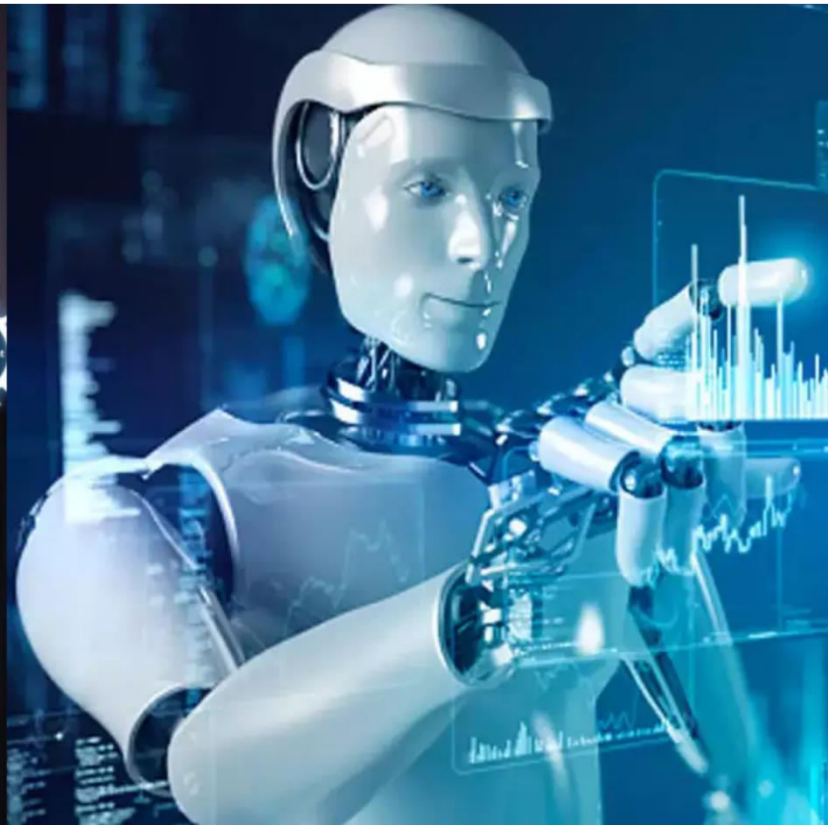
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404729|

Real-Time Twitter Spam Detection

Aman Lekharajani, Sakshi Duralkar, Anuj Ghongade, Kauntey Gawande, Prof. H. D. Misalkar

U.G. Student, Department of Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera, India

U.G. Student, Department of Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera, India

U.G. Student, Department of Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera, India

U.G. Student, Department of Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera, India

Associate Professor, Department of Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera, India

ABSTRACT: In today's digital era, social media platforms like Twitter have become pivotal channels for communication, news sharing, and public engagement. However, the open nature of these platforms has made them highly vulnerable to spam, phishing, and malicious content, which can compromise user experience and digital security. This project presents a robust and scalable Near Real-Time Twitter Spam Detection System that employs machine learning techniques to classify tweets as spam or non- spam efficiently.

The system is developed using Django for the backend, MySQL for secure user authentication and data storage, and Python for implementing machine learning models. A dataset comprising approximately 16,000 labelled tweets was used for training and testing. Four classification algorithms Random Forest, Logistic Regression, Naive Bayes, and Gradient Boosting were applied to assess their detection performance. The dataset was split into various training/testing ratios (50:50, 45:55, 30:70, and 25:75) to compare model accuracy under different conditions. The TF-IDF (Term Frequency-Inverse Document Frequency) technique was used for feature extraction, enabling effective transformation of textual data into numerical form for analysis.

The system allows users to register and log in securely, submit tweets for analysis, select data split ratios, and view algorithm-wise accuracy results. A visual comparison using Matplotlib graphs is also generated, helping users interpret model performance immediately. Among the evaluated models, Random Forest consistently demonstrated superior accuracy and time efficiency across most testing scenarios.

KEYWORDS: Spam Detection, Machine Learning, NLP, Real-time Data Processing, Supervised Learning, Accuracy Comparison, Data Visualization.

I. INTRODUCTION

In the modern digital landscape, social media platforms like Twitter have emerged as essential tools for real-time communication, news broadcasting, brand engagement, and public discourse. With over hundreds of millions of active users globally, Twitter enables the rapid sharing of information. However, its open and unfiltered nature also renders it vulnerable to exploitation by spammers, bots, and malicious users. These actors inundate the platform with irrelevant or harmful content such as phishing links, promotional spam, clickbait, and misinformation, severely degrading the quality of user experience and posing substantial security and trustworthiness risks.

To address this growing concern, the project titled "Near Real-Time Twitter Spam Detection" proposes an intelligent, scalable, and deployable web-based system capable of automatically classifying tweets as spam or non-

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404729|

spam (ham) using state-of-the-art machine learning techniques. The solution places a strong emphasis on real-time processing, high classification accuracy, and user interactivity, making it suitable for practical deployment scenarios.

The core architecture of the system is built using the Django web framework, allowing for a secure, interactive, and modular interface where users can input tweet data for analysis. The backend leverages Natural Language Processing (NLP) techniques to preprocess tweet text removing noise, normalizing content, and extracting linguistic features followed by TF-IDF vectorization to convert the text into numerical format suitable for machine learning.

A labelled dataset comprising approximately 16,000 tweets, annotated as either spam or non-spam, serves as the foundation for training and evaluation. Four supervised machine learning algorithms were selected based on their effectiveness in text classification tasks and comparative diversity in performance characteristics:

- Random Forest
- Logistic Regression
- Naive Bayes
- Gradient Boosting

To assess the models under varying data conditions and training scenarios, the dataset was split into different training-to-testing ratios: 50:50, 45:55, 30:70, and 25:75.

Each model was trained and evaluated across these splits to observe the impact of data distribution on classification performance. The system records and visualizes the accuracy scores for each model under each configuration, enabling transparent performance comparison. The results are displayed via interactive graphs and dashboards on the web interface.

The system also supports real-time tweet input: users can enter any tweet text via the interface and instantly receive the classification output, supported by all four trained models. This real-time functionality highlights the system's applicability in live environments, such as social media moderation tools, brand monitoring dashboards, and misinformation detection services.

The overall project achieves several key objectives:

- Demonstrates a practical application of supervised learning to a real-world problem.
- Evaluates model performance across varied data distributions.
- Provides a user-friendly, web-based platform for real-time tweet analysis.
- Offers insights into the comparative strengths and limitations of popular classification algorithms.
- Contributes to the broader goals of digital safety, content moderation, and AI-driven automation.

This project not only demonstrates the practical application of machine learning for social media moderation but also provides a deployable web-based tool capable of performing real-time tweet classification. The results and visual insights make it easier to identify which algorithm performs best in terms of both speed and accuracy, offering valuable contributions to the field of automated spam detection and content filtering on social media.

II. LITERATURE SURVEY

The problem of detecting hashtag-oriented spam on Twitter is discussed by S. Sedhai and A. Sun in [1]. The authors introduced a large dataset comprising 14 million tweets to aid in research on spam detection. The dataset captures a wide variety of spam types and reflects different deceptive strategies used by spammers. This resource is considered valuable for building and evaluating effective spam detection techniques.

Analyzing the spread of fake content on Twitter, especially during critical events, was studied by A. Gupta, H. Lamba, and P. Kumaraguru in [2]. The authors collected and analyzed tweets from the Boston Marathon bombing incident, identifying patterns of misinformation. They emphasized the challenge of real-time fake content detection and highlighted the importance of early detection mechanisms.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404729|

In [3], F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida discussed identifying spammers using machine learning techniques. They proposed a model combining tweet content, user behavior, and network features to accurately detect spam accounts. Their classifier achieved high accuracy, indicating the robustness of their feature selection.

Spam detection using traditional classifiers was proposed by A. McCord and M. Chuah in [4]. The authors trained Naive Bayes, Decision Trees, and SVMs on features extracted from tweets and user profiles. Their study concluded that traditional ML classifiers can be highly effective when combined with appropriate feature engineering.

In [5], C. Chen, J. Zhang, Y. Xiang, W. Zhou, and G. Min focused on detecting offensive language in social media posts to protect adolescents. Their classification system uses NLP techniques to identify harmful content. The paper highlights the need for context-aware models due to the subtlety and variability of offensive language online. Spam and promotion campaign detection was addressed by J. Zhang, L. Zhu, J. Liu, and H. Li in [6]. The authors proposed a hybrid method using content-based and network-based features. Their ML approach effectively classified tweets and identified spammers, aiming to improve the user experience on social platforms.

K. Lee, J. Caverlee, and S. Webb in [7] uncovered social spammers by combining social honeypots and machine learning. Their hybrid approach proved effective in detecting unsolicited content and maintaining the platform's integrity. The study demonstrated the value of blending behavioral and social interaction cues.

In [8], S. Ghosh, G. Korlam, and N. Ganguly analyzed the structure of spammer networks within Twitter. By using network analysis techniques, the authors identified patterns that differentiate spammer communities from legitimate users. Their findings provide insights useful for developing enhanced spam detection systems.

The detection of spam in Twitter networks was explored by S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd in [9]. They developed a framework by analyzing tweet content, user behavior, and interactions. The study emphasized the necessity of multi-dimensional approaches to reduce spam prevalence on the platform.

Twitter's dual role as a social network and a news media platform was studied by H. Kwak, C. Lee, H. Park, and S. Moon in [10]. The authors analyzed how this duality influences spam propagation and proposed methods to distinguish between legitimate news and spam. Their insights help in designing better detection systems that consider Twitter’s unique characteristics.

III. METHODOLOGY

The process of detecting spam tweets in real-time using machine learning involves a series of well-defined steps. this methodology is designed to systematically collect, process, analyze, and classify tweets to distinguish between spam and legitimate content. the approach is modular, allowing for scalability, comparison across models, and real-time implementation.

The project begins with data collection, where tweets are gathered using the twitter api. this is achieved by defining a set of keywords, hashtags, and real-time trending topics. the dataset is composed of both spam and non-spam tweets to ensure a balanced classification environment. additionally, previously published benchmark datasets are also incorporated to enhance the reliability and coverage of the training data.

Once the data is collected, it undergoes preprocessing to make it suitable for analysis. raw tweets often contain noise in the form of urls, user mentions, hashtags, emojis, and special characters. these elements are removed during preprocessing. further normalization steps include converting all text to lowercase, tokenizing the text into individual words, and applying lemmatization or stemming to reduce words to their base forms. this step ensures uniformity in the textual data and reduces dimensionality for the subsequent feature extraction process.

Following preprocessing, the next step involves feature extraction, which is critical for enabling machine learning algorithms to understand and differentiate spam tweets from legitimate ones. both textual and behavioral features are extracted. textual features include n-grams and term frequency-inverse document frequency (TF-IDF) values, which

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404729|

capture the relevance and context of words in the dataset. behavioral features such as the user's tweet frequency, follower-following ratio, and tweet metadata (e.g., presence of urls, mentions, and hashtags) are also incorporated. these features are chosen based on their significance in prior research and their proven effectiveness in spam classification tasks. With the feature set prepared, various machine learning algorithms are trained to classify tweets. in this project, five well-known algorithms are selected: logistic regression, random forest, naive bayes, gradient boosting, and support vector machine (svm). to evaluate the stability and performance of each model under different conditions, five distinct train-test split ratios are used: 50:50, 60:40, 70:30, 80:20, and 90:10. this results in a total of 25 model combinations, which are compared to determine the best-performing configuration.

The performance of each model is evaluated using standard classification metrics, namely accuracy, precision, recall, and f1-score. a confusion matrix is also generated for each combination to analyze the distribution of true positives, false positives, true negatives, and false negatives. to better visualize the results, accuracy comparison graphs are generated using the matplotlib library, helping in the interpretation of model effectiveness under varying training conditions. To provide a user-friendly interface and support real-time detection, a web application is developed using the Django framework. this application allows users to input tweet text or fetch live tweets, select a preferred algorithm, and receive instant classification results (spam or not spam). the frontend also displays accuracy comparison graphs, enabling users to understand the performance of each model.

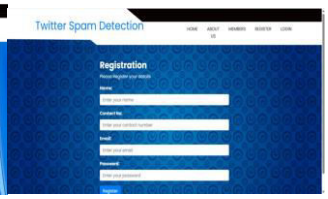
Lastly, an ensemble-based approach is implemented, combining predictions from multiple algorithms to improve the overall reliability of spam detection. this ensemble method helps to reduce the risk of individual model bias and improves generalization. the final outcome identifies the most accurate and robust model for real-time twitter spam detection.

IV. EXPERIMENTAL RESULTS

Figures shows the results of tweet detection from an tweet enter by the user. Website include: Home page, About page, Register page, login page, Homepage after login, tweet input page, output page with pop message, accuracy graphs, multiple algorithm option.

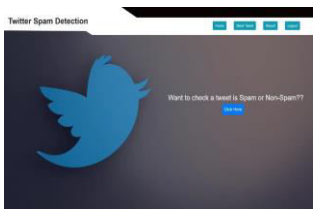
(a)

(b)

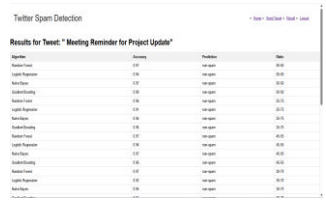


(c)

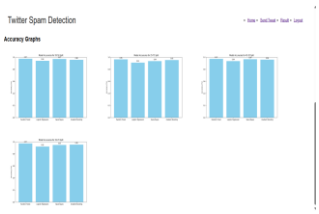
(d)



(e)



(f)



(g)

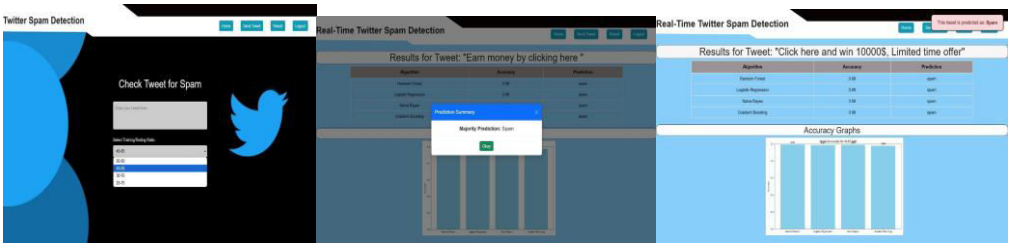
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404729|



(g) (h) (i)

V. CONCLUSION

The Twitter Spam Detection project presents a robust and scalable solution for identifying spam content on social media using supervised machine learning techniques. The core idea revolves around building a secure, user-friendly web-based application that enables registered users to input tweets, choose from multiple classification models, and view real-time classification results. The system supports both dataset uploads and individual tweet analysis, with configurable training-to-testing data split ratios (such as 50:50 or 30:70). Tweets undergo preprocessing steps including cleaning, tokenization, and normalization, followed by feature extraction using TF-IDF vectorization.

REFERENCES

[1] S. Sedhai and A. Sun, "HSpam14: A Collection of 14 million Tweets for Hashtag- Oriented Spam Research," Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Vol. 2015, No. 8, pp. 223-232, 2015.

[2] A. Gupta, H. Lamba, and P. Kumaraguru, "1.00 per RT #BostonMarathon #PrayForBoston: Analyzing Fake Content on Twitter," Proceedings of the 22nd International Conference on World Wide Web, pp. 1003-1010, 2013.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter," Proceedings of the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), 2010.

[4] A. McCord and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers," Proceedings of the 8th International Conference on Autonomic and Trusted Computing, pp. 175-186, 2011.

[5] C. Chen, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust, and 2012 International Conference on Social Computing, pp. 71-80, 2012.

[6] J. Zhang, L. Zhu, J. Liu, and H. Li, "Detecting Spam and Promoting Campaigns in Twitter," Proceedings of the 2012 International Conference on Distributed Computing Systems, pp. 153-160, 2012.

[7] K. Lee, J. Caverlee, and S. Webb, "Uncovering Social Spammers: Social Honeypots+ Machine Learning," Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 435-442, 2010.

[8] S. Ghosh, G. Korlam, and N. Ganguly, "Spammers' Networks within Online Social Networks: A Case-study on Twitter," Proceedings of the 3rd International Conference on Social Computing, pp. 188-195, 2011.

[9] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting Spam in a Twitter Network," First Monday, Vol. 15, No. 1, pp. 25-30, 2010.

[10] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?," Proceedings of the 19th International Conference on World Wide Web, pp. 591-600, 2010.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details