BDM 3035: BIG DATA ANALYTICS CAPSTONE PROJECT

Submitted to: MEYSAM EFFATI

Submitted on: 2024-06-11

SMS SPAM CLASSIFIER MILESTONE REPORT 1

GROUP F

ALWIN KANNYAKONIL SCARIA - C0894287

ANISHA SUSAN MATHEW - C0907393

ASHNA VIJI ALEX - C0901082

JOBIN PHILIP – C0895950

MOHAMED AFTHAB - C0891945

TABLE OF CONTENTS

INTRODUCTION

Mobile communication has become essential to our everyday lives in the current digital era. Short Message Service, or SMS, is still a popular way to communicate on a personal, business, and promotional level. However, the widespread use of SMS has also resulted in a sharp rise in spam messages, which can be annoying as well as a possible danger to users' privacy and security. It will take creative solutions to effectively distinguish between spam and legitimate messages to address this challenge.

We are pleased to present our AI-powered SMS Spam Detection System, a cutting-edge tool that will completely transform the way consumers handle their SMS correspondence. Our system is a powerful and easy-to-use tool for managing and filtering SMS spam by using machine learning and natural language processing (NLP) techniques.

Our AI-driven SMS Spam Detection System seeks to provide both individual and business users with a host of advantages. Professionals will benefit from improved communication as a result of it preventing crucial messages from getting lost in spam and increasing efficiency and productivity. It will assist individual users in keeping their inboxes organized so they can concentrate on important discussions and steer clear of potentially hazardous content.

Our system's primary function is to reliably identify SMS messages as either spam or non-spam (ham). Advanced machine learning models trained on large datasets of labeled SMS messages are used to achieve this. In the constantly changing world of SMS spam, our system stays effective by continuously learning and adjusting to new spam patterns.

Our approach is comprehensive and methodical, starting with data collection and preprocessing to ensure that the raw text data is ready for analysis. Key preprocessing steps include text cleaning, tokenization, stop word removal, and normalization techniques such as stemming and lemmatization. These steps are crucial for transforming raw SMS data into a format suitable for machine learning, enabling our models to perform with high accuracy and efficiency.

To summarize, our AI-based SMS Spam Detection System represents a significant advancement in the fight against SMS spam. It is designed to enhance user experience, improve communication efficiency, and protect against the risks associated with spam messages. We are confident that our solution will set a new standard in SMS spam detection, providing users with a reliable and intelligent tool to manage their SMS communications effectively.

PROGRESS REPORT

APPROACH AND METHODOLOGY

To guarantee the development of an efficient and successful spam detection system, the AI-based SMS Spam Detection System's approach and methodology are important. The actions that were followed are described in depth in this report, with a focus on the data collection and preparation methods that are necessary to create an accurate learning model.

SUMMARY OF TASKS COMPLETED

Data Collection

The first step of the research will be to gather a large dataset of SMS messages with labels. We used the SMS Spam Collection Dataset, which is a well-known dataset in the field. A large and varied collection of SMS texts, both spam and non-spam (ham), is offered by this dataset. This dataset is significant since it can serve as a strong basis for the machine learning model's construction and training. The dataset's diversity and comprehensiveness guarantee that the model will encounter an extensive range of spam and ham messages, facilitating its ability to integrate unique features.

Exploratory Data Analysis (EDA)

We carried out extensive Exploratory Data Analysis (EDA) in the project's initial phases to fully comprehend the underlying patterns and structures in the SMS dataset. Our EDA's main objectives were to investigate the distribution of spam and non-spam messages, examine the text's properties, and identify any anomalies or outliers that might have an impact on the model's performance.

Key Highlights of EDA:

- Data Distribution: We visualized the distribution of spam and non-spam messages to understand the class balance within the dataset.
- Text Length Analysis: Analyzed the distribution of message lengths to identify any unusual patterns that could affect preprocessing decisions.
- Word Frequency Analysis: Conducted word frequency analysis to identify the most common words in both spam and non-spam messages, which provided insights into potential features for the model.

Data Preprocessing

Data preprocessing is a critical step to prepare the raw text data for machine learning. The preprocessing includes the following steps:

a. Text Cleaning

Column Renaming: The dataset's unnamed columns will be removed first. We will rename the columns v1 and v2 to class and text, respectively. The SMS content will be in the text column, while the class column will indicate if the message is spam (1) or ham (0). By ensuring consistency and clarity throughout the dataset, this renaming makes it simpler to figure out and manipulate the data.

Noise Removal: The text's redundant components must be removed in the following step. Punctuation, special characters, and numbers that fail to assist the reader in understanding the text will be removed. By removing unnecessary data, this procedure helps in keeping the model less confused during training.

b. Tokenization

Definition: The process of splitting the text into separate words, or tokens, is known as tokenization. Tokenization would result in the following sentence, for instance: "This is a sample SMS" ["This", "is", "a", "sample", "SMS"].

Benefits:

- **Simplifies Analysis**: Makes it easier to handle text data at the word level. Tokenization breaks down the text into manageable units, facilitating more precise and focused analysis.
- **Enables Frequency Analysis**: Helps in determining the frequency of each word, which is important for feature extraction. By analyzing word frequencies, we can identify common patterns and features that distinguish spam from non-spam messages.

c. Stop Words Removal

Definition: Stop words are commonly used words that do not carry significant meaning, such as "and", "the", "is".

Benefits:

- **Reduces Dimensionality**: By removing these words, the dataset becomes less sparse, making it more manageable. This dimensionality reduction simplifies the analysis and improves the efficiency of the model.
- **Focus on Meaningful Words**: Helps in concentrating on the words that carry more semantic weight and are likely to contribute more to the classification task. By eliminating stop words, we can focus on the words that have a greater impact on the meaning and context of the messages.

d. Stemming and Lemmatization

Stemming: Words are shortened to their root or base form through stemming. For instance, "running" turns into "run". This procedure involves trimming word ends using heuristic rules to reduce the complexity and simplify the text data.

Lemmatization: Lemmatization breaks down words into their dictionary form, like stemming but more advanced. For instance, "better" turns into "good". This approach ensures that every word is handled consistently by breaking words down into their most basic forms using a dictionary.

Benefits:

- **Normalization**: Reduces multiple spellings of the same word to one. By ensuring that words with comparable meanings are handled as identical features, this normalization improves the model's accuracy.
- **Improved Accuracy**: improves the model's capacity to identify similar words accurately by treating them as having the same feature. Stemming and lemmatization improve the model's capacity to identify patterns and generate precise predictions by reducing the variances in the text data.

KEY ACHIEVEMENTS AND MILESTONES

1. **Effective Data Collection**
2. **Robust Preprocessing Pipeline**
3. **Exploratory Data Analysis**
4. **Integration of New Tools and Methodologies**
5. **Project readiness for subsequent phases**

DEVIATIONS FROM ORIGINAL PLAN

Dataset Shift

- Initial Plan: Initially, the project utilized the Enron spam dataset.
- Deviations: Transitioned to a more recent dataset that includes up-to-date spam and non-spam SMS.
- Reasons:
    - Data Relevance and Currency: Even though the Enron spam dataset was large, it wasn't relevant to the latest spam trends. The updated dataset guarantees that the model is trained on more relevant data by providing examples of recent spam and non-spam SMS.
    - Complexity Handling: The Enron dataset presented issues with quality and complexity, requiring the creation of a more manageable dataset to enable more seamless early development stages.

Scope Expansion

- Initial Plan: Initially focused on traditional spam detection.
- Deviations: Expanded the project's scope to address newer forms of spam and phishing attacks.
- Reason: Recognized the evolving nature of cyber threats, necessitating the integration of additional preprocessing tools and methodologies.
- Impact: To meet the enlarged scope and guarantee the model's flexibility and responsiveness to the most recent spam and phishing detection methods, the project's initial phases' timeline was extended.

MODIFIED TIMELINE TABLE

Name	Start Date	End Date	Progress %	Color
Data Collection and Advanced Text Preprocessi…	Jun 03, 2024	Jun 11, 2024	100	
Feature Extraction	Jun 12, 2024	Jun 24, 2024	20	
Model Training	Jun 25, 2024	Jul 11, 2024	0	
Model Evaluation and Selection	Jul 12, 2024	Jul 25, 2024	0	
Deployment and Real-time Integration	Jul 26, 2024	Aug 08, 2024	0	
Bug fixes, Deployment testing, and Fixing Unkn…	Aug 08, 2024	Aug 12, 2024	0	

NEXT STEPS

Upcoming Tasks and Activities

1. **Feature Engineering**
   - **Task Description:** Extract meaningful features from the preprocessed SMS data, such as term frequency-inverse document frequency (TF-IDF), n-grams, and other relevant text-based features.
2. **Model Selection and Training**
3. **Model Evaluation and Validation**
4. **Hyperparameter Tuning**

5. **Implementation of a Real-time Detection System**
6. **Testing and Validation in Real-world Scenarios**
7. **Documentation and Reporting**
8. **Final Presentation and Submission**

Expected Outcomes and Goals for the Next Phase

- **Expected Outcomes:** Creation of a feature-rich dataset that enhances the performance of machine learning models by providing a better representation of SMS messages.

CHALLENGES FACED

Data Quality and Relevance

- **Obstacle**: There have been errors in the model training and testing because the original dataset—the Enron spam dataset—was out of date and did not accurately reflect the trends in SMS spam today.
- **Solution**: We then switched to the SMS Spam Collection Dataset, which offered up-to-date samples of both spam and non-spam messages. This dataset was more recent and pertinent. By doing this, we made sure that the data used to train our model reflected the most recent spam trends.

Data Preprocessing Complexity

- **Obstacle**: Handling text data carefully during the preprocessing stages—tokenization, cleaning, and normalization—was necessary. Significant challenges were also presented by variations in SMS formats and language usage.
- **Solution**: Our preprocessing pipeline was effectively implemented with well-known NLP libraries such as NLTK. Tokenization, normalization (lemmatization and stemming), and text cleaning (removing punctuation, special characters, and numbers) were all automated processes in this pipeline. Additionally, we included unique preprocessing scripts to manage SMS format variations.

LESSONS LEARNED

We discovered during this project how crucial it is to use current and relevant datasets to ensure model relevance and accuracy. We found that to properly prepare data for machine learning, thorough preprocessing is essential. This includes text cleaning, tokenization, and normalization. SMOTE and other techniques dealt with imbalanced datasets, which greatly enhanced our model's performance in minority classes. The need for agile methodologies was highlighted by our ability to adjust to changing spam threats due to the flexibility in the project scope.

CONCLUSION

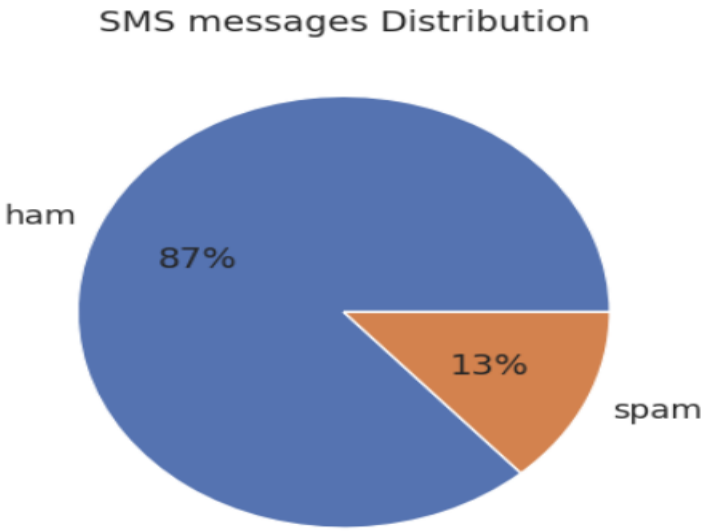
In conclusion, we have successfully completed significant stages of our AI-based SMS Spam Detection System project, including data collection, preprocessing, exploratory data analysis, and initial model development. After tackling important issues with dataset relevance, preprocessing complexity, and model performance, we were able to create an expandable and reliable preprocessing pipeline as well as efficient machine-learning models. The process that lies ahead of us includes feature engineering, more model selection, and tuning, deploying the system in real-time, thorough testing, and documentation. By putting in these efforts, we can make sure that our system is ready for user testing and future deployment—that is, that it can reliably and effectively detect SMS spam.

REFERENCES

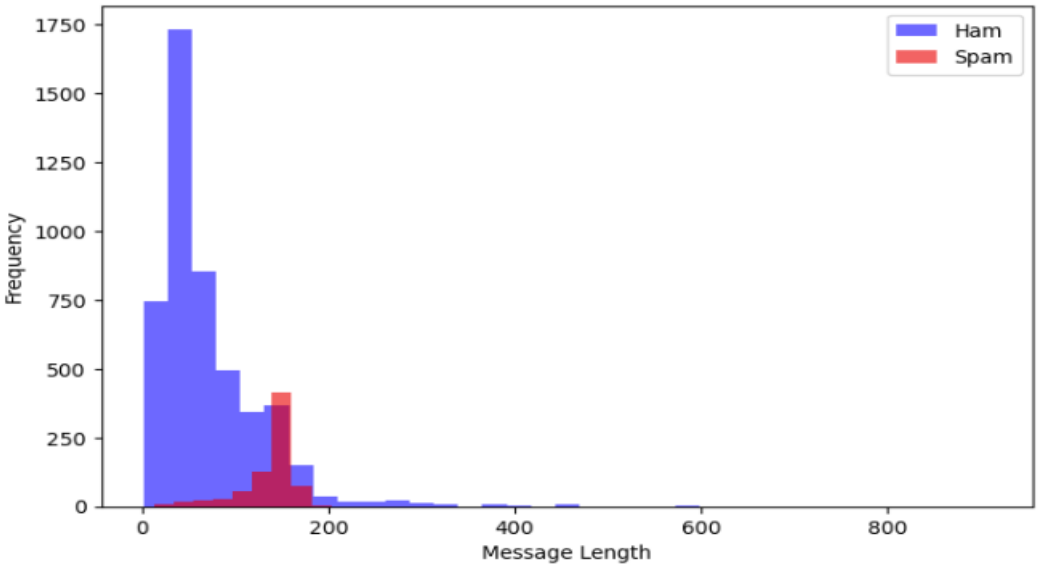
1. Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). Contributions to the Study of SMS Spam Filtering: New Collection and Results. Proceedings of the 11th ACM Symposium on Document Engineering. https://www.dt.fee.unicamp.br/~tiago/smsspamcollection/
2. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
4. 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.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30.
6. SpaCy Documentation. (n.d.). Industrial-Strength Natural Language Processing in Python. https://spacy.io/
7. Flask Documentation. (n.d.). Flask: Web Development, One Drop at a Time. https://flask.palletsprojects.com/
8. React Documentation. (n.d.). A JavaScript Library for Building User Interfaces. https://reactjs.org/

APPENDICES

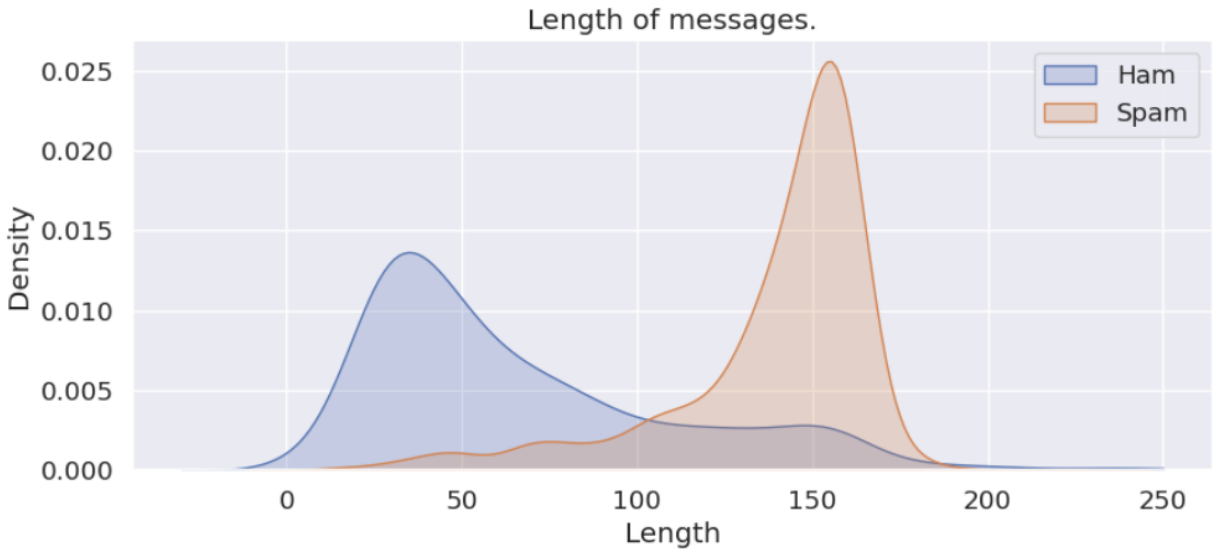
Appendix A: Exploratory Data Analysis (EDA) Visualizations



This visualization helps to understand the distribution of 'ham' and 'spam' messages in the dataset, providing insights into the balance between the two classes and helping to inform subsequent modeling decisions.



This visualization provides a quick and effective way to understand the distribution of message lengths between 'Ham' and 'Spam' messages, which is crucial for building and evaluating a classification model based on these features.



This visualization provides a deeper insight into the distribution of message lengths for 'Ham' and 'Spam' messages, which is essential for understanding the data and potentially for building predictive models.

Appendix B: DATA PREPROCESSING

Text preprocessing functions:

```
def clean_text(words):
    """The function to clean text"""
    words = re.sub("[^a-zA-Z]"," ", words)
    text = words.lower().split()
    return " ".join(text)

def remove_stopwords(text):
    """The function to removing stopwords"""
    text = [word.lower() for word in text.split() if word.lower() not in stop_words]
    return " ".join(text)

def stemmer(stem_text):
    """The function to apply stemming"""
    stem_text = [porter.stem(word) for word in stem_text.split()]
    return " ".join(stem_text)
```