Spam Detection Visualization Using Power BI

Monika Barde[1], Neha Sahu[2], Pranay Narnaware[3], Yashika Deshmukh[4], Ankeshvar Mawase[5]

Abstract

Spam detection is a crucial task in data management, where identifying and filtering unwanted content is essential for enhancing the quality of user experience and system performance. This project presents a visualization approach for spam detection using Power BI, which leverages data analytics to provide an interactive and intuitive platform for understanding and managing spam data. By integrating datasets of email content or messages, Power BI dashboards facilitate real-time monitoring and detection of spam patterns. This study explores the implementation of Naive Bayes for spam detection, leveraging Power BI as a data analysis and visualization tool. Power BI's interactive interface is used to preprocess and visualize the data, allowing the integration of Naive Bayes classification models for effective spam filtering. This combination of Naive Bayes and Power BI offers an efficient, user friendly framework for spam detection, suitable for real-time monitoring and decision-making.

Keyword: Naïve Bayes Classification, Machine Learning, Power BI, Spam Massages, Python

1. Introduction:

Spam detection has become an integral part of maintaining the integrity and usability of digital communication platforms, whether in email systems, social media, or messaging applications. With the increasing volume of unsolicited and malicious messages, there is a growing need for automated solutions that can effectively filter out spam from legitimate content. Traditional spam detection methods often rely on machine learning models or rule-based systems to classify and flag spam messages. However, the complexity and variety of spam tactics require sophisticated tools for real-time analysis and monitoring. Power BI, a powerful business intelligence tool developed by Microsoft, offers a unique solution to visualize and interpret large datasets through interactive and intuitive dashboards. When applied to spam detection, Power BI enables users to visualize patterns and trends in spam data, identify anomalies, and assess the effectiveness of different spam filtering techniques. Technologies used in our project is Power BI for visualization and data interaction, Python for data processing, machine learning, and model integration, Naive Bayes for classification Text Processing tools for converting text to features, Power BI integration with Python for executing the spam detection model. his combination of technologies enables efficient spam detection workflows.

Objective of Research:

The primary objective of this research is to explore how Power BI can be utilized as a powerful tool to enhance the process of spam detection through data visualization and analytics. The specific objectives include:

To Visualize Spam Detection6+ Data: The research aims to create interactive and insightful visualizations of spam detection data, enabling users to easily interpret and explore key patterns in spam behaviour.

To Enhance Real-Time Monitoring of Spam Activity: The research seeks to implement real time dashboards that allow for the continuous monitoring of spam activity.

To Identify Patterns and Trends in Spam: Another key objective is to identify recurring patterns and trends in spam data, including content characteristics (e.g., keywords, links), sender behaviours, temporal fluctuations, and geographical distributions.

2. Literature Reviews

Literature reviews provide a comprehensive overview of the current state of knowledge and understanding in the field related to future research. The recent surveys and research articles regarding spam detection techniques have been taken from different perspectives into account for the literature survey.

Kutub Thakur, MD Liakat Ali, and Muath A. Obaidat (November 2023) provide a systematic review on deep-learning-based phishing email detection. This review explores various deep learning techniques used for phishing email detection, assessing their effectiveness and identifying areas for future research [6].

C. T. Goncalves, M. J. A. Goncalves, and M. I. Campante (November 2023) discuss the development of integrated performance dashboards using Power BI. Their research highlights how business intelligence tools aid decision-making by transforming data from multiple sources into dynamic visual dashboards [7].

F. J. Martino, Rocio Alaiz, V. G. Castro, and Eduardo Fidalgo (May 2022) present an in-depth analysis of spammer strategies and the dataset shift problem in spam email detection. Their study reviews state-of-the-art machine learning techniques for spam filtering and examines the challenges posed by evolving spam tactics and dataset variability [8].

Yuhao Zhu (2022) provides a review of Naïve Bayesian spam filtering. This study evaluates the Naïve Bayes classifier’s efficiency in spam detection, comparing unigram and bigram models, with the latter achieving a higher accuracy of 79.36% [9].

3. Methodology

The methodology used for creating a dashboard and Spam detection can be approached in a variety of ways, depending on the nature of the data and tools available for spam detection using Power BI with Naïve Bayes can be broken down into several systematic steps.

Data Collection and Preparation: Gather and clean the data, and preprocess text, Feature Extraction: Convert text into numerical features, Feature Selection: Identify the most relevant features that best represent the spam versus non-spam categories Model Building: Use Naïve Bayes for classification, Leverage Power BI for visualization and model integration, Visualization and Reporting: Create insightful visualizations and interactive reports, Real-Time Detection.

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combination of values in a given dataset. In this research, Naïve Bayes classifier use bag of words features to identify spam e-mail and a text is representing as the bag of its word. The bag of words is always used in methods of document classification, where the frequency of occurrence of each word is used as a feature for training classifier.

$$P(spam\ |word) = \frac{P(spam).P(word|spam)}{P(spam).\ P(word|spam)+P(non-spam).\ P(word|non-Spam)}$$

Where;

(i) P (spam word) is probability that an e-mail has particular word given the e-mail is spam.

(ii) P(spam) is probability that any given message is spam.

(iii) P (word spam) is probability that the particular word appears in spam message.

(iv) P (non − spam) is the probability that any particular word is not spam.

(v) P (wordnon − spam) is the probability that the particular word appears in non-spam message

To achieve the objective, the research and procedure is conducted in three phases. The phases involved are as follows:

(i) Phase 1: Pre-processing

(ii) Phase 2: Feature Selection

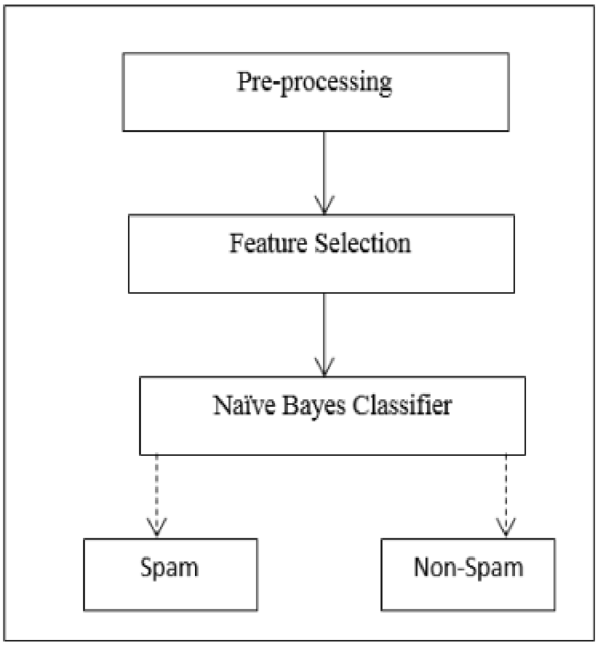


Fig 4.1: Process of E-mail spam filtering based on Naive Bayes Algorithm

(iii) Phase 3: Naïve Bayes Classifier

Pre-processing

Today, most of the data in the real world are incomplete containing aggregate, noisy and missing values. Pre-processing of e-mails in next step of training filter, some words like conjunction words, articles are removed from email body because those words are not useful in classification.

Feature selection

After the pre-processing step, we apply the feature selection algorithm, the algorithm which deploy here is Best First Feature Selection algorithm.

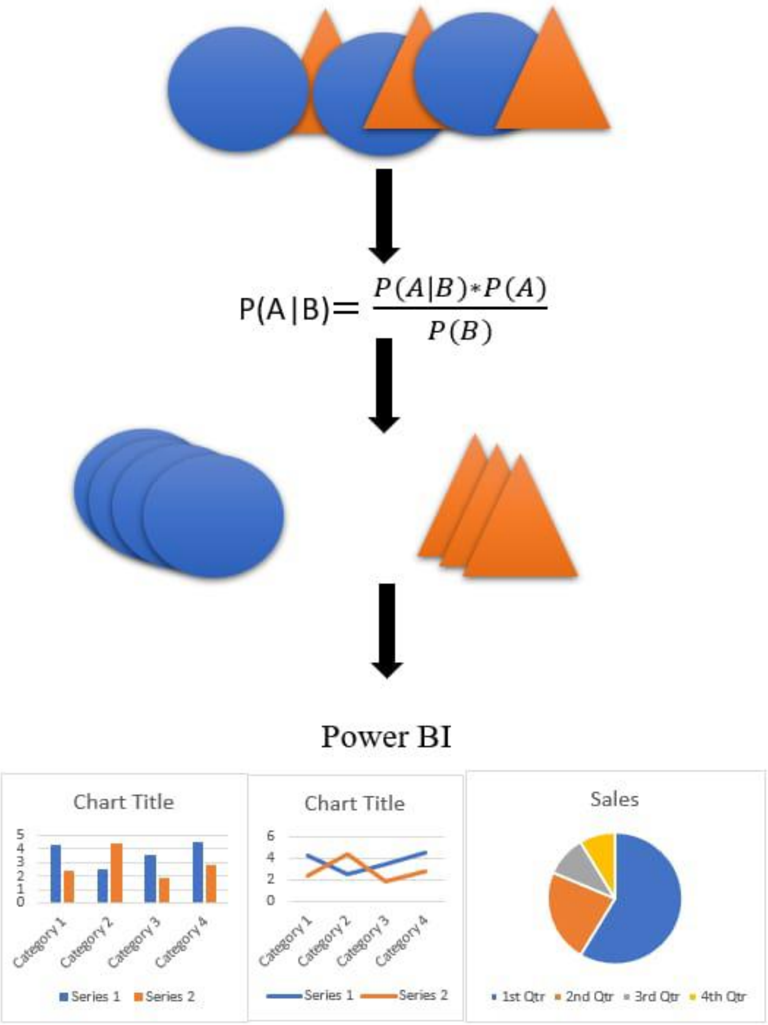
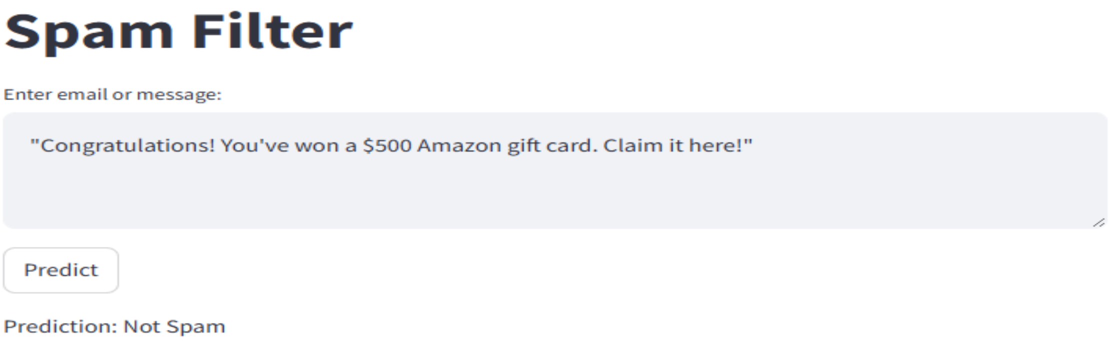


Fig 4.2: Data classifier

4. Result:

The results obtained from implementing spam detection using Power BI with an Naives bayes model can vary depending on the quality of the dataset and the tuning of the model Spam vs. Non-Spam Distribution



Spam Filter

Enter email or message:

"Congratulations! You've won a $500 Amazon gift card. Claim it here!"

Prediction: Not Spam

Fig 5.1: Expected output

5. Conclusion

In conclusion, Power BI serves as a powerful tool for analyzing and visualizing spam detection data, offering valuable insights that help identify patterns, trends, and behaviours associated with spam messages. By leveraging its data transformation capabilities, integration with machine learning models, and rich visualization options, organizations can gain a deeper understanding of spam activity, optimize spam detection strategies, and improve overall communication security. Power BI enables the analysis of large volumes of data from various sources, such as email logs, social media interactions, and message bodie Spam Detection and Visualization using Power BI offers a powerful approach to enhancing email security and mitigating the risks associated with spam. By leveraging machine learning algorithms, we can accurately identify and filter out spam emails, thereby protecting users from potential threats and improving productivity.

References

1. Banerjee, Kallal, Siddharth Das, and Soumen Nath. "Data Visualization Approach for Business Strategy Recommendation Using Power BI Dashboard." International Journal of Research in Management, vol.6 Jan. 2024.
2. Dharroa, Deepak, and Shailesh Gawai. "Classifying SMS as Spam or Ham: Leveraging NLP and Machine Learning Techniques." International Journal of Safety and Security Engineering, vol. 14, Mar. 2024.

3. Kumar, Stendra, Raj Kumar, and Ashish Saini. "Supervised Learning-Based E-Mail/SMS Spam Classifier." June 2024.
4. Metre, K. V., A. Mathur, R. P. Dahake, Y. Bhapkar, J. Ghadge, P. Jain, and S. Gore. "An Introduction to Power BI for Data Analysis." International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 1s, June 2024, pp. 142-147.
5. Asmitha, M., and Kavita C. R. "Exploration of Automatic Spam/Ham Message Classifier Using NLP." International Conference for Convergence in Technology, June 2024.
6. Thakur, Kutub, MD Liakat Ali, and Muath A. Obaidat. "A Systematic Review on Deep-Learning-Based Phishing Email Detection." Department of Professional Security Studies, New Jersey City University, Nov. 2023.
7. Goncalves, C. T., M. J. A. Goncalves, and M. I. Campante. "Developing Integrated Performance Dashboards Visualizations Using Power BI as a Platform." CEOS.PP, ISCAP, Polytechnic of Porto, LIACC, University of Porto, Nov. 2023.
8. Martino, Francisco Janez, Rocío Alaiz Rodríguez, Víctor González Castro, Eduardo Fidalgo, and Enrique Alegre. "A Review of Spam Email Detection: Analysis of Spammer Strategies and the Dataset Shift Problem." Artificial Intelligence Review, May 2022.
9. Zhu, Yuhao. "Naive Bayesian Spam Filtering." Engineering and Technology, vol. 38, 2022.