms). These techniques help reduce noise in the data and improve the accuracy of the feature extraction and classification steps.

4. EXPERIMENTAL RESULTS

The proposed system was evaluated using a labeled dataset of YouTube comments, categorized as spam

and non-spam. After preprocessing and feature extraction using TF-IDF, the Logistic Regression classifier was trained and tested. The model achieved a high accuracy rate, with promising precision and recall scores, indicating its effectiveness in correctly identifying spam without misclassifying genuine comments. The confusion matrix showed minimal false positives and negatives, proving the reliability of the model in real-world YouTube environments. Overall, the system demonstrated strong performance in detecting spam comments automatically and efficiently.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = the\ AP\ of\ class\ k$$

$$n = the\ number\ of\ classes$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

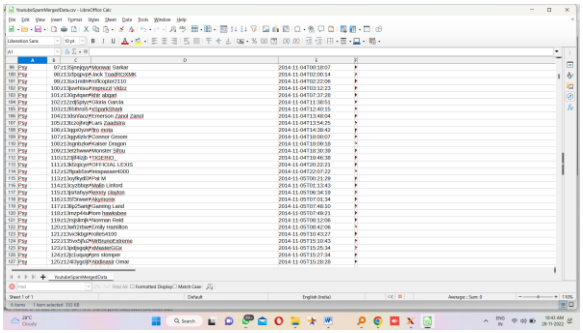

Fig.1. dataset

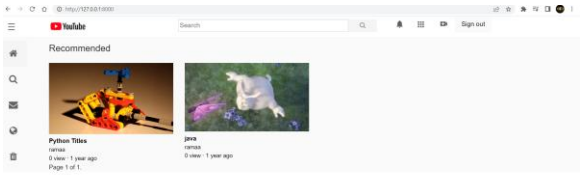

Fig: home page

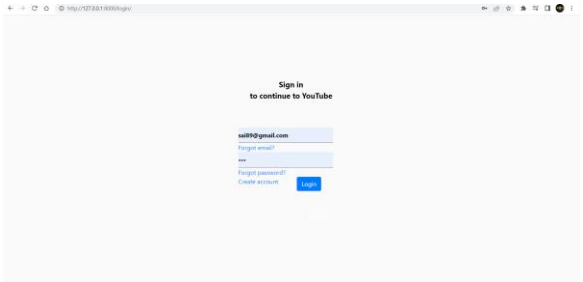

Fig: login form

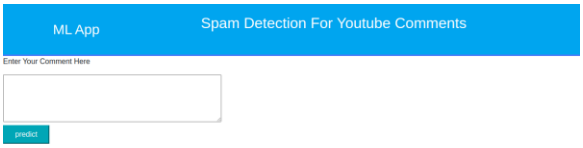

Fig.2. predicted results

5. CONCLUSION

The major goals of this study are to increase the accuracy of ham comments and steer clear of YouTube spam comments. This study will provide a fresh reference point for people interested in YouTube spam comments to use when making next comparisons. YouTube spam comments are assembled using social network data.

6. FUTURE SCOPE

Since there was no one approach that yielded the best results for all datasets, we may infer that an ensemble of classification techniques can surpass single classifiers in the future. Given their short duration and abundance of idioms and abbreviations, we want to preprocess the conversations using text normalisation and semantic indexing techniques. About TubeSpam, we want to build browser plug-ins that would eliminate spam from the video-sharing website straight at the source.

REFERENCES

[1] G. Mishne, D. Carmel, and R. Lempel, “Blocking blog spam with language model disagreement,” in Proc. of the 1st AIRWeb, Chiba, Japan, 2005, pp. 1–6.

[2] R. M. Silva, T. A. de Almeida, and A. Yamakami, “Artificial Neural Networks for Content-based Web Spam Detection,” in Proc. of the 2012 ICAI, Las Vegas, NV, EUA, 2012, pp. 1–7.

[3] C. Romero, M. Valdez, and A. Alanis, “A comparative study of machine learning techniques in blog comments spam filtering,” in Proc. of the 6th WCCI, Barcelona, Spain, 2010, pp. 63–69.

[4] T. C. Alberto and T. A. Almeida, “Aprendizado de maquina aplicado ´ na detecc¸ao autom ˜ atica de

coment ´ arios indesejados," in ´ Anais do X Encontro Nacional de Inteligencia Artificial e Computacional (ENIAC'13) ˆ , Fortaleza, Brazil, 2013.

[5] Z. Li and H. Shen, "Soap: A social network aided personalized and effective spam filter to clean your e-mail box," in INFOCOM, 2011 Proceedings IEEE, April 2011, pp. 1835–1843.

[6] T. Almeida, J. Almeida, and A. Yamakami, "Spam filtering: How the dimensionality reduction affects the accuracy of naive bayes classifiers," Journal of Internet Services and Applications, JISA'11, vol. 1, no. 3, pp. 183–200, 2011.

[7] J. M. Gomez Hidalgo, T. Almeida, and A. Yamakami, "On the Validity ´ of a New SMS Spam Collection," in Proc. of the 11st ICMLA, vol. 2, Miami, FL, EUA, 2012, pp. 240–245.

[8] T. P. Silva, I. Santos, T. A. Almeida, and J. M. Gomez Hidalgo, ´ "Normalizac¸ao Textual e Indexac ˜ ¸ao Sem ˜ antica Aplicadas na Filtragem ˆ de SMS Spam," in Proc. of the 11st ENIAC, Sao Carlos, Brazil, 2014, ˜ pp. 1–6.

[9] G. Mishne and N. Glance, "Leave a reply: An analysis of weblog comments," in Proc. of 3rd WWE, Edinburgh, UK, 2006, pp. 1–8.