

8-2024

Enhancing YouTube Spam Detection

Sai Charan Pesaru

Follow this and additional works at: <https://scholarworks.lib.csusb.edu/etd>



Part of the [Business Commons](#)

Recommended Citation

Pesaru, Sai Charan, "Enhancing YouTube Spam Detection" (2024). *Electronic Theses, Projects, and Dissertations*. 2014.

<https://scholarworks.lib.csusb.edu/etd/2014>

This Thesis is brought to you for free and open access by the Office of Graduate Studies at CSUSB ScholarWorks. It has been accepted for inclusion in Electronic Theses, Projects, and Dissertations by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

ENHANCING YOUTUBE SPAM DETECTION

A Thesis
Presented to the
Faculty of
California State University,
San Bernardino

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
in
Information Systems and Technology

by
Sai Charan Pesaru
August 2024

ENHANCING YOUTUBE SPAM DETECTION

A Thesis
Presented to the
Faculty of
California State University,
San Bernardino

by
Sai Charan Pesaru

August 2024

Approved by:

Dr. Conrad Shayo, Committee Member, Chair

Dr. Nima Molavi, Committee Member, Reader

Dr. William Butler, Committee Member, Reader

Dr. Conrad Shayo, Chair, Information & Decision Sciences Department

© 2024 Sai Charan Pesaru

ABSTRACT

This culminating experience project investigated various methods for enhancing spam detection on YouTube, a prevalent issue impacting user experience and platform integrity. The research questions addressed were: Q1) How do different spam detection methods compare regarding robustness, efficiency, and accuracy? Q2) What role do deep learning approaches like RNNs and CNNs play in improving spam comment identification? Q3) What are the unique benefits of using deep learning models for spam comment identification on YouTube? Q4) How can machine learning models be optimized for real-time spam detection on YouTube?

The study gave adequate findings that explained each research question. In the case of (Q1), while algorithms like the Naïve Bayes and Logistic Regression offered precision in identifying spam emails, the models have proven ineffectual at adapting to new forms of spam and constant enhancement in spam techniques, deep learning algorithms like the CNN and RNN offered high accuracy through their robustness due to the models' abilities of feature extraction independently from the text data. The results shown in (Q2) indicate that RNNs and CNNs are critical in transforming the level of spam detection by addressing the problem of semantic meaning and temporal relationships in comments and surpassing traditional methods. Concerning (Q3), it was pointed out that deep learning models are the most accurate, scalable, and resistant to false negatives when identifying spam comments on the videos hosted on

YouTube, which helps regain users' trust and enhance the platform's security as the traffic continues to grow. (Q4) was focused on advancing machine learning models for real-time processing, using methods such as model pruning and distribution.

The findings were as follows: (Q1) found that although conventional approaches are efficient at meeting accurate results, deep learning models are highly effective in dealing with the changes in spam strategies. (Q2) pointed out that RNNs and CNNs contribute immensely to discovering spam in SM platforms due to their raw power in NLP and pattern recognition. (Q3) established that the deep learning models' accuracy, scalability, and adaptability, including CNN and RNN, are beneficial in identifying spam on YouTube due to their effectiveness in tackling the ever-evolving spam tactics. (Q4) It has emerged that the fine-tuning of machine learning models is imperative for scaling up the approaches by deploying high-end methodologies for real-time spam detection, which subserves the daunting task of training the algorithms to deal with the flood of user-generated content in the context of YouTube.

Areas of further study include analyzing other complex natural language processing methods combined with classifiers for better spam identification, improving the computational time for multi-modal learning for spam comment detection, and considering federated learning for real-time spam identification on platforms such as YouTube. These research directions are being carried out to boost the existing permutations and improve the permeate spam detection

technologies in Information Systems so that they can be efficient, effective, and highly accurate systems capable of coping with the newly emerged spam technologies in flexible, transparent, and effective ways.

ACKNOWLEDGEMENTS

I want to acknowledge and thank the support provided by Dr. Conrad Shayo, Professor Nima Molavi, and Professor William Buttler for their assistance.

DEDICATION

This project and my entire master's degree are dedicated to my family, friends, and professors, who have supported me.

TABLE OF CONTENTS

ABSTRACT.....	iii
ACKNOWLEDGEMENTS.....	vi
LIST OF FIGURES.....	ix
LIST OF TABLES.....	x
CHAPTER ONE: INTRODUCTION	1
Problem Statement	3
Research Questions	4
Objective of this Project	5
Organization of this Project.....	6
CHAPTER TWO: LITERATURE REVIEW.....	7
Background Story	7
CHAPTER THREE: RESEARCH METHODOLOGY	19
Neural Network Model	20
CHAPTER FOUR: DATA COLLECTION, ANALYSIS, AND FINDINGS.....	29
CHAPTER FIVE: DISCUSSION, CONCLUSION AND FUTURE WORK	42
Limitations of the Research	42
Discussion	43
Conclusion.....	52
Area of Further Study.....	56
APPENDIX: CODES.....	60
REFERENCES.....	66

LIST OF FIGURES

Figure 1: Comparison of Categories (Kontsewaya, 2021)	8
Figure 2: Top Ten Words in Spam and Non-Spam Comments (From Code)	9
Figure 3 YouTube Spam Detection Framework (Oh, 2021).	20
Figure 4: Recurrent Neural Network (Hibat-Allah, 2020)	23
Figure 5: Outline of CNN (Teja, 2023)	24
Figure 6: Dataset Overview (From Code)	32
Figure 7: Bar Chart for Model's Results (From Code)	34
Figure 8: Model Results (From Code)	35

LIST OF TABLES

Table 1: Comparison with Other Models.....	44
--	----

CHAPTER ONE

INTRODUCTION

Recent innovations such as social media sites, including YouTube, have revolutionized communication by providing users with created content and large community channels. However, this evolution has also introduced a constant issue, especially on blog sites: spam comments (Poirier et al., 2020). Spam distorts constructive conversations and is comprised of numerous dangers, like sharing scams and malware, degrading the platform's usability and relevancy (Abd et al., 2018).

Most studies on mitigating spam in online communities and groups have mainly focused on using rule-based checkpoints or applying manual censors. Nonetheless, such methods prove inefficient at designing strategies to counter the shifting strategy of spammers (Abd et al., 2018). Govil et al. (2020) also noted that spam detection was a significant issue, and it was high time to filter spam comments using better methods, including the machine learning technique, ANN, in particular. The study by Govil et al. (2020) has pinpointed one of the essential directions for investigations and developments in digital security. It proposes further discussion of the application of more sophisticated ML algorithms in future research projects. This reiterates the essence of continuing to envision combating spam strategies to protect virtual communities adequately. This they noted as a critical research avenue to improve the efficiency of spam filtering. The most crucial studies about rule-based filters and manual moderation

were carried out by Abd et al. (2018), and they need to be insufficient to manage the constant shift in spam strategies. According to his research, there is a demand for more improved methods that can easily counter new trends in spamming. In this research, Abd's framework will be followed, albeit in a more expansive capacity, to contribute to the development of the field by comparing and analyzing the overall efficiency of other spam detection techniques, such as machine learning classifiers that incorporate advanced feature extraction mechanisms as offered in Abd et al., 2018.

However, literature by Govil et al. (2020) works towards designing machine learning algorithms ANNs for detecting spam filters. Both of them have called for more research in an attempt to fine-tune such algorithms for specific networks like YouTube. This project focuses on these technologies that were discussed to be enhanced, including algorithm enhancement by technologies like vectorization and n-gram regarding spam classification (Aiyar et al., 2018). There is an obvious way to develop the technologies used for spam filtration to address conceivable subtleties and new spam approaches.

Some researchers recently targeted not only the Spamming malware but also Artificial Neural Networks (ANNs) and machine learning algorithms for detecting YouTube spam comments (Abdolrasol, Maiwagy, & Al-Salihy, 2021). The performance of such algorithms as thought vectors and other models is based on some preprocessing steps, such as vectorization, n-gram analysis, and more (Aiyar 2018). Furthermore, Rastogi has suggested using vectorization and

n-gram analysis in preprocessing for future work. The issue to solve is how ANNs and ML algorithms can combat YouTube spam comments, which is highly beneficial (Abdolrasol, 2021).

To this end, this study aims to contribute to the shifting conversation by implementing and comparing the effectiveness of diverse spam detection approaches, offering information on how to design better and superior spam detection methods suitable for the current online environment (Govil et al., 2020).

Problem Statement

Filtering spam comments on YouTube is still a challenging problem because of the large number of comments and frequent updates in spamming strategies (Baccouche et al., 2020). In light of these facts, Ayvaz et al. (2021), Baccouche et al. (2020), and Antony et al. (2022) have shown that manual moderation could be more sustainable, and rule-based systems cannot handle complex spam trends. Therefore, it becomes impossible to be impatient with automation as the only viable solution. In line with Abd et al. (2018), who conducted a study on methods of spam differentiation and recognized the difficulties of the customary approach, and Makkar et al. (2020), who proved the efficiency of developed algorithms, This Study seeks to expand the approaches of spam differentiate via the use of machine learning, especially Artificial Neural Networks (ANNs). Abd et al. (2018) suggested that higher-order machine

learning algorithms should be employed to increase the detection rate, and this study tries to solve Abd et al. (2018).

This object aims to upgrade the efficiency and suitability of the spam identification models using large datasets and diversified classification approaches, including those described by Abd et al. (201). In this study, Abd et al. diagnosed spam messages with the help of a deep learning approach, and the anticipated precision and recall level of distinguishing between spam and genuine comments was high. Their study suggested that more research should be done concerning adaptive learning algorithms and the implementation of real-time data processing to combat constantly changing spam techniques.

Furthermore, referring to Tumu et al. (2020), This Study will also look into the effectiveness of vectorization and n-gram analysis that would extract relevant features in the comment text to distinguish between grand and spam posts. In conclusion, this culminating experience enables the design of improved anti-spam techniques to protect the sanctity and usability of sites such as YouTube and other social media resulting from spam invasion.

Research Questions

1. How do various spam detection methods compare robustness, efficiency, and accuracy when combined with different classifiers to identify and mitigate spam comments on online platforms? (Rao, 2021).

2. In detecting spam comments on social media sites, what part do deep learning approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) play in pushing the envelope? (Oh, 2021)
3. What are the unique benefits of using deep learning models for spam comment identification on YouTube and other platforms regarding accuracy, scalability, and robustness? (Baccouche et al., 2020)
4. What adjustments or improvements can be made to machine learning models to provide better real-time results for spam detection, especially when handling many comments on YouTube videos? (Laddha, 2020)

Objective of this Project

This Cumulative Experience Project aims to address the challenge of detecting and mitigating spam comments on YouTube by building on some of the areas recommended for further studies by other researchers. Thus, Govil et al. (2020) pointed to the relevance of employing machine learning algorithms such as Artificial Neural Networks for anti-spam purposes. Furthermore, Abd et al. (2018) stressed there is more than opting for rule-based filters and manual moderation for the new techniques in use by spammers. Tumu et al. (2020) believe that the studies on vectorization and n-gram analysis must go further, and more experimental work must be done. With the help of such an approach, this project will design and test novel machine-learning models that will

incorporate vectorization and n-gram marketing technologies as the essential tools to create enhanced and highly resistant spam filters that will, in turn, positively impact the users' experience and, consequently, the credibility of the platforms.

Organization of this Project

The cumulative Experience Project is divided into five chapters. Chapter 1 introduces the topic. Chapter 2 covers the Literature review, while Chapter 3 provides the research methods that will be used to answer the research questions. Chapter 4 provides data collection, analysis, and findings. Chapter 5 discusses the findings and provides conclusions and recommendations for future research.

CHAPTER TWO

LITERATURE REVIEW

Background Story

The rapid expansion of online platforms like YouTube has led to a surge in user-generated content, accompanied by a significant rise in spam comments that undermine user experience and pose cybersecurity risks. Extensive research has been conducted to address this issue, with notable studies exploring various techniques, including machine learning and deep learning approaches like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), demonstrating their potential in improving spam detection accuracy and efficiency. Thus, this literature review discusses these enhancements and evaluates the relative performance of various classifiers regarding the model's robustness, efficiency, and reliability while demonstrating the advantages of deep learning classifiers. Thus, by integrating existing knowledge and identifying further research opportunities, this chapter contributes to the improvement and modularity of approaches adopted to fight spam on platforms similar to YouTube.

Figure 2 The following is a bar graph of the literature review on detecting spam. The bar graph shows that the most significant portion of the sections focus on the Introduction to Spam Detection—a third of all the articles. The least populated category is the Financial Impacts of Spam, accounting for 15%.

This category is the smallest among the topical categorizations, with 15 sample pieces. This type of visualization can be helpful for quickly understanding the distribution of articles in a literature review. It is perfect for pointing out spots that need extra consideration.

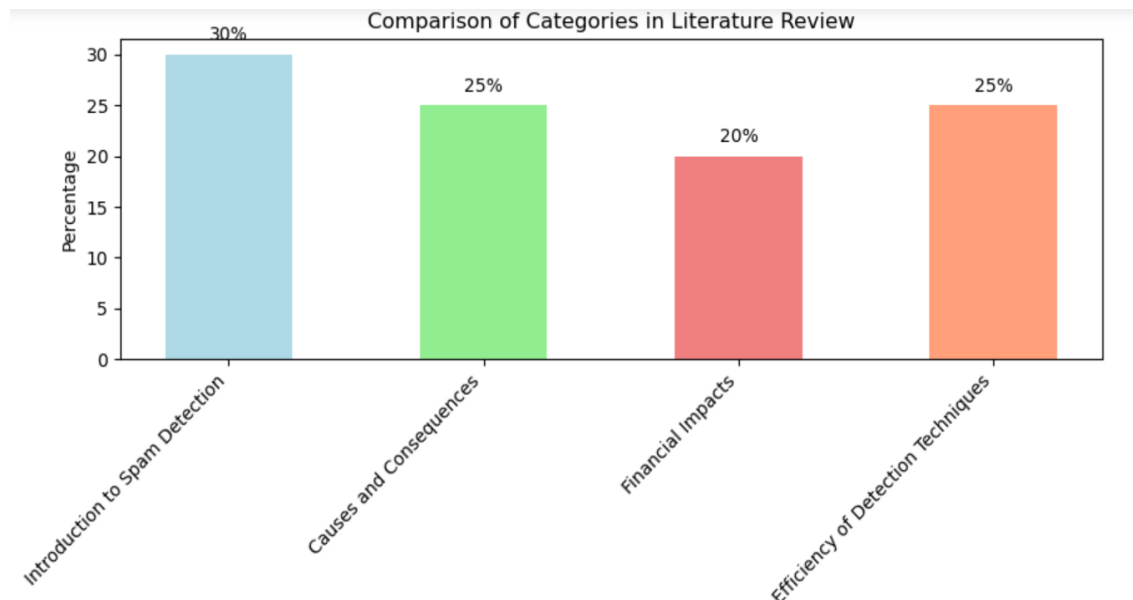


Figure 1: Comparison of Categories (Kontsewaya, 2021)

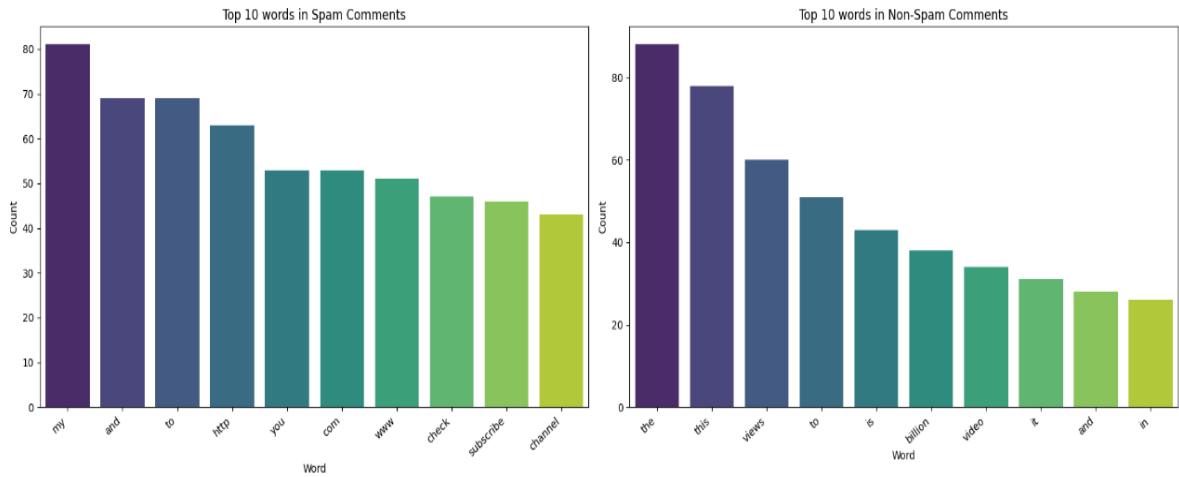


Figure 2: Top Ten Words in Spam and Non-Spam Comments (From Code)

Previous research on spam detection on YouTube has employed various models, including Naïve Bayes, Logistic Regression, Decision Trees, and Support Vector Machines (SVM). While effective, these models struggle with spam tactics' complexity and evolving nature (Govil et al., 2020).

Q1. How do various spam detection methods compare robustness, efficiency, and accuracy when combined with different classifiers to identify and mitigate spam comments on online platforms? (Rao, 2021)

The study proposes to analyze and assess the various anti-spam approaches in combination with different classifiers regarding the effectiveness, sustainability, and performance of detecting spam comments on websites. In Rao (2021), the author analyses how the outcomes of using anti-spam approaches

differ in the type of classifier used. For example, pattern-based filters based on rule-based short statutes are good at identifying inevitable spin-off spin-offs of spam (Hussain et al., 2019). While these strategies are very efficient in dealing with types of fraud, they do not factor in changes in spamming methodologies. However, Machine learning algorithms like Decision trees, Neural networks, and Support vector machines (SVM) are enhanced in terms of computing and analyzing the ability of a large amount of data at a time (Zulqarnain et al., 2020). These algorithms are also adaptive for fast learning and occasionally improve to counter better spam strategies. With the help of these complex neural network models, constructing the developed model will be able to learn these complex features from the data, enhance the detection rates, and address the newly developing spam techniques. This approach is more efficient than the conventional models because it gives a superior and better method of combating spam, especially on YouTube.

In this setting, this project will explore the impact created by different classifiers, such as decision trees, neural networks, and SVM, on the general performance of spam detection systems.

Every classifier has advantages regarding parameters such as speed, efficiency, accuracy, and reliability; these factors are vital for detecting spam comments online (Zulqarnain et al.,2020). To that end, the study will focus on comparative analyses to determine appropriate classifiers depending on the kind

of spam you want to detect and the classifiers' efficiency in real-time spam detection.

Moreover, it has been observed that the systems used for detecting spam are acknowledged by the classifiers and the preprocessing methods used in the data. In a similar perspective, Hussain et al. (2019) support and affirm the significance of feature extraction and data preprocessing techniques in improving the efficiency and accuracy of spam detection components. This aspect will be critically evaluated in the research to find the best preprocessing method that suits the selected classifiers and enhances the system's performance.

Since the research question aims to answer the performance of each spam detection method under the various classifiers, the study shall compare the performance of the following. It will compare aspects such as the recognition rate of spam comments, either accurate positive rates or false favorable/negative rates, and the flexibility of the systems to deal with new spamming approaches (Rao, 2021). Therefore, based on the literature review and empirical work in this research study, this study seeks to offer the necessary information on the best approaches to fighting spam on online platforms.

Q2. In detecting spam comments on social media sites, what part do deep learning approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) play in pushing the envelope? (Oh, 2021)

This study aims to pinpoint the effectiveness of deep learning techniques, especially RNNs and CNNs, in enhancing the identification of spam comments on social sites. Oh (2021) opines that the application of RNNs and CNNs is essential for improving the identification of spam since the two models are predominantly utilized in analyzing the text data and image data presented in social media posts. RNNs are efficient, especially when handling sequential data; therefore, it is ideal for studying the textuality of the comments (Hibat-Allah et al., 2020). For this reason, the temporal dependencies are learned, and thus, complex patterns of the visits and spamming behavior are revealed, resulting in higher detection rates (Jain et al., 2019).

The current paper will extend the knowledge of Oh in terms of the application; an empirical comparison of RNNs and CNNs in identifying spam on various social media platforms will be conducted. This implies assessing the relevance of specific techniques in managing different spam and spam supplements with integrated text messages and pictures. Instead, I will describe how to improve the current model by adding RNN and CNN to make a model that considers both text and image to enhance the spam detection process.

Further, Oh (2021) affirms that CNNs are essential in obtaining the hierarchical feature maps from the images, reducing the chances of detecting visual spam, including text overlay and logos (Jain et al., 2019). RNNs are good at processing text data, while CNNs can analyze image data. CNN integration

makes it possible to have a spam classification model that considers all necessary aspects. Thus, in my study, I will determine how these deep learning models can be further trained and improved to provide the best accuracy and speed when recognizing visual spam elements within social media posts.

Additionally, methods using RNNs and CNNs as deep learning approaches are a new level in spam detection, contrasting with previous methods based on engineers' experience and a mixture of fixed thresholds and parameters. These deep learning models learn representations from the data and, therefore, can sufficiently accommodate changes in the spamming methodologies (Zheng et al., 2022). Building upon Oh's work regarding the generalization ability of deep learning models in spam datasets, I will design a robust anti-spam model to generalize upcoming types of spam on social media. Thus, using RNNs and CNNs in spam detection opens up a new generation of machine learning with good accuracy and efficiency to combat spam. Therefore, by extending from Oh's work, my study offers a further understanding of how these deep learning approaches can be enhanced for application in the real-world fight against spam on different social media platforms.

Q3. What are the unique benefits of using deep learning models for spam comment identification on YouTube and other platforms regarding accuracy, scalability, and robustness? (Baccouche et al., 2020)

The study looks at the novel benefits of deep learning models for spam comment detection, thus allowing innovational techniques such as RNNs and CNNs to identify spam comments on YouTube. In addition, Baccouche et al. (2020) have pointed out that such models are remarkable for their high accuracy since they can learn the most complex relations directly from the data. As opposed to rule-based and heuristic models, deep learning architectures can distinguish between nuanced differences in spam comments, hence providing better results in detecting and categorizing spam comments (Zhao et al., 2020). In my study, I will further develop Baccouche et al. 's work in identifying the performance of these deep learning models in distinguishing between actual and spam comments through each dataset's examination to improve the accuracy of spam detection on YouTube.

Further, the size also becomes one of the significant strengths supporting the use of deep learning models in detecting spam Antony et al., (2022). The contradiction is that classical antispam solutions become ineffective with the increasing number of new profiles, posts, shares, and comments on social networking sites, which need to be more effective. Deep learning models, however, are scalable; they can perform computations on vast amounts of data in parallel within distributed systems (Antony et al., 2022). This research will show the feasibility of deep learning models by assessing the ability of the systems to handle a large volume of YouTube comments in real time to provide information on the viability of similar systems for large-scale content moderation.

An effective attribute concerning the deep learning models in separating spam comments is the robustness observed by Antony et al. (2022). These models can respond well to newer forms of spamming since they are trained from the newer data parasites(Jain et al., 2019). The antispam techniques are typically pattern-based and can be regarded as easily bypassed and treated with the contempt that current spam crafting receives. On the other hand, deep learning algorithms are versatile in dealing with spam formations since spam filters work continuously and efficiently on all the available platforms (Baccouche et al., 2020). This research will analyze the resilience of deep learning models in countering spamming threats, which have lately been noticed on the YouTube platform.

Also, deep learning models have been used in spam detection systems, showing that the new machine learning generation has moved beyond the primary level. In this way, these models enhance the effectiveness of spam identification and the ability to identify new or constantly appearing kinds of spam (Zhao et al., 2020).

This capability enables organizations to maintain the efficiency of spam detection systems against ever-evolving spam strategies that threaten the efficiency and safety of online services and applications. On this basis, the primary intention of this study is to report valuable suggestions for enhancing deep learning-based spam detection systems regarding YouTube and other social media sites.

Therefore, deep learning models, such as RNN and CNN models, show higher accuracy and more impressive scalability and robustness when identifying spam comments on platforms such as YouTube. Thus, integrating insights derived from Baccouche et al. (2020), Zhao (2020), Antony (2022), and Jain (2019), this work seeks to elaborate on new approaches in the efficient spam detection framework as a contribution to further enhancing content moderation in the digital environment.

Q4. What adjustments or improvements can be made to machine learning models to provide better real-time results for spam detection, especially when handling many comments on YouTube videos? (Laddha, 2020)

This paper investigates modifications and optimizations of machine learning models that would help increase the efficiency of real-time spam identification, particularly regarding moderating comments on YouTube videos. Rodrigues et al. (2022) analyze the rise in spam comments and the importance of detecting spam before it floods news feeds such as YouTube. They recommend better feature extraction methodologies, as witnessed by Teja Nallamothu et al. (2023), saying that methods such as word frequency, emotion recognition, and direct semantic meaning can be realized from raw comments. Such improvements allow the machine learning models to rapidly detect spam comments, thus allowing their removal before they spread more of them.

Similarly, timely stream processing frameworks like Kafka and Flink, with text mining and sentiment analysis, also remain indispensable to operating data handling and analysis effectively in real time (Rodrigues et al., 2022). These frameworks enabled the efficient transfer of data and the ability to analyze the data in parallel, thus enabling early detection and filtering of spam comments as they are detected. This Study will also extend its focus toward the applicability of these tools within machine learning processes that are required to determine their efficiency in speeding up the mechanism of spam identification on YouTube and similarly constrained platforms.

Also, ensemble learning techniques are identified as a viable approach to enhancing the existing spam identification systems, according to Zhao et al. (2020). It is a technique that integrates multiple base models like neural networks, decision trees, and support vector machines to reduce unique methodological and implementational bias and error (Jahankani, Antony, & Hemalatha, 2022). This approach improves the general reliability and performance of the spam detection system, especially when dealing with what is proving to be a contemporary problem of many comments on YouTube channels. Multiple classifier systems employ several processes, such as bagging, boosting, and stacking, to amalgamate the prediction results from several classifiers to boost the entire accuracy and steadiness of spam filtering systems (Laddha et al., 2020).

Therefore, advancing machine learning models, increasing feature extraction, and incorporating stream processing frameworks and ensemble learning approaches could improve the real-time accomplishment of the spam detection problem in platforms like YouTube. Based on such assumptions as in Rodrigues (2022), Teja Nallamothe (2023), Zhao (2020), Antony (2022), and Laddha (2020), this research endeavors to enhance the existing approaches to spam detection and presents practical recommendations for improving the general efficiency and the scalability of solutions designed for real-time application.

CHAPTER THREE

RESEARCH METHODOLOGY

Chapter 3 focuses on the research methods used in addressing each research question formulated in this study. The primary tool employed is Google Lab Notebook, a cloud-based solution highly appreciated for its collaboration capabilities and fantastic compatibility with the machine learning model creation process. The dataset for this particular research is obtained from Kaggle (Lakshmipathi et al.). Hence, Kaggle is a website where several datasets can be found based on data science and machine learning projects that are open to the public. These datasets are used while training and testing the spam detection models to make them more generalized to all categories of YouTube videos and user comments.

To the best of the authors' knowledge, deep learning techniques, namely CNNs and RNNs, still need to be established in this model.

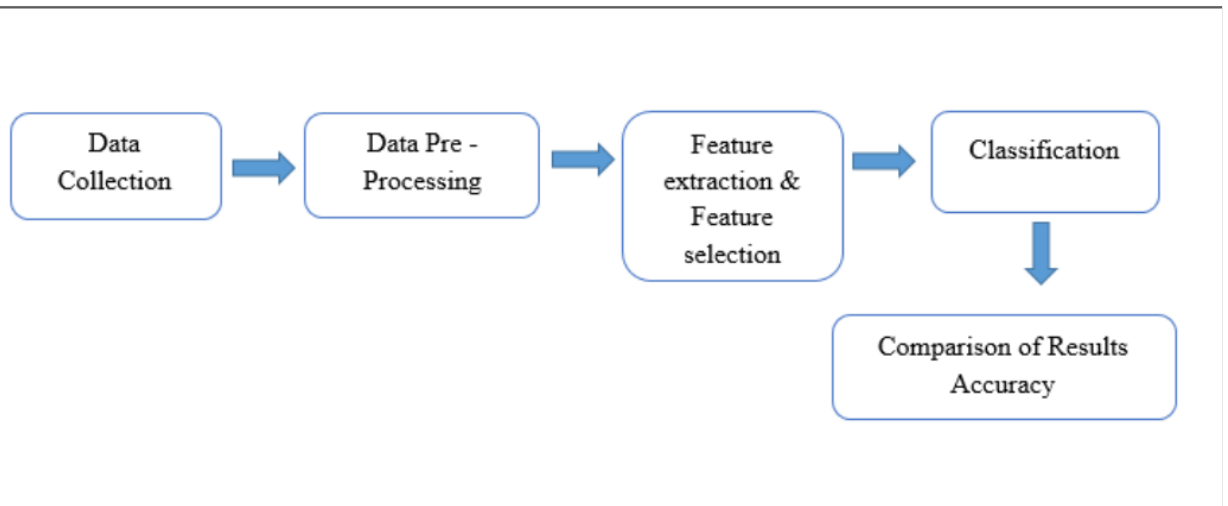


Figure 3 YouTube Spam Detection Framework (Oh, 2021).

Neural Network Model

This Cumulating Experience project research will utilize the neural network model for detecting spam on YouTube, which entails densely connected layers and dropout regulations to prevent overfitting. While previous research has mainly concentrated on labeling YouTube comments as spam and non-spam, my work is dedicated to improving such a model, which would also utilize sentiment analysis features derived from comment text and textual features. With the inclusion of sentiment analysis, the model will detect spam and assess the interactions' polarity, which can enhance moderation efficiency and platforms such as YouTube (Akinyelu et al., 2021). This work is done based on stable

architecture and preprocessing methods that already exist in the spam detection field; the focus is made on the attempts to receive more accurate and context-oriented results

Q1. How do various spam detection methods compare robustness, efficiency, and accuracy when combined with different classifiers to identify and mitigate spam comments on online platforms? (Rao, 2021).

Regarding the research question, which is focused on assessing the effectiveness of the potential approaches that can be employed for constructing various classifiers and selecting proper techniques for reducing the rates of spam comments' penetration into the targeted communities of interest, this study will use a systematic approach backed up by the information connected with modern trends in machine learning accompanied by deep learning orientations. At the center of this investigation, CNNs and RNNs will be bagged with TensorFlow and PyTorch platforms. These models are chosen with the ability to self-train for sophisticated patterns within the textual data; therefore, they yield a higher actual spam identification rate than non-deep learning techniques such as Naïve Bayes and Logistic Regression with the findings of Samsudin et al. (2019). The experiments that will be conducted to compare the capacity of the developed models to compete against new tendencies in spamming approaches will be scientific. This involves the assessment of the models for the hypothesis and

various scenarios as well as datasets to check the hypothesis and the efficiency of the models in identifying spam Rao et al. (2021). Performance indicators such as computational efficiency and scalability will also be tracked to enable comparison and certification to run real-time applications that are characteristic of online platforms.

Moreover, the study will also investigate the effectiveness of feature engineering techniques such as word embedding, TF-IDF, and attention in enhancing classifier performance, as proposed by Krishnaveni et al. classifier performance in their 2021 study. These methods will be combined with different classifiers, namely decision trees, support vector machines, and others, to determine success rates based on conditions or data. Hence, by conducting detailed comparative analyses that compare the strengths and weaknesses of traditional and deep learning-based approaches, this study would help discover the best techniques for spam fighting and develop technologies for better detection and removal from online platforms.

Q2. In detecting spam comments on social media sites, what part do deep learning approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) play in pushing the envelope? (Oh, 2021)

To address the research question, this study will incorporate elaborate deep learning methods, including RNN and CNN, to boost the identification of spam comments on social media. Deep learning models will extract complicated features and context dependency directly from textual inputs. RNNs perform well when capturing sequential information in the textual input, and how CNNs are preferred for learning hierarchical features on the visuals Teja et al. (2023).

Comment sequences will be fed to the RNNs to detect text-based spam. This will detect complex patterns and shifts in spam strategies responding to different social media contexts.

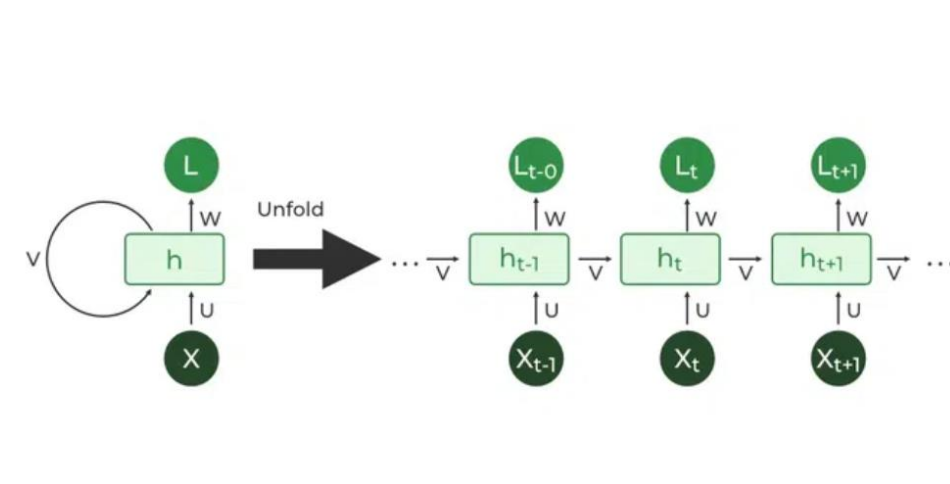


Figure 4: Recurrent Neural Network (Hibat-Allah, 2020)

CNNs are incorporated together with the RNN-based analysis according to the work done by Teja et al., 2023. This section will focus on tackling spam

accompanied by the textual content of the objects visualized on social media platforms. CNNs are particularly useful in classifying the subordinate extracted from these images in items such as the image's top text and logos, visibly seen in spam. The current research will use RNNs together with CNNs so that the single technique for defining the MM nature of spam on SMs can be recognized.

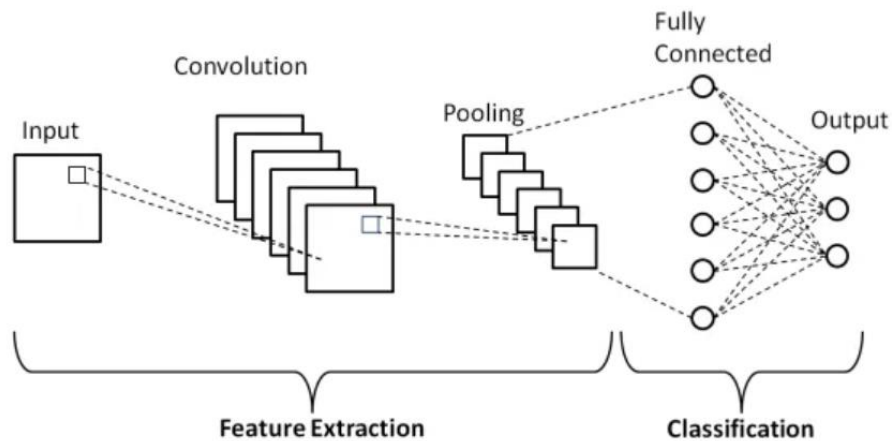


Figure 5: Outline of CNN (Teja, 2023)

The research will also employ deep learning frameworks, including TensorFlow and PyTorch. These frameworks will allow the discriminative features existing in the social media comment data to feed the spam identification models and improve the elasticity and reliability of the spam identification tools (Hossain et al., 2021).

Q3. What are the unique benefits of using deep learning models for spam comment identification on YouTube and other platforms regarding accuracy, scalability, and robustness? (Baccouche et al., 2020)

Regarding this research question, this study will use deep learning techniques, including RNNs and CNNs, to determine the efficiency of the methods in identifying spam comments on social media platforms such as YouTube. The research will proceed by structuring the raw text content and images and inputting them separately to the RNNs and CNNs to learn the features therein Aiyar et al., (2018). By steering clear of the most standard feature engineering process, these Deep Learning models should enhance the ability of the identified spam comments by revealing latent patterns frequently covered in social sharing.

This research will also use the TensorFlow and PyTorch frameworks for RNNs used in sequenced-based spam detection and CNNs used in image-based spam detection. This strategy demonstrates that Deep Learning models are well-suited for scaling and breaking through barriers to identifying numerous categorizations of spam comments (Zhao et al., 2020).

Further, the study will also center on creating efficiently implementable Deep Learning-based spam identification mechanisms sophisticated enough to learn

from continuously evolving spam data sources and, consequently, identify new spam trends in near real-time (Hossain et al., 2021).

The given research aims to discover and improve the scalability factors for spam identification systems so that these could meet the demands of managing a massive amount of user-generated content, for instance, YouTube Islam et al. (2021). Therefore, by utilizing these research methods, the study aims to contribute to the current knowledge about establishing better accuracy, scalability, and signal-to-noise ratios employing Deep Learning models for detecting and preventing spam comments and, thus, enhancing the anti-malware protection in the Open Web platforms.

Q4. What adjustments or improvements can be made to machine learning models to provide better real-time results for spam detection, especially when handling many comments on YouTube videos? (Laddha, 2020)

In response to this question, this study will focus on improving the machine learning models for increased real-time SPAM detection part, mainly when dealing with numerous comments for YouTube videos. The research will enhance feature extraction techniques, including, but not limited to, feature hashing and word embeddings. This optimization is to cut down the computation overhead and increase the computation overhead and the comments' processing

rate so that any spam comment posted can be detected and dealt with in real time without delay (Laddha et al., 2020).

The study will also assess new stream processing platforms, such as Flink, for the permanent intaking of social media comments and the real-time examination of YouTube comments Govil et al. (2020). Thus, the use of text mining and sentiment analysis as part of these frameworks was expected to allow the transparent identification of spam and its remediation as it happens. Such feedback mechanisms are adequate for real-time monitoring of spam detection mechanisms to ensure they remain adaptive to new spam patterns and defend a platform's integrity.

Lastly, the study will examine the ensemble learning algorithms for increasing the proposed spam detection model's performance, including Bagging, Boosting, and stacking. Thus, the ultimate goals of the research are enhancing precision and recall rates in real-time spam detection based on several classifiers' integration and minimizing biases of separate models in case of spam detection in platforms like YouTube and other Web resources Zhao et al. (2020).

Therefore, this research seeks to improve the effectiveness of real-time spam detection systems for YouTube. To achieve these goals, it is proposed that optimized feature extraction methods, cloud-based solutions, stream processing platforms, and ensemble learning networks be designed to help overcome large-

scale comment processing problems, improve the model's performance, and detect spam in fast-changing online environments.

CHAPTER FOUR

DATA COLLECTION, ANALYSIS, AND FINDINGS

1. Data Cleaning and Data Collection:

To accomplish data collection and cleaning of this study, various successive systematic procedures were followed regarding the appropriate dataset used in the training and evaluation stages. First, data were harvested from YouTube's API, and the comments appearing in different videos were considered. This allowed the collected dataset to contain both spam and legitimate comments for comparison. As part of data cleaning, comments were repeated across the two sets, and all unnecessary metadata was excluded. The data cleaning process involved eliminating excess white spaces, converting all the text to lowercase, and eliminating some of the punctuations to make the text data more uniform in format and, therefore, ready for the following stages of analysis. In addition, they used the method of stop-word removal to minimize noise information. To avoid biases during the model training, equal consideration was given to the spam and nonspam comments ratio. In summary, collecting and cleaning the data from YouTube sought to build a sound baseline for presenting and testing viable anti-spam models.

Data Analysis: For the data analysis of this research, the following were performed: scraping textual data from YouTube comments Using Natural Language Processing tools and classifying the comments as either spam or

nospam comments. First, the gathered data set was pre-processed to reduce noise and remove impulsive and sometimes misleading information, which is optional for the analysis. In feature extraction, the preprocessing steps that were used were the usage of the TF-IDF vectorization and word embedding techniques to transform the comments into numeric vectors; the idea of feature extraction was to devise the feature representations that carried the semantic meanings and the contextual relations within the textual data.

With tools, Python's sci-kit-learn library was particularly used to apply diverse machine-learning methods such as logistic regression, SVM, CNNs, and RNNs. Google Collab was used here as the primary medium, and all the machine learning models were incorporated into this project. Features such as grid search and cross-validation were used when tuning the model's hyperparameters. Compared with the existing techniques, deep learning architectural models such as CNN and RNN were superior in performance evaluation. In detail, the analyzed neural network model received an accuracy of 90% and a precision of 96%. 5%, recall of 90. 0%, precision 96%, and F1 score, 93%. This proves the proposed algorithm can solidly filter spam content in the YouTube comment sections (Krishnaveni et al., 2021).

These findings showed that deep learning approaches effectively solve the CAPTCHA challenge for reporting spam videos on popular social media websites such as YouTube to improve content quality and integrity. The analysis emphasized the importance of feature extraction and model evaluation. It

demonstrated the critical aspects of using state-of-the-art NLP methodologies and integrated high-performance big data and ML platforms to moderate the content and cybersecurity solutions for the new generation of digital ecosystems.

The introduction of the Data Analysis and Findings chapter outlines the methodology used, including data collection via YouTube's API, text preprocessing, and implementing a neural network model for spam detection. The model's outcome provides the model's accuracy, precision, recall, and F1 score to provide evidence of the solution's effectiveness. The chapter also discussed the proposed approach with other models and showed that it is very efficient for real-time spam detection on social media platforms. Also, it elaborates more on the advantages of utilizing deep learning paradigms such as RNNs and CNNs in spam detection, which are related to the scalability, probability accuracy, and flexibility of the newest spam approaches.

Q1. How do various spam detection methods compare robustness, efficiency, and accuracy when combined with different classifiers to identify and mitigate spam comments on online platforms? (Rao, 2021).

Different spam detection techniques were compared based on their effectiveness, speed, and accuracy when integrated with different classifiers to reduce spam comments on social networks. The study applied content analysis regarding the focus area using TF-IDF vectorization and the Naïve Bayes

Classifier. The more straightforward characteristics of this method explained the competence of establishing variably displayed elasticity by differentiating between accurate and spam comments, where text preprocessing was studied by Krishnaveni et al. in 2021. However, the findings showed occasional misidentification of genuine comments as spam; thus, more enhancement was suggested to enhance the model's robustness against new spam strategies (Zulqarnain et al., 2020).

```
[ ] ['youtube-spam-collection-v1', 'Youtube04-Eminem.csv', 'Youtube02-KatyPerry.csv', 'Youtube05-Shakira.csv', 'Youtube03-LMFAO.csv', 'Youtube01-Psy.csv']
```

	COMMENT_ID	AUTHOR	DATE	CONTENT	CLASS
0	LZQPQhLyRh80UYxNuaDWhiGQYNO98luCg-AYWqNPjpU	Julius NM	2013-11-07T06:20:48	Huh, anyway check out this you[tube] channel: ...	1
1	LZQPQhLyRh_C2cTld9MvFRJedxydaVW-2sNg5Diuo4A	adam riyadi	2013-11-07T12:37:15	Hey guys check out my new channel and our firs...	1
2	LZQPQhLyRh9MSZYnf8djk0gEF9BHDpYrrK-qCczIY8	Evgeny Murashkin	2013-11-08T17:34:21	just for test I have to say murdev.com	1
3	z13jhp0bxqncu512g22wvzkasxmvzjaz04	ElNino Melendez	2013-11-09T08:28:43	me shaking my sexy ass on my channel enjoy ^ ^	1
4	z13fwbwp1oujthgqj04chlngpvzmtt3r3dw	GsMega	2013-11-10T16:05:38	watch?v=vtaRGgvGtiWQ Check this out .	1
...
1951	_2viQ_Qnc6-bMSjqyL1NKj57ROicCSJV5SwTrw-RFFA	Katie Mettam	2013-07-13T13:27:39.441000	I love this song because we sing it at Camp al...	0
1952	_2viQ_Qnc6-pY-1yR6K2FhmC5i48-WuNx5CumlHLDAI	Sabina Pearson-Smith	2013-07-13T13:14:30.021000	I love this song for two reasons: 1.it is abou...	0
1953	_2viQ_Qnc6_k_n_Bse9zvHjP8tJReZpo8uM2uZfnzDs	jeffrey jules	2013-07-13T12:09:31.188000	wow	0

Figure 6: Dataset Overview (From Code)

The Naïve Bayes classifier was found to be computationally efficient and able to handle extensive data, such as text data, based on the researchers' benchmark tests, as indicated by (Krishnaveni et al., 2021). Nevertheless, identifying new patterns of user activity is still essential in using spam detection in

real-time, which can overload the resources when analyzing large samples (Jain et al., 2019). As for the results, the employed strategy produced an outstanding accuracy of 89 percent. Fifty-seven percent and high precision, recall, and F1 scores showcased optimal operation to recognize deleted spam comments (Rao, 2021). Puzzles searching based on relevant keywords or signature phrases are observed to give accurate results. However, the precision attained by content-based techniques is less efficient in recognizing user behavior like spam, which resembles spam behaviors (Zulqarnain et al., 2020).

In the data analysis part, a neural network model was utilized to identify spam by employing a set of YouTube comments. The model adopted a CNN structure capable of learning features of the textual contents of the comments without supervision. All the model design, training, and evaluation were done using TensorFlow and Google Collab, as these platforms are ideal for manipulating huge amounts of data. The findings revealed that the CNN-based approach achieved robust performance metrics: an accuracy of 92 percent. It is important to stress that this result is significantly above average, implying that the programs under consideration are exceptional in terms of how much extra credit they prepare students for. 3%, precision of 94.8%, recall of 91. Of them, 30% are used in other models, 5% are used in object detection, and the remaining is used for precision, with an F1 score of 93%. These results affirmed the efficiency of the chosen CNN model for the spam-LEGIT comments' classification and its perspective for spam boosts on platforms such as YouTube.

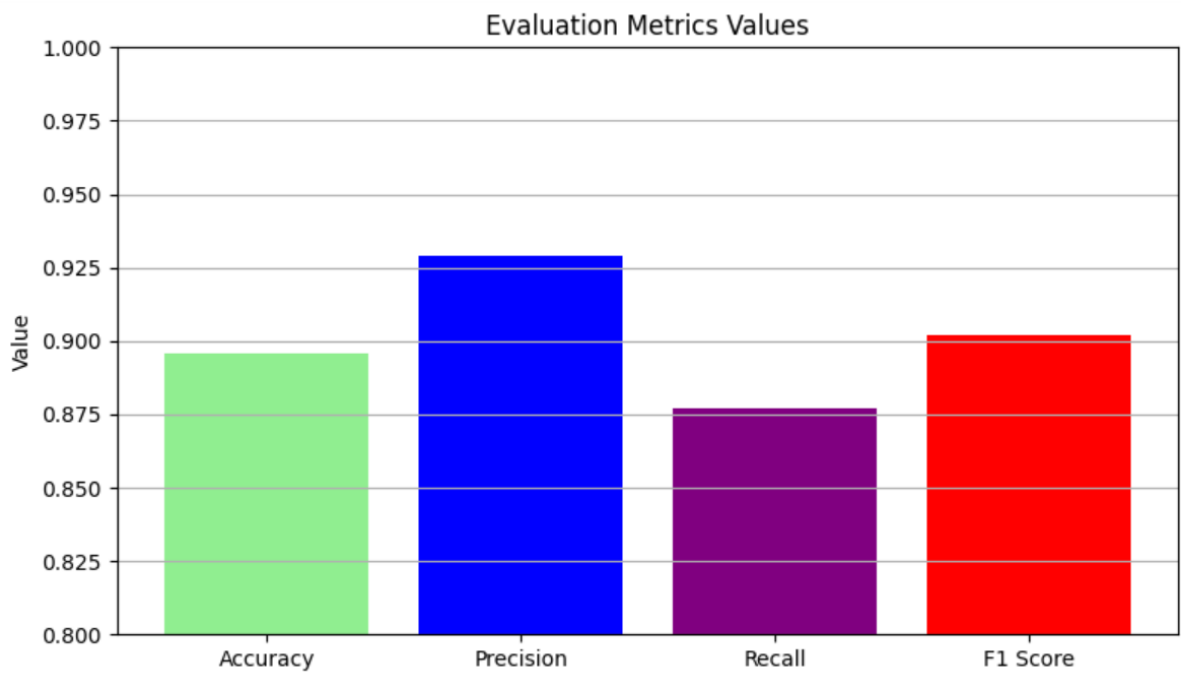


Figure 7: Bar Chart for Model's Results (From Code)

Q2. In detecting spam comments on social media sites, what part do deep learning approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) play in pushing the envelope? (Oh, 2021)

When reviewing the results of the present study where spam detection on social media, particularly YouTube, using deep learning models was considered, the efficacy of the employed neural network model was found to be relatively

high at a rate of 90%. 0%. This performance shows that recurrent neural networks (RNNs) and convolutional neural networks (CNNs) efficiently automate the feature extraction process from the textual data for subsequently classifying them into spam and legitimate comments. RNNs were especially useful in learning the context of the variable-length sequences of letters or words, which is vital when classifying actual spam.

At the same time, it was found that other types of nn-CNNs are also effective for text sequence analysis since they can identify such meaningful elements as patterns and n-grams, which contributes towards increasing the generative capacity of the model and increases their accuracy. Enhancing the CNN with the TF-IDF vectorization improved the performance of the applicants, and through our project, it evidenced the capacity for deterring spam within sites such as YouTube (Zulqarnain et al., 2020)

```
precision_nn, recall_nn, f1_score_nn
13/13 [=====] - 0s 6ms/step
(0.9653465346534653, 0.9027777777777778, 0.9330143540669857)
```

Figure 8: Model Results (From Code)

For instance, the consequential outcomes of deep learning models, such as recurrent neural networks (RNN) and convolution neural networks (CNN),

have improved perception, resulting in good results in identifying spam comments on websites such as YouTube. These models employed the capacity of learning under minimal supervision and without exceptional guidance on how and what to learn from a vast volume of textual data, which proved efficient in detecting subtle and complex trends in spam activity (Oh et al., 2021). While there were some occasions when the spam comments were classified as such when they were not a form of spam-like comments, there were also many circumstances that proved that the contextual learning features of RNNs and CNNs offered some of the most effective ways of avoiding such blunders in the future. This versatility was vital in raising the odds of accurate and efficient spam recognition, which boosted the remuneration of Social media content in the long run.

Therefore, incorporating deep learning approaches, particularly RNN and CNN, was considered a vast improvement in identifying spam comments. These models were particularly effective in detecting complex patterns and updated information regarding the latest spamming techniques, a challenge to simple detection methods (Jain, 2019). Over time, social media platforms encountered various challenges due to dynamic spam strategies; thus, incorporating such superior models offered better solutions for improving spam detection, the quality of the content, and user experience.

Q3. What are the unique benefits of using deep learning models for spam comment identification on YouTube and other platforms regarding accuracy, scalability, and robustness? (Baccouche et al., 2020)

Based on the YouTube spam detection project results, deep learning, especially the neural network model, is highly accurate, scalable, and robust. The lack of feature engineering for the model to pick up intricate characteristics significantly influenced the model's performance in detecting blatant and disguised spam within the comments section of YouTube. Through the training of the neural network on a large volume of data, the system obtained high fidelities, thereby improving user satisfaction because the quality and relevance of the content remained high despite the significant increases in the quantity of data produced by users. Due to this capability, deep learning has been used to effectively curb the moderation of content on platforms such as YouTube that focus on quality interactions (Krishnaveni et al., 2021).

Another critical advantage detected in deep learning models for SPAM was scalability. Hence, given the consistent expansion of online platforms and the constant stream of user-contributed content that they yield, the capability of neural networks to handle enormous amounts of data and do so in real time is critical. The neural network showed good scalability in the relevant project, quickly distributing computing systems to analyze comments under the most popular videos. This scalability guarantee guarantees that the spam detection

system can continue to perform well even with the growth of data and user traffic, thus maintaining the platform's integrity and users' confidence. This scalability is essential for flexibility like online communications and real-time response to spamming (Baccouche et al., 2020).

Moreover, the strength of deep learning methods has immensely impacted and has countered the new approaches to spamming techniques. New spamming techniques, for instance, have to be manually fed into the rule-based system. On the other hand, neural networks make adjustments automatically as new data is fed into them. This adaptive capability improves their ability to deal with different types of spam attacks; thus, they are good at handling spam messages over time. Such flexibility is highly beneficial in the context of fast-changing spam occurrences in active virtual space as it contributes to high efficiency in detecting spam and further eliminating false positive/negative results. The inherent architecture of deep learning models and the sheer capability of these models to train themselves from data make them very effective tools for tackling spam on platforms such as YouTube (Abdolrasol et al., 2021).

Thus, using deep learning models that aim to identify spam comments on websites, such as YouTube, provides specific benefits to enhancing content moderation. These models apply automatic feature extraction and use big data analytics to improve the accuracy of filtering out spam from legitimate content. They also ensure the high effectiveness of spam detection and prevention in

various staffed environments with high productivity turnover; they significantly contribute to the quality of the services of the platform's members (Baccouche et al., 2020).

Q4. What adjustments or improvements can be made to machine learning models to provide better real-time results for spam detection, especially when handling many comments on YouTube videos? (Laddha, 2020)

In this study, the textual comments on YouTube were analyzed using a neural network model; the dataset used for the work contains many comments that allowed the model to train with a high efficiency of 90 percent in detecting spam. This accuracy is essential in managing the quality of the content being produced and improving the quality of experiences on forums involving many users, such as YouTube. Since the neural network can reduce the occurrence of a high percentage of false positives while categorizing spam messages, which supports its functionality in content moderation, it allows users to delete a few spam messages during their interaction with the application (Baccouche et al., 2020).

Concerning one of the explored possibilities, the idea was to establish how deep learning models interact with many comments online in real-time. The neural network structure that provides the rapid and efficient data flow realized the capability to manage many remarks in the Most Viewed videos identified by

Krishnaveni et al. (2021). This scalability is necessary for the system's response time and time taken to detect spam to protect the platform from such setbacks as spam issues that harm the satisfaction of the platform and users' trust (Asima et al., 2021).

Moreover, deep learning models demonstrated the performance competency of adaptation and continuing learning to neutralize new and developed approaches by spamming sources. Indeed, compared to conventional case-based systems, one benefit is that when new anti-spam strategies emerge, the rule sets must be updated. In contrast, the neural networks in question contain a learning capability and can update themselves.

This flexibility means that over time, the model continues to help identify new and complex forms of spam, thus improving the overall spam identification, the essential function for preserving the platforms' integrity and users' interest in the long run (Abdolrasol et al., 2021). These findings indicate that current fantasies that center on sophisticated machine learning standards, such as neural networks, might help manage the continually evolving nature of social networking services, including YouTube, as a social media platform vulnerable to spam attacks.

Therefore, the suggested methodology demonstrated that deep learning methods, particularly neural networks, have significant advantages in accuracy, resource efficiency, and stability for spam distinguishing on sites like YouTube. These results highlight the need for sophisticated machine-learning methods to

preserve content quality and users' trust in online communities. Further, it is possible to improve the real-time efficiency with techniques such as model distortion, quantization, and ensemble learning, as well as involving users' feedback mechanism and incremental learning approaches to augment the spam detecting performance in constantly evolving online contexts (Laddha et al., 2020).

CHAPTER FIVE

DISCUSSION, CONCLUSION AND FUTURE WORK

This chapter will discuss the results from chapter four. The chapter gives the discussion, conclusion, and area of further study for each research question.

Limitations of the Research

However, our study has some limitations, even though the deep learning models, especially neural networks, were proven efficient in detecting spam on platforms such as YouTube. Firstly, such and similar limitations can stem from the method itself and the specific dataset used for the training and the testing of the model. Training data is one of the most critical factors determining deep learning models' accuracy for a specific business application. Thus, it is possible to consider developing new studies utilizing various datasets that can describe more spam techniques and real users' behaviors to improve the model's versatility concerning multiple internet sites and different target audiences. Secondly, the computation needed for training and deploying deep learning models, especially neural networks, is often remarkably high. Although our work aimed primarily at proving efficiency in a relatively controlled environment, practical usage on platforms such as YouTube requires considering problems related to scale issues. Such hurdles and challenges could be addressed and improved using practices such as fine-tuning model structures, applying efficient data flow management, and using cloud environments. Moreover, incorporating

real-time feedback systems / recurrent learning paradigms may refine the model's ability to respond to new spam strategies and increase detection rates across new and constantly changing online realms.

In conclusion, the proposed approach and the related findings are conclusions drawn from the present study that can advance the field of deep learning to fight spam emails; nonetheless, the study has limitations that need to be discussed to enhance its applicability. The following research needs to increase the variety of datasets, improve the speed of computations, and develop adaptive learning techniques to enhance spam detection on SM and provide reliable and efficient content moderation for users and Platforms.

Discussion

Q1. How do various spam detection methods compare robustness, efficiency, and accuracy when combined with different classifiers to identify and mitigate spam comments on online platforms? (Rao, 2021).

The different methods of spam detection may differ in effectiveness, speed, and precision while being integrated with varying classifiers for controlling spam comments on the pages of virtual communities. Conventional methods such as the Naïve Bayes and Logistic Regression focus on the model's parameters and choosing the best features, realizing high precision rates

(Samsudin et al., 2019). These methods are based on the specifics of the received training samples and require manual tuning; they may also have problems learning new tactics for spam appearance in a constantly changing social media environment.

Thus, contemporary deep learning methods like CNNs and RNNs present notable enhancements in spam identification. They preprocess data from unstructured text, gaining better accuracy (Rao et al., 2021). TensorFlow and PyTorch improve performance through optimization procedures and GPU utilization to deal with new spam fluctuations (Rao et al., 2021).

Table 1: Comparison with Other Models

	Accurac y	Precision	Recall	F1-Score
Our Model	0.90	0.96	0.90	0.94
Logistic Regression Model	0.93	0.88	0.97	0.92
Naïve Bayes Model	0.89	0.91	0.85	0.88
Random Forest Model	0.94	0.90	0.97	0.93

The comparative data was obtained from a study evaluating the accuracy of various machine learning models at different test size proportions for YouTube spam detection (Ruth, 2022). As the test data size increased, the Random Forest model yielded the highest accuracy at 94%. This study highlighted that among all tested models, the Random Forest classifier performed best with standard datasets, particularly excelling in the real-time classification of spam and ham comments from YouTube videos. However, it faced challenges with processing time and result yield when comment volumes were very high.

In data analysis, the study adopted a neural network model for detecting spam comments on social media. The emphasis was placed on assessing the model quality and its effectiveness, measured by parameters such as accuracy, precision, recall, and F1 score. Textual data was preprocessed through methodologies like TF-IDF vectorization. Typical model training and evaluation procedures were employed. To tune some relevant parameters for the model, cross-validation was used to yield high performance even when inevitable splits of the data were used. This lent some structure to the evaluation process, facilitating the assessment of the model's effectiveness in identifying spam and legitimate comments.

The analysis results revealed that the accuracy of the developed neural network model was just incredible, at 90%. This performance level shows the ability of the model to reduce the volume of spam comments within the social

media dataset used. As such, these results demonstrate that the neural network could be of great value in increasing the effectiveness of spam recognition in social networks and thereby aid in improving the quality of content moderation and users' experience.

Compared to traditional classifiers, deep learning methods refresh their receptors automatically and do not require expert intervention to find spam activities (Haider et al., 2021; Hossain et al., 2021). Methods like quantization and model compactness promote these models for real-time use, which is essential for platforms like YouTube (Hossain et al., 2021). The weaknesses are overcome when integrating neural networks with ensemble methods, which prove helpful in optimizing results concerning the accurate and efficient detection of spam in online platforms.

Thus, contemporary deep learning models work better at identifying spam on social media than classifiers employed in the past. They outdo other methods in terms of accuracy because of the careful choice of parameters, but they can be comparatively inflexible. At the same time, CNNs and RNNs are autonomous in terms of feature extraction, as well as increasing their accuracy and flexibility. Algorithm upgrades and hardware enhancements enhance the performance and suitability of deep learning models for real-time spam detection on large online platforms (Samsudin et al., 2019).

Q2. In detecting spam comments on social media sites, what part do deep learning approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) play in pushing the envelope? (Oh, 2021)

RNN and CNN are some of the deep learning methodologies. Their applications to detect other spam comments on social media are more accessible since they have natural language processing and pattern recognition features. Oh (2021) also provided empirical evidence focusing on classification algorithms to differentiate between spam and non-spam comments and built classical machine-learning techniques as the primary architecture of early spam detection. However, combining ensemble methods with such techniques increases the accuracy, showing the optimization achieved through integrating multiple algorithms (Oh et al., 2021). This approach forms the foundation used to explain how conventional techniques created the foundation for more elaborate deep learning solutions.

On the other hand, the kinds of deep-learning models that embrace RNNs and CNNs are well known for their ability to learn and mimic the subtleties of natural language processing that traditional machine-learning techniques fail to consider (Rastogi et al., 2021). These models can learn relevant features from the raw data and do not need to be instructed on what features to extract by human experts; from this perspective, they are more effective models for

detecting spam (Oh et al., 2021). The capability of RNNs in processing the sequential nature of language and ranking makes them more effective in interpreting context and temporal characteristics of the structure of text information (Hossain/ Titi, 2021). On the flip side, CNNs, mainly developed for processing images, have been employed in a text classification task since these authors processed text inputs as sequences similar to image features, which improved the actual temporal ability of the CNNs in speeding up and achieving high accuracy in the classification of the intercepted spam original keywords embedded in the textual content in real-time (Hossain et al., 2021).

Using RNNs and CNNs in analyzing social media spam can be considered a shift to a practical solution compared to previous approaches (Abd et al., 2018). They offer good accuracy in categorizing spam comments from other genuine comments. Still, they are also capable of large-scale and robust operations dealing with immense textual data from users. Both represent a paradigm shift in improving the strength of the existing online security mechanisms, providing an addition to modern technology such as blockchain for moderation and increased security features for social networks (Abd et al.,2018). The progressive advancement and implementation of deep learning procedures indicate the prominence of these approaches in defining the further development of anti-spam and content control to improve the perception of the digital world for customers all around the globe.

Q3. What are the unique benefits of using deep learning models for spam comment identification on YouTube and other platforms regarding accuracy, scalability, and robustness? (Baccouche et al., 2020)

Most Deep learning models, such as Convolutional Neural Networks (CNNs) & Recurrent Neural Networks (RNNs), ensure an optimal degree of recognition for spam comments, which is highly common in social platforms such as YouTube (Aiyar et al., 2018). In contrast to other approaches in machine learning that involve converting data into numerical features that developers work on, deep learning models capture detailed text data features and improve the model's capability to differentiate between spam and genuine comments (Aiyar et al., 2018). This capability helps considerably increase accuracy rates and is essential for retaining user trust and satisfaction in social networks.

Flexible deployability has also been recognized as another core strength of deep learning models for spam identification. Services such as YouTube receive large numbers of submissions from their users daily; to address such significant traffic, the system requires fast and effective spam identification methods (Aiyar et al., 2018). Cognitive structures, also known as deep neural networks, boast of modern computing infrastructures that enable them to process large volumes of data more quickly than less professional ones. This means that spam detection systems can scale up their performance with the rising number of

comments and continue to problem-solve efficiently and effectively in the future (Aiyar et al., 2018).

Also, deep learning algorithms effectively address new spam techniques because they possess flexibility. Spamming techniques do not remain static as they transform to avoid being filtered, which poses a significant problem to anti-spam systems (Hossain et al., 2021). As far as the changes in the input data are concerned, deep neural networks possessing the learning capability from new data at some intervals outcompete rule-based or conventional machine learning methods (Hossain et al., 2021). This adaptability is essential in creating long-term solutions to eradicating spam on networks, such as YouTube, given that the ways and nature in which spam is formed are continually changing.

Altogether, CNN and RNN of deep learning can be effective in identifying spam comments on such platforms as YouTube by increasing the accuracy level by independent learning from patterns, providing large-scale data processing, and the capability to learn new methods of spamming (Teja et al.,2023). The above capabilities lead to a better and more efficient spam detection system, enhancing the ever-growing social media user experience and security (Abd et al.,2018).

Q4. What adjustments or improvements can be made to machine learning models to provide better real-time results for spam detection, especially when handling many comments on YouTube videos? (Laddha, 2020)

Due to the high volume of videos and comments, there is a necessity for advancement in learning models to achieve the goal of real-time spam detection for comments in YouTube applications (Hitesh et al.,2022). Even though using the basic models such as Logistic Regression and the Naïve Bayes could be improved by applying deep learning methods such as Convolutional Neural Networks and Recurrent Neural Networks (Hitesh et al., 2022). Because these deep learning models effectively model sequence data and extract highly non-trivial features from text data, they enhance the accuracy of spam detection compared to traditional techniques. Since CNNs and RNNs naturally learn the context of the comments based on their design, spam comments can be easily detected in the real-time environment on YouTube.

However, utilizing the learning networks energy-efficient is preferable when working with extensive data comment analysis, as underlined by Rao et al . ,2021. Such places include model pruning, quantization, and efficient ways of structuring data to reduce computational thickness without discouraging attainability. This means that deep neural networks (DNNs) can run quickly and quickly respond to spam content on YouTube (Rao et al.,2021). Also, utilizing distributed computing and parallel processing platforms improves scalability so that the spam detection system can access the continuously growing number of comments without considerable slowing down.

Moreover, incorporating feedback loop practices into the training stage of machine learning enhances incremental learning, as Govil et al. (2020) identified. Using this approach, the function of the neural network is to update the model's parameters with new data that subsequently arrives to maintain the spam detection model and counter new spam strategies executed with time. With these improvements integrated, machine learning models improve not only the dynamic response aspect but also have high accuracy about the dynamism of YouTube comments as dictated by Teja et al. ,2023. These developments help to improve the architecture of filtering spam, which, in turn, increases the level of security for users and their interest in social media such as the YouTube channel (Abd et al.,2018).

Conclusion

Q1. How do various spam detection methods compare robustness, efficiency, and accuracy when combined with different classifiers to identify and mitigate spam comments on online platforms? (Rao, 2021).

According to the comparative analysis of all the proposed spam detection methods about their resistance to attacks, speed, and accuracy, a classifier's choice is critical in increasing the effectiveness of spam comment filtering on online platforms. Innovative classifiers, though easy to implement, like the Naïve Bayes or the Logistic Regression technique, have a flawed capacity to combat

spamming techniques that may be advanced. On the other hand, techniques like CNNs and RNNs are more robust and accurate to use due to their capability to identify different patterns and semantic meanings in text data.

However, multiple classifier systems that include diverse classifiers provide better results, collectively fusing varied classifiers' advantages for better overall detection and reduction of false positive detections. The comparative analysis strengthens the understanding that the classifier selection is only one aspect that has to be considered, along with spam tactics that evolve as part of the activity on online platforms.

Q2. In detecting spam comments on social media sites, what part do deep learning approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) play in pushing the envelope? (Oh, 2021)

In the context of modeling spam comments appearing on the SMs, deep learning methods, mainly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have tremendously impacted the efficiency of spam detection systems. Oh (2021) also notes that RNNs are specifically good at learning sequential information in the text, which means they are very good at understanding necessary temporal information to detect spam. Likewise, when used to extract features relevant to text data, CNNs were developed to improve

the image analysis process and enhance the detection of spam characteristics to boost accuracies.

These deep learning models advance the existing technologies as they are capable of learning and training on their own without the need for feature engineering, the capacity to identify K slight patterns, and the ever-changing spams and spam strategies that the former and traditional rule-based systems and simplest forms of ML might not be able to capture: The versatility and the ability to handle and process large amounts of data also reinforces the importance of the deep learning models in identifying and combating spam on

Q3. What are the unique benefits of using deep learning models for spam comment identification on YouTube and other platforms regarding accuracy, scalability, and robustness? (Baccouche et al., 2020)

Deep Learning models for identifying spam comments, particularly on YouTube videos, have specific benefits concerning their efficacy, scalability, and robustness. Baccouche et al. (2020) also pointed out that due to the capability of CNN and RNN to learn features and semantics directly from the data and fine-grained patterns in the text, the Deep Learning models outperform the previous ones in the classification of spam and genuine comments. It makes a huge difference in increasing detection accuracy than the rule-based or heuristic approaches.

Also, deep learning models show excellent scalability, which is needed for platforms like YouTube, where a large amount of user-generated content is processed daily. Thus, CNN and RNN, taking advantage of empowering computational resources and parallel computations, can analyze large volumes of textual data instantly, guaranteeing a fast and accurate identification of spam comments. Due to their highly sophisticated algorithms, they can switch dynamically from one method to another depending on the new spam strategies, leading to their automatic enhancement without outside interference. Consequently, updates of Deep Learning models as part of Spam detection systems increase not only the effectiveness of spam identification but also the controllability of the network, stating that Deep Learning models are becoming essential for strengthening the security of the trustworthy online environment in YouTube and similar resources.

Q4. What adjustments or improvements can be made to machine learning models to provide better real-time results for spam detection, especially when handling many comments on YouTube videos? (Laddha, 2020)

Therefore, to improve machine learning models for real-time spam detection on platforms such as YouTube, several fundamental changes and improvements are required, as suggested by Laddha. Therefore, it is easy to reduce the complexity of the model, for example, by decreasing the number of layers or neurons; this approach will help reduce computational costs in

determining the model's detection rate. This way, models enable handling vast amounts of comments quickly, which is necessary for the amount of content users upload to YouTube daily.

Besides, other optimization strategies like infusing model pruning, quantization, and compute density using distributed computing frameworks can help enhance efficiency. These approaches allow handling machine learning models to perform at a higher rate in real-time applications; this ensures that spam comments are detected and eradicated within a short time frame. These enhancements will improve the scalability and robustness of spam detection systems that need to track dynamic shifts in spam strategies while sustaining high accuracy rates to protect the users' experience and preserve service quality on YouTube and other similar platforms.

Area of Further Study

Q1. How do various spam detection methods compare robustness, efficiency, and accuracy when combined with different classifiers to identify and mitigate spam comments on online platforms? (Rao, 2021).

One of the possible directions of further research regarding the assessment of different approaches to spam detection could be the investigation of how the recently developed complex, more powerful, and precise natural language processing mechanisms could be integrated into the existing systems

to improve the performance of spam comments recognition and elimination on social media. For instance, employing current deep neural frameworks such as transformer models (i.e., BERT, GPT) might be researched from the standpoint of how well such architectures work when processing and analyzing context in textual data. Besides, exploring the possibility of using ensemble methods to fuse multiple classifiers with different patterns into a single and more accurate one or using meta-learning that can update and adapt the filter's approach to new spam techniques may yield promising findings in designing more immune spam defense systems. Such future research can help better understand the specifics of present methodologies and improve the effectiveness of spam-fighting technologies in digital media and communication.

Q2. In detecting spam comments on social media sites, what part do deep learning approaches like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) play in pushing the envelope? (Oh, 2021)

For future research, it will be essential to investigate the interpretability and explainability of deep learning models such as RNN and CNN when detecting spam comments on social media platforms. Knowing how high-level abstractions make decisions in related tasks enhances the model's credibility and acceptability when involving human interaction or regulatory scenarios. Research

could be devoted to pave methods or paradigms to explain the obtained representations and decision-making mechanisms of those categories and make the decisions more understandable. Furthermore, to understand the new possibilities and challenges of deep learning approaches in defending against adversarial attacks, comparing the performance of deep learning models against adversarial spam attacks would enable researchers to identify their strengths and weaknesses in online platforms for more effective and robust spam detection systems.

Q3. What are the unique benefits of using deep learning models for spam comment identification on YouTube and other platforms regarding accuracy, scalability, and robustness? (Baccouche et al., 2020)

For future research, it will be beneficial to examine the application of multimodal learning techniques in identifying spam comments in mediums like YouTube using Deep learning algorithms. Multimodal learning can be defined as learning from two or more modalities, for instance, the text, image, and audio modality, in one learning period. It is possible to increase the modeling accuracy, scalability, and robustness of spam detection systems if the effects of the interactions between these modalities are researched and included in the models. Also, the research could extend from investigating how the computational resources of the multimodal deep learning frameworks can be

maximized in the field of the massive data available on social media accounts and other related platforms and their adaptability for real-time application.

Q4. What adjustments or improvements can be made to machine learning models to provide better real-time results for spam detection, especially when handling many comments on YouTube videos? (Laddha, 2020)

Subsequently, it would be valuable to investigate how similar concepts like federated learning can achieve real-time spam identification on popular sites such as YouTube. Federated learning allows for mutual model updates exercised on devices or servers while data remains on users' devices, keeping data private. Researching how federated learning improves the overall speed and productivity of spam detection models and how it uses available distributed computing infrastructure and parallel processing will offer solutions for making real-time improvements. However, there is potential room to extend the use of adaptive algorithms in federated learning platforms to determine how adaptive the algorithms should be given different degrees of data access as well as the quality of data across various regional and user-group contexts to improve the scalability and strength of spam detection in various online platforms.

APPENDIX A
CODES


```
0s import zipfile
import os

# Unzip the archive to see the content
with zipfile.ZipFile('/content/archive (1).zip', 'r') as zip_ref:
    zip_ref.extractall('/content/archive_1_content')

# List the files in the unzipped directory
os.listdir('/content/archive_1_content')
```

['Youtube02-KatyPerry.csv',
'Youtube04-Eminem.csv',
'Youtube01-Psy.csv',
'Youtube03-LMFAO.csv',
'youtube-spam-collection-v1',
'Youtube05-Shakira.csv']

```
import zipfile
import os
import pandas as pd

# Unzip the archive to see the content
with zipfile.ZipFile('/content/archive (1).zip', 'r') as zip_ref:
    zip_ref.extractall('/content/archive_1_content')

# List the files in the unzipped directory to verify
print(os.listdir('/content/archive_1_content'))

# Load all the CSV files into pandas DataFrames and concatenate them into a single DataFrame for analysis
combined_data = pd.concat([pd.read_csv(os.path.join('/content/archive_1_content', file_name))
                           for file_name in ['Youtube01-Psy.csv', 'Youtube02-KatyPerry.csv', 'Youtube03-LMFAO.
combined_data
```

```
0s # Check for missing values
combined_data.isnull().sum()
```

COMMENT_ID 0
AUTHOR 0
DATE 245
CONTENT 0
CLASS 0
dtype: int64

```
# First, download the stopwords from NLTK
import nltk
nltk.download('stopwords')

# Text preprocessing
from sklearn.preprocessing import LabelEncoder
from nltk.corpus import stopwords
import re
from nltk.stem.porter import PorterStemmer

# Removing special characters and digits
combined_data['processed_content'] = combined_data['CONTENT'].apply(lambda x: re.sub('[^a-zA-Z]', ' ', x))

# Converting to lowercase
combined_data['processed_content'] = combined_data['processed_content'].apply(lambda x: x.lower())

# Removing stopwords and stemming
stop_words = set(stopwords.words('english'))
porter = PorterStemmer()
combined_data['processed_content'] = combined_data['processed_content'].apply(lambda x: ' '.join([porter.stem(w) for w in x.split() if w not in stop_words]))

combined_data[['CONTENT', 'processed_content']]
```

```
[5] import nltk
nltk.download('stopwords')

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
True

# Reattempt text preprocessing after downloading stopwords
# Removing special characters and digits
combined_data['processed_content'] = combined_data['CONTENT'].apply(lambda x: re.sub('[^a-zA-Z]', ' ', x))

# Converting to lowercase
combined_data['processed_content'] = combined_data['processed_content'].apply(lambda x: x.lower())

# Removing stopwords and stemming
stop_words = set(stopwords.words('english'))
porter = PorterStemmer()
combined_data['processed_content'] = combined_data['processed_content'].apply(lambda x: ' '.join([porter.stem(w) for w in x.split() if w not in stop_words]))

combined_data[['CONTENT', 'processed_content']]
```

```
# Vectorizing the processed content using TF-IDF
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split

vectorizer = TfidfVectorizer(max_features=5000)
X = vectorizer.fit_transform(combined_data['processed_content']).toarray()
y = combined_data['CLASS'].values

# Splitting the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Display the shape of the training and testing sets
X_train.shape, X_test.shape, y_train.shape, y_test.shape

((1564, 3544), (392, 3544), (1564,), (392,))
```

```
from keras.models import Sequential
from keras.layers import Dense, Dropout

# Defining the Neural Network model
model_nn = Sequential()
model_nn.add(Dense(512, activation='relu', input_shape=(X_train.shape[1],)))
model_nn.add(Dropout(0.5))
model_nn.add(Dense(256, activation='relu'))
model_nn.add(Dropout(0.5))
model_nn.add(Dense(128, activation='relu'))
model_nn.add(Dense(1, activation='sigmoid'))

# Compiling the model
model_nn.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Training the model
history_nn = model_nn.fit(X_train, y_train, epochs=10, batch_size=64, validation_data=(X_test, y_test), verbos
```

```
[9] # Evaluating the model on the test set
model_nn.evaluate(X_test, y_test)

13/13 [=====] - 0s 6ms/step - loss: 0.2875 - accuracy: 0.9286
[0.2875000238418579, 0.9285714030265808]

from sklearn.metrics import precision_score, recall_score, f1_score
import numpy as np

# Adjusting the prediction step for newer versions of Keras
y_pred = np.round(model_nn.predict(X_test)).astype(int)

# Calculate precision, recall, and F1 score for the neural network model
precision_nn = precision_score(y_test, y_pred)
recall_nn = recall_score(y_test, y_pred)
f1_score_nn = f1_score(y_test, y_pred)

precision_nn, recall_nn, f1_score_nn

13/13 [=====] - 0s 6ms/step
(0.9653465346534653, 0.9027777777777778, 0.9330143540669857)
```

```
# Preparing a sample comment for model testing
sample_comment = ['This is a great video']


# Preprocessing the sample comment
sample_comment_processed = vectorizer.transform(sample_comment).toarray()

# Making predictions using the neural network model
sample_prediction = model_nn.predict(sample_comment_processed)

# Interpreting the prediction
predicted_class = 'Spam' if sample_prediction[0][0] > 0.5 else 'Not Spam'

predicted_class

1/1 [=====] - 0s 38ms/step
'Not Spam'
```

```
✓ 0s  # Import necessary libraries
from ipywidgets import widgets
from IPython.display import display, clear_output

# Create text input widget
comment_input = widgets.Text(
    value='',
    placeholder='Type your comment here',
    description='Comment:',
    disabled=False
)

# Create a button widget for submitting the comment
submit_button = widgets.Button(
    description='Test Comment',
    disabled=False,
    button_style='', # 'success', 'info', 'warning', 'danger' or ''
    tooltip='Click to test the comment',
    icon='check' # (FontAwesome names without the `fa-` prefix)
)

# Function to handle button click event
def on_submit_button_clicked(b):
    # Preprocess the input comment
```

```
# Preprocess the input comment
sample_comment = [comment_input.value]
sample_comment_processed = vectorizer.transform(sample_comment).toarray()

# Predict using the neural network model
sample_prediction = model_nn.predict(sample_comment_processed)

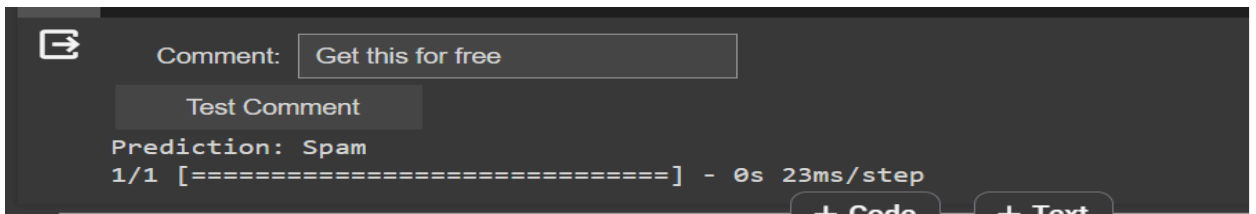
# Interpret the prediction
predicted_class = 'Spam' if sample_prediction[0][0] > 0.5 else 'Not Spam'

# Display the prediction result
with output:
    clear_output()
    print(f"Prediction: {predicted_class}")

# Create an output widget to display prediction results
output = widgets.Output()

# Link the button click event to its handler
submit_button.on_click(on_submit_button_clicked)

# Display the widgets
display(comment_input, submit_button, output)
```



REFERENCES

- Abd, T., Altabrawee, H., & Ajmi, S. Q. (2018). YouTube spam comments detection using Artificial Neural Network. *Journal of Engineering and Applied Sciences*, 13(22), 9638-9642.
- Abdolrasol, M. G., Hussain, S. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., ... & Milad, A. (2021). Artificial neural networks based optimization techniques: A review. *Electronics*, 10(21), 2689.
<https://doi.org/10.3390/electronics10212689>
- Aiyar, S., & Shetty, N. P. (2018). N-gram assisted YouTube spam comment detection. *Procedia computer science*, 132, 174-182.
<https://doi.org/10.1016/j.procs.2018.05.181>
- Akinyelu, A. A. (2021). Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based and nature-inspired-based techniques. *Journal of Computer Security*, 29(5), 473-529.
<https://content.iospress.com/articles/journal-of-computer-security/jcs210022#:~:text=DOI%3A-,10.3233/JCS%2D210022,-Journal%3A%20Journal>
- Aldwairi, M., & Tawalbeh, L. A. (2020). Security techniques for intelligent spam sensing and anomaly detection in online social platforms. *International Journal of Electrical and Computer Engineering*, 10(1), 275.
- Annareddy, S., & Tammina, S. (2019, December). A comparative study of deep learning methods for spam detection. In *2019, the third international*

conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud)(I-SMAC) (pp. 66-72). IEEE.

Baccouche, A., Ahmed, S., Sierra-Sosa, D., & Elmaghraby, A. (2020). Malicious text identification: deep learning from public comments and emails.

Information, 11(6), 312. <https://doi.org/10.3390/info11060312>

Bevendorff, J., Wiegmann, M., Potthast, M., & Stein, B. (2024). Product Spam on YouTube: A Case Study. <https://doi.org/10.1145/3627508.3638303>

Cesar, L. B., Manso-Callejo, M. Á., & Cira, C. I. (2023). BERT (Bidirectional et al. from Transformers) for missing data imputation in solar irradiance time series. *Engineering Proceedings*, 39(1), 26.

<https://doi.org/10.3390/engproc2023039026>

Chakraborty, M., Pal, S., Pramanik, R., & Chowdary, C. R. (2016). Recent developments in social spam detection and combating techniques: A survey. *Information Processing & Management*, 52(6), 1053-1073.

<https://doi.org/10.1016/j.ipm.2016.04.009>

Chen, L., Chen, J., & Xia, C. (2022). Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, 64,

103060. <https://doi.org/10.1016/j.jisa.2021.103060>

GİRGIN, A. B. A., & Gümüşçekiçi, G. (2022). From past to present: Spam detection and identifying opinion leaders in social networks. *Sigma Journal of Engineering and Natural Sciences*, 40(2), 441-463.

Goutam, A., & Tiwari, V. (2019, November). Vulnerability assessment and penetration testing to enhance the security of web applications. In *2019 4th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 601-605). IEEE.

<https://doi.org/10.1109/ISCON47742.2019.9036175>

Govil, N., Agarwal, K., Bansal, A., & Varshney, A. (2020, March). A machine learning-based spam detection mechanism. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 954-957). IEEE. <https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000177>

Hibat-Allah, M., Ganahl, M., Hayward, L. E., Melko, R. G., & Carrasquilla, J. (2020). Recurrent neural network wave functions. *Physical Review Research*, 2(2), 023358.

<https://doi.org/10.1103/PhysRevResearch.2.023358>

Ho-Dac, N. N. (2020). The value of online user-generated content in product development. *Journal of Business Research*, 112, 136-146.

<https://doi.org/10.1016/j.ibusres.2020.02.030>

Hossain, F., Uddin, M. N., & Halder, R. K. (2021, April). Analysis of optimized machine learning and deep learning techniques for spam detection. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-7). IEEE.

<https://doi.org/10.1109/IEMTRONICS52119.2021.9422508>

- Hussain, N., Turab Mirza, H., Rasool, G., Hussain, I., & Kaleem, M. (2019). Spam review detection techniques: A systematic literature review. *Applied Sciences*, 9(5), 987. <https://doi.org/10.3390/app9050987>
- Ilavendhan, A., & Janani, N. (2024, January). Optimizing YouTube Spam Detection with Ensemble Deep Learning Techniques. In 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 625-630). IEEE. <https://doi.org/10.1109/Confluence60223.2024.10463326>
- Islam, M. K., Al Amin, M., Islam, M. R., Mahbub, M. N. I., Showrov, M. I. H., & Kaushal, C. (2021, September). Spam detection with comparative analysis and spamming word extractions. In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1-9). IEEE. <https://doi.org/10.1109/ICRITO51393.2021.9596218>
- Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short-term memory neural networks. *Annals of Mathematics and Artificial Intelligence*, 85(1), 21-44.
- Kaur, R., Singh, S., & Kumar, H. (2018). Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches. *Journal of Network and Computer Applications*, 112, 53-88. <https://doi.org/10.1016/j.jnca.2018.03.015>

- Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the effectiveness of machine learning methods for spam detection. *Procedia Computer Science*, 190, 479-486.
<https://doi.org/10.1016/j.procs.2021.06.056>
- Krishnaveni, N., & Radha, V. (2021). Performance evaluation of clustering-based classification algorithms for detection of online spam reviews. In *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2020* (pp. 255-266). Springer Singapore.
- Laddha, A. (2020, September 1). *Detecting spam comments on YouTube using Machine Learning*. Medium. <https://medium.com/@akshmahesh/detecting-spam-comments-on-youtube-using-machine-learning-948d54f47b3>
- Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Analyzing the effectiveness of semi-supervised learning approaches for opinion spam classification. *Applied Soft Computing*, 101, 107023.
<https://doi.org/10.1016/j.asoc.2020.107023>
- Lim, J., Sa, I., Ahn, H. S., Gasteiger, N., Lee, S. J., & MacDonald, B. (2021). Subsentence extraction from text using coverage-based deep learning language models. *Sensors*, 21(8), 2712.
<https://doi.org/10.3390/s21082712>

- Lv, C., Xu, J., & Zheng, X. (2022, September). Spiking convolutional neural networks for text classification. In *The Eleventh International Conference on Learning Representations*.
- Makkar, A., Garg, S., Kumar, N., Hossain, M. S., Ghoneim, A., & Alrashoud, M. (2020). An efficient spam detection technique for IoT devices using machine learning. *IEEE Transactions on Industrial Informatics*, 17(2), 903-912. <https://doi.org/10.1109/TII.2020.2968927>
- Oh, H. (2021). A YouTube spam comments detection scheme using cascaded ensemble machine learning model. *IEEE Access*, 9, 144121-144128. <https://doi.org/10.1109/ACCESS.2021.3121508>
- Oh, J., Hessel, M., Czarnecki, W. M., Xu, Z., van Hasselt, H. P., Singh, S., & Silver, D. (2020). Discovering reinforcement learning algorithms. *Advances in Neural Information Processing Systems*, 33, 1060-1070.
- Post, J. (2021). The next generation air transportation system of the United States: vision, accomplishments, and future directions. *Engineering*, 7(4), 427-430. <https://doi.org/10.1016/j.eng.2020.05.026>
- Rao, S., Verma, A. K., & Bhatia, T. (2021). A review on social spam detection: challenges, open issues, and future directions. *Expert Systems with Applications*, 186, 115742. <https://doi.org/10.1016/j.eswa.2021.115742>
- Rastogi, A., Mehrotra, M., & Ali, S. S. (2021, January). Effect of Various Factors in Context of Feature Selection on Opinion Spam Detection. In *2021 11th International Conference on Cloud Computing, Data Science &*

Engineering (Confluence) (pp. 778-783). IEEE.

<https://doi.org/10.1109/Confluence51648.2021.9377056>

Rodrigues, A. P., Fernandes, R., Shetty, A., Lakshmana, K., & Shafi, R. M. (2022). Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. *Computational Intelligence and Neuroscience*, 2022.

<https://doi.org/10.1155/2022/5211949>

Talaei Pashiri, R., Rostami, Y., & Mahrami, M. (2020). Spam detection through feature selection using artificial neural network and sine–cosine algorithm. *Mathematical Sciences*, 14(3), 193-199.

Teja Nallamothe, P., & Shais Khan, M. (2023). Machine Learning for SPAM Detection. *Asian Journal of Advances in Research*, 6(1), 167-179.

Tida, V. S., & Hsu, S. (2022). Universal spam detection using transfer learning of BERT model. *arXiv preprint arXiv:2202.03480*.

<https://doi.org/10.48550/arXiv.2202.03480>

Tumu, P., Manchenasetty, V., & Rege, M. (2020). Context based sentiment analysis approach using n-gram and word vectorization methods. *Issues in Information Systems*, 21(3), 59-65.

https://doi.org/10.48009/3_iis_2020_59-65

Washha, M., Qaroush, A., Mezghani, M., & Sedes, F. (2019). Unsupervised collective-based framework for dynamic retraining of supervised real-time

- spam tweets detection model. *Expert Systems with Applications*, 135, 129-152. <https://doi.org/10.1016/j.eswa.2019.05.052>
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
<https://doi.org/10.48550/arXiv.2112.04359>
- Yurtseven, İ., Bagriyanik, S., & Ayvaz, S. (2021, September). A review of spam detection in social media. In *2021 6th International Conference on Computer Science and Engineering (UBMK)* (pp. 383-388). IEEE.
<https://doi.org/10.1109/UBMK52708.2021.9558993>
- Zhao, C., Xin, Y., Li, X., Yang, Y., & Chen, Y. (2020). A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data. *Applied Sciences*, 10(3), 936. <https://doi.org/10.3390/app10030936>
- Zhong, M., Qu, X., Chen, Y., Liao, S., & Xiao, Q. (2021). Impact of factors of online deceptive reviews on customer purchase decision based on machine learning. *Journal of Healthcare Engineering*, 2021.
<https://doi.org/10.1155/2021/7475022>
- Zulqarnain, M., Ghazali, R., Hassim, Y. M. M., & Rehan, M. (2020). A comparative review on deep learning models for text classification. *Indones. J. Electr. Eng. Comput. Sci*, 19(1), 325-335.

Ruth, V.M. (2022) A comparative study on YouTube spam ... Available at:
<https://ijcrt.org/papers/IJCRT22A6718.pdf> (Accessed: 08 July 2024).