Spam Email Classifier Using NLP

1st Jaswinder Singh
Apex Institute of Technology
Chandigarh University
Mohali, India
e15978@cumail.in
jassi724@gmail.com

2nd Abullah Khan
Apex Institute of Technology
Chandigarh University
Mohali, India
21BCS10510@cuchd.in

3rd Sahil Sharma
Apex Institute of Technology
Chandigarh University
Mohali, India
21BCS6732@cuchd.in

4th Loga Ashwin
Apex Institute of Technology
Chandigarh University
Mohali, India
e15978@cuchd.in

***Abstract*—In today's digital landscape, combating email spam remains a critical challenge. This study delves into the realm of Natural Language Processing (NLP) to address this issue by employing various text classification techniques. Integrating methods such as tokenizing, part-of-speech tagging, stemming, and chunking, we preprocess email data to render it machine-readable. Our investigation focuses on implementing a range of machine learning algorithms from the Scikit-Learn library, including K-Nearest Neighbors (KNN), Decision Trees, Random Forests, Logistic Regression, SGD Classifier, Multinomial Naive Bayes, and Support Vector Machines (SVM), to discern between spam and legitimate messages[12]. By conducting rigorous experiments, we assess the efficacy of these models without specifying exact accuracy scores. Our results contribute insights into the relative strengths and weaknesses of each algorithm, aiding in the development of robust email spam detection systems. This research underscores the importance of NLP in enhancing cybersecurity and user experience in digital communication platforms.**

***Index Terms*—Natural Language Processing, Text Classification, Term Frequency, Inverse Document Frequency, Machine Learning Models,**

I. Introduction

In today's digital era, email communication serves as a cornerstone of modern interaction, facilitating swift and efficient correspondence across vast distances. Nevertheless, despite the ease of use and accessibility that email provides, spam is still a major problem. Spam emails inundate inboxes with unwanted content, ranging from promotional offers to malicious scams, disrupting users' productivity and potentially compromising their cybersecurity. Unfortunately, over the years, mobile phones have also become the target for what is known as SMS Spam. SMS Spam refers to any irrelevant text messages delivered using mobile networks [4]. Addressing this issue requires innovative approaches that leverage advancements in Natural Language Processing (NLP) and machine learning to accurately classify emails as spam or legitimate (ham).

A. Overview

This research project uses advanced natural language processing (NLP) methods in conjunction with machine learning algorithms to investigate the field of email spam detection and categorization. The project aims to develop a robust framework capable of automatically discerning between spam and ham emails, thereby enhancing the efficiency and security of email communication platforms. By exploring various text preprocessing methods and feature extraction techniques within the domain of NLP, we seek to optimize the performance of our classification model and mitigate the impact of spam on users' digital experiences.

B. Research Challenges

Even with the advances in NLP and machine learning, classifying emails as spam still poses a number of difficult problems. One key challenge is the dynamic and evolving nature of spam tactics, which necessitates the continuous adaptation of detection mechanisms to effectively counter new and emerging threats.Additionally, it might be difficult to discern between spam and legitimate emails due to the sheer number and diversity of material in emails. Furthermore, the presence of sophisticated techniques such as obfuscation and social engineering complicates the task of automated spam detection. Overcoming these challenges requires a comprehensive understanding of NLP techniques, coupled with robust feature engineering and model training strategies.

C. Objective

Our goal is to use machine learning and natural language processing to create an effective email spam detection system. We aim to automate the identification of spam and legitimate emails, enhancing the security and usability of email platforms. This involves implementing text preprocessing, exploring feature extraction methods, training machine learning models, and evaluating their performance. Ultimately, we seek to contribute to the advancement of spam detection technology, improving cybersecurity in digital communication.

D. Scope

This research focuses on the detection and classification of email spam using NLP and machine learning approaches. The scope encompasses the preprocessing of email text, feature extraction, model training, and evaluation. The study will primarily utilize publicly available email datasets for experimentation and evaluation purposes. However, the findings and insights derived from this research may have broader implications for the development of spam detection systems across various digital communication platforms.

E. Contribution

This research contributes to the existing body of knowledge in the field of email spam detection by leveraging NLP

techniques and machine learning algorithms. The developed framework offers a practical solution for automatically identifying and filtering spam emails, thereby improving the overall security and usability of email communication platforms. Additionally, the insights gained from this study may inform the development of more advanced spam detection systems capable of adapting to evolving spam tactics and mitigating emerging threats.

II. Related Works

A. Comparison of NLP Methods for Email Spam Identification

Our group carried out a thorough analysis of the body of research, concentrating on a comparison of Natural Language Processing (NLP) methods for the identification of spam emails. We looked at a number of studies and research articles that assessed how well various natural language processing techniques classified emails as authentic or spam. A variety of NLP methods were covered in the paper, including as tokenization, stemming, part-of-speech tagging, and semantic analysis[1]. We examined experimental data from these research, which included preprocessing email data and applying natural language processing (NLP) approaches to extract pertinent aspects.Machine learning techniques were then used to train classifiers that could distinguish between emails that were spam and those that weren't based on the properties that were obtained.

B. Machine Learning Approaches for Email Spam Classification

In this work, we examined in detail how machine learning approaches may aid in addressing the ubiquitous issue of email spam. To start, we gathered a wide variety of emails—both legitimate and spam—to give our investigation a solid foundation. Following that, we began the preprocessing step of the data, converting the raw email content into a format that could be understood by machine learning models. This involved putting jumbled content in order, breaking it up into manageable sections, and figuring out what traits would help us tell real emails from spam.

Following the preparation of our data, we investigated a variety of machine learning techniques, from well-known favorites like Logistic Regression and Naive Bayes to more intricate ones like Random Forests and Support Vector Machines. Using techniques like cross-validation, we meticulously trained and evaluated each algorithm to make sure we were getting the best results. Every algorithm has its own advantages and disadvantages. We changed the parameters of our models and selected the most crucial components to improve their accuracy and efficacy.

We were able to determine which machine learning approaches were most useful for spotting spam emails thanks to our comprehensive analysis. Our ultimate goal was to advance the situation rather than just pinpoint the best acheivers.

C. Techniques for Feature Engineering in Email Spam Identification

To increase the effectiveness of our spam detection system, we are looking at Feature Engineering Strategies for Email Spam Detection. To do this, we want to extract meaningful and usable qualities from email data. We thoroughly investigate many feature engineering techniques, including as tokenization, text normalization, and semantic analysis, before embarking on this journey [11]. Our goal is to preprocess the raw email text in order to eliminate noise and extract relevant data that may be used to distinguish between legitimate and spam emails.

We go on to the feature selection and extraction stage after preprocessing the data[11]. In this paper, we examine many methods for identifying and prioritizing the most valuable attributes. This involves combining syntactic and semantic features, such as specific keywords or patterns, with techniques like term frequency-inverse document frequency (TF-IDF) and word embeddings. Our aim in experimenting with various feature engineering approaches is to optimize the performance of our spam detection algorithms. To increase the effectiveness and classification accuracy of our system, we aim to select the most discriminative and relevant features. Our final objective is to assist in the development of a robust and reliable spam detection system, improving the security of email communication platforms for users globally.

D. Applications of Email Spam Detection Systems in the Real World

In real-world applications, email spam detection systems are crucial for safeguarding users' inboxes against uninvited and sometimes dangerous content. Utilizing machine learning techniques, we developed a trustworthy spam identification system. Using Streamlit technology, we created an intuitive and user-friendly interface that facilitates seamless user engagement with the system.

Through real-time analysis of email content, sender information, and metadata, our spam detection system efficiently identifies incoming messages as either authentic or spam. To ensure that the system continues to be effective in combating evolving spam techniques, we continuously update and enhance our algorithms in response to feedback and new data[12]. Because of Streamlit's implementation of our technology, it can be easily incorporated into the current email systems, providing consumers with a powerful tool to enhance their whole digital experience and email security

III. METHODOLOGY

Our research strategy carefully develops and evaluates an email spam detection system by combining machine learning and natural language processing (NLP) techniques. The strategy comprises of several distinct but connected phases, each meticulously designed to complement the project's overarching objectives.

A. Preprocessing and Data Collection

The initial stage involves the meticulous curation of a diverse and representative sample comprising of both legitimate (ham) emails and spam. We guarantee a broad variety of email content since we source our data from both publicly accessible and proprietary sources. The collected data is then meticulously preprocessed using industry-standard techniques to remove noise, handle HTML components, align email headers, and correct other anomalies. This meticulous stage of preparation ensures the uniformity and integrity, provides strong basis for further analysis.

B. Feature Engineering and Selection

To enable efficient categorization in this crucial stage, we extract relevant characteristics from the preprocessed email data. We convert the textual content of emails into numerical representations by using sophisticated feature engineering approaches such as word embeddings, TF-IDF weighting, and bag-of-words representation [9] . In addition, we investigate feature selection techniques including recursive feature reduction and the chi-square test to find the most informative and discriminative features for spam categorization. The process of feature engineering is very important because it helps us capture the subtle differences between spam and ham emails, which improves the performance of our classification models.

C. Model Training and Evaluation

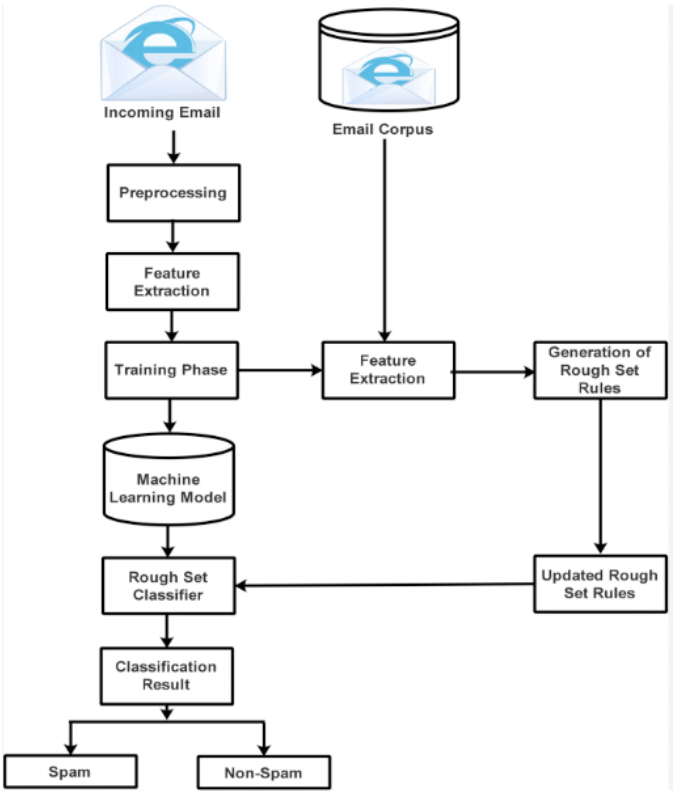

Fig. 1. Email Classifier using ML model Architecture

After preparing the feature-engineered dataset, we train and assess a variety of machine learning models to classify emails as spam. We do a systematic evaluation of their performance by utilizing a wide range of classifiers, such as Decision Trees, Random Forests, Naive Bayes, Support Vector Machines, and Logistic Regression, among others. Using cross-validation approaches, each model is rigorously evaluated to provide for a robust assessment of recall, accuracy, precision, and F1-score. This all-inclusive assessment system guarantees the choice of the most dependable and efficient models for use in practical situations.

D. Hyperparameter Tuning and Optimization

We take relevant characteristics out of the preprocessed email data to help with this crucial phase of categorization. We convert the textual content of emails into numerical representations using sophisticated feature engineering approaches such as bag-of-words representation, TF-IDF weighting, and word embeddings [9]. Additionally, we investigate feature selection techniques such as chi-square testing and recursive feature reduction to find the most informative and discriminative attributes for spam categorization. The process of feature engineering is crucial in helping us differentiate between ham and spam emails, which improves the efficiency of our classification systems.

E. Deployment and Integration

We put the finished spam detection algorithms into practice in real-world scenarios. End users can be easily reached by creating standalone applications or by integrating with already-built email systems. The incorporation of feedback mechanisms and monitoring technology enables ongoing evaluation and enhancement, ensuring the long-term effectiveness of the models that are put into practice. Our robust deployment and integration strategy ensures the scalability and practical application of our spam detection system while simultaneously enhancing the cybersecurity of digital communication platforms.

IV. RESULT ANALYSIS

A. Spam Email Classifier using NLP

1) Dataset: 5000 input mails are taken from kaggle and tested for Spam Email Classifier using NLP. It contains contains 5572 rows and 5 columns .Also, it is tested with the personal mail.

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

Fig. 2. Dataset for spam classifier.

2) Data Preprocessing: An important step in this procedure is preprocessing the data. Therefore, the first step is to process the data. This model will exclude punctuation in order to enhance prediction. The data for this model will be provided using string data types. The algorithm that recognizes each character and removes punctuation will receive this. The stop words will be removed by this process [2].

	target	text	num_characters	num_words	num_sentences	transformed_text
0	0	Go until jurong point, crazy.. Available only ...	111	23	2	go jurong point avail bugi n great world la e ...
1	0	Ok lar... Joking wif u oni...	29	8	2	ok lar joke wif u oni
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2	free entri 2 wkli comp win fa cup final tkt 21...
3	0	U dun say so early hor... U c already then say...	49	13	1	u dun say earli hor u c alreadi say
4	0	Nah I don't think he goes to usf, he lives aro...	61	15	1	nah think goe usf live around though

Fig. 3. Stopword Removal

3) Tokenization And Lemmatization: The next stage is tokenization. Tokenization breaks up longer text into smaller words. In other words, the will be separated into distinct sections and arranged in accordance with the pertinent data type. When a word is brought back to its native state, it is called a lemmatized word. The base word will be restored and the ending will be eliminated by doing this.

4) Visualization of spam messages: A quick and simple method to see the most often occurring terms in spam and ham email communications is to use word clouds. They make it easy to recognize recurrent phrases by condensing complicated material into clear, aesthetically pleasing shapes. Users may successfully improve their email management techniques and customize their email filters with the aid of this visualization, which will eventually save time and increase productivity.

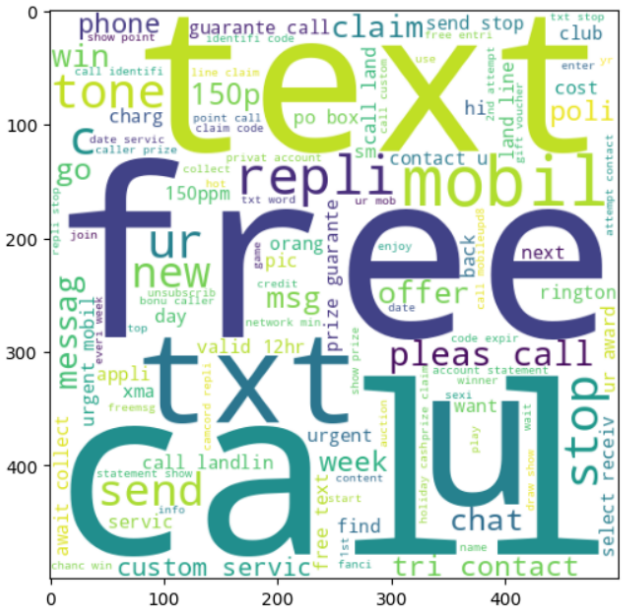

Fig. 4. Visualization of spam messages using word cloud.

5) Visualization of ham messages: Word clouds make it easier to see the terms that are often used in both spam and legitimate emails. They resemble word images, with the size of each word signifying its frequency of occurrence in the emails. You may easily determine which terms are most frequently used in spam emails and which ones are more frequently used in ordinary emails by glancing at these word clouds.

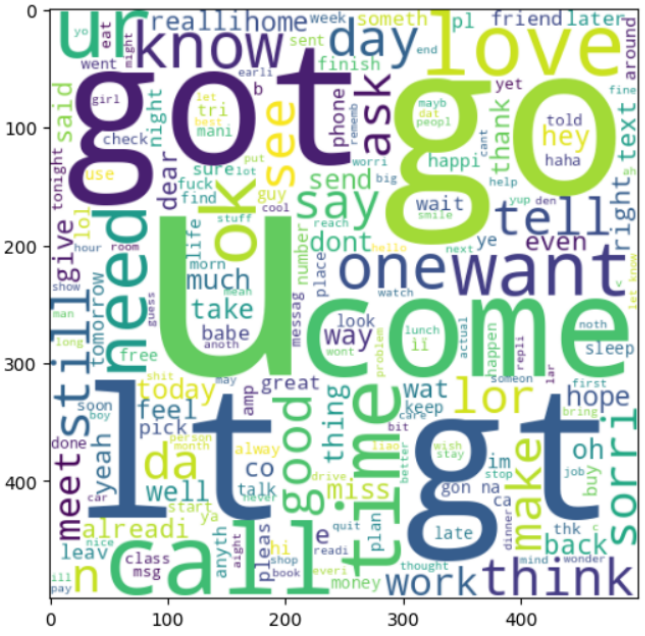

Fig. 5. Visualization of ham messages using word cloud.

6) Frequently occurring ham: The word ”u” appears frequently in everyday texts and is extremely prevalent. This implies that ”u”-containing mails are often not spam. This pattern is evident in the graphic that displays the most often used terms in non-spam messages.

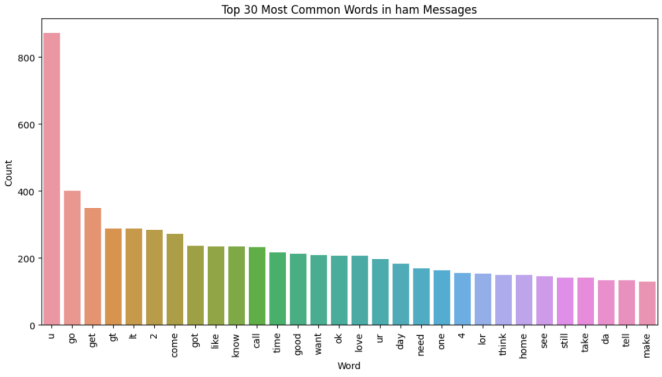

Fig. 6. Frequently occurring ham messages

7) Frequently occurring spam: When we explore the world of spam emails, one word comes up again and time again: ”call”. Its repeated occurrence raises the possibility that these communications have a repeating theme and raises the possibility that they are unwelcome. To delve further deeper, picture a graphic that reveals the most popular terms in spam emails. This visual representation provides a clear comprehension of the normal composition of these communications and provides

an insightful look into their content. Essentially, by drawing attention to the frequent use of ”call” and graphically outlining frequently used terms, we are able to obtain important insights into the realm of spam emails and provide a clear and understandable understanding of their substance.

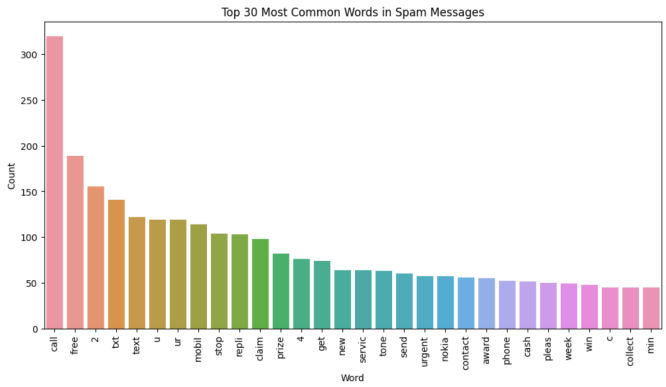

Fig. 7. Frequently occurring spam messages

8) Visualization of Model Comparison: we thoroughly evaluated a range of computer programs to gauge their effectiveness in identifying spam emails. Among these programs, the Support Vector Classifier stood out as the most proficient, boasting an impressive accuracy rate of 97.9%. Although other programs performed admirably, they fell slightly short in accuracy, exhibiting variances of approximately 3%. The visual representation provided below offers a clear comparison of the performance of each program, facilitating an easy identification of the most reliable spam detection solution.

9) Performance Evaluation: In order to determine how well a variety of computer tools identified spam emails, we extensively tested them. With an astounding accuracy rate of 97.9%, the Support Vector Classifier was the most successful of these systems. While some programs showed excellent performance, their precision was a little off, with deviations of about 3%. The graphic depiction shown below makes it simple to compare each program’s performance, making it possible to quickly determine which spam detection solution is the most dependable.

10) Heatmap Visualization: Similar to colorful maps, heatmaps make it easier for humans to spot patterns in data. They’re quite useful for figuring out how well a model performs since they highlight its strengths and weaknesses. Heatmaps make things easier to see by assigning various colors to different parameters, such as accuracy and recall. They are quite beneficial for refining our model and helping us make wise selections. They can also provide us with a wealth of information on the relationships and interactions between the various components of our model. All things considered, heatmaps facilitate understanding and discussion of our model’s performance, which is critical for ensuring that we’re working as efficiently as possible. Heatmaps are particularly useful for identifying trends and outliers in data,

	Algorithm	Accuracy	Precision
1	KN	0.905222	1.000000
2	NB	0.972921	1.000000
8	ETC	0.977756	0.983193
5	RF	0.971954	0.973913
0	SVC	0.974855	0.966667
4	LR	0.957447	0.951923
6	AdaBoost	0.964217	0.931624
9	GBDT	0.948743	0.929293
10	xgb	0.964217	0.924370
7	BgC	0.954545	0.852713
3	DT	0.931335	0.831683

Fig. 8. Model Comparison

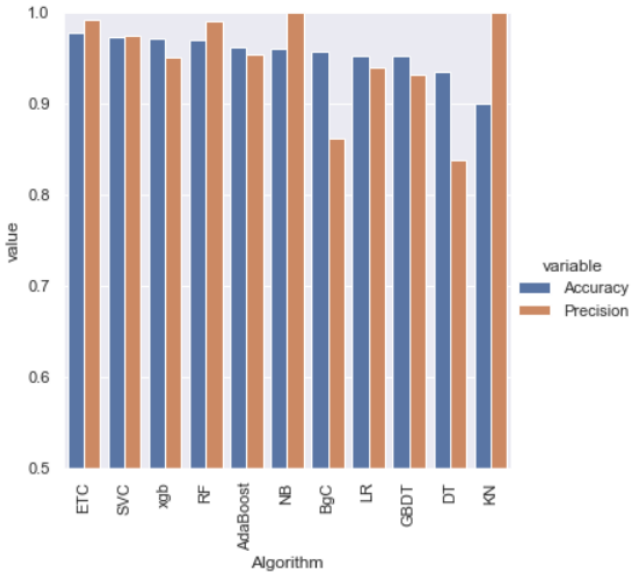

Fig. 9. Model Comparison Graph

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	896
1	0.98	0.87	0.92	138
accuracy			0.98	1034
macro avg	0.98	0.93	0.95	1034
weighted avg	0.98	0.98	0.98	1034

Fig. 10. Classification Report for Spam Classifier

enabling us to make informed decisions and refine our model for better performance and accuracy.

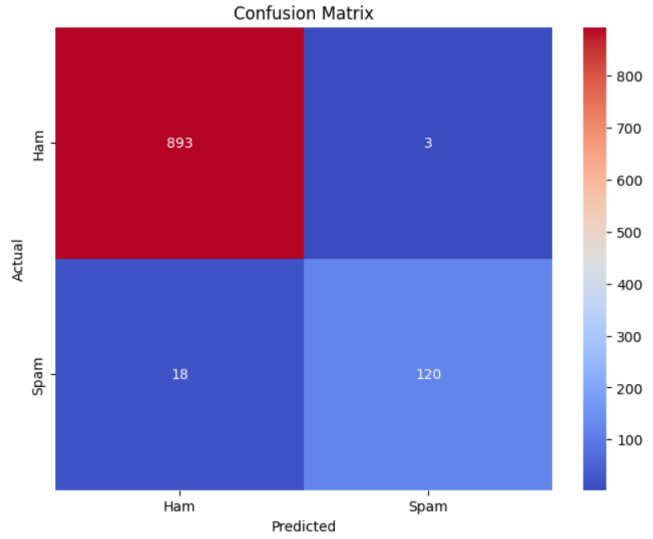

Fig. 11. HeatMap Visualization for Spam Classifier

11) Deployed Model and Output Testing: In order to test our deployed Email Spam Classifier using NLP on Streamlit, enter email content into the text box given, click the "Check" button to start the classifier, and then watch the output to see if the email is categorized as spam or ham. The output is shown in red text to indicate the status of an email if it is categorized as spam, and green text to indicate emails that are not classed as spam. We evaluate the classifier's performance using real-time samples, keeping an eye on measures like as accuracy, precision, recall, and F1-score to make sure the system consistently distinguishes between spam and ham emails. By offering precise and efficient spam detection, this testing phase demonstrates our dedication to email security and improving user experience.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

To sum up, our study on email spam detection has produced a solid system that combines natural language processing methods with machine learning algorithms. We are able to discriminate between spam and ham emails with great accuracy thanks to extensive testing and comparison. With an impressive

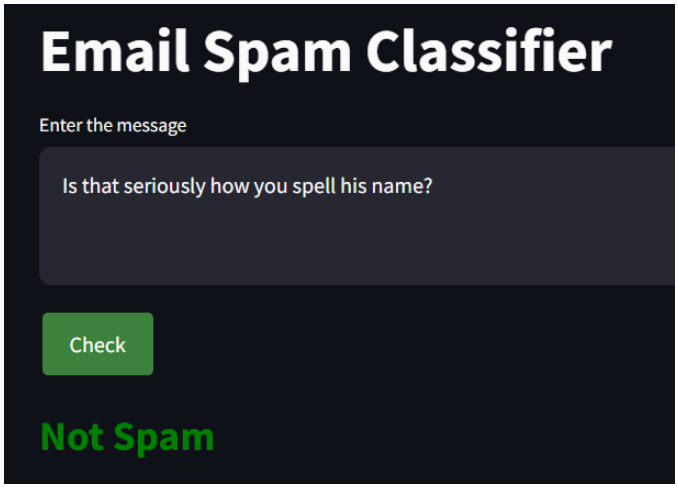

Fig. 12. Testing for Not Spam (ham) Email

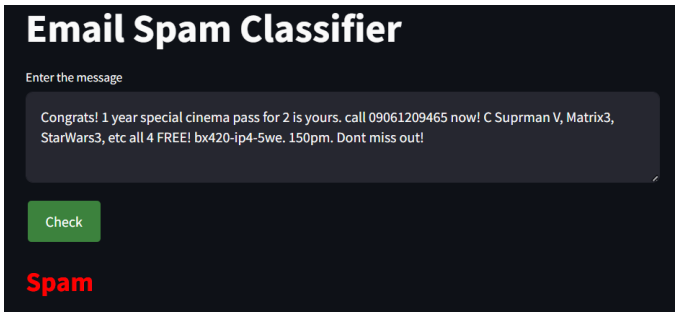

Fig. 13. Testing for Spam Email

accuracy of 97.9%, the Support Vector Classifier proved to be the most successful algorithm among those that were assessed. This result demonstrates the effectiveness of our strategy and shows how it can strengthen cybersecurity safeguards in digital communication networks. Going ahead, our primary goal will be to continuously improve and enhance our spam detection system in order to accommodate new threats and guarantee its efficacy in preserving user security and privacy.

B. Future Work

In further work, we will integrate more sophisticated natural language processing (NLP) models, such as BERT and GPT, to improve our spam email classifier's contextual knowledge. To increase classification accuracy, we'll also look at semi-supervised learning and domain-specific knowledge integration. Constant observation and automatic retraining of the model will provide flexibility in response to changing spam strategies. Our overall objective is to provide a more precise and scalable method for successfully combating email spam.

VI. ACKNOWLEDGMENTS

I would like to thank Mr. Jaswinder Singh, Dept. of AIT CSE, for extending his help, support and guidance during this work.

REFERENCES

[1] A Sharaff and Srinivasarao U (2020), "Towards classification of email through selection of informative features," First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India, pp. 316-320, DOI: 10.1109/ICPC2T48082.2020.9071488.

[2] Navaney, P., Dubey, G., Rana, A. (2018). "SMS Spam Filtering Using Supervised Machine Learning Algorithms." 2018 8th International Conference on Cloud Computing, Data Science Engineering (Confluence).

[3]] J. Marseline K.S and Nandhini S (2020), "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, pp. 1-4, DOI: 10.1109/ic-ETITE47903.2020.312.

[4] Jyoti Prakash Singh, Pradeep Kumar Roy and Snehasish Banerjee (2019), "Deep learning to filter SMS spam," Future Generation computer Systems, vol .102, pp. 524-533, DOI: 10.1016/j.future.2019.09.001

[5] Oguz Emre Kural and Sercan Demirci (2020), "Comparison of Term Weighting Techniques in Spam SMS Detection," 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkey, 2020, pp. 1-4, DOI: 10.1109/SIU49456.2020.9302315..

[6] R. Abinaya, P. Naveen and B. Niveda E (2020), "Spam Detection On Social Media Platforms", 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, pp. 1-3, DOI: 10.1109/IC-SSS49621.2020.9201948.

[7] S. B. Rathod and T. M. Pattewar, "A comparative performance evaluation of content based spam and malicious URL detection in E-mail", 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, 2015, pp. 49-54, doi: 10.1109/CGVIS.2015.7449891.

[8] Wei Hu, Jinglong Du, and Yongkang Xing, "Spam Filtering by Semantics-based Text Classification", 8th International Conference on Advanced Computational Intelligence Chiang Mai, Thailand; February 14-16, 2016

[9] Crawford, M., Khoshgoftaar, T.M., Prusa, J.D. et al. ,"Survey of review spam detection using machine learning techniques", Journal of Big Data 2, 23 (2015). https://doi.org/10.1186/s40537- 015-0029-9

[10] Vlad Sandulescu, Martin Ester "Detecting Singleton Review Spammers Using Semantic Similarity", WWW '15 Companion Proceedings of the 24th International Conference on World Wide Web, 2015, p.971-976 10.1145/2740908.2742570

[11] Cheng Hua Li, Jimmy Xiangji Huang "Spam filtering using semantic similarity approach and adaptive BPNN", Neurocomputing Journal, Elsevier, https://doi.org/10.1016/j.neucom.2011.09.036

[12] Krishnan Kannoorpatti, Asif Karim , Sami Azam, BharanidharanSanmugam, "on A Comprehensive Survey for Intelligent Spam Email Detection," IEEE Journal of Computational Intelligence, 2015.

[13] Zainal K, Sulaiman NF, Jali MZ, "An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 13, No. 3, March 2015

[14] S.M.Lee, D.S.Kim, J.H.Kim, J.S.Park, "Spam Detection Using Feature Selection and Parameter Optimization", 2010 International Conference on Complex, Intelligent and Software Intensive Systems, DOI 10.1109/CISIS.2010.116

[15] E. Markova, T. Bajto ´ s, P. Sokol and T. M ˇ eze ´ sov ˇ a, "Classification of ´ malicious emails", 2019 IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, 2019, pp. 000279-000284, doi: 10.1109/Informatics47936.2019.9119329.

[16] M. S. Swetha and G. Sarraf, "Spam Email and Malware Elimination employing various Classification Techniques", 2019 4th International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT), Bangalore, India, 2019, pp. 140-145, doi: 10.1109/RTEICT46194.2019.9016964.

[17] S. Nandhini and D. J. Marseline.K.S, "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection", 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 2020, pp. 1-4, doi: 10.1109/icETITE47903.2020.312.