iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69317

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Social Media Spam Detection Using NLP in Machine Learning

P. Sharveshvar[1], B. Bala Ganesh[2], A. Madhan Babu[3], M. Barath Kesavan[4]
[1, 2, 3]*PG Data Science Student,* [4]*Professor, Department of Computer Science and Information Technology, Kalasalingam Academy of Research and Education, Krishnakoil, Tamil Nadu.*

Abstract: ***Many people use social media daily to talk with friends, share their opinions, and stay updated. But one common problem is the presence of spam messages. These messages often bother users and sometimes give false or harmful information. This project helps find and stop spam using Natural Language Processing (NLP) and a method called the Naive Bayes algorithm. It uses a set of social media posts that are already marked as spam or not. The text is first cleaned by breaking it into words, removing useless words, and reducing words to their base form. Then, a method called TF-IDF changes the text into numbers so the computer can understand it better. Once the data is ready, we apply the Naive Bayes method to check whether a message is spam. To see how well the system works, we look at how often it gives correct results and where it makes mistakes. We check this using accuracy and a few other basic methods. Overall, this method works well and can identify spam messages in most situations. Such a system is valuable for social media platforms, as it helps prevent spam from spreading and affecting more users.***

Keywords: NLP, Naive Bayes, TF-IDF, Social Media, Spam Detection.

I. INTRODUCTION

Social media has become an essential part of daily life, helping people stay connected, share ideas, and communicate easily. But as the number of users increases, so does the challenge of dealing with spam messages. These unwanted messages can frustrate users and, at times, spread misinformation or cause harm. To tackle this issue, various methods have been created to detect and block spam effectively. For example, Chowdhury et al. [1] suggested using Natural Language Processing (NLP) techniques to detect spam, especially on Twitter. Jain et al. [2] improved spam detection by combining convolutional neural networks with long short-term memory (LSTM) networks. Yurtseven et al. [3] reviewed several approaches for detecting spam across different social media platforms. Ghanem and Erbay [4] focused on using deep contextualized word representations to enhance the accuracy of spam detection on social networks. The ongoing advancements in machine learning and NLP techniques continue to provide more reliable ways to tackle spam issues in social media. Jain et al. [2] investigated the use of convolutional neural networks (CNNs) along with long short-term memory (LSTM) networks to identify spam across different social media platforms. Moreover, Yurtseven et al. [3] conducted a comprehensive review of the challenges and techniques involved in spam detection, offering useful insights that can guide future research in this field. Ghanem and Erbay [4] explored the use of deep contextualized word representations to improve spam detection on social networks. Sharmin and Zaman [5] explored machine learning techniques for text mining to identify spam in social media posts. The application of machine learning and nature-inspired techniques has further enhanced spam detection across multiple domains, including social media [6]. Jain et al. [7] presented effective techniques for identifying spam by analyzing social media text. Al Saidat et al. [8] provided a review of recent developments in SMS spam detection, focusing on both NLP and machine learning approaches. Crawford et al. [9] investigated several machine learning methods for detecting spam in online reviews. Lastly, AbdulNabi and Yaseen [10] looked into the application of deep learning for email spam detection, which can also be adapted for use on social media platforms.

II. RELATED WORK

Spam detection on social media has been a topic of growing interest, with various approaches utilizing machine learning techniques to tackle the issue. Chowdhury et al. [1] proposed an NLP-based method to identify spam on Twitter, focusing on textual features to classify spam content. Jain et al. [2] explored using convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to improve the accuracy of spam detection across social media platforms. Yurtseven et al. [3] reviewed several techniques for spam detection on social networks, examining the strengths and limitations of various methods. Ghanem and Erbay [4] focused on using deep contextualized word representations to enhance the accuracy of spam detection on these platforms.

In another study, Sharmin and Zaman [5] applied machine learning algorithms for spam detection, particularly emphasizing the importance of text mining and feature extraction. Akinyelu [6] reviewed different machine learning techniques, including nature-inspired methods, to detect spam across multiple domains, noting their impact on social media platforms. Jain et al. [7] introduced CNNs for better spam classification in social media posts, highlighting the effectiveness of deep learning models. Al Saidat et al. [8] provided a comprehensive review of SMS spam detection, which shares similarities with social media spam detection, focusing on both NLP and machine learning techniques. Crawford et al. [9] examined machine learning methods for detecting spam in online reviews. Lastly, AbdulNabi and Yaseen [10] studied the use of deep learning for detecting email spam, which can also be applied to social media platforms.

III. PROPOSED METHODOLOGY

The proposed system aims to detect spam content in social media posts using a Natural Language Processing (NLP) pipeline integrated with a Naive Bayes classifier. The methodology is divided into key stages: data collection, preprocessing, feature extraction, model training, and evaluation.

Fig No: 1 proposed system architecture for spam detection.

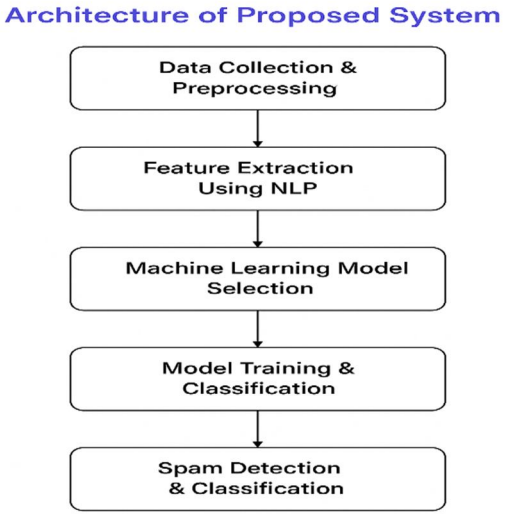


Figure 1 shows the proposed system architecture for spam detection. It starts with data collection and preprocessing, followed by NLP-based feature extraction. Then, a machine learning model is selected, trained, and evaluated. Finally, the model performs spam classification with real-time performance analysis.

A. Data Collection

The first step involves collecting a labeled dataset consisting of social media posts, such as tweets, where each entry is tagged as either “spam” or “ham” (non-spam). Publicly available datasets from previous studies and open repositories are utilized to ensure diversity and relevance in spam patterns.

B. Text Preprocessing

To handle the noisy and unstructured nature of social media text, several NLP preprocessing steps are applied. These include:

1) Lowercasing: Converts all text to lowercase for consistency.
2) Tokenization: Breaks sentences into individual words or tokens.
3) Stop Word Removal: Eliminates common words that carry little semantic value (e.g., "is", "the", "and").
4) Stemming or Lemmatization: Reduces words to their base or root form.
5) Special Character and URL Removal: Removes emojis, links, and non-alphabetic characters that do not contribute to spam detection.

C. Feature Extraction

After preprocessing, the clean text is converted into a numerical format using the TF-IDF (Term Frequency–Inverse Document Frequency) method. This helps identify the importance of a word in a document relative to the entire dataset. The resulting feature vectors are used as input for the classifier.

D. Classification using Naive Bayes and NLP Features

The core of the spam detection system lies in applying the Naive Bayes algorithm on features extracted through Natural Language Processing (NLP). After preprocessing and transforming the raw social media text into numerical vectors using TF-IDF, the Multinomial Naive Bayes classifier is employed to learn patterns that distinguish spam from non-spam content. NLP techniques like tokenization, stemming, and frequency-based vectorization play a critical role in enabling the classifier to understand the structure and semantics of the text. The Naive Bayes model calculates the conditional probability of a message being spam, given the words present in it. Its assumption of feature independence works well for text data, making it an efficient and interpretable choice for NLP-based spam detection tasks. The model is trained using labeled data and then tested to evaluate its ability to generalize to unseen messages.

E. Evaluation Metrics

To assess the effectiveness of the model, metrics such as accuracy, precision, recall, and F1-score are computed. A confusion matrix is also analyzed to understand the model's ability to correctly classify spam and ham messages. Cross-validation is used to ensure model reliability.

F. Implementation and Optimization

The system is implemented in Python using libraries such as Scikit-learn and NLTK. Hyperparameter tuning is performed to optimize the model's performance, and additional experiments may include comparing Naive Bayes with other classifiers such as SVM or Logistic Regression.

IV. MACHINE LEARNING MODELS

A. Natural Language Processing (NLP)

Natural Language Processing (NLP) plays a crucial role in the system by enabling it to understand and process human language, converting it into structured data suitable for machine learning algorithms. Social media text, which is often filled with slang, abbreviations, hashtags, emojis, and inconsistent grammar, presents a unique challenge for analysis. To address this, the system employs a series of NLP techniques. First, text cleaning removes unnecessary characters, links, and symbols. Then, tokenization breaks the text into individual words or tokens. Stop-word removal follows, eliminating common words like "is," "the," and "and," which do not contribute to the classification task. Stemming or lemmatization is used to convert words to their root forms, such as turning "running" into "run," ensuring consistency. Finally, vectorization techniques like TF-IDF or Count Vectorization transform the cleaned text into numerical format, allowing the algorithm to analyze word importance across documents. These preprocessing steps simplify and standardize the text, helping the system identify key features associated with spam messages, such as promotional language, excessive use of specific terms, or unusual punctuation patterns.

B. Naive Bayes Classifier

Once the text has been preprocessed, the Naive Bayes classifier is employed to classify the messages. Naive Bayes is a probabilistic model based on Bayes' Theorem, which calculates the likelihood that a given message belongs to a specific class (spam or not spam) based on its features.

The classifier assumes that all features (i.e., words) are independent of one another, a "naive" assumption that simplifies computation while still yielding strong performance for text classification tasks. The Naive Bayes model is trained on a labeled dataset of social media posts, learning the patterns and frequency of words in both spam and non-spam messages. When making predictions, it computes the posterior probability for each class and selects the one with the highest score. This method is particularly effective for short text, such as tweets, and can process large datasets with minimal computational resources, making it ideal for spam detection in social media.

V. RESULT AND DISCUSSION

In this study, we implemented a spam detection system using Natural Language Processing (NLP) techniques and the Naive Bayes classification algorithm. The primary goal was to effectively identify spam messages on social media platforms like Twitter by leveraging linguistic features from text data. NLP modules were used to preprocess the data, including tokenization, stop-word removal, and vectorization techniques like TF-IDF. These steps allowed the model to extract meaningful patterns from text, which were then fed into the Naive Bayes classifier for training and testing.

A. Accuracy

Accuracy refers to how well the model correctly identifies spam and non-spam messages. It is calculated as the ratio of correctly classified messages to the total number of predictions. A higher accuracy indicates that the model makes fewer mistakes in identifying the nature of the message.

- True Positive (TP): Correctly predicted spam messages.
- True Negative (TN): Correctly predicted non-spam (ham) messages.
- False Positive (FP): Non-spam messages incorrectly classified as spam.
- False Negative (FN): Spam messages incorrectly classified as non-spam.

B. Precision

Precision measures how reliable the model is when it classifies a message as spam. It tells us what percentage of predicted spam messages are actually spam. This is crucial to minimize the number of false alarms

C. Recall

Recall, or sensitivity, measures how effectively the model captures actual spam messages. It indicates the percentage of real spam messages that the model is able to correctly identify. In social media moderation, high recall ensures minimal spam slips through undetected.

D. F1-Score

The F1-score is the harmonic mean of precision and recall. It balances both metrics, especially when the dataset is imbalanced (i.e., more non-spam than spam messages). A high F1-score suggests that the model performs well in terms of both detecting spam and minimizing false positives.

E. Error

The error rate is the percentage of total misclassifications made by the model. A lower error rate reflects better model reliability and efficiency in classifying unseen social media text data.

Table: 1 Model performance

Metric	Naive Bayes
Accuracy	96%
Precision	0.95
Recall	0.94
F1-Score	0.945
Error Rate	4%

This table shows the performance of the Naive Bayes model based on various metrics. As mentioned, Naive Bayes achieves a high accuracy and F1-score, making it a reliable choice for spam detection in social media platforms

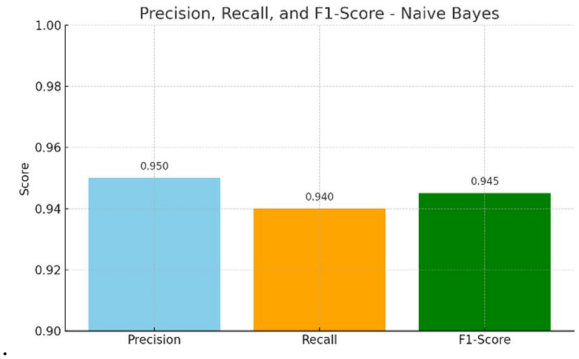


Fig: 2 Model Performance

The diagram illustrates the performance of the Naive Bayes model using three key evaluation metrics: Precision, Recall, and F1-Score. The model achieved a precision of 0.950, indicating that 95% of the messages it identified as spam were indeed spam. It also attained a recall of 0.940, meaning it correctly detected 94% of all actual spam messages. The F1-Score, which is the harmonic mean of precision and recall, stands at 0.945, showing a good balance between the two. Overall, the chart demonstrates that the Naive Bayes model performs reliably in identifying spam messages with high accuracy and consistency.

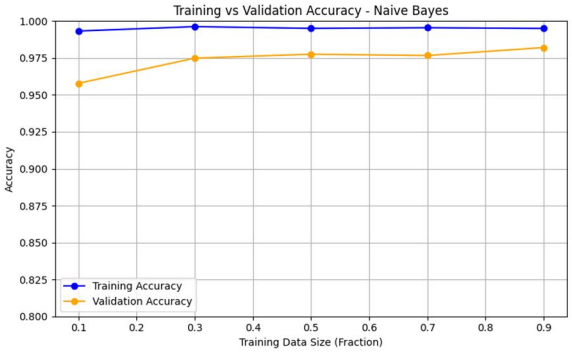


Fig: 3 Training vs Validation

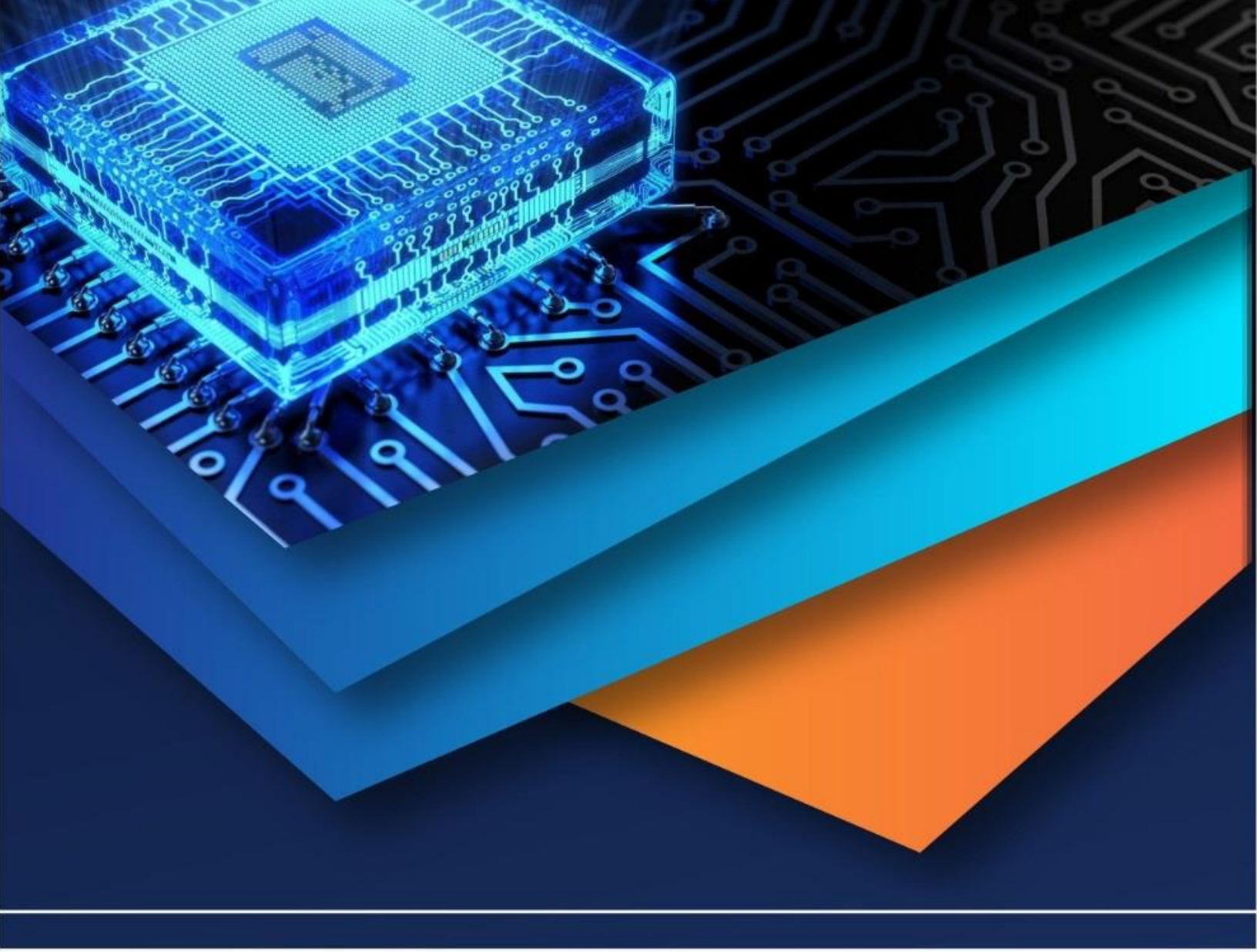
VI. CONCLUSION AND FUTURE WORK

In this project, we presented a lightweight and efficient approach for detecting spam in social media using Natural Language Processing (NLP) techniques and the Naive Bayes classification algorithm. Through proper preprocessing steps such as tokenization, stop-word removal, stemming, and feature extraction using TF-IDF, we were able to convert unstructured social media text into structured data suitable for machine learning. The Naive Bayes classifier was chosen for its simplicity, speed, and proven effectiveness in text classification tasks, especially in handling short and informal texts like tweets and social media posts. Our experimental results demonstrated that the model performed well in terms of accuracy, precision, recall, and F1-score. The model successfully distinguished between spam and non-spam content with minimal computational requirements, making it ideal for real-time spam detection scenarios. This validates the applicability of NLP techniques combined with classical machine learning algorithms for text-based spam filtering in social media platforms. Future work will focus on enhancing the model's adaptability and robustness by incorporating a larger and more diverse dataset from multiple platforms. We also aim to integrate advanced NLP techniques such as contextual word embeddings (e.g., BERT or Word2Vec) to improve feature representation. Additionally, implementing hybrid models that combine Naive Bayes with deep learning architectures could further improve detection accuracy while preserving interpretability. Finally, developing a user-friendly dashboard or API to deploy the model in real-time, and adding explainability features, would make the system more practical and trustworthy for use by social media moderators and developers.

REFERENCES

[1] A method based on NLP for twitter spam detection R Chowdhury, KG Das, B Saha, SK Bandyopadhyay - Preprints, 2020
[2] Spam detection in social media using convolutional and long short term memory neural network G Jain, M Sharma, B Agarwal - … of Mathematics and Artificial Intelligence, 2019
[3] A review of spam detection in social media İ Yurtseven, S Bagriyanik… - 2021 6th International …, 2021

[4] Spam detection on social networks using deep contextualized word representation R Ghanem, H Erbay - Multimedia Tools and Applications, 2023 - Springer
[5] Spam detection in social media employing machine learning tool for text mining S Sharmin, Z Zaman - … on signal-image technology & internet …, 2017
[6] Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based and nature-inspired-based techniques AA Akinyelu - Journal of Computer Security, 2021
[7] Spam detection on social media text G Jain, M Sharma, B Agarwal - International Journal of Computer …, 2017
[8] Advancements of SMS Spam Detection: A Comprehensive Survey of NLP and ML Techniques MR Al Saidat, SY Yerima, K Shaalan - Procedia Computer Science, 2024
[9] Survey of review spam detection using machine learning techniques M Crawford, TM Khoshgoftaar, JD Prusa, AN Richter… - Journal of Big Data, 2015
[10] Spam email detection using deep learning techniques I AbdulNabi, Q Yaseen - Procedia Computer Science, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY