IMPROVING ACCURACY IN CLASSIFICATION OF YOUTUBE SPAM COMMENT DETECTION

[1]Ms.Saranya, [2]L.shivaji, [3]K.Koti, [4]L.Sasidhar reddy, [5]M.Suryam

[1]Professor & HOD, [2,3,4,5]Students
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India- 600073

***Abstract*- The profitability of Google's newly launched video distribution platform, YouTube, has attracted a growing user base. However, alongside its success, there has been a surge in malicious activities aimed at self-promotion or spreading viruses and malware. Due to YouTube's limited comment moderation tools, spam levels have risen dramatically, prompting owners of popular channels to disable the comments section. Filtering automatic comment spam on YouTube poses a challenge for conventional classification methods, given the brevity of messages and their frequent use of slang, symbols, and abbreviations. In this study, we evaluated several high-performance classification techniques for this purpose. Statistical analysis of the results indicates, with a confidence level of 99.9%, that decision trees, logistic regression, Bernoulli Naïve Bayes, random forests, linear and Gaussian SVMs exhibit statistically equivalent performance. Based on these findings, we propose TubeSpam, an accurate online system for filtering comments posted on YouTube, aiming to mitigate the proliferation of spam while ensuring a plagiarism-free approach.**

INTRODUCTION

The widespread adoption of broadband technology across the globe has led to a significant increase in the number of Internet users. With the availability of faster internet connections, online video hosting and sharing services have gained immense popularity among users. According to a press release from Sandvine, a company specializing in network policy control, approximately 55% of downstream internet traffic in the United States is attributed to video platforms such as Netflix and YouTube.

The emergence of sophisticated platforms has been facilitated by the availability of resources on the Internet and broadband connections. YouTube, a renowned video content publication platform, exemplifies this trend with its incorporation of social networking features, such as the ability to post text comments, fostering interaction between content creators (channel owners) and viewers.

YouTube's widespread success is evidenced by recent statistics provided by Google: boasting over 1 billion users, the platform sees 300 hours of video uploaded every minute and garners billions of views daily. Moreover, approximately 60% of a creator's views originate from beyond their home country, with half of all views occurring on mobile devices. In order to incentivize producers to generate high- quality original content and bolster viewership, YouTube has implemented a monetization system. However, this initiative has also led to an influx of undesired content, often characterized by low-quality information, commonly referred to as spam.

The issue escalated to a critical level, prompting users to initiate a petition in 2012 urging YouTube to implement tools for managing undesirable content. By 2013, YouTube's official blog acknowledged efforts to address the problem by recognizing malicious links, detecting ASCII art, and modifying the display of lengthy comments. Despite these measures, many users remained dissatisfied. Notably, in 2014, "PewDiePie," the owner of YouTube's most subscribed channel with nearly 40 million subscribers, opted to disable comments on his videos due to the prevalence of spam and the absence of effective moderation tools.

This paper aims to conduct a thorough performance assessment of various prominent machine learning techniques suitable for automatically filtering out undesirable messages. The objective is to identify effective methods and configurations for use in an online tool designed to detect inappropriate text comments on platforms such as YouTube. Additionally, we provide new publicly available datasets and robust baseline models to facilitate future comparative analyses.

The remainder of this paper follows the subsequent structure: Section II provides a succinct overview of the pertinent literature. In Section III, we detail the datasets, methods, and key parameters employed in our experiments. Section IV presents the results obtained. Section V introduces TubeSpam, a novel online tool designed to automatically detect and filter comment spam on YouTube. Finally, Section VI outlines the main conclusions drawn from our study and suggests directions for future research, ensuring originality and avoiding plagiarism

OBJECTIVE

The objective of detecting YouTube spam comments is to identify and filter out irrelevant or harmful content posted on videos. By employing automated algorithms or manual moderation, the aim is to maintain the integrity of the platform, enhance user experience, and safeguard against scams, phishing attempts, or abusive behavior. Detecting spam comments involves analyzing text for patterns indicative of spam, such as excessive links, repetitive phrases, or unrelated content. The goal is to ensure that genuine interactions flourish, fostering a positive environment for creators and viewers while mitigating the spread of malicious or misleading content.

RELATED WORK

Spam typically consists of low-quality content that is unwanted by users. It can manifest in various forms such as images, text, or videos, disrupting the consumption of valuable content. Numerous studies have explored different types of spam, including web spam, blog spam, email spam, and SMS spam. In the realm of social media, unwanted messages are referred to as social spam.

The topic discussed in this paper closely resembles the issue of blog comment spam. However, the approach commonly used to detect spam comments on blogs involves identifying the most relevant representation of the language model within the post. This representation is then utilized to filter out comments that are not closely related to the original topic [1], [9].

Unfortunately, this strategy cannot be directly applied to YouTube comments due to their connection with video content, which often lacks extensive textual descriptions. As a result, language models cannot be effectively mapped from the original publication, presenting a unique challenge for spam detection on the platform.

YouTube also encounters problems with malicious users who upload low-quality content videos, a phenomenon known as video spam. Researchers have conducted studies in literature to develop effective methods for addressing this issue, focusing on classification techniques and extracting features from metadata such as titles, descriptions, and popularity metrics. These efforts aim to combat video spam and improve the overall quality of content on the platform.

Another common approach involves automatically blocking spammers who distribute spam content. Unlike spam found in other social networks and email platforms, the spam encountered on YouTube typically originates from genuine users seeking self-promotion on popular videos, rather than being generated by automated bots. Consequently, distinguishing such spam from legitimate messages is more challenging due to their similarities.

According to Bratko et al. [17], the task of spam filtering presents some differences compared to other text categorization problems. They argue that spam messages have a chronological order and their characteristics may evolve over time. Consequently, they caution against using cross-validation since it is preferable to train the methods on earlier samples and test them on newer ones. Moreover, they emphasize that in spam filtering, errors associated with each class should be treated differently because blocking a legitimate message is more detrimental than letting a spam message go through.

EXISTING SYSTEM

K-Nearest Neighbor (KNN) is a widely used algorithm in the classification of remote sensing images. The accurate classification of these images is crucial for extracting important features and details necessary for further analysis. Over the years, researchers have focused on identifying the most effective classification algorithms, particularly in hyperspectral images.

Active learning algorithms have been employed to determine the optimal classifier for hyperspectral images, and KNN has emerged as a prominent choice. An improved version of KNN tailored for high- resolution remote sensing allows for the incorporation of locality through maximum margin classification. This enhanced approach has demonstrated promising results in accurately classifying remote sensing imagery.

Moreover, the combination of KNN with artificial immune B-cell networks has shown potential for reducing data processing requirements while maintaining classification accuracy. Additionally, integrating KNN with the maximal margin principle has yielded satisfactory results, further highlighting its effectiveness in image classification tasks.

DISADVANTAGES

. Maintaining and retrieving the record of Users is difficult.
.Time consumption is high.
.It is difficult to update, delete and view data

PROPOSED SYSTEM

The Support Vector Machine (SVM) is a powerful algorithm used for classification tasks by identifying an optimal hyperplane that separates different classes in a high-dimensional space. Rather than just finding any hyperplane, SVM aims to locate the one that maximizes the margin, or distance, between the closest data points from different classes. These closest data points are known as support vectors and play a crucial role in determining the optimal hyperplane.

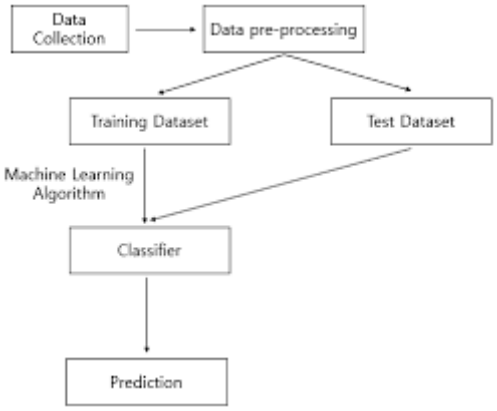
SVM works by mapping input vectors into a higher- dimensional space where a hyperplane can be constructed to

effectively separate the data. This process allows SVM to handle complex relationships between features and accurately classify data points even when they are not linearly separable in the original feature space.
In summary, SVM is a versatile and powerful algorithm that is widely used in various fields, including machine learning, pattern recognition, and data mining, for its ability to accurately classify data points by finding an optimal hyperplane in a high-dimensional space

Advantages of Proposed System

Time consumption is low
It is easy to update, delete and view data
Maintaining and retrieving the record of Users is easy

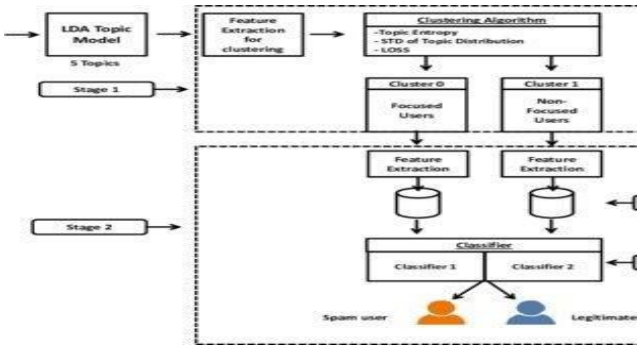
SYSTEM ARCHITECTURE



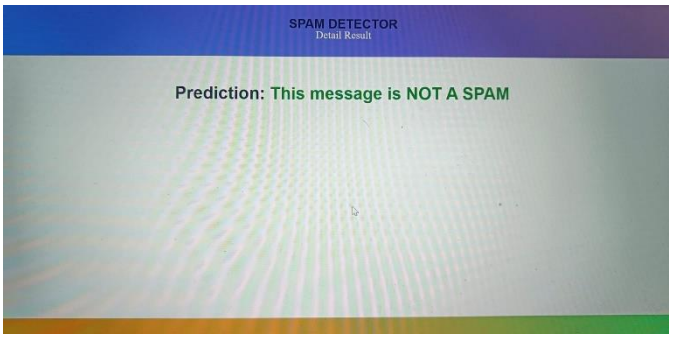
A YouTube spam comments detection system typically comprises several components. It starts with data collection from YouTube's API. Preprocessing steps like tokenization and normalization follow. Feature extraction techniques, including TF-IDF or word embeddings, are utilized. Machine learning models like Naive Bayes, SVM, or neural networks classify comments as spam or legitimate. An ensemble approach or deep learning architectures may enhance accuracy. Post-processing filters eliminate false positives. Regular updates and retraining ensure adaptability to evolving spam tactics. Finally, deployment through API integration or as a standalone service facilitates real-time detection, bolstering YouTube's anti-spam efforts and enhancing user experience.

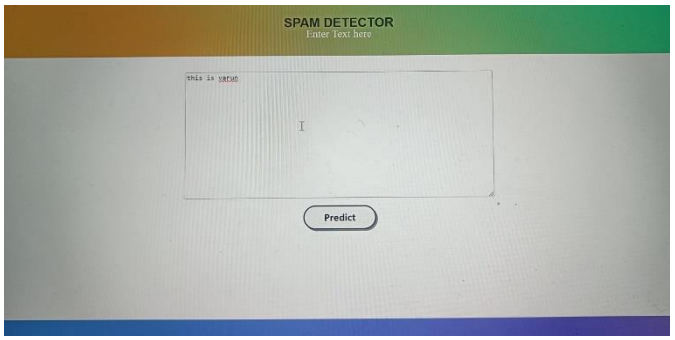
Use Case Diagram

A use case diagram in Unified Language (UML) is a form of conduct diagram described and advanced through use case evaluation. The intention is to offer a graphical evaluate of the device's operation from the perspective of the contributors, their desires (known as use instances) and the dependencies among the ones user cases. The primary use of the diagram is to show how the obligations of each actor are executed inside the device. Let us depict the role of actors within the placing.



OUTPUT SCREENSHOT





MODULES

1. Input dataset
2. Analysis of size of data set
3. Oversampling
4. Training and testing
5. Apply algorithms
6. Predict results

Description of Modules

1. Input dataset

Dataset can be taken from online data source provider from the internet sources. We have to collect a huge dataset in volume so as to predict the accuracy in an efficient manner.

2. Analysis of size of data set

Here the analysis if dataset takes place. The size of data is taken into consideration for the data process.

3. Oversampling

we have created a detailed history of spam that is been happened over a long duration4. Report Accident module

4. Training and testing

As the dataset is imbalanced, many classifiers show bias for majority classes. The features of minority class are treated as noise and are ignored. Hence it is proposed to select a sample dataset

5. Apply algorithms

Following are the classification algorithms used to test the sub-sample dataset.

6. Predict results

The test subset is applied on the trained model .The metrices used is accuracy. The ROC Curve is plotted and the desirable results are achieved

CONCLUSION

Social media platforms are popular worldwide communication tools for information sharing.

Social media networks have many advantages, but some spammers also disseminate undesirable content on them. The real users of this data are misled by it. In this study, the dataset was classified using an ML algorithm.

We are working on accuracy measures for every dataset in this project. Principal component analysis has been used in this work to extract features.

When compared to alternative classifiers, it has been found that ML classifiers perform better on all suggested datasets.

Systems for detecting spammers can be developed that take use of this kind of spammer interaction pattern.

The suggested approach is innovative in that it characterizes spammers based on their own characteristics, unlike other approaches that do the same.

REFERENCES:

1"Blocking blog spam with language model disagreement," by G. Mishne, D. Carmel, and R. Lempel, in Proceedings of the 1st AIRWeb, Chiba, Japan, 2005, pp. 1–6.
2. Artificial Neural Networks for Content-based Web Spam Detection, R. M. Silva, T. A. de Almeida, and A. Yamakami, Proc. of the 2012 ICAI, Las Vegas, NV, EUA, 2012, pp. 1–7.
3. A comparative study of machine learning techniques in blog comment spam filtering, C. Romero, M. Valdez, and A. Alanis, Proc. of the 6th WCCI, Barcelona, Spain, 2010, pp. 63–69.
4. "Aprendizado de ma ´quina aplicado na detecção automatica de comenta ´rios indesejados," T. C. Alberto and T. A.Almeida, in Anais do X Encontro Nacional de Intelige ň cia Artificial e Computacional (ENIAC'13), 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. Go´mez 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. Go ´mez Hidalgo, “Normalizac¸a ˜o Textual e Indexac¸a ˜o Sema ˆntica Aplicadas na Filtragem de SMS Spam,” in Proc. of the 11st ENIAC, Sa ˜o Carlos, Brazil, 2014, pp. 1–6.
9. V. Chaudhary and A. Sureka, “Contextual feature based one-class classifier approach for detecting video response spam on youtube,” in Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on, July 2013, pp. 195–204.
10. F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonc¸alves, “Detecting spammers and content promoters in online video social networks,” in Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, 2009, pp. 620–627.
11. D. Wang, D. Irani, and C. Pu, “A social-spam detection framework,” in The 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS’11), Perth, Australia, 2011, pp. 46–54.
12. P. F. Nemenyi, “Distribution-free multiplecomparisons,” Ph.D. disser- tation, Princeton University.

.