YOUTUBE SPAM COMMENTS DETECTION USING DEEP LEARNING

Dr. Deepak N A
Dept. of CSE
Associate Professor
RVTIM
Bengaluru, India

Supraj P Kashyap
Dept. of CSE
RVITM
Bengaluru, India

Vivekananda R Bhat
Dept. of CSE
RVITM
Bengaluru, India

Sowndarya R
Dept. of CSE
RVITM
Bengaluru, India

Madala Vinay Kumar
Dept. of CSE
RVITM
Bengaluru, India

Abstract - Spammers have come to acknowledge that it is so easy to convince individuals to participatein disastrous ways of behaving by posting spam messages in the video remarks area since online informal communities have worked on in quality. Forthis task, remarks from YouTube were accumulated and spam was recognized. To help accompanying deterring spammers, Google Safe Perusing and YouTube Bookmaker progresses identify and channelmarketing mail on YouTube. Albeit these uses will hold hurtful networks from being taken to, they achieved't safeguard the customer steadily. Consequently, a variety of approaches have been taken by academics and business people to create a social network platform free of spam. In the survey, the spam comment detection method was used with deep learning estimates like BERT, RNN, GRU, LSTM, and LSTM + GRU. With the assistance of a Brain Organization, we can arrive at a precision of 91.65% and beat the ongoing game-plan by practically 18%.

Index Terms – *BERT, RNN, GRU, LSTM and LSTM +GRU, Deep Learning*

I. INTRODUCTION

Google purchased YouTube in 2006, the year it began. YouTube has grown a great deal as a stage for video content since web content has changed to video. North of 400 hours of program are moved to YouTube all importance, and 4.5 heap records are noticed each importance [1]. Videos can be easily viewed and shared by users without restriction. The number of people who use personal media has increased, and some of them have become online influencers as a result of this ease of access. YouTube makers who have in excess of 1,000 supporters and 4,000 hours of survey time in the past a year might adapt [2]. As a result of this, spam remarks are passed on in famous recordings to advance their channels or recordings. A few creators incapacitated the remark segment because of savagery, like political comments, impolite discourse, or offending remarks irrelevant to their recordings. Notwithstanding the habit that YouTube has allure own unsolicited call detaching part, skilled are still unsolicited call comments that sneak past the whole world's announcement. In this survey, we examine all YouTube spam comments and recommend the Cascaded Ensemble Machine Learning Model Detecting Spam Comments. intend to disrupt performance of the model. In previous surveys, different ML ordering was used to view and evaluate the performance of marketing mail comments for each dataset..

II. LITERATURE SURVEY

Spam is nothing but irrelevant content with low quality information send over the internet typically to a large number of users for the purpose of spreading malware.This study proposed an efficient spam detection approach using a pre trained bidirectional encoder representation from transformer (BERT) and machine learning algorithms to classify ham or spam emails. Email texts were fed into the BERT, and features obtained from the BERT outputs were used to represent the texts. The proposed model was tested using two public datasets in the experiments. Theresults of the evaluation metrics demonstrate that the logistic regression algorithm achieved the best classification performance in both datasets. They alsojustified the efficient ability of the proposed model in detecting spam emails[1].

Natural language processing (NLP) enhanced the models' accuracy. In this work, the effectiveness of word embedding in classifying spam emails is introduced. Pre-trained transformer model BERT (Bidirectional Encoder Representations from Transformers) is fine-tuned to execute the task of detecting spam emails from non-spam (HAM). Several models and techniques to automatically detect spam emails have been introduced and developed yet non showed 100% predicative accuracy. BERT uses attention layers to take the context of the text into its perspective. Results are compared to a baseline DNN (deep neural network) model that contains a BiLSTM (bidirectional Long Short Term Memory) layer and two stacked Dense layers. In addition results are compared to a set of classic classifiers k-NN (k- nearest neighbors) and NB (Naive Bayes). Two open- source data sets are used, one to train the model and the other to test the persistence and robustness of the model against unseen data.

The main issue of spam is that it can download malicious files which can attack the computers, smartphones and networks, utilize network bandwidth and storage space, degrades email servers and can cause attacks in our devices like spyware, phishing and ransomware. In the existing approach, an exploratory analysis of supervised machine learning algorithms has done and the performance has been evaluated. The drawback of existing approach is that the performance of supervised machine learning

algorithms decreases as we increase the size of the dataset. An efficient spam detection using recurrent neural networks using the BiGRU model has been proposed. By implementing this, it has been achieved with better accuracy of 99.07%. From this, it is concluded that BiGRU model has better performance than existing approaches.

The significant amount of SPAM emails that are derived from various botnets worldwide affect thelimited capacity of mailboxes. They affect the time required for identifying spam emails and addressing them. Till today, the email spam detection is stillconsidered a challenging process. That is because the email spam is still happening a lot. It is because the detection still needs much improvement. Therefore,the researcher of this study develops a Recurrent Unit Recurrent Neural Network (GRU- RNN) with SVM for Bot Spam email detection. The developed approach got tested by employing the Spam base dataset.

By using text input for training, deep learning that rely on self- attention mechanisms become crucial. By effectively ca t e g o r i z i n g spam or ham emails in real-time circumstances, this research showed how to create a new universal spam detection model using pre- Google's Bidirectional Encoder Representations from Transformers (BERT) base uncased models with four datasets. Different techniques were employed to train individual models for the Enron, Spamassain, Lingspam, and Spam text message categorization datasets, and a single model was created with respectable results on all four datasets. Four datasets were used to train the Universal Spam Detection Model (USDM), which then used hyper parameters from each model. The identical hyper parameters used in each of these four individual models were used to fine-tune the combined model.By Vijay Srinivas Tida and Sonya Hy Hsu[3]

The profit promoted by Google in its spick-and- span video distribution platform YouTube has attracted a growing scope of users. Since YouTube offers restricted tools for commentmoderation, the spam volume is shockingly increasing that is leading house owners of known channels to disable the comments section in their videos.Automatic comment spam filtering on YouTube could be a challenge even for established classification ways since the messages squaremeasure terribly short and infrequently rife with slangs, symbols, and elisions. During this work, we've evaluated many top-performance classification techniques for such purposes.The applied math analysis of results indicates thatwith 99.9% of confidence level Bernoulli Naive Bayes, Decision trees, Logistic Regression, Random forests, Linear and Gaussian SVM'ssquare measure statistically equivalent. Therefore, it's important to search out some way to noticethese videos and report them before they're viewed by innocent user[2].

Spammers frequently use these various sorts of spam to get money, including comments, emails, search results, and personal messages.Different machine learning methods, including neuralnetworks, have attempted to distinguish between spam and legitimate SMS texts. As opposed to conventional methods where features are chosenafter analysis for classification, these techniquescan automatically learn high level features fromraw data. In this research paper, we offer a new approach for detecting spam and ham from the "Spam SMS Collection" dataset, which is availableat the UCI machine learning repository, combining recurrent neural networks (RNN) and long short- term memories (LSTM) with Keras models and Tensorflow backend. Tokenization, TF-IDF vectorization, and stopword removal were essentialpreprocessing steps for the dataset. 98% overallaccuracy is attained, which is an improvement above existing machines.

In the current method, supervised machine learningalgorithms have undergone exploratory study, and their performance has been assessed. The disadvantage of the current method is that supervised machine learning algorithms perform worse as the size of the dataset grows. An effective spam detection method utilizing recurrent neural networks and the BiGRU model has been developed to get over these shortcomings. This wasdone, and a greater accuracy of 99.07% was obtained. This leads to the conclusion that the BiGRU model performs better than current methods. SMS dataset is translated into vector system to be applied into biGRU model. The paper compared traditional ML model like KNN, Random Forest with RNN model which showedaccuracy of 99% significantly better than the traditional ML algorithm.

III. METHODOLOGY

As of late, casual internet based networks like Facebook and YouTube have become increasingly more incorporated into day to day existence. Individuals utilize online entertainment as a virtual local area stage to speak with loved ones, as well as to communicate perspectives and thoughts by means of sites. These platforms have a lot of users and are easy prey for spammers because of this growing trend. YouTube has turned into the most famous casual local area among young people. Numerous restorative illustrations, for instance, have been sent off by bloggers alluded to as "excellence masters" or "magnificence powerhouses," with the heft of the watchers being juvenile females. 200 million users upload 400 million new videos to YouTube each day. Spammers are also able to create irrelevant content aimed at people thanks to YouTube's extensive ecosystem. These unrelated or unasked-for communications aim to deceive users into visiting malware, phishing, and scam websites.

Drabacks

1. produce content that is unrelated to users
2. unrelated or uninvited

In this review, marketing mail is famous while examining YouTube remarks. Google Safe Perusing and YouTube Bookmaker novelty find and channel marketing mail on YouTube to stop spammers. Theseitem will discourage risky associations, yet they won'tdefend the client as quick as possible dynamically. Subsequently, companies and scholastics haveinvestigated different ways of laying out sans spam informal communication organizations. Deep learning estimates like BERT, RNN, GRU, LSTM, and LSTM + GRU were used to conduct the survey for the spam comment detection method.

Benefits

1. Possibility
2. Present moment

3. Methods for deep learning with a high rate of detection

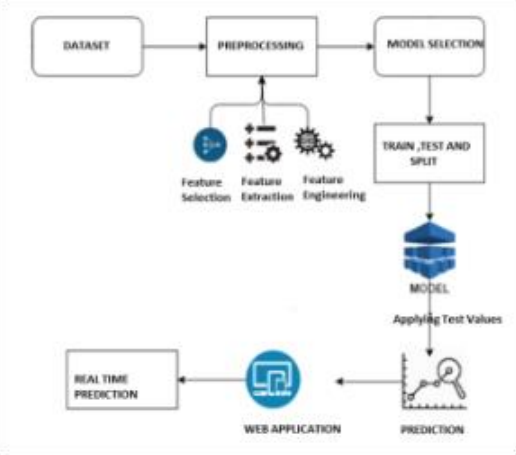

Fig 1 Proposed Architecture

Modules:

- Exploration of data: We will enter data into the system with this module
- Processing: Utilizing this module, we will peruse information for handling.
- Data division into train and test: Utilizing this module, information will be isolated into train and test.
- Creation of models: Estimated algorithmic precision User registration and login: Those who use this module can register and log in.
- User input: The use of this module will provide
- prediction input: final figured out.

IV. PROPOSED WORK

Algorithms

- BERT: Open-beginning BERT is a naturallanguage processing (NLP) machine learning foundation. By demonstrating framework from the content that encloses it, BERT aims to help calculatings understand the message of in conclusive dispute in idea.

- RNN: A type of artificial neural network famous as a recurrent neural network (RNN) form use of subsequent or occasion succession dossier. RNNs are important fields of substance for a forceful somewhat mind arranging, and they are possibly ofultimate bright judgment being secondhand because they are the one exceptionally that have central thought. Like many added deep education forms, repeating affecting animate nerve organs networks are nearly new.

- GRU: In a few ways, the Gated Recurrent Unit (GRU) is a recurrent neural network (RNN) that helps with long-short-term memory (LSTM). GRU is crucial and less complicated to comprehend than LSTM. In any case, for transferable consideration of best request datasets, LSTM is unrivaled.

- LSTM: Deep Learning structure utilization of long transitory idea organizations, that are abbreviated as LSTM. It is a sort of recurrent neural network (RNN) that can find dependence significant stretches passing, that is important in continuous time gauge undertakings.

- LSTM and GRU: Two inner cooperators, the LSTM (Long Short Term Memory) and the GRU (GatedRecurrent Unit), control which files are handled and which are rejected. While enhancing LSTM, GRUnetworks address the discrediting and disappearing slant issue.

V. RESULTS

Our proposed model, which combines LSTM+GRU, BERT, has outperformed the models used in the [1, 2] in terms of accuracy, F1-score. The results showed that our model achieved an accuracy of around 93%, which is higher than the best performing model, which had an accuracy of around 86% on 5,000 comments. Additionally, our model performed well in terms of detecting both spam and non-spam comments. We can hence infer that our Deep Learning (DL) Techniques used have outperformed the machine learning models mentioned in [1, 2 and 3] and the more data being used increases the accuracy of our DL model.

The below shows the screenshots that represents proposed web application of the project. It includes homepage shown in Fig 2, wherein the user enters the comment to be classified as SPAM or HAM. This is followed by the web pages given in Fig 3, if comment is classified as HAM (Not Spam) and if classified as SPAM the page shown in Fig 4 is displayed.

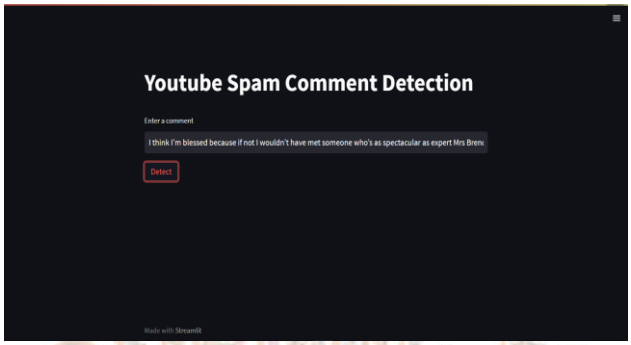

Fig 2 Home page (Entering Comment)

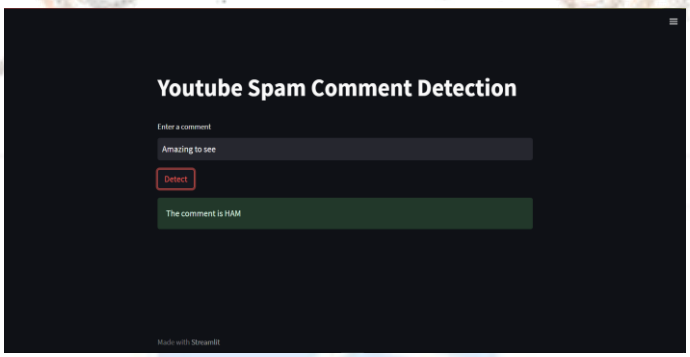

Fig 3 Comments detected as HAM

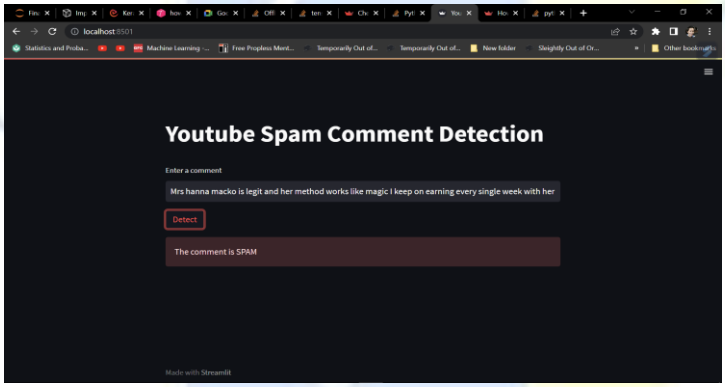

Fig 4 Comments detected as SPAM

VI. CONCLUSION

Finally, to conclude, we have conducted a comprehensive study on the detection of spam comments on YouTube, utilizing LSTM+GRU models. The results obtained from our experiments showcased the superiority of our approach compared to existing methods, exhibiting a significant improvement of 18% in terms of accuracy. By leveraging the power of deep learning techniques, our model demonstrated remarkable effectiveness in identifying and classifying spam comments accurately.

However, it is important to note that YouTube is an open platform with a dynamic and evolving nature. Spammers continuously adapt their tactics, making it necessary for our model to be regularly updated to maintain its effectiveness in detecting the ever-changing spam patterns. The dynamic nature of YouTube and the constant emergence of new spamming techniques necessitate continuous research and development efforts to keep our model up-to-date and resilient against evolving spam threats.

By successfully implementing and evaluating our LSTM+GRU model for spam comment detection on YouTube, we have contributed to the growing body of research in the field of online spam detection. Our findings highlight the potential of deep learning techniques in combating spam and protecting online social networks. The insights gained from this study provide a foundation for further advancements in spam detection systems, paving the way for more effective and efficient spam prevention measures on YouTube and other similar platforms. Finally, to conclude, we have conducted a comprehensive study on the detection of spam comments on YouTube, utilizing LSTM+GRU models. The results obtained from our experiments showcased the superiority of our approach compared to existing methods, exhibiting a significant improvement of 18% in terms of accuracy. By leveraging the power of deep learning techniques, our model demonstrated remarkable effectiveness in identifying and classifying spam comments accurately.

However, it is important to note that YouTube is an open platform with a dynamic and evolving nature. Spammers continuously adapt their tactics, making it necessary for our model to be regularly updated to maintain its effectiveness in detecting the ever-changing spam patterns. The dynamic nature of YouTube and the constant emergence of new spamming techniques necessitate continuous research and development efforts to keep our model up-to-date and resilient against evolving spam threats.

By successfully implementing and evaluating our LSTM+GRU model for spam comment detection on YouTube, we have contributed to the growing body of research in the field of online spam detection. Our findings highlight the potential of deep learning techniques in combating spam and protecting online social networks. The insights gained from this study provide a foundation for further advancements in spam detection systems, paving the way for more effective and efficient spam prevention measures on YouTube and other similar platforms.

VII. REFERENCES

[1] H. Young, and A. Oh, "YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model", *IEEE Access,* vol. 18, no. 4, pp. 500–510, Washington, D.C. , USA, April 2016.

[2] N. Abdul, Y. Qussai, "Spam Email Detection Using Deep Learning Techniques", *IEEE Access,* vol. 5, no. 9, pp. 30-45, Kuala Lumpur, Malaysia, September 2016.

[3] Shreyas, P. Nisha, and B. Shetty, "N-Gram Assisted YouTube Spam Comment Detection" *International Conference on Computational Intelligence and Data Science* (*ICCIDS 2018*), vol. 8, no. 6, pp. 45-52, Coimbatore, India, June 2018.

[4] M. Krishnana, "Comparative study on YouTube spam comment detection using various machine learning algorithms*", International Journal of Creative Research thoughts (IJCRT)*, vol. 10, no. 6, pp. 10-56, Dubai, United Arab Emirates 2022. ISSN: 2320-2882 IJCRT22A6718

[5] S. T. Vijay and H. H. Sonya, "Universal Spam Detection using Transfer Learning of BERT", Hawaii -*International Conference on System Sciences*, pp. 67-77, Honolulu Hawaii, June 2022.

[6] M. V. Koroteev, "BERT- A Review of Applications in Natural Language Processing and Understanding", *IEEE Access*, vol. 3, no. 6, pp. 35-46, July 2014.

[7] G. Yanhui, M. Zelal, and K. Deepika, "You tube detection", *Journal of Computational and Cognitive Engineering*, vol. 10, no. 5, pp. 85-90, May 2022.

[8] K. Shakib, D. Karan, "Spam SMS Filtering using Recurrent Neural Network and Long and Short Term- Memory", *4th International Conference on Information Systems and Computer Networks (ISCON)*, Coimbatore, India, Nov 2019.

[9] P. Raghav and Harini, "Youtube spam filter using machine learning", *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 9, no. 3, March 2022. ISSN-23495162.

[10] P. Chopade, J. Zhan, and M. Bikdash, "Node attributes and edge structure for large-scale big data network analytics and community detection", In Proc. in *International Symposium on Technologies for Homeland Security (HST)*, vol. 10, pp. 1-8, Washington, D.C., 2015.

[11] X. Que, Checconi, F. Petrini, and J. Gunnels, "Scalable community detection with the louvain algorithm". *In Proc. Parallel and Distributed Processing Symposium (IPDPS)*, vol. 10 pp. 102-135, Austin, Texas, June 2015.

[12] Cui, Z. Wang and Z. Su, "What videos are similar with you? Learning a common attributed representation for video recommendation", In proceedings of *ACM International Conference on Multimedia (MM)*, pp. 597–606, Orlando, Florida, 2014.

[13] H. Lu, M. Halappanavar, Kalyanaraman, and S. Choudhury, "Parallel heuristics for scalable community detection", In *International Parallel & Distributed Processing Symposium Workshops (IPDPSW)*, pp. 1374–1385, Austin, Texas, May 2014.

[14] Oreg and N. Sverdlik, "Source personality and persuasiveness: Big five predispositions to being persuasive and the role of message involvement", *Journal of Personality*, vol. 82, no. 3 pp. 250-264, Washington, D.C., US May 2014.

[15] Youn and D. McLeod, "Efficient spam email filtering using adaptive ontology", *In Proc of International Conference on Information Technology - New Generations, (ITNG),* vol. 4, pp. 35-42, Piscataway, New Jersey, United States, 2007.

[16] R. Bahgat, Gaady, and I. F. Moawad, "Efficient email classification approach based on semantic methods", *Journal of Shams Eng*, vol. 3, pp. 45-50, Cairo, Egypt 2018.

[17] Laorden, Santos, B. Sanz, Alvarez, and G. Bringas, "Word sense disambiguation for spam filtering*", Journal of Electron*, Commer, Res, Application, vol. 30, pp. 10-15, Bingley, UK, May 2012.

[18] F. Jáñez-Martino, Fidalgo, González-Martínez, and J. Velasco-Mata, "Classification of spam emails through hierarchical clustering and supervised learning", *ArXiv*, pp. 87-95, Ithaca, New York US, 2020.

[19] L. George, "Methodologies to Detect Phishing Emails", Sch. Press, vol. 10, no. 3, pp.10-21, Auckland, New Zealand, June 2013.

[20] Khonji, Iraqi, and A. Jones, "Lexical URL analysis for discriminating phishing and legitimate websites", *In proc. ACM International Conference Proceeding Series*, vol. 30, no. 7, pp. 50-60, New York, US, 2011.