Gaurav Kadam, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Threat detection of Spam And URL Using Machine Learning And NLP

Gaurav Kadam , Soham Patel, Piyush Bande
Department of Artificial intelligence Vishwakarma University
Pune, India

Abstract- Spam Messages are very irritating and an upscaled issue in online communication devices and system , it leads to security issue and fraudness in todays world. This paper presents a simple and normal solution that uses Natural language processing techniques to overcome the security issues and help to maintain security. The system developed in this paper represents the combination of NLP and ML that uses advanced text preprocessing, TF-IDF feature extraction, and classification using models or classifiers such as naive bias Logistic Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM) networks. The system predefined outcomes showcases that two model get higher accuracy then others which are Logistic Regression and LSTM and very rigid for outliers. Our Spam detection also forms a user Interface with the help of Streamlit which is further explained in the given paper.

Keywords—Natural Language Processing, Spam Detection, TF-IDF, LSTM, Machine Learning, Text Classification,Natural Language Processing (NLP), Online security, spam filtering, Text mining,Information filtering, Random Forest, Logistic Regression, Spam, Ham.

I. INTRODUCTION

Reform in existing financial decision systems is clearly un- derway due to the latest advancements in artificial intelligence. The establishment of an explainable multi-stage ensemble of 1D convolutional neural networks by N. Pavitha and S. Sugave (2024) has established a new level of transparency and added credibility to current credit decision systems by providing enhanced levels of trustworthiness through the ability to use multiple learning phases, whereas machine learning models use adaptive hyperparameter tuning methods that optimize performance while providing additional efficiency, which high- lights the need for optimization and explainability as key factors in today's intelligent systems[1][2]. The introduction of spam detection helps in rectifying the errors and solving the underlying problem , with the generation of mobile phones the messages of spam have been increased vastly and become a major problem . The message has fraud links or malicious texting to mislead one person. The NLP Combined with deep learning methods has been stated to improve the quality for detecting the spam messages through neural language models (Bengio etal., 2003; Yih et al., 2011; Mikolov et al., 2013)and performingcomposition over the learned word vectors for classification (Collobert et al., 2011) [7].

To identify The Spam Messages one should know the structureless and structured information that can obtained by databases or text fields. The access of this information is crucial and have benefits of lin- guistic analysis of text, as contrary of shallower "word basis" analysis.[14] Using NLP is very easy and only requires simple steps it simply means that a language that humans use for everyday communication (for ex English Hindi Marathi etc). In comparison with AI language ,programming and etc NLP uses fragmentation to generate the words and remove it , it is hard to follow up the words and note it down. NLP is important and is used in scientific economic, social, and cultural development. NLP trend is at peak and its theories and methods are deployed in variety of new and upcoming technologies , For this reason it is important for a wide range of people to have a working knowledge of NLP[4]This research paper develops a sure shot efficient SMS Spam detection architecture using NLP

that one not only recognise the already known spam but also adapt to new one smooth and efficiently. For reaching this , our study combines the knowledge of NLP And ML that even machines can easily understand and get to know the human language more effectively and in a thoughtful way .

We tried classical machine learning Methods which resulted in statical reasoning combining with Deep learning approaches Like LSTM long short term memory networks along wth our NLP for enhancing the quality of the prediction , these all models are uniquely suited together that they can easily capture the patterns and flow of eat spam messages and also predict the new spam based on the older datasets . Altogether these all methods forms a strong base detection framework designed to be both intelligent and rigourus in the field of spam and ham detection

II. LITERATURE REVIEW

The issue of SMS SPAM is not new to the modern world over the years researchers have made several advancement in making the ideal software or automated system that identifies spam and unwanted useless messages . The early researches depended heavily on traditional machine learning methods , one of the major know as Naive Bais was most commonly used and used multiple times due to its easiness and good computational ability to classify the spam messages with high accuracy . The NLP used in these are as follows Words and morphemes : these are helpful units of representation, but difficult to understand normally.

- Unicode is a character representation system for the different alphabets in the world's languages.
- Each character is internally represented by a unique identifier known as a code point. Also, code points are able to be stored in a file using encoding techniques such as UTF-8, that is avariable-length encoding.
- The standard way to generate tokens in a data driven approach is known as Byte-Pair Encoding or BPE. This is the first step for most large language models.
- BPE tokens may be considered to be the size of a whole word or even a morpheme, although there is a possibility for them to be as small as individual letters.[3]

There are many research paper on Natural Language Pro- cessing that are written to help researchers to gather all the information needed to help assemble the perfect spam filtering system architecture . In happening days many resources are been deployed to develop system that can do task and have phenomenal computational abilities , the prime examples adhering these models of assisting user to guide the right way through Artificil Intelligence are Apple's Siri, Google Assistant, Cortana, Amazon Echo and ReQall that all these models work on basis of NLP Algorithms , among these SIRI was the first model developed by apple [3][16]. The nature of NLP is probabilistic so it was a great choice for comparing the spam from legitimate messages that has high accuracy of patterns and frequencies . Some Machine learning model like SVM simple vector machine started having more attention in community for having more accuracy and powerful for classification. It was working by finding the balanced decisions border and proved to be effective when combined with structured, well-preprocessed text data.

This paper mainly focuses on developing a way to pass these rules and limitation by applying techniques such as advanced deep learning and machine learning . The architecture is build on BERT transformers and Bidirectional Long Short- Term Memory (BiLSTM) networks, Deep Neural Networks (DNN),and it also at the end compares the outcome with traditional methods such as naive bias and random forest to obtain wording similarities and context knowledge, while BiLSTM captures the sequential patterns in email text[20]. Moreover the promises given by the traditional approach of ML algorithm and same with the deep learning , still some iff and buts remains in most of the researches that gives the basic idea to researchers to where to focus while having a sense of direction of results. The hybrids models that are ix of both comes handy when solving or fixing issues its combination of interpretability and efficiency of classical algorithms with the contextual depth and sequential learning power of LSTM network simulataniously .

III. METHODOLOGY

A. Dataset

The dataset used in the system is SMS Spam Collection dataset, containing labeled messages as one to one input and output as spam or ham, The Dataset also contains traditional spam words and contains words to recognize rending patterns in day to day life.

B. Text Preprocessing

The Text preprocessing is a essential step in Natural Lan- guage Processing (NLP) that makes sure that raw text data is transformed into a clean and structured format suitable for further deployment . The following steps are taken to achieve this work:

- Removal of URLs and Special Characters: Links and urls in the text are removed or replaced with classic to- kenisation to destroy the irrelevant information that is not related to spam detection. This ensures less noisy dataset with enhanced quality of features for model training and obtaining of meaningful words only is acquired while ongoing process
- Lowercasing of Text: The he entire text is converted into lowercase to maintain consistency across the dataset. The standardised text and the complexity of the code is reduced which helps improving computational power effectively and lets us treat similar words equally during training. This step is essential for avoiding duplicate words that has same spelling but different meaning for e'xample FREE' and 'free' by equally selecting all words.
- Tokenization and Lemmatization: Tokenization breaks the clean text into separate words, making it easier to handle and analyze. Lemmatization is done next using the NLTK WordNet Lemmatizer, which changes words into their base or root forms. This helps reduce different forms of the same word (like changing "winning" to "win"), improving generalization and allowing the model to learn real patterns more easily.
- Retention of Contextual Words: In this method, stop words are kept instead of removed like in traditional preprocessing techniques. Keeping stop words maintains the contextual meaning of sentences that can be vital in differentiating between spam and real messages.
- Keyword Enhancement for Spam Detection: The key- word enhancement technique is used that highlights the common spam words like "urgent," "free," "win," and "prize." These words are changed by adding a special token to them, which makes them more important when features are extracted.

C. Feature Extraction

- TF-IDF Vectorization: TF-IDF is a method that turns text into numerical feature vectors by assessing how important words are in a document compared to the whole dataset. This way, it brings attention to key terms while downplaying words that show up all the time.
- Maximum Features (5000): We're focusing on the top 5,000 key terms using TF-IDF scores. This helps us cut down the complexity, speeds up our computations, and stops the model from getting too caught up with less important words.

- N-grams (1,2): Both unigrams and bigrams are used to capture solo words as well as pairs of same words. This enhances the model's ability to understand context and improves classification performance.

D. Machine Learning Models

- Naive Bayes: Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence be- tween features. It is efficient for text classification tasks and performs well on high-dimensional data like TF-IDF vectors.
- Logistic Regression: Logistic Regression is a straight- forward classification model that predicts probabilities through a sigmoid function. It's quite effective for binary classification, especially when you add class balancing to tackle imbalanced datasets.
- Support Vector Machine (SVM): On the other hand, SVM, or Support Vector Machine, is a robust classifier that identifies the best hyperplane to differentiate data points in high-dimensional space. This makes it partic- ularly

useful for text classification since it can manage sparse data efficiently.

- Random Forest: Then there's Random Forest, which is an ensemble learning approach. It creates several decision trees and merges their predictions. This method boosts accuracy and helps avoid overfitting by averaging the results from the various trees.

E. Deep Learning Model

- Embedding Layer: The embedding layer takes the input text and turns it into dense vector representations that show how words relate to one another. It changes the high-dimensional, sparse input into a more manageable lower-dimensional vector space, making learning more efficient.
- LSTM Layer: Then, the LSTM layer comes in to handle sequences of those word embeddings. It helps capture long-term dependencies in the text, which is super important for understanding context and the order of words—key for accurately spotting spam.
- Dense Output Layer with Sigmoid Activation: Finally, the dense layer gives us the output by using a sigmoid activation function for binary classification. This means it provides a probability between 0 and 1, telling us whether a message is spam or not.

F. Training Traction

The Machine learning contains EPOCH it means that the a system runs an epoch it completes pass through one of the parts of the dataset where every value is gone through the model and updated on basis of the error and loss calculated . Basically one model requires ,many ephos for training and improving accuracy by continuous adjustment of error to achieve minimum loss

Fig. 1. Epoch Trained

IV. SYSTEM FLOWCHART

The work architecture of the system of spam detection system is shown in Fig. 2. The system starts by installing required datasets and libraries , after this it does convert categorical data into numerical form.

In the preprocessing stage, the text data is cleaned by removing URLs, special characters, and converting all text to lowercase. Tokenization and lemmatization are going to ap- plied to normalize the text and improve consistency. similarly, important spam-related keywords are deployed to enhance detection accuracy.

The processed text is then passed through a feature extrac- tion phase using TF-IDF vectorization with n-grams, which captures both word importance and contextual relationships. The transformed data is used to train multiple machine learning models, including Naive Bayes, Logistic Regression, SVM, and Random Forest, and the best-performing model is selected.

In parallel, a deep learning approach using an LSTM network is implemented to capture sequential patterns in the text. The model processes tokenized and padded sequences to improve classification accuracy.

Finally, the trained model is deployed using a Streamlit- based application, where user input is processed and classified as either spam or ham.

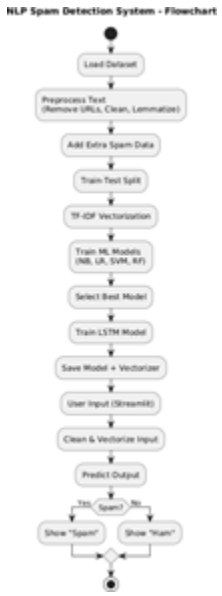

Fig. 2. Flowchart of the Proposed Spam Detection System

V. RESULTS AND DISCUSSION

The differences between the machine learning and deep learning for spam detection are clearly visible cause of the advancements done to achive desirable result. Logistic Re- gression sets a solid foundation with good accuracy and precision.On the other hand, Naive Bayes does a great job with recall, detecting real spam messages . SVM support vector machine has a nice balance among the other classical methods and always does well in accuracy , precision, recall, and the F1-score. But when you compare these classical models to the Long Short-Term Memory (LSTM) network, their limitations stand out. The LSTM goes above and beyond all the traditional models in every metric—offering better accuracy by grasping contextual relationships between words, higher precision by cutting down on false positives, stronger recall by spotting sneaky spam, and a top-notch F1-score that shows a much better balance between precision and recall overall. Sure, classical models are faster and easier to interpret, but the LSTM's ability to learn from sequences really boosts detection performance, making it the stronger choice for tackling real- world spam

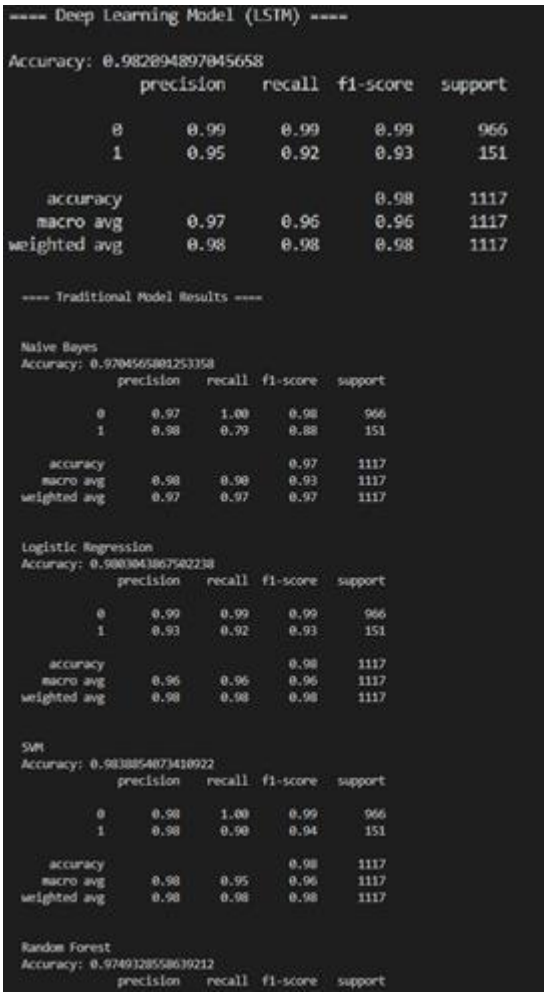

Fig. 3. Comparison of Models

VI. SYSTEM IMPLEMENTATION

The trained model is saved using joblib and deployed using a Streamlit web application for real-time spam detection.

Fig. 4. System Interface

VII. ADVANTAGES

- Hybrid approach improves accuracy
- Handles class imbalance
- Real-time prediction capability

VIII. LIMITATIONS

- Limited dataset size
- Higher computational cost for LSTM

IX. FUTURE WORK

Future improvements include:

- Using transformer models like : T5 (Text-to-Text Transfer Transformer), BART (Bidirectional and AutoRegressive Transformers), Pegasus.
- Expanding dataset
- Multilingual spam detection

X. CONCLUSION

To conclude this paper represents a detailed and complex understanding of NLP- based SMS spam detection system that combines both classical and advanced machine learning and deep learning methods . This system is soley designed to aim for removing spam and securing the information in relation to real world. The system has several steps and can continue to develop more by accurate models for more precise outcomes

.This study takes a close look at the pros and cons of traditional methods like Logistic Regression, Naive Bayes, and Support Vector Machines, comparing them to the advanced capabilities of Long Short-Term Memory (LSTM) networks. It gives a balanced view of what each method does well and where they might not measure up. The findings show that while the classic models are solid, easy to understand, and efficient in terms of computing, the LSTM network really outshines them in key areas like accuracy, precision, recall, and F1-score. This is mainly due to its special knack for grasping the context and sequence of human language. Looking past just the data, what we found really highlights how the new system can be applied in the real world. It seems like it's not only solid from a tech standpoint but also practical for use in live messaging situations where spam tactics are always changing. In the end, this research is an important step toward creating smarter, more adaptive spam filters — ones that go beyond just following strict rules and leverage the power of modern AI to keep users safe from the constant threat of SMS spam.

REFERENCES

1. Pavitha, N., Sugave, S. (2024). Explainable multistage ensemble 1D convolutional neural network for trustworthy credit decision. International Journal of Advanced Computer Science and Applications, 15(2), 351–358.
2. Pavitha, N., Sugave, S. (2023). Optimizing machine learning models: An adaptive hyperparameter tuning approach. International Journal of Intelligent Systems and Applications in Engineering, 11, 344–354.
3. D. Jurafsky and J. H. Martin, "Speech and Language Processing," 3rd ed., Pearson, 2023.
4. S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly, 2009.
5. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.
6. T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," arXiv, 2013.
7. J. Ramos, "Using TF-IDF to Determine Word Relevance," 2003.

8. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.
9. Y. Kim, "Convolutional Neural Networks for Sentence Classification," EMNLP, 2014.
10. A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI, 1998.
11. C. Cortes and V. Vapnik, "Support Vector Networks," Machine Learning, 1995.
12. L. Breiman, "Random Forests," Machine Learning, 2001.
13. I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.
14. A. Ge´ron, "Hands-On Machine Learning with Scikit-Learn and Tensor- Flow," O'Reilly, 2019.
15. J. Brownlee, "Deep Learning for Natural Language Processing," Ma- chine Learning Mastery, 2017.
16. R. M. Shete and S. W. Mohod, "Using Natural Language Processing for Detection of Events and Spam Control from User Data Stream in Social Sites," IJERT, vol. 4, no. 4, pp. 1–5, 2015.
17. R. D. Warkar and I. R. Shaikh, "Detection of Spam Comments Using NLP Algorithm," International Journal of Engineering and Computer Science, vol. 7, no. 1, pp. 23386–23389, 2018.
18. S. Shankar, "Advanced Detection of Spam and Email Filtering using Natural Language Processing Algorithms," IJARIIT, vol. 4, no. 4, 2018.
19. N. Choudhary and N. Dubey, "Spam Detection Approach Using Mod- ified Pre-processing with NLP," International Journal of Computer Sciences and Engineering, vol. 7, no. 10, pp. 158–161, 2019.
20. T. Lakshman et al., "SMS Spam Detection in Machine Learning using Natural Language Processing," IJARIIT, vol. 9, no. 5, 2023.
21. A. Srivastava and P. Singh, "Spam Detection Using Natural Language Processing," Journal of Applied Science and Education, vol. 4, no. 2, pp. 1–7, 2024.
22. A. S. P., G. Prasad, and E. P. Kumar, "Email Spam Detection using Natural Language Processing and Deep Learning," IJSRET, vol. 5, no. 5, pp. 33–35, 2025.
23. I. Dutse, M. Liptrott, and I. Korkontzelos, "Detection of Spam Posting Accounts on Twitter," Neurocomputing, 2018.
24. A. Jain et al., "Identifying Spam Posts on Social Media using Machine Learning," arXiv, 2018.
25. B. Feng et al., "Spam Detection in Mobile Social Networks through Deep Learning," IEEE Network, 2018.
26. J. Fattahi and M. Mejri, "SpaML: A Bimodal Ensemble Learning Spam Detector using NLP Techniques," arXiv, 2020.
27. Y. Liu et al., "Opinion Spam Detection using Deep Learning Tech- niques," Neurocomputing, 2019.