Project Report

Spam Email Classification using PySpark

Course Title: Big Data & IOT Lab

Course Code: CSE 413

SUBMITTED TO:

Md. Zahim Hasan

Designation: Lecturer

Department of CSE

Daffodil International University

SUBMITTED BY:

Khadiza Akter Mafia	Oly Rani Pal	Jui Saha
ID: 201-15-13652	ID: 201-15-14100	ID: 201-15-14115
Section : I-55	Section : I-55	Section : I-55

1. Objective:

The primary objective of this project is to develop a spam email classification system using PySpark. The project involves data preprocessing, implementing a machine learning model for classification, designing a user interface, and displaying the output through web development.

2. Introduction:

In the contemporary digital landscape, email communication remains a ubiquitous and indispensable means of information exchange. However, the proliferation of unsolicited and malicious emails, commonly known as spam, poses a significant threat to the efficiency and security of email systems. Spam emails not only inundate users with irrelevant content but also often harbor malicious intents, including phishing attacks and malware distribution. Consequently, the need for robust spam email classification systems has become paramount to safeguard users and maintain the integrity of communication platforms.

This project addresses the imperative challenge of spam email classification, aiming to develop an effective and scalable solution using PySpark, a powerful distributed computing framework. PySpark's capabilities make it well-suited for processing large volumes of email data, enabling the creation of a sophisticated machine learning model for accurate spam detection.

.3. Motivation:

In recent years, the proliferation of spam emails has become a pervasive issue, posing significant challenges to individuals, businesses, and organizations alike. The surge in the volume and sophistication of spam not only threatens the efficiency of communication systems but also jeopardizes data security and user experience. As a response to this pressing concern, the motivation behind undertaking the project on Spam Email Classification using PySpark is multi-faceted.

1. Email Security Enhancement:

- The primary motivation is to enhance email security by developing a robust and accurate system capable of distinguishing between legitimate and spam emails. By implementing advanced machine learning techniques facilitated by PySpark, we aim to create a scalable solution that can effectively identify and filter out unwanted and potentially harmful email content.

2. User Experience Improvement:

- Spam emails not only inundate inboxes but also contribute to a decline in user experience. Sorting through a flood of irrelevant and potentially fraudulent emails consumes valuable time and resources. This project aims to alleviate this burden by providing a reliable and efficient spam email classification system, ultimately improving the overall user experience in managing email communication.

3. Scalability Challenges:

- Conventional spam filtering methods may face scalability challenges as email datasets grow exponentially. PySpark, with its distributed computing capabilities, offers a scalable framework that can handle large volumes of data efficiently. The motivation is to leverage PySpark to develop a solution that can seamlessly scale with the increasing demands of modern email systems.

4. Adaptation to Evolving Spam Patterns:

- The dynamic nature of spam necessitates adaptive solutions. Traditional rule-based systems may struggle to keep pace with the evolving tactics employed by spammers. By utilizing machine learning algorithms within PySpark, we seek to create a classification model that can adapt to changing spam patterns, ensuring a proactive defense against emerging threats.

5. Open Source Collaboration and Contribution:

- PySpark, being an open-source framework, provides an opportunity for collaborative development and community contribution. The motivation includes not only addressing a critical issue in email security but also actively participating in the open-source community to share insights, methodologies, and advancements in spam email classification.

By addressing these motivations, this project aspires to make a meaningful contribution to the realm of email security, offering a scalable, efficient, and adaptive solution to combat the ever-growing challenges posed by spam emails. The integration of PySpark ensures not only the effectiveness of the model but also the potential for widespread adoption and continuous improvement in the fight against spam.

4. Dataset:

The dataset for this project is sourced from Kaggle, consisting of 5573 labeled emails. It has two colamn named Categories and Message. The dataset is split into spam and non-spam categories, providing a diverse set of examples for training and evaluating the classification model.

5. Methodology:

The project implementation is conducted in a Google Colab notebook, and the code can be accessed here. The notebook details the step-by-step process, including data preprocessing, model implementation, and UI design.

5.1 Data Preprocessing:

The data preprocessing phase involves cleaning the dataset, handling duplicates, and performing text preprocessing. Tokenization, stop word removal, and TF-IDF feature extraction are applied to prepare the data for model training.

5.2 Machine Learning Model Implementation:

PySpark's MLlib is utilized to implement a classification model. The chosen algorithm, logistic regression, is trained on the preprocessed data. The model is evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

3. UI Design and Web Development:

The user interface is designed using a web framework (Flask in this case). The backend integrates with PySpark for email classification. The frontend allows users to upload emails, triggering the classification process. The classification results are displayed in a user-friendly format.

6. Result:

After implementing a lot of model we got thw best result by using SVC. Its Accuracy Rate is 0.98 and precision rate is 0.98 as well, which is a pretty good accuracy rate. The web interface provides a seamless experience for users to interact with the system and interpret the results.

7. Discussion:

7.1 Challenges:

As we used a huge dataset, there were a lot of unusual value. Removing all these unusual was bit challenging. But at the end we successfully completed all the task.

7.3 Limitations:

Acknowledge the limitations of the implemented solution, such as potential biases in the dataset or constraints in the model's generalization.

8. Conclusion:

The project successfully develops a spam email classification system using PySpark, combining data preprocessing, machine learning, and web development. The integration of PySpark ensures

scalability, while the user interface enhances accessibility. The classification results indicate the effectiveness of the implemented solution in identifying spam emails.

9. Future Work:

Despite the project's success, there are opportunities for future enhancements. Potential areas for improvement include exploring different machine learning algorithms, conducting more extensive hyperparameter tuning, and refining the user interface for a more intuitive experience. Additionally, addressing any limitations observed during the project could contribute to further system optimization.

10. References:

Online resources – Kaggle.