Proceedings of the 7th International Conference on Intelligent Computing and Control Systems (ICICCS-2023)
IEEE Xplore Part Number: CFP23K74-ART; ISBN: 979-8-3503-9725-3

Machine Learning based Spam Comments Detection on YouTube

Hema Valpadasu
Dept. Of CSE
MLR Institution of Technology
(of JNTUH Affiliation)
Hyderabad, India
Hemav1248@gmail.com

P. Chakri
Dept. Of CSE
MLR Institution of Technology
(of JNTUH Affiliation)
Hyderabad, India
pagillachakri@gmail.com

Puli Harshitha
Dept. Of CSE
MLR Institution of Technology
(of JNTUH Affiliation)
Hyderabad,India
puliharshitha25@gmail.com

P. Tarun
Dept. Of CSE
MLR Institution of Technology
(of JNTUH Affiliation)
Hyderabad, India
tarunpasupuleti94@gmail.com

2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS) | 979-8-3503-9725-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICCS56967.2023.10142608

***Abstract*—YouTube provides only some tools for the modification of the comments in the comment section. Because of this, the volume of spam comments increasing rapidly. Using Machine Learning, the comments can be detected and prevented. There are a lot of approaches in ML to detect spam. It is often seen in applications like YouTube where people watch a lot of videos for so many purposes it can be for entertainment or learning and it provides a way for users to interact with the creators through the comment section. There exists a way where people post scam comments which are quite harmful. These comments can be dangerous and they can include links to other pages which can hack any information or data or steal any confidential details when a link is composed on that comment, in some cases it can also redirect to the page where it attracts people to earn money while playing a game and it is a scam and most people have actually lost their money by clicking on such type of links displayed via comment or messages. The purpose of this project is to detect all the spam comments that are being posted on the internet to avoid any scams or any unrelated information. In this study, Naive Bayes classification algorithm is used. The detection accuracy of this proposed system is 92.78%.**

Keywords—Spam comments, Machine learning, YouTube, Naive Bayes.

I. INTRODUCTION

There are a lot of content creators on YouTube platform. Every creator has their own content to post or stream. They get a lot of following for their content in the form of the SUBSCRIBERS and they get more VIEWS for that content. So, YouTube provided a comment section for every post that the creator posted on YouTube in order to know the opinions of the VIEWERS.Some viewers like the posted video and some might not like it. So, these users might post some negative or cursed comments. But there some other category of comments which are unwanted and unasked electronic messages known as spam comments. These spam comments are sent in a heavy or large amount. The dangerous threat of the spam is when the involved spam comment which leads to the inappropriate websites. So, by using the concept called machine learning we can predict and detect the spam. The algorithm that have been used for detecting is Naive Bayes algorithm which predicts which comments are spam and which not and various algorithm have been used for spam comment detection but however Naive Bayes is best suitable as it is faster compared to other algorithms and perform better probabilistic calculations.

1.1 What is a Spam Comment?

Comments that have the express intent of collecting personal data from readers, deceiving readers into leaving YouTube, or engaging in any of the banned actions listed above. Leave many, duplicate, or repetitive remarks that are not targeted.

II. EASE OF USE

Machine Learning has been expanding in various sectors including health, business, retail, finance and education. This project gives a clear cut understanding of the spam comments that are being posted in applications like YouTube using machine learning algorithms and techniques. This provides a better way to know which comments are spam and which are not automatically without the need of humans. We often came across situations facing like this receiving spam messages or comments on the internet and some which can harmful that can lead to another page which can potentially steal our data and some might scam our money. We detect comments that are spam by using machine learning algorithm called Naive Bayes which is a supervised classification algorithm. We classify the comments by spam and not spam, we classify them by grouping certain objects which are similar and share common characteristics.

III. SCOPE

This project works on detecting the spam comments that are being posted on applications like YouTube. We can see not

all comments that are posted have to be 100% real some can be fake that is can be a scam. In a application like YouTube, there provides a way for users to interact with the owners of the video by posting a comment. And which can be an open platform for anyone to post comments. Some of which can be spam that is unrelated to the information that is being posted. Some spam comments include link composed of messages when any person clicks on that link can redirect to new page that can be harmful can possibly contain virus or can steal confidential information or can steal money. These projects work by posting comments via a website and a button detect that predicts whether a comment is spam or not. The algorithm that have been used is Naive Bayes algorithm which and is considered to be the fasted and better for any kind of probabilistic calculations and can is easy to build. All the data cleaning and processing is done and exploratory data analysis is done to clearly understand the data and then the data is split for training and testing purposes and finally the data is fit into the model for model building for detection.

IV. Survey and Research

4.1 Using K-Nearest Neighbor and Support Vector Machine, to detect spam comments on YouTube.

The project to provide a framework for detecting YouTube videos using K-Nearest Neighbor and Support Vector Machine (SVM) (KNN). This study involves five (5) phases, including data collection, pre processing, feature selection, classification, and detection. Use of Weka and Rapid Miner is made for the experiments.

4.2 Spammer Detection: A Study of Spam Filter Comments on YouTube Videos.

The goal of this study is to identify users who leave spam comments, those who do so with the aim of promoting themselves, or users whose comments are irrelevant to the given video.

4.3 Analysis and Classification of User Comments on YouTube Videos

We group user remarks on the video-sharing website YouTube according to how closely they relate to the information provided in the description of the uploaded video. Positive and unfavorable comments are further separated based on polarity analysis.

4.4 Detection of Spam in YouTube Comments Using Different Classifiers

In this study, we used datasets of YouTube comments from five well-known singers to identify spam using both normal and artificial neural network-based classifiers. The suggested method compares the classifiers' deduced results and proposes the top classifiers for identifying spam comments.

4.5 Tube Spam: Comment Spam Filtering on YouTube

Since there are few tools available on YouTube for comment moderation, the amount of spam is shockingly rising, forcing renowned channel owners to turn off the comments feature on their videos. According to the statistical analysis of the data, decision trees, logistic regression, Bernoulli Naive Bayes, random forests, linear SVM, and Gaussian SVM are statistically equivalent with a degree of confidence of 99.9%.

3.6 Detection of Spam in YouTube Comments Using Different Classifiers

These days, thousands of videos are shared on YouTube every minute, and users instantly begin to like and comment. Millions of comments are left on some famous and viral videos; some of these comments are positive and healthy, while others are spam, abusive, and occasionally include a URL for commercial advertising or a site redirect. In this study, we used datasets of YouTube comments from five well-known vocalists to identify spam using both normal and artificial neural network-based classifiers. The suggested method compares the classifiers' deduced results and recommends the top classifiers for identifying spam comments.

V. Existing System

Google safe browsing and YouTube book marker are some tools used by the YouTube to protect the user from spammers. These tools can block malicious links but cannot save the user in the early stage,in real-time. This is because they support the SVM and K-nearest algorithms for detecting the spam. These algorithms cannot predict the accuracy rate. In some cases NN algorithm is also used.

Algorithms used are

1. SVM - Support Vector Machine.
2. K-nearest neighbour algorithm.
3. NN- Neuaral Network algorithm.

5.1 Limitations of Existing System.

The limitations of the current algorithms are listed below since they are all used in the codes of the existing systems stated above.

1. **SVM limitations.**
   i.Large data sets are not a good fit for the SVM method.
   ii.When the data set has more noise, such as when the target classes overlap, SVM does not work very well.
2. **KNN limitations.**
   i. The prediction step may take a while if there are a lot of data.
   ii. High RAM is necessary because all of the training data.
3. **NN limitations.**
   i . Computationally expensive.
   ii. Duration of development.

VI. Proposed System

The spam comments can be detected using machine learning techniques. The spam comments are the potential in spreading malware throughout the system. These exploit the machines of the user. The naive Bayes algorithm is used to predict and detect spam comments. Using this algorithm will give accurate results and takes less time in predicting the spam. The algorithm that we've been using is a naive Bayes classifier which works on any classification problem. Classification problems include yes or no; true or false, it can be a binomial classification or multinomial classification.

A. Equation

The Naive Bayes algorithm works on the basis of the Bayes theorem which is

$$P(A/B) = P(B/A)*P(A)/P(B).$$

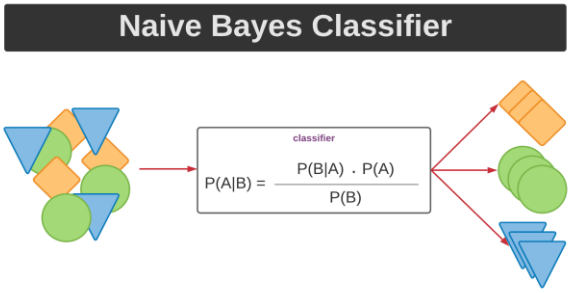
Here,

P(A/B) is the posterior probability of class A with respect to target B.

P(B/A) is the likelihood which is the probability of the predictor given class.

P(A) is the prior probability of class.

P(B) is the probability of Evidence.



This equation predicts the result based on the occurrence of any condition by taking its probabilistic value under any certain condition. It can be explained if there is data on whether we can play tennis or not and along with the data there is provided many features and our dependent value would be whether we can play tennis or not. The features include outlook, temperature, humidity, and wind. By taking the probability of the features and including that in a condition that describes the feature we can be able to predict if we can play tennis or not.

B. Accuracy Values of algorithms.

Classifier	Accuracy
Naïve Bayes	92.78%
Decision Tree	90.38%
Logistic	88.32%
Support Vector Machine	74.40%
Random Forest	73.54%
Random Tree	52.92%
k-Nearest Neighbor	56.70%

As seen in the above table, the Naive Bayes algorithm has the greatest accuracy value, with a percentage of 92.78%. So, the Naive Bayes is employed.

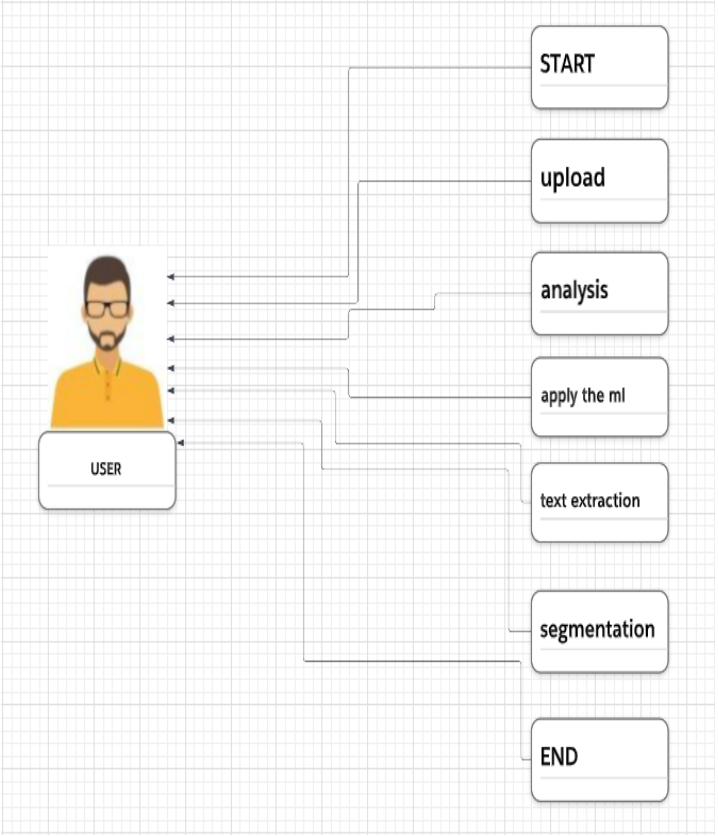
6.1 Selection of the Algorithm.

Initially, all the algorithms mentioned in the above table are tested for their accuracy. So, when the accuracy values were taken, Naive Bayes algorithm got the highest accuracy with a value of 92.78% . So, Naive Bayes algorithm is considered.

VII. DIAGRAMS

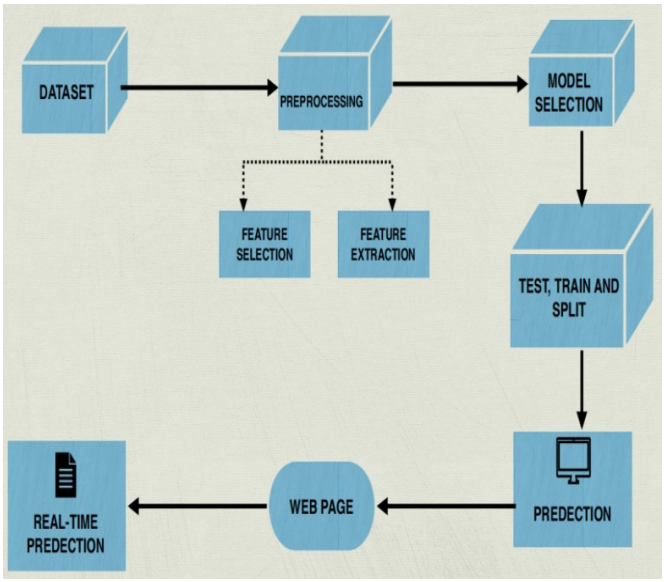
A. Use-Case Diagram

In the Unified Modelling Language (UML), a use case diagram is a specific kind of behavioral diagram that results from and is defined by a use-case analysis. So, the system is originally started by the user. The user will upload the input or comment when the system starts. Regarding the current data set, this input comment is examined. Then the algorithm for machine learning is used. The text is then separated into segments. This split information is then analyzed to determine whether or not the input comment is spam.



B. System Architecture Diagram

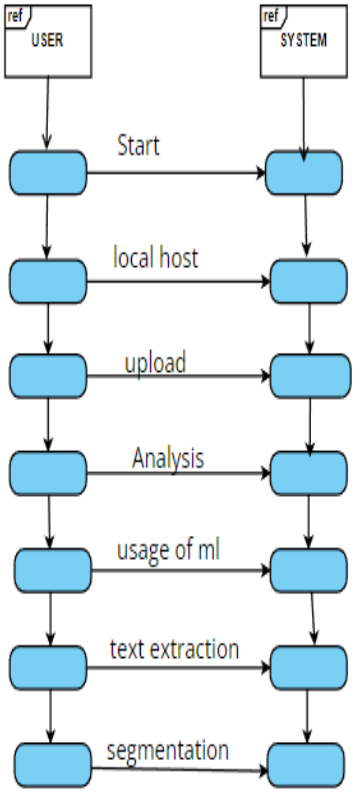
An architecture diagram shows the relationships between the system's parts and how the system functions. The data set made up of YouTube comments is first uploaded. The preprocessing of this data set comes next. The operations of feature extraction and feature selection are carried out during preprocessing. A model is chosen following this phase. The system is then trained using the uploaded data set, and testing is also done at this point. The system's front end is used to submit a new input comment at that point. Then, based on the comments in the data set, this new input comment is analyzed and classified as spam or not. The system's front end page will then display these results.



C. Sequence Diagram

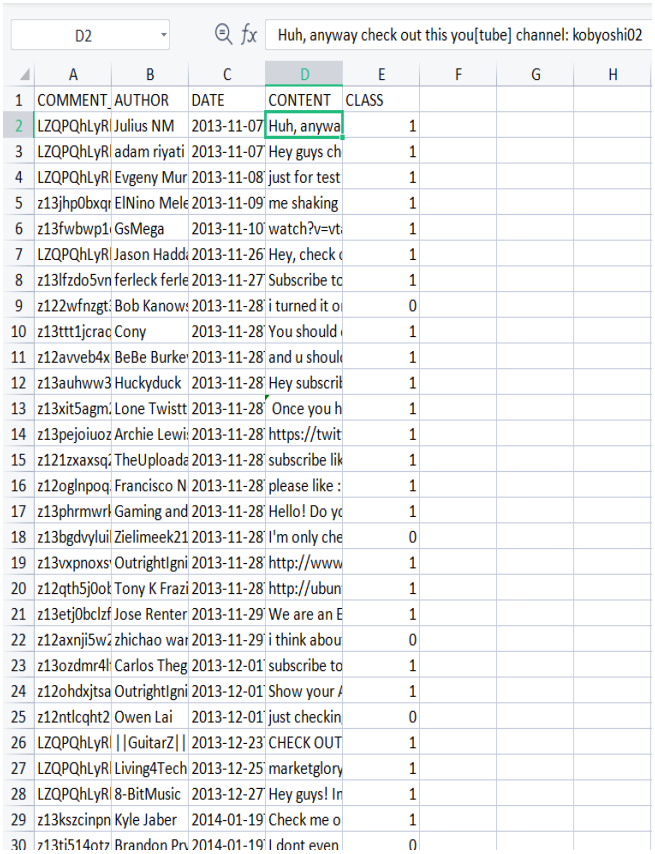
Because it illustrates the interactions between a group of items and the order in which they take place, a sequence diagram is a sort of interaction diagram. Software engineers and business experts use these diagrams to comprehend the specifications for a new system or to describe an existing procedure.

A type of interaction diagram used in the Unified Modelling Language (UML), a sequence diagram demonstrates how and when processes interact with one another. It is a Message Sequence Chart construct. Event diagrams, event situations, and timing diagrams are other names for sequence diagrams.



VIII. DATA SET

A collection of YouTube comments that have been classified as either 1 (spam) or 0 (genuine) serves as the data-set used for training and testing the algorithm. The data-set only contains about 2000 tuples, which is a small number. All this labeling is done manually.

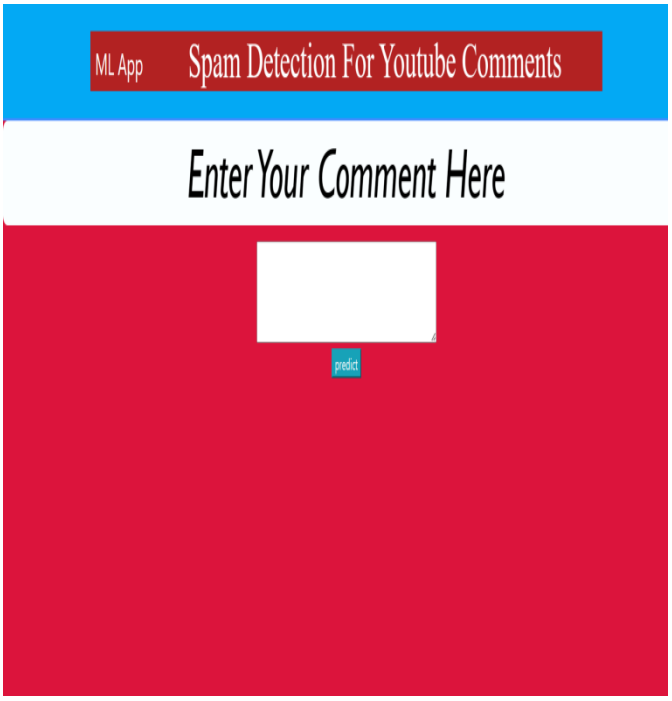


D2 | Huh, anyway check out this you[tube] channel: kobyoshi02

	A	B	C	D	E	F	G	H
1	COMMENT_	AUTHOR	DATE	CONTENT	CLASS			
2	LZQPQhLyR	Julius NM	2013-11-07	Huh, anywa	1			
3	LZQPQhLyR	adam riyati	2013-11-07	Hey guys ch	1			
4	LZQPQhLyR	Evgeny Mur	2013-11-08	just for test	1			
5	z13jhp0bxqr	ElNino Mele	2013-11-09	me shaking	1			
6	z13fwbwp1	GsMega	2013-11-10	watch?v=vt	1			
7	LZQPQhLyR	Jason Hadd	2013-11-26	Hey, check	1			
8	z13lfzdo5vn	ferleck ferle	2013-11-27	Subscribe to	1			
9	z122wfnzgt	Bob Kanows	2013-11-28	i turned it o	0			
10	z13ttt1jcrac	Cony	2013-11-28	You should	1			
11	z12avveb4x	BeBe Burke	2013-11-28	and u shoul	1			
12	z13auhww3	Huckyduck	2013-11-28	Hey subscri	1			
13	z13xit5agm	Lone Twistt	2013-11-28	Once you h	1			
14	z13pejoiuoz	Archie Lewi	2013-11-28	https://twit	1			
15	z121zxaxsq	TheUploada	2013-11-28	subscribe lik	1			
16	z12oglnpoq	Francisco N	2013-11-28	please like :	1			
17	z13phrmwrl	Gaming and	2013-11-28	Hello! Do yo	1			
18	z13bgdvylui	Zielimeek21	2013-11-28	I'm only che	0			
19	z13vxpnoxs	OutrightIgni	2013-11-28	http://www	1			
20	z12qth5j0ol	Tony K Frazi	2013-11-28	http://ubun	1			
21	z13etj0bclzf	Jose Renter	2013-11-29	We are an E	1			
22	z12axnji5w	zhichao wa	2013-11-29	i think abou	0			
23	z13ozdmr4l	Carlos Theg	2013-12-01	subscribe to	1			
24	z12ohdxjtsa	OutrightIgni	2013-12-01	Show your	1			
25	z12ntlcqht2	Owen Lai	2013-12-01	just checkin	0			
26	LZQPQhLyR	||GuitarZ||	2013-12-23	CHECK OUT	1			
27	LZQPQhLyR	Living4Tech	2013-12-25	marketglory	1			
28	LZQPQhLyR	8-BitMusic	2013-12-27	Hey guys! In	1			
29	z13kszcinpn	Kyle Jaber	2014-01-19	Check me o	1			
30	z13ti514otz	Brandon Pr	2014-01-19	I dont even	0			

IX. Front End Design

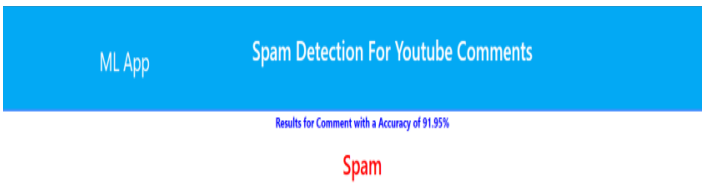
The layer above the back end is the front end and it includes all software or hardware that is part of a user interface. Human or digital users interact directly with various aspects of the front end of a program, including user-entered data, buttons, programs, websites, and other features. It is often afford as User Interface. In this project, the front end consists of a text block and a button saying detect. We have to enter the comment in comment box and click on detect button.This gives whether the comment is spam or not.



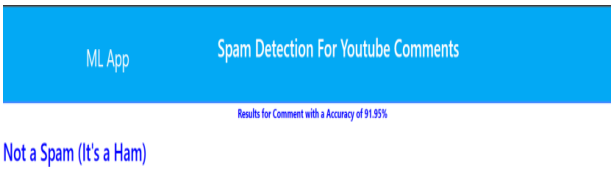
X. Results

This project's findings indicate if a comment is spam or ham. As a result, when the input is given and the programme is told to make a prediction, it will deliver the result as spam if the comment is deemed spam and as not spam if it is deemed a ham comment.

The comment *Hey, check out my new website!! This site is about kids stuff. kidsmediausa . com* is classified as spam. So, the output will be show as spam.



And, the comment *i think about 100 millions of the views come from people who only wanted to check the viewsï»¿* is classified as Not Spam. So, the output will be delivered as Not spam.



XI. CONCLUSION

10.1 Conclusion.

Several methods are employed to categorize comments as spam or not spam. This strategy is 18% more effective than the previous strategy. Every user on YouTube has access to its open platform. There may be a shift in the spammers' behaviour over time.

10.2 Future Scope.

This project aims to eliminate unwanted spam comments from YouTube and enhance ham comments with high accuracy. The project's output enhances the findings for future comparison and serves as a baseline for anyone interested in YouTube spam comments. Information on spam comments on YouTube is gathered from social networking sites. A data mining technique will be used to compare the accuracy of the results utilizing this data.

REFERENCES

[1] Sah, U. K., & Parmar, N. (2017). An approach for Malicious Spam Detection in Email with comparison of different classifiers.

[2]Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015, December). Tubespam: Comment spam filtering on youtube. In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on (pp. 138-143). IEEE.

[3] Alsaleh, M., Alarifi, A., Al-Quayed, F., & Al-Salman, A. (2016). Combating comment spam with machine learning approaches. Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015, 295–300. https://doi.org/10.1109/ICMLA.2015.192

[4] Scheltus, P., Dorner, V., & Lehner, F. (2013). Leave a Comment! An In-Depth Analysis of User Comments on YouTube. Wirtschaftsinformatik, 42.

[5] A. Kantchelian, J. Ma, L. Huang, S. Afroz, A. Joseph, J. D. Tygar, Robust detection of comment spam using entropy rate, in: Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence, AISec '12, ACM, New York, NY, USA, 2012, pp. 59-70. doi:10.1145/2381896.2381907.

[6] S. Aiyar and N. P. Shetty, "N-gram assisted Youtube spam comment detection", Proc. Comput. Sci., vol. 132, pp. 174-182, Jan. 2018.

[7] A. Kantchelian, J. Ma, L. Huang, S. Afroz, A. Joseph and J. D. Tygar, "Robust detection of comment spam using entropy rate", *Proc. 5th ACM Workshop Secur. Artif. Intell. (AISec)*, pp. 59-70, 2012.

[8] A. Madden, I. Ruthven and D. Mcmenemy, "A classification scheme for content analyses of Youtube video comments", *J. Documentation*, vol. 69, no. 5, pp. 693-714, Sep. 2013.

[9] A. Severyn, A. Moschitti, O. Uryupina, B. Plank and K. Filippova, "Opinion mining on Youtube", Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Long Papers), vol. 1, pp. 1-10, 2014.

[10] M. Z. Asghar, S. Ahmad, A. Marwat and F. M. Kundi, "Sentiment analysis on Youtube: A brief survey", *arXiv:1511.09142*, 2015, [online] Available: http://arxiv.org/abs/1511.09142.

[11] T. C. Alberto, J. V. Lochter and T. A. Almeida, "TubeSpam: Comment spam filtering on Youtube", *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, pp. 138-143, Dec. 2015.

[12] A. U. R. Khan, M. Khan and M. B. Khan, "Naïve multi-label classification of Youtube comments using comparative opinion mining", *Proc. Comput. Sci.*, vol. 82, pp. 57-64, Jan. 2016.