Project Report: Team 6

Trần Thanh Sơn[1, 3]**, Lê Hải Nam**[1, 3]**, Vũ Kiên Trường**[2, 3]

[1]Khoa Khoa học và Kỹ thuật thông tin, Trường Đại học Công nghệ Thông tin

[2]Khoa Khoa học máy tính, Trường Đại học Công nghệ Thông tin

[3]Đại học Quốc gia Thành phố Hồ Chí Minh

{21522557, 21522357, 21522734}@gm.uit.edu.vn

Abstract

This project builds Spam or Ham Classifiers designed to enhance email communication by effectively distinguishing between legitimate (ham) and unwanted (spam) messages. Leveraging machine learning algorithms and natural language processing techniques, the classifier demonstrates high precision and recall rates. The model's adaptability, efficiency, and potential applications in real-world email systems are highlighted, emphasizing its role in improving user experience and safeguarding against phishing threats..

1 Exploratory Data Analysis

In this project, we utilized a dataset comprising email messages with the objective of classifying them into two categories: spam and ham. The dataset consists of approximately 20,000 messages, with spam messages constituting around 30% of the total. To conduct Exploratory Data Analysis (EDA), we gathered information on the length of each message, the distribution of spam and ham classes, and the frequency of word occurrences in both types of messages.

2 Data Preprocessing

Preprocessing steps included the removal of irrelevant words (HTML tags, punctuation and special characters), convert text to lowercase, handling missing values, dropping duplicated rows, tokenize the text, remove stop words, and stemming (using PorterStemmer). Throughout the study, notable observations were made, such as the length disparity between spam and ham messages and the prevalence of certain words in spam. These findings will inform the development of an effective classification model for accurately identifying and filtering spam messages.

Figure 1: Wordcloud of Spam Email Terms: Visualizing frequent terms in spam emails

3 Feature & Model Selection

Feature selection is pivotal in our model development, where we employed both the TF-IDF and Count Vectorizer to capture relevant information from the text data. The chosen features were identified during exploratory data analysis, emphasizing the importance of word n-grams. For model selection, we conducted a grid search over hyperparameters for various models, including Naive Bayes (NB), Logistic Regression (LR), XGBoost (XGB), and Support Vector Machine (SVM). Each model was coupled with its respective hyperparameter grid, optimizing for the best performance on a weighted F1 score during cross-validation. The final choice balanced computational efficiency with predictive accuracy, ensuring a robust spam or ham classification model.

4 Result & Discussion

Model	Acc.	**F1**	Recall	Precision
SVM_tfidf	94.25	**90.00**	91.01	89.01
XGB_count	94.42	**89.98**	88.07	91.97
LR_count	94.37	**89.67**	85.91	93.77
LR_tfidf	93.81	**89.18**	89.80	88.58
SVM_count	94.08	**89.15**	85.57	93.05
XGB_tfidf	93.73	**88.52**	84.96	92.39
NB_tfidf	93.51	**88.24**	85.57	91.08
NB_count	92.83	**87.89**	91.62	84.46

Table 1: Performance Metrics of Different Models, ordered by F1 measure

Model	**CPU Time**	**Wall Time**
NB - TF-IDF	2.38 s	20.9 s
NB - Count	3.22 s	22.2 s
LR - TF-IDF	5.11 s	1 min 9 s
LR - Count	2.88 s	1 min 11 s
XGBoost - TF-IDF	5 min 55 s	1 h 12 min 9 s
XGBoost - Count	16 min 43 s	1 h 23 min 10 s
SVM - TF-IDF	1.56 s	20.3 s
SVM - Count	1.5 s	29.1 s

Table 2: CPU and Wall Times for Different Models and Vectorization Methods

In this study, we evaluated various machine learning models for spam and ham classification (see Table 1). Highlights include SVM_tfidf achieving the highest F1 score (90.00%), indicating robust performance. XGB_count demonstrated excellent accuracy (94.42%) and a balanced F1 score (89.98%), while LR_count maintained high accuracy (94.37%) and precision (93.77%). Notably, SVM_tfidf stands out as the top performer, excelling in accuracy, precision, and recall. XGB_count closely follows, demonstrating a reliable ability to classify spam messages accurately. Logistic Regression with Count Vectorizer also performs well, especially in terms of accuracy and precision, with considerations for precision versus recall. These findings emphasize the significant impact of vectorizer and model choice on performance, showcasing the effectiveness of TF-IDF vectorization and the versatility of XGBoost and SVM in handling classification

tasks

Experimental Results

The computational efficiency of each model, assessed through CPU times and wall times, provides valuable insights into their performance. Table 2 presents the CPU times and wall times for various models using both TF-IDF and Count Vectorization.

These times reflect the computational resources required by each model, providing valuable insights into their efficiency. Notably, the Naive Bayes models exhibit relatively low CPU and wall times, while XGBoost models, particularly with TF-IDF, demand more computational resources. These considerations are crucial for applications where real-time processing or resource constraints play a significant role.

5 Conclusion

In conclusion, the SVM model with TF-IDF vectorization or XGBoost with Count Vectorizer are recommended choices based on their strong overall performance. However, the final selection should consider the specific goals and constraints of the application, such as the importance of precision or recall in the context of spam detection. Further fine-tuning or ensemble methods could be explored to optimize performance further.

Saab, S. A., Mitri, N., & Awad, M. (2014). Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering. *17th IEEE Mediterranean Electrotechnical Conference*, Beirut, Lebanon, 13-16 April 2014. Retrieved from https://ieeexplore.ieee.org/document/6820574

References

Saab, S. A., Mitri, N., & Awad, M. (2014). Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering. *17th IEEE Mediterranean Electrotechnical Conference*, Beirut, Lebanon, 13-16 April 2014. Retrieved from https://ieeexplore.ieee.org/document/6820574