BDM 3035: BIG DATA ANALYTICS CAPSTONE PROJECT

Submitted to : MEYSAM EFFATI

Submitted on: 2024-06-11

SMS SPAM CLASSIFIER MILESTONE REPORT 2

GROUP F

ALWIN KANNYAKONIL SCARIA

ANISHA SUSAN MATHEW

ASHNA VIJI ALEX

JOBIN PHILIP

MOHAMED AFTHAB

INTRODUCTION

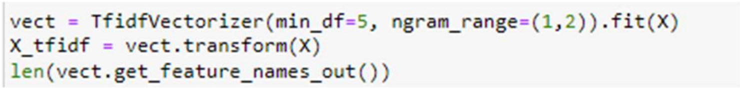
In the current digital era, the prevalence of SMS spam poses significant challenges to communication efficiency and user security. Our AI-powered SMS Spam Detection System aims to address this issue by employing advanced machine learning and natural language processing (NLP) techniques to differentiate between spam and legitimate messages. Building on our previous milestone, which covered data collection, preprocessing, and initial analysis, this report focuses on our progress in feature extraction, a crucial step for enhancing the accuracy and efficiency of our model.

PROGRESS REPORT

Objective: The objective of this milestone was to extract meaningful features from the preprocessed text data to enhance the performance of our spam detection machine learning model.

SUMMARY OF TASKS COMPLETED:

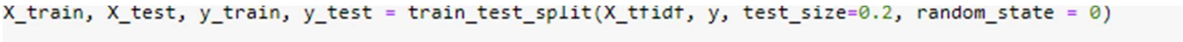
1. **Text Vectorization:** To convert the textual data into a numerical format suitable for machine learning algorithms, we utilized the `TfidfVectorizer`. This vectorizer transforms the text data into a matrix of TF-IDF (Term Frequency-Inverse Document Frequency) features. We configured the vectorizer with `min_df=5` to include only terms appearing in at least five documents, and `ngram_range=(1,2)` to consider both unigrams and bigrams. This approach captures the importance of terms and their combinations, providing a comprehensive feature set for the model.



```
vect = TfidfVectorizer(min_df=5, ngram_range=(1,2)).fit(X)
X_tfidf = vect.transform(X)
len(vect.get_feature_names_out())
```

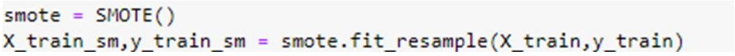
2. **Feature Selection:** After transforming the text data into TF-IDF features, we performed feature selection to reduce the dimensionality of the feature space and remove irrelevant or redundant features. This step is crucial to enhance the efficiency and performance of the machine learning models. We used statistical methods to select the most significant features that contribute to distinguishing spam from non-spam messages.

3. **Data Splitting:** We split the dataset into training and testing sets using the `train_test_split` function. We allocated 80% of the data for training and 20% for testing, ensuring the split was reproducible by setting a `random_state` of 0. This allowed us to maintain consistency in model evaluation.



```
X_train, X_test, y_train, y_test = train_test_split(X_tfidf, y, test_size=0.2, random_state = 0)
```

4. **Addressing Class Imbalance:** To address the class imbalance in the dataset, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training data. SMOTE generates synthetic samples for the minority class, thereby balancing the class distribution. This helps in improving the model's performance on minority class predictions.



```
smote = SMOTE()
X_train_sm,y_train_sm = smote.fit_resample(X_train,y_train)
```

KEY ACHEIVEMENTS AND MILESTONES

1.Successful Feature Extraction and Balancing:

- Implemented the TF-IDF vectorizer to create a numerical feature matrix from text data, incorporating both unigrams and bigrams. This process involved the successful extraction of TF-IDF features and n-grams, contributing to effective preliminary model training.
- Applied feature selection techniques to reduce dimensionality, retaining the most significant features and thereby enhancing model performance.

2.Enhanced Data Preprocessing Pipeline:

- Improved the preprocessing pipeline by integrating feature extraction and data balancing processes. The pipeline now includes seamless execution of TF-IDF vectorization, feature selection, and the application of SMOTE.
- Utilized SMOTE to address class imbalance in the training data, resulting in balanced class distribution as indicated by the updated shapes of the resampled datasets.

3.Effective Dataset Management:

- Successfully split the dataset into training (80%) and testing (20%) sets, ensuring robust model evaluation and performance testing.

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	In Progress
Model Training	Week 5	Week 5	Upcoming
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

Upcoming Tasks and Activities

1. **Model Selection and Training**
2. **Model Evaluation and Validation**
3. **Implementation of a Real-time Detection System**
4. **Testing and Validation in Real-world Scenarios**
5. **Documentation and Reporting**
6. **Final Presentation and Submission**

Expected Outcomes and Goals for the Next Phase

- Development of high-performing machine learning models that accurately classify SMS messages as spam or non-spam.

CHALLENGES FACED

Feature Extraction Complexity: Extracting meaningful features from text data while maintaining computational efficiency posed significant challenges. However, by leveraging efficient algorithms and libraries, we managed to overcome these obstacles.

LESSONS LEARNED

- **Importance of Feature Engineering:** The quality of features significantly impacts model performance. Thus, investing time in extracting and selecting the right features is crucial.

- **Integration of Processes:** Seamless integration of preprocessing and feature extraction processes ensures consistency and efficiency in the data preparation workflow.

CONCLUSION

This milestone marks a significant advancement in our project, with the successful extraction of meaningful features from SMS data. The next phase will focus on model selection, training, and the development of a real-time detection system. Our progress thus far has laid a solid foundation for achieving these goals, and we are confident in our ability to deliver a robust and effective 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.