YouTube Spam Filter Using Machine Learning

Prof. Swati Patil[1], Shrushti Pawar[2], Prathamesh Paware[3], Ashvini Kawade[4], Sahil Gargate[5]
[1] *Assistant Professor, Department of Computer Engineering, JSPM's Jaywantrao Sawant College of Engineering, Pune*
[2-5] *Student, Department of Computer Engineering, JSPM's Jaywantrao Sawant College of Engineering, Pune*

***Abstract*— The profit promoted by Google in its spick-and-span video distribution platform YouTube has attracted a growing scope of user community. However, such success has attracted malevolent people who want to promote their videos or bear viruses and malware. Since YouTube offers restricted tools for comment moderation, the spam volume is shockingly increasing that's leading homeowners of known channels to disable the comments section in their videos. Automatic comment spam filtering on YouTube might be a challenge even for established classification ways since the messages unit terribly short and sometimes rife with slangs, symbols, and elisions. We've tested a number of high-performance classification algorithms for this purpose during this project. The math analysis of results indicates that with 99.9% of confidence level Bernoulli Naive Bayes, Decision trees, Logistic Regression, Random forests, Linear and Gaussian SVM's area unit statistically equivalent. Therefore, it is vital to look out how to note these videos and report them before they're viewed by innocent user.**

Index Terms: **Machine learning, Random Forests, Logistic Regression, Bernoulli Naïve Bayes, Decision trees, linear and Gaussian SVMs.**

I.INTRODUCTION

In previous years of the pandemic, YouTube, a web video entertaining and also a social media platform is gaining recognition by people in whole world. People of all ages can enjoy the numerous types of video material available on this sitejust about all age groups are drawn to it, which makes YouTube a straightforward target for spammers. as an example, there are educational videos available for us to look at. This comprehensive and attractive environment provided by YouTube creates a chance for various spammers to form unrelated content geared toward users. These uninvited spam comments/messages are aimed to attack users by tempting them into clicking malicious sites which contain malware, phishing, and scams. YouTube has an amazing feature called "Remarks," which allows users to express their feelings about a video in the form of comments..

II. LITERATURE SURVEY

1. Spam is usually related to unwanted content with caliber info. They frequently appear as images, texts, or videos, obstructing the representation of engaging content. There unit of measurement several pieces of research related to spam in literature, like internet spam, blog spam, e-mail spam, and SMS spam.
2. On social networking sites, unwanted messages square measure named as social media spam, journal comment spam is that the most similar state of affairs. However, the most-known strategy to note a journal spam comment sometimes is to hunt out the only outline of language model in post-distribution, abuse that representation to channel less associated remarks to its unique subject.
3. Such a method can't be implemented on YouTube, since the comments square measure is related to video content with little or regardless of description, thus language models cannot be properly mapped from initial publication.
4. YouTube additionally faces malicious users that publish caliber content videos, it's noted as video spam. There unit of measurement some studies within the literature to hunt out economic ways during which to handle this activity through classification approach and have extraction from data, like title, description, and recognize numbers.

5. Another popular method is to automatically block spammers - people that spread spam. Unlike spam distributed through other social media platforms and email, spam on YouTube is typically made by genuine people promoting themselves through popular videos. As a result of their closeness to valid messages, such messages are more difficult to detect. Automatic spam filtering can also help with other duties. When spam samples were deleted before training a classifier, Severyn et al. found a substantial improvement in performance in the opinion identification test.
6. The spam filtering function differs slightly from similar text categorization problems, according to Bratko et al. They say that unwanted messages follow a chronological order, and that their properties may change as a result.
7. It also explains why cross validation isn't advised because older samples should be used to train the techniques, while fresh samples should be used to test them. Furthermore, faults associated with each class should be treated differently in spam filtering, because a blocked valid message is worse than unblocked spam.

III. PROBLEM STATEMENT

To design and develop an ML-based YouTube Spam Filter which is capable of processing data of comments of varied YouTube channels and classifying them into spam and legit comments supported the dataset on which various machine learnings models are applied

- Goals and Objectives:

The main objectives of developing this method are:
1. Training system on a dataset of hottest YouTube channels to label spam comments.
2. Testing system for the test dataset.
3. Testing system on actual spam data with a presentable program.

- Motivation:

Recently, YouTube used a monetization system to reward producers, to stimulate them to create high-quality original content and increase value to be seen. After the deployment of this technique, the platform was flooded by undesired content, usually of low-quality information referred to as spam. Among different forms of undesired content, YouTube is experiencing problems managing the massive volume of unwanted text comments posted by users that aim to advertise their videos or to disseminate malicious links to steal private data. The YouTube spam is directly associated with the attractive profit provided by the monetization system.

IV. PROPOSED SYSTEM

1. The task is divided into three stages: the initial, middle, and final stages.
2. Data Exploration, Data Cleaning, and Data Transformation are the first three stages.
3. Data modeling is included in the middle stage.
4. Data analysis is performed using three models: the KNN Algorithm, regression toward the mean, and SVM.
5. Data exploration is similar to initial data analysis, but instead of using traditional data management tools, it uses visual exploration to learn what's in a large dataset and also the properties of the data.
6. Data cleaning is the process of removing incorrect or incorrect records from a recordset, table, or database. It entails identifying incomplete, incorrect, inaccurate, or insignificant sections of the data and then restoring, changing, or deleting them.
7. Data transformation is that the process of conversion of knowledge from one format to a different, usually from one format of a source system into the desired format of a destination system.

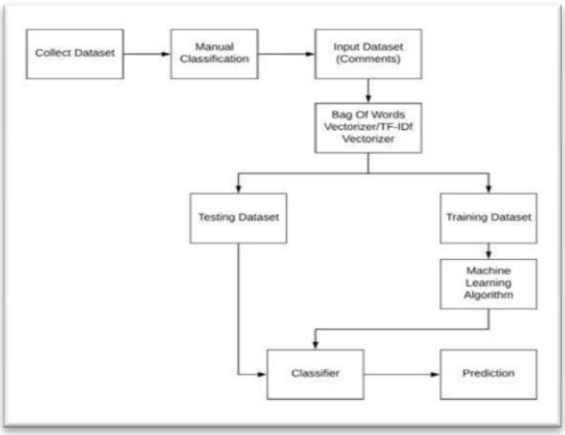

Fig. System Architecture

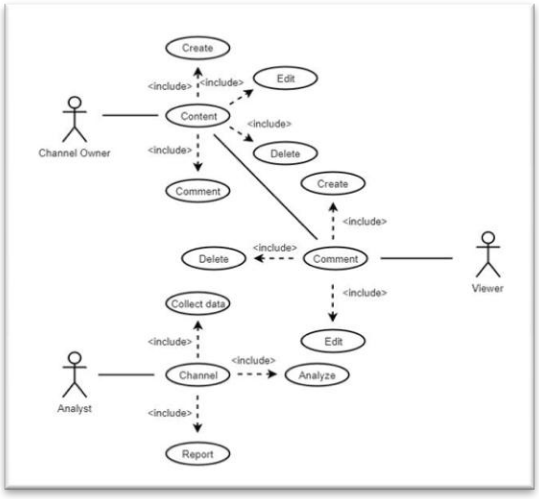

Fig. Use Case Diagram

Advantages:

1. When consideration of independent predictions catches on, Of course, the Naive Bayes classifier performs well as compared to other models.
2. Naive Bayes requires a limited amount **of** coaching data for the estimation of the test data. So, the training period is a smaller amount.
3. Naive Bayes is additionally easy to implement.

Applications:

1. Social Network.
2. Spam detection.

Outputs:

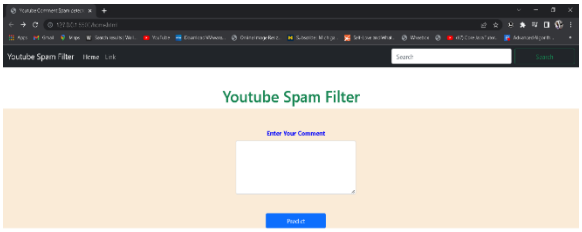

Fig. Enter comment

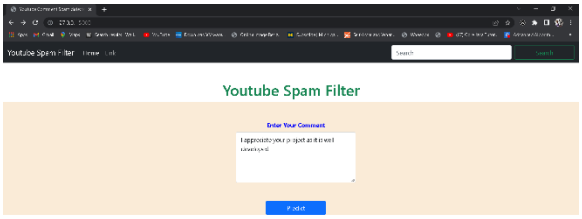

Fig.comment entered

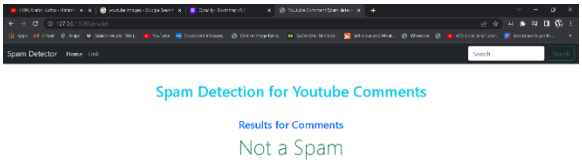

Fig. spam not detected

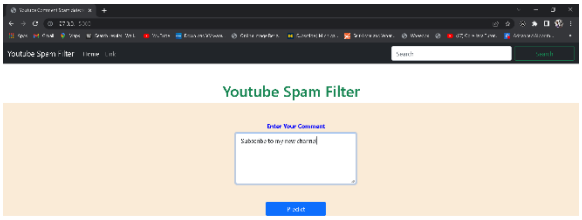

Fig.comment entered

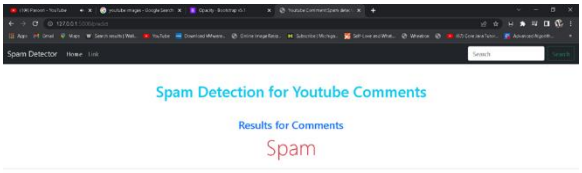

Fig. spam detected

V. CONCLUSION

Various strategies are used to categorize YouTube comments as spam and not spam (ham). This method has been tested with real-time YouTube comments and has produced an overall result that is eighteen times more accurate than the current method. Because the YouTube API is an open platform for all users, it will modify spammers' behavior over time. The YouTube spam function will not be constant on the planet; it will change rapidly.

REFERENCE

[1] Chao Chen, Jun Zhang, Yi Xie, and Yang Xiang, "An overall performance assessment of machine

learning-based streaming unwanted mail tweets detection," IEEE Transactions on Computational Social Machines, Vol. 2 No. 3, 2014.

[2] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "aiding the detection of fictitious money owing in large-scale social online services," in Proc. Symp. Netw. Syst. Des. (NSDI), 2013, pp. 197–210.

[3] "Spam filtering in Twitter using sender-receiver dating," in Proc. 14th Int. Conf. recent Adv. Intrusion Detection, 2010, pp. 301–317.

[4] "Detecting spammer on Twitter," by F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, in the 7th Annu. Collab. Electron. Unsolicited letter anti-abuse messaging Conf., Redmond, WA, united states of america, 2016.

[5] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honey pots + system studying," in Proc 33rd Int. ACM SIGIR Conf. Res. increase. Inf. Retrieval, pp 435-442, 2011.

[6] Nathan Aston, Jacob Liddle, and Wei Hu*, "Twitter Sentiment in Fact Streams using Perceptron," Journal of Computer and Communications, Vol. 2 No. 11, 2015.