Proceedings of the International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS 2023)
IEEE Xplore Part Number: CFP22DN7-ART; ISBN: 979-8-3503-0085-7

Leveraging the Ensemble Model in Spam detection for Youtube Comments

N Saikiran
Asst. Professor, Dept. of CST
Madanapalle Institute of Technology & Science (A), Madanapalle
Andhra Pradesh, India
nlmsaikiran@gmail.com

P Kiran
Asst. Professor, Dept. of CSE(AI)
Madanapalle Institute of Technology & Science (A), Madanapalle
Andhra Pradesh, India
kiranpalakeeti@gmail.com

Srikanth Kama
Asst. Professor, Dept. of CSE
Malla Reddy Engineering College (A), Telangana, India
srikanthkama09@gmail.com

Dammati Ganga
Asst. Professor, Dept. of CSE
Chalapathi Institute of Technology
Andhra Pradesh, India
gangadammati2587@gmail.com

Bakka Rajeev Gandhi
Asst. Professor, Dept. of CSE
Chalapathi Institute of Technology
Andhra Pradesh, India
brajeevg@gmail.com

Dr. M Sambasivudu
Assoc. Professor, Dept. of CSE
Malla Reddy College of Engineering and Technology (A), Telangana, India
m.sambasivudu@mrcet.ac.in

2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-0085-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSAS57918.2023.10331815

***Abstract*— YouTube (YuTe) spam replies have lately shown an explosive increase, and this research presents a method to identify them. YuTe uses its trojan technology, although it hasn't been effective so far. For this reason, the authors of this research looked at prior research on YuTe spamming remark filtering and ran trials on remark information from four different types of famous Telugu music videos (melody, classical, rock, and duet) to see how well E-GARCH+ARIMA performed as classification techniques. Therefore, this article will use the Hybrid time-series predictive ARIMA (Autoregressive Integrated Moving Average)+ E-GARCH (Exponential Generalized Auto-Regressive Conditional Heteroskedasticity) to anticipate spam identification in Yute music videos. When compared to both standalone simulations and cutting-edge hybrid methods, the aforementioned hybrid ARIMA+E-GARCH paradigm fared better across the board for RMSE, MAE, and MSE.**

Keywords— Comments, Spam, YuTe, Hybrid Model, Filtering.

I. INTRODUCTION

ML and DL strategies are leveraged to predict anything in this modern era including vulnerabilities in IoT devices [1] and Wireless Sensor Networks [2], spam detection, churn prediction, credit card fraud detection, etc. Google bought YuTe, the biggest multimedia website, in 2006. With the current trend toward multimedia information on the internet, YuTe's popularity has skyrocketed as a multimedia provider. YuTe now receives over 500 hours of video content per day and has 5.5 billion unique viewers every day. Visitors may freely view and post recordings. As a result of this widespread availability, more and more individuals are creating their content, with some going on to establish influential figures in the digital sphere.

As long as a YuTe developer has 1,000 members and 4,000 hours of media consumption in the past year, YuTe developers are eligible to receive payment. As a result, fraudulent remarks are being generated so that producers may advertise their films in high-traffic movies. Some producers have disabled remarks because they've received too much hate mail, especially if it's partisan, or if it's personal and has nothing to do with the movie's subject matter. Although YuTe has its built-in spam detection mechanism, some spamming remarks still make it through.

In this study, the authors of this research survey existing research on spamming remarks on YuTe and then present the Ensemble ML Model aware YuTe Illegal Remarks Identification Strategy to boost the model's efficacy. Prior research has implemented numerous ML methodologies to each input data to identify spam submissions and evaluate their effectiveness. In this article, the authors of this research consequently suggest an ensemble ML approach, which integrates the outputs of multiple models into a single prediction. To better understand how to identify spam material and users, researchers have turned their attention to several different areas. Many investigations have centered on internet spamming. Despite YuTe's meteoric recent resurgence, it's become a focus for fraudsters who upload subpar videos or push questionable products. As the number of fraudsters attempting to disrupt the YuTe network rises, studying how to identify and stop fraudsters represents a potentially fruitful area of study. When dealing with YuTe remarks, users can't employ the same approach since the data has unique characteristics. Fewer verbal explanations and details are represented by YuTe comment capabilities. Users add nothing to the video and have no real connection to it. To locate YuTe spamming, therefore, a novel strategy is required. Some research uses a categorization system to determine if a given group of remarks or a certain video is spam.

The sophistication of fraudsters' methods for tricking customers into clicking on dangerous links has expanded with the popularity of social media sites on the Internet. This is achieved by flooding the remarks sections of different online networking platforms with irrelevant content. This research focuses on detecting fraud in YuTe remarks using YuTe

reviews as the information. Technologies like Gmail Secure Surfing, which can identify and prevent unrelated advertising on YuTe, are now used to combat fraudsters. These technologies do a good job of preventing potentially dangerous connections from being accessed, but the technologies aren't enough to keep people safe in the physical world. Thus, a variety of strategies have been used to provide a spam-free setting. Some of them rely entirely on contributions from other users, while others get inspiration from videos posted on YuTe.

Individuals are more likely to interact with others via well-known online connecting and entertainment platforms than connecting in person. As a result, digital networking platforms are seeing unprecedented growth in the present day. In the same vein as other YuTe's popularity continues to soar, and the site itself is growing at a dizzying rate. YuTe's success is predicated largely on the user-generated material and its subsequent viral dissemination. Businesses and prominent people are capitalizing on this trend by setting up their profiles and using the site's high traffic to spread their message. Nevertheless, YuTe has grown increasingly vulnerable to many forms of undesirable and harmful advertising as a result of its increased prominence. There is presently no method for YuTe to deal with its content fraudsters. Trash is defined as merely bulk remarks or mails. Movie reply spamming is a kind of movie manipulation in which malevolent people intentionally upload videos that do not respond to the actual footage or do not include the intended material to artificially raise the views of the actual footage.

The following is the outline for this article: Relevant works are discussed in Section 2. In Section 3, the authors of this research detailed the design methodology and the procedures for the method, while in Section 4, the authors of this research go into detail about the trials and the outcomes. After that, in Section 5, the authors of this research draw to a close.

II. RELATED WORKS

R. Abinaya et al., In [3], an ML system was employed to identify fake positive and negative brand attitudes on the web. Authors zeroed in on fake testimonials that may be posted under many user names. It was determined that there were identical comments by comparing each comment to every other post. The following words of a sentence were predicted using ML. Simply said, ML mechanically disintegrates phrases into their parts.

To differentiate fraudsters from legitimate users on YuTe, R. Chowdhury et al., [4] provide a categorization of sociable and anti-social behavior of individuals and movie qualities. To determine who and what videos had a lively relationship, the authors compiled a randomized sample of individuals who submitted at the designated time. H. Oh [5] used an ensemble ML technique for spam detection in YuTe.

S. Kanodia et al., [6] used the Markov decision approach for spam detection in videos of Yute's Platform. Everyone's daily lives now often include some kind of interaction with an online connecting platform. Individuals use online networking to connect with others, exchange thoughts and information, educate themselves, have fun, and keep up with current happenings. The greatest widely used platform for uploading, distributing, and watching videos online is without a doubt YuTe. Due to YuTe's immense prominence, fraudsters have begun uploading videos with the express intent of clogging up the platform with irrelevant material and frustrating legitimate users. These spamming movies might include inappropriate material or have nothing to do with the title. As a result, it is crucial to figure out how to recognize and denounce such movies before users may be watched by unwitting individuals.

N. Aggarwal et al., [7] When it comes to online movie-uploading platforms (that also include personal connecting capabilities), YuTe is among the most well-known and substantial. A meaningful portion of submitted movies on YuTe are offensive or otherwise break YuTe's social rules. Many videos on YuTe infringe on copyrights, promote hatred and radicalism, are obscene, or otherwise breach users' confidentiality. The low bar to entry for publishing and the ability to remain anonymous are major factors.

S. Sharmin et al., [8] Recently, several malicious attacks have been made against internet community platforms. While YuTe allows us to freely express users' ideas to the world, the widespread public is at risk because of the abuse of this potent tool. YuTe, for instance, has been utilized as a marketing platform for different artists to publish pop songs, feature films, etc., and visitors may leave their judgment on it. The remarks area is a popular place for unscrupulous people to spread ransomware and spyware by posting connections to fake sites, harmful advertising, and other forms of bogus content. Consequently, it is important to identify these potentially dangerous remarks so that online networking services may remain operational without a hitch.

V. Chaudhary et al., [9] YuTe is among the most popular online video-sharing and social networking platforms. Many kinds of movie contamination, including the posting of harmful, copyright-violating, and spamming video or material, have occurred from YuTe's tremendous ubiquity, obscurity, and cheap publishing hurdle. Movie responses are a frequent and well-liked functionality on YuTe, where viewers may submit their videos in response to those shared by others. Hundreds upon millions of audio answers are posted to some of the most famous movies on YuTe.

M. Alsaleh et al., [10] Users of platforms such as YuTe with a commenting option may engage with and append to the published material in meaningful ways. The reality that these remarks are becoming an integral part of the blog's information, read by a large number of users, and are rarely moderated makes them a prime target for fraudsters looking to promote their products, disperse ransomware, conduct malicious scams, or propagate their ideological or religious beliefs. Due to the sheer number of trash remarks received, human screening and validation are impractical, making automated junk identification approaches a must for combating trash online. S. Vinothkumar et al., [11]. Used Ml approaches for spam detection in Yute's videos.

E. Elakkiya et al., [12]. Online research has become more popular than traditional printed media in previous

decades. People who don't have the leisure to check the media regularly often use online media to stay informed. Recently, there has been a rise in the frequency with which junk emails are sent to anyone in the world. Users will be tricked into downloading software that will collect their data. It must guard against junk because it spreads malware and gives hackers entry to sensitive information. Individuals who aren't savvy to these scams fall for junk emails since people look and feel like real business correspondence.

Y. Tashtoush et al., [13] used data mining techniques to detect Arabic junk in YuTe. With the rise of YuTe's popularity as a viable generator of money, so too has the emergence of a growing number of marketers who use the platform to push malicious software or just to get views and subscribers for their stations. Several YouTubers have shut down their channels or disabled remarks in response to these kinds of abuse since the platform does not provide sufficient tools to combat it. Due to the many different Arabic languages and the vast number of equivalents authors include, screening out spamming remarks in Arabic is a difficult task.

B. Coskun et al., [14] used a network-based method to detect junk messages on online media platforms. C. Radulescu et al., [15] used Natural Language processing techniques (NLP) for detecting junk in Yute's platform. Embedding and other information pretreatment technologies are often used in conjunction with ML algorithms to handle the traditional NLP issue of textual spamming identification.

F. Concone et al., [16] used a novel approach named SPADE to detect spam in Yute's platform. R. Thapa et al., [17] used brain-inspired algorithms for detecting junk remarks in YuTe's platform. F. Benevenuto et al., [18] proposed an active learning approach for detecting spam in YuTe's videos. H. Oh [19]. Used several ML and DL approaches for detecting junk remarks in YuTe's videos.

III. Proposed Methodology

A. ARIMA

The AM (u) technique uses the linear information regarding previous instances for a feature to estimate expected patterns. The MM (v) concept calculates the predicted dispersion in spam detection by integrating previously estimated shortfalls successively in the past. ARMA (Autoregressive Moving Average) may be utilized for asymmetrical and constantly changing chronological data when paired with the MM (v) and AM (u) techniques. It looks at the tactical connections between prior spam information and earlier spam data. Consequently, the ARIMA approach integrates a prediction paradigm when confronting uncertain phenomena like spam data. It is possible to employ uncertain input with the ARIMA (u, e, v) concept because it blends AM(u) and MM(v) techniques with minimal transformation to stabilize availability. The Autoregressive Method AM(u) and the Moving Average Method MM(v) are combined to form ARIMA.

B. GARCH

The GARCH concept is used for parameter responses containing semi-components, such as spam detection. The uniform pattern of distribution of the data is assessed using the GARCH framework. It does effectively with data that has a large SD. The MM and AM components are both included in the GARCH model, a variation of the ARCH concept. The constant polynomial is all that is shown, relying on the previous remaining erroneous data. Uncertainty in GARCH is dependent on previous recurrence. To foresee and characterize turbulence fluctuations, the GARCH formulation includes the constant functionality as an ingredient of the previous error squared and its corresponding sign.

C. E-GARCH

The GARCH framework is modified by the EGARCH concept. Nelson developed the E-GARCH concept in 1991 to solve the issues with GARCH's handling of financial historical intervals permitting uneven effects on both positive and negative possessions. This is how an E-GARCH (u,v) is described:

$$X_t = \mu + A_t$$

The rate at which things happen of the temporal pattern at time 't' is x_t. The mean of the GARCH hypothesis is 'μ'. 'A_t' is the simulation's leftovers at interval 't'. The dependent position variation (i.e. uncertainty) at time frequency is denoted by the symbol 't'. The pattern of the ARCH component paradigmatic is 'u', and the properties of the ARCH ingredient hypothesis are α_o, α_1, α_2,...,α_u. The pattern of the GARCH ingredient hypothesis is 'v', and the features of the E-GARCH ingredient hypothesis are certainly β_1, β_2,...,..β_v. The representative of the homogenized specimens is [ϵ_t].

The E-GARCH framework differs from the GARCH framework in several ways. For instance, the measured dependent variances were used to lessen the likelihood that the prototypical characteristics would be restricted. A unilateral EGARCH architecture is EGARCH. The EGARCH package outputs an additional integration component that provides the organizational framework of an EGARCH (U, V) concept and retains its operational configurations. Important components of an EGARCH paradigm include the ones that follow: delayed and documented dependent variances are the building blocks of the GARCH formula. The extent of predictability is indicated by the delayed normalized inventiveness values that make up the parameter 'U' in ARCH mathematical formulation. The constrained proportion is made up of normalized improvements that have stalled. The ARCH and exponential categories that use parameter 'V' are the most significant.

The parameter 'U' is the largest nonnegative latency in the GARCH exponential, whereas the parameter 'V' is the most significant positive integer disparity in the ARCH and boosted constants. Additional simulating components include a discovery midpoint demonstrating offsetting, an alternative deconstruction representing the eternal, and the distribution of findings. All equations are unknown and illustrative till the composite component approach is used. Utilize

approximations to make educated guesses about precedents with fully or largely uncertain constituent ratings. For completely specified architectures (attributes where all constant readings are determined), predicting or anticipating responses using recreation or estimation is appropriate. The EGARCH concept provides an extra imbalanced perspective by considering the magnified consequences of an overall change on the contingent deviation. As a consequence, rather than only a huge affordability gain, a substantial reduction in dimensionalities may have had a bigger effect on uncertainty. A common analytical model using explicable parameters is EGARCH.

D. ARIMA+ E-GARCH Paradigm

The significance of employing the continuous function for spam detection is illustrated by a straightforward explanation of the ARIMA and E-GARCH techniques. The consistent probability is often referred to by the term covariance. The prognostic initiative's spam data input is the initial phase. Using all the data you possess, create an interval frame and, in the second step, transform it into a definite scheduled interval. Phase 3's estimation of the pattern and attributes makes use of the suggested hybrid ARIMA+ E-GARCH architecture. Step 4 completes the process by making spam detections and looking at discrepancies. The suggested technique's whole process is shown in Fig. 1. In this section, the blend of the ARIMA+ E-GARCH architecture is discussed.

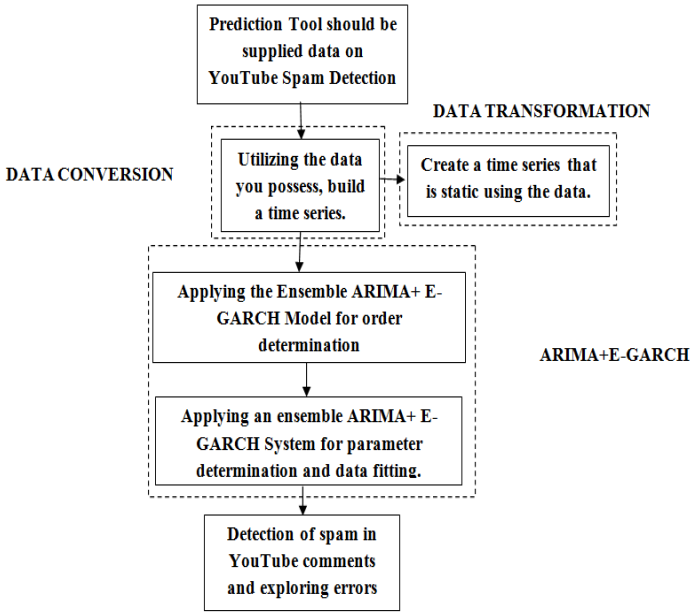

Fig. 1. The framework of the proposed ensemble ARIMA+E-GARCH Model

IV. RESULTS AND DISCUSSION

Fig. 2 displays the inability analysis for the most recent hybrid approaches and ARIMA+E-GARCH. The suggested strategy, which is highlighted in yellow in Table I, has a reduced failure ratio for RMSE, MSE, and MAE of 1.115, 2.129, and 2.102. The detection rate of the ARIMA+E-Garch model is 95%.

TABLE I. AN EXAMINATION OF THE PROPOSED HYBRID APPROACH'S FAILURE RATES WITH THOSE PRESENT-DAY HYBRID TECHNIQUES

Study	Method	RMSE	MAE	MSE
N. Aggarwal et al., (2020). [7]	MLP+CNN	6.7 to 23.28	22.9	13.8
V. Chaudhary et al., (2020). [9]	DPS+MLP	1.52	2.489	3.129
Suggested Hybrid Method	ARIMA+E-GARCH	1.115	2.129	2.102

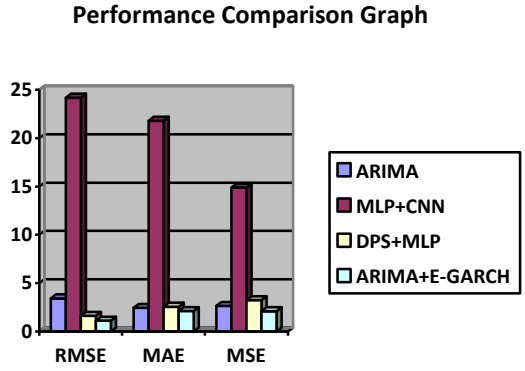

Fig. 2. Chart comparing the proposed model to the latest techniques in terms of effectiveness

V. CONCLUSION

To combat the exponential increase of spam comments on YuTe, the authors suggested a method in this research that uses an Ensemble ML Model. Identification trials using ARIMA+E-GARCH were done, and related research on spamming comments filtering on YouTube was analyzed. According to the data, the ARIMA+E-GARCH paradigm that this study proposes performs the best in practice. In the upcoming, the authors of this research want to investigate spam detection using GARCH modeling such as Quadrant GARCH (QGARCH), and GARCH-in-mean (GARCH-M).

REFERENCES

[1] K. Vivek, M. R. Kale, V. S. K. Thotakura and K. Sushma, "An Efficient Triple-Layered and Double Secured Cryptography Technique in Wireless Sensor Networks," 2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 2021, pp. 117-122, doi: 10.1109/DISCOVER52564.2021.9663674.

[2] Rao, K.V., Latha, D.S., Sushma, K., Vivek, K. (2022). A Study on Defensive Issues and Challenges in Internet of Things. In: Satyanarayana, C., Samanta, D., Gao, XZ., Kapoor, R.K. (eds) High Performance Computing and Networking. Lecture Notes in Electrical Engineering, vol 853. Springer, Singapore. https://doi.org/10.1007/978-981-16-9885-9_48.

[3] R. Abinaya, B. Niveda E. and P. Naveen, "Spam Detection On Social Media Platforms," 2020 7th International Conference on Smart Structures and Systems (ICSSS), 2020, pp. 1-3, doi: 10.1109/ICSSS49621.2020.9201948.

[4] R. Chowdhury, M. N. Monsur Adnan, G. A. N. Mahmud and R. M. Rahman, "A data mining based spam detection system for YouTube," Eighth International Conference on Digital Information Management (ICDIM 2013), 2013, pp. 373-378, doi: 10.1109/ICDIM.2013.6694038.

[5] H. Oh, "A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model," in IEEE Access, vol. 9, pp. 144121-144128, 2021, doi: 10.1109/ACCESS.2021.3121508.

[6] S. Kanodia, R. Sasheendran, and V. Pathari, "A Novel Approach for Youtube Video Spam Detection using Markov Decision Process," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 60-66, doi: 10.1109/ICACCI.2018.8554405.

[7] N. Aggarwal, S. Agrawal, and A. Sureka, "Mining YouTube metadata for detecting privacy-invading harassment and misdemeanor videos," 2014 Twelfth Annual International Conference on Privacy, Security and Trust, 2014, pp. 84-93, doi: 10.1109/PST.2014.6890927.

[8] S. Sharmin and Z. Zaman, "Spam Detection in Social Media Employing Machine Learning Tool for Text Mining," 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2017, pp. 137-142, doi: 10.1109/SITIS.2017.32.

[9] V. Chaudhary and A. Sureka, "Contextual feature-based one-class classifier approach for detecting video response spam on YouTube," 2013 Eleventh Annual Conference on Privacy, Security and Trust, 2013, pp. 195-204, doi: 10.1109/PST.2013.6596054.

[10] M. Alsaleh, A. Alarifi, F. Al-Quayed and A. Al-Salman, "Combating Comment Spam with Machine Learning Approaches," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 295-300, doi: 10.1109/ICMLA.2015.192.

[11] S. Vinothkumar, S. Varadhaganapathy, R. Shanthakumari, D. Ramkishore, S. Rithik and K. P. Tharanies, "Detection Of Spam Messages In E-Messaging Platform Using Machine Learning," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 283-287, doi: 10.1109/CCiCT56684.2022.00060.

[12] E. Elakkiya, S. Selvakumar and R. L. Velusamy, "CIFAS: Community Inspired Firefly Algorithm with fuzzy cross-entropy for feature selection in Twitter Spam detection," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225321.

[13] Y. Tashtoush, A. Magableh, O. Darwish, L. Smadi, O. Alomari and A. ALghazoo, "Detecting Arabic YouTube Spam Using Data Mining Techniques," 2022 10th International Symposium on Digital Forensics and Security (ISDFS), 2022, pp. 1-5, doi: 10.1109/ISDFS55398.2022.9800840.

[14] B. Coskun and P. Giura, "Mitigating SMS spam by online detection of repetitive near-duplicate messages," 2012 IEEE International Conference on Communications (ICC), 2012, pp. 999-1004, doi: 10.1109/ICC.2012.6363989.

[15] C. Rădulescu, M. Dinsoreanu, and R. Potolea, "Identification of spam comments using natural language processing techniques," 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP), 2014, pp. 29-35, doi: 10.1109/ICCP.2014.6936976.

[16] F. Concone, G. L. Re, M. Morana and S. K. Das, "SpADe: Multi-Stage Spam Account Detection for Online Social Networks," in IEEE Transactions on Dependable and Secure Computing, 2022, doi: 10.1109/TDSC.2022.3198830.

[17] R. Thapa, B. Lamichhane, D. Ma and X. Jiao, "SpamHD: Memory-Efficient Text Spam Detection using Brain-Inspired Hyperdimensional Computing," 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2021, pp. 84-89, doi: 10.1109/ISVLSI51109.2021.00026.

[18] F. Benevenuto, T. Rodrigues, A. Veloso, J. Almeida, M. Goncalves and V. Almeida, "Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 3, pp. 688-701, June 2012, doi: 10.1109/TSMCB.2011.2173799.

[19] H. Oh, "Corrections to “A YouTube Spam Comments Detection Scheme Using Cascaded Ensemble Machine Learning Model”," in IEEE Access, vol. 10, pp. 40860-40860, 2022, doi: 10.1109/ACCESS.2022.3166635.