BDM 3035: BIG DATA ANALYTICS CAPSTONE PROJECT

Lambton College

Submitted to: MEYSAM EFFATI

Submitted on: 2024-07-15

SMS SPAM CLASSIFIER MILESTONE REPORT 3

GROUP F

ALWIN KANNYAKONIL SCARIA

ANISHA SUSAN MATHEW

ASHNA VIJI ALEX

JOBIN PHILIP

MOHAMED AFTHAB

INTRODUCTION

Following the completion of feature extraction and data preprocessing, the next crucial step in our SMS Spam Classifier project is model development. This milestone focuses on selecting appropriate machine learning models, training them, and working on optimizing their performance for accurately classifying SMS messages as spam or non-spam.

PROGRESS REPORT

Objective:

1. Model Selection: Choosing the best-suited machine learning algorithms for the classification task.

2. Model Training: Training the selected models using the preprocessed dataset.

SUMMARY OF TASKS COMPLETED:

1. Model Selection

We considered several machine learning models known for their effectiveness in text classification tasks:

- Logistic Regression: A simple yet powerful algorithm for binary classification.

- Naive Bayes: Particularly effective for text data due to its probabilistic approach.

- Support Vector Machines (SVM): Known for high accuracy in text classification.

- Random Forest: Offers robustness against overfitting due to ensemble learning.

- Gradient Boosting Machines (GBM): Useful for capturing complex patterns in the data.

2. Model Training

Each selected model was trained on the training dataset, which was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate class imbalance. The models were evaluated based on various metrics such as accuracy, precision, recall, and F1-score.

KEY ACHEIVEMENTS AND MILESTONES

1. Model Selection and Training:

- Successfully trained multiple models, achieving a baseline performance.

- Initial performance metrics indicated good potential for improvement with further tuning.

2. Evaluation Metrics:

- Established evaluation metrics and benchmarks for model comparison.

CHALLENGES FACED

- Hyperparameter Tuning Complexity: The high dimensionality of hyperparameter space made the tuning process computationally intensive. Efficient parallelization are under progress which is essential to manage this complexity.

- Class Imbalance: Despite the use of SMOTE, handling the nuances of class imbalance remained challenging, particularly in ensuring robust performance on the minority class.

LESSONS LEARNED

- Model Diversity: Different models offer unique strengths and weaknesses, and ensemble methods can help leverage these differences for improved performance.

- Importance of Hyperparameter Tuning: Fine-tuning model parameters is critical to achieving optimal performance.

TIMELINE TABLE

Task	Original Timeline	Revised Timeline	Status
Data Collection	Week 1	Week 1	Completed
Data Preprocessing	Week 2	Week 2	Completed
Feature Extraction	Week 3	Week 3	Completed
Model Selection	Week 4	Week 4	Completed
Model Training	Week 5	Week 5	Completed
Model Evaluation	Week 6	Week 6	Upcoming
Real-time System	Week 7	Week 7	Upcoming
Final Report and Demo	Week 8	Week 8	Upcoming

NEXT STEPS

1. Model Evaluation and Validation:

- Finalize the evaluation of all models, including additional metrics like Area Under the Curve (AUC) for better insight into model performance.

- Perform cross-validation to ensure robustness.

2. Implementation of a Real-Time Detection System:

- Develop and integrate a real-time system for spam detection, focusing on scalability and efficiency.

3. Documentation and Reporting:

- Prepare comprehensive documentation for the final report, detailing methodologies, findings, and recommendations.

4. Final Presentation and Submission:

- Consolidate all project components and present the findings and final model to stakeholders.

CONCLUSION

This milestone has marked a significant step forward in our project, with successful model training .As we move towards the final stages of the project, we are well-prepared to develop a robust and efficient SMS spam detection system.

REFERENCES

1. Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). Contributions to the Study of SMS Spam Filtering: New Collection and Results.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.