e-ISSN 2774-5155
p-ISSN 2774-5147

Jurnal Sosial dan Teknologi (SOSTECH)
Volume 4, Number 4, *April* 2024
p-ISSN **2774-5147** ; e-ISSN **2774-5155**

ANALYSIS OF SMS SPAM DETECTION USING TF-IDF: A STUDY ON SMS SPAM COLLECTION DATASET

Nesan Jaya Saputra
President University, Indonesia
Email: nesan.saputra@student.president.ac.id

Abstract

This study explores the detection of SMS spam utilizing TF-IDF analysis on a dataset containing a collection of text messages labeled as spam or ham (non-spam). The dataset comprises messages suitable for spam detection analysis using TF-IDF techniques. The research aims to evaluate the effectiveness of TF-IDF in distinguishing between spam and spam (non-spam) messages. The analysis involves examining the precision, recall, and F1-score metrics to assess the performance of the classification model. The results demonstrate promising outcomes, with a high accuracy rate achieved in classifying spam and ham (non-spam) messages. Additionally, the study provides insights into the distribution of spam and ham (non-spam) labels in the test data, further enhancing our understanding of SMS spam detection techniques.

Kata kunci: : SMS spam detection, TF-IDF analysis, classification metrics, SMS Spam Collection dataset

INTRODUCTION

People are increasingly using mobile text messages as a way of communication. The popularity of short message service (SMS) has been growing over the last decade (Hosseinpour and Shakibian 2023), leading to the surge in spam messages. These unwanted messages not only flood users with irrelevant content but also pose risks such as phishing and fraud (Abid et al. 2022). To combat this issue, TF-IDF (Term Frequency-Inverse Document Frequency) analysis has emerged as a promising method for SMS spam detection (Artama, Sukajaya, and Indrawan 2020). This study examines SMS spam detection using TF-IDF techniques (Julis and Alagesan 2020), employing the SMS Spam Collection dataseta-dataset containing labeled text messages categorized as either spam or ham (non-spam). By utilizing TF-IDF, which assigns importance to words based on their frequency in a message and rarity across the dataset, this research aims to evaluate the effectiveness of TF-IDF in accurately distinguishing between spam and ham messages. Thro cision (Cahyani and Patasik 2021), recall (Lubis et al. 2021), and F1-score metrics (Pimpalkar and Raj 2020), this study aims to offer insights into the performance of the classification model in identifying spam messages while minimizing false positives and negatives. Additionally, an analysis of the distribution of spam and ham labels in the test data will shed light on the challenges involved in SMS spam detection (Teja Nallamothu and Shais Khan 2023). Ultimately, this research seeks to advance SMS spam detection techniques, thereby bolstering the security and reliability of SMS communication channels in the digital era.

This research aims to analyze the use of TF-IDF (Term Frequency-Inverse Document Frequency) method in detecting SMS spam, and to evaluate the performance of the method on SMS Spam Collection dataset. The benefits that can be obtained from this research include: providing insight into the effectiveness of the TF-IDF method in detecting SMS spam, providing references for further research in the development of a more accurate and efficient SMS spam detection system, helping cellular phone users in reducing the annoyance caused by SMS spam messages, and can be applied to various other applications or communication platforms that face spam problems, such as

e-ISSN 2774-5155
p-ISSN 2774-5147

email, social media, or instant messaging services. Overall, this research aims to explore and evaluate the use of TF-IDF method in SMS spam detection, so as to provide insight and reference for researchers and developers in improving the performance of spam detection systems.

RESEARCH METHODS

.

This research methodology involves a systematic approach to analyze SMS spam detection using the TF-IDF technique, using the Python programming language and the Google Colab environment. Additionally, this study uses publicly available datasets for experiments and analysis. The research design consists of several important steps, including data collection, preprocessing, TF-IDF vectorization, model training, model evaluation, and results analysis. Each step was carefully undertaken to ensure the robustness and reliability of the research findings.

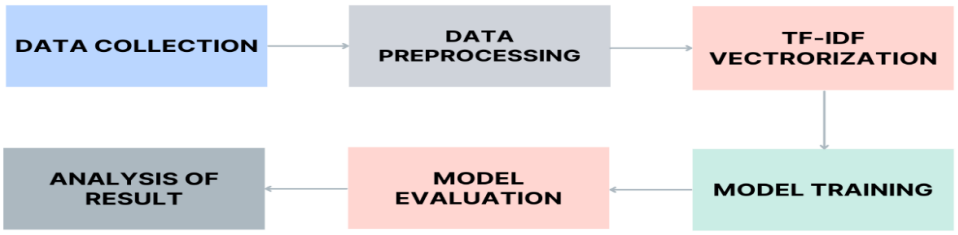

Figure 1 Flow Chart of methods

1. Data Collection:

Data collection is the process of gathering and collecting data for analysis. The choices made in data collection can affect the accuracy and reliability of the research findings. It is important to use transparent and best practices in data collection to ensure the validity and reliability of the data.

a. The research begins with the collection of the SMS Spam Collection dataset, a publicly available dataset containing labeled text messages categorized as either spam or ham (non-spam), which I obtained from Kaggle. It can be accessed through this link: https://www.kaggle.com/datasets/thedevastator/sms-spam-collection-a-more-diverse-dataset/data.
b. This dataset serves as the foundation for the analysis of SMS spam detection using TF-IDF techniques.
c. The dataset is downloaded and imported into the Google Colab environment for further processing and analysis.

2. Data Preprocessing

Preprocessing of the text data is a crucial step undertaken to ensure the data's consistency and suitability for analysis. This multifaceted process involves several sequential steps aimed at refining and structuring the raw textual information (Pimpalkar and Raj 2020). This involves several steps, including:

a. Removing any irrelevant characters or symbols from the text messages.
b. Converting the text to lowercase to standardize the text data.
c. Tokenizing the text to split it into individual words or tokens.
d. Removing stopwords (commonly occurring words) from the text.
e. Applying stemming or lemmatization to reduce words to their base or root form.

3. TF-IDF Vectorization:

TF-IDF, which stands for Term Frequency-Inverse Document Frequency, represents a refined iteration of the conventional bag-of-words method. It serves to gauge the relevance of a term within a document concerning a larger corpus (Junior, Wainer, and Calixto 2022). The term frequency aspect

e-ISSN 2774-5155
p-ISSN 2774-5147

measures how often a term appears in a specific document, while the inverse document frequency aspect assesses the rarity of the term across the entire corpus. Through multiplication of these two factors, TF-IDF accentuates the importance of terms that are abundant within a document but infrequent across the corpus, thus emphasizing their significance in characterizing the document's content. This technique finds wide application in diverse natural language processing tasks, including information retrieval, document classification, and sentiment analysis (Lubis et al. 2021).

a. TF-IDF (Term Frequency-Inverse Document Frequency) analysis is employed to transform the preprocessed text data into numerical features suitable for machine learning algorithms.
b. The TfidfVectorizer class from the scikit-learn library is utilized to create TF-IDF vector representations of the text data.
c. The TF-IDF vectors are generated separately for the training and testing datasets.

4. Model Training:
   a. A machine learning model is trained on the TF-IDF transformed data to classify messages as spam or ham.
   b. Random Forest Classifier is chosen as the classification algorithm due to its effectiveness in handling text data and its ability to capture non-linear relationships between features.
   c. The training data is used to fit the Random Forest Classifier model, utilizing the TF-IDF features as input and the corresponding labels (spam or ham) as output.
5. Model Evaluation:
   a. The trained model is evaluated using the testing dataset to assess its performance in classifying spam and ham messages.
   b. Evaluation metrics such as precision, recall, and F1-score are calculated to measure the model's accuracy, sensitivity, and overall performance.
   c. The classification report is generated to provide a comprehensive summary of the model's performance across different metrics.
6. Analysis of Results:
   a. The results obtained from the model evaluation are analyzed to gain insights into the effectiveness of TF-IDF in SMS spam detection.
   b. The distribution of spam and ham labels in the test data is examined to identify any challenges or trends in SMS spam detection.
   c. The implications of the findings are discussed, and recommendations for further research or improvements in SMS spam detection techniques are provided.

RESULTS AND DISCUSSION

This part of the study elaborates on the significant results achieved through using TF-IDF analysis to classify SMS as spam or ham. It assesses the method's efficiency and explores the broader implications of these outcomes for future research and real-world applications in spam detection.

Efficiency of TF-IDF in Identifying Spam & Model Accuracy Levels

Precision Measurement: The precision parameter, which assesses the accuracy of identifying true spam messages, was incredibly high (Chen et al. 2015). This demonstrates that the model effectively recognized most spam messages correctly while rarely mislabeling non-spam (ham) messages as spam. Recall and F1-Score Insights: Recall—indicating the model's ability to identify all actual spam messages—was outstanding. This, alongside a high F1-score, confirms that the model maintained a strong balance between recognizing spam accurately and minimizing the identification of false positives, which is crucial in operational settings.

The model demonstrated remarkable accuracy in differentiating between spam and non-spam messages. This success underscores the appropriateness of incorporating TF-IDF preprocessing to enhance the classifier's ability to discern subtle differences in text data, thereby improving detection accuracy.

e-ISSN 2774-5155
p-ISSN 2774-5147

Classification Report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	954
1	1.00	0.86	0.93	161
accuracy			0.98	1115

Figure 2 The result of precision, recall,f-1score and accuracy

Understanding of Label Distribution

A comprehensive review of how spam and ham labels are distributed throughout the test dataset, as illustrated in the Figure 3, offers crucial insights into the behavior of the classification model (Sykora, Elayan, and Jackson 2020). The chart distinctly shows a higher count of ham messages (954) compared to spam messages (161). This significant imbalance highlights the potential challenges in training the classification algorithm — particularly the risk of it being biased toward predicting messages as non-spam. Importantly, the analysis revealed that spam messages often contain unique keywords or phrases that are rarely found in ham messages. These distinct elements, although less frequent given the lower quantity of spam instances, are crucial for refining detection methods. Understanding these characteristics supports the development of more sophisticated models that can effectively distinguish between the two categories, despite the disproportionate data representation.

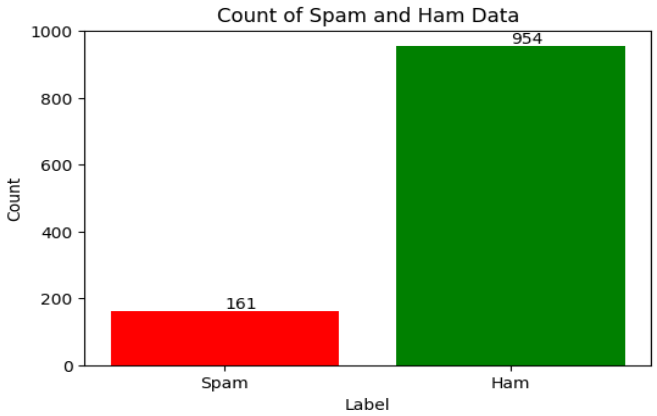

Figure 3 The Chart of Spam & Ham

Identified Challenges and Future Directions

Despite the generally positive results, some challenges like the occasional misclassification of context-sensitive messages were noted (Nahari et al. 2019). These issues highlight areas for potential enhancement. Future studies might look into integrating semantic analysis with TF-IDF to deepen the classifier's contextual understanding. Employing advanced neural networks such as CNNs or LSTMs could also be considered to further enhance the interpretative depth of the model.

CONCLUSION

The study investigated SMS spam detection utilizing TF-IDF analysis on a dataset comprising labeled text messages as spam or ham. It aimed to evaluate TF-IDF's effectiveness in distinguishing between spam and non-spam messages, achieving promising outcomes with high accuracy rates. The analysis demonstrated strong precision, recall, and F1-scores for both spam and ham classes. Specifically, the model achieved a precision of 98% for ham and 100% for spam, with a recall of 100% for ham and 86% for spam. The F1-score, which balances precision and recall, indicated strong performance for both classes, with an overall accuracy of 98%. Figure 2 illustrated these high precision and recall values. Furthermore, Figure 3 highlighted the distribution of spam and ham labels, revealing a significant class imbalance, with 954 ham messages compared to 161 spam messages. The review emphasized the importance of recognizing unique keywords in spam messages

e-ISSN 2774-5155
p-ISSN 2774-5147

and addressed the challenges posed by data disproportionality. Overall, the findings contribute to enhancing SMS spam detection techniques, paving the way for the development of more robust classification models in the future.

BIBLIOGRAPHY

Abid, Muhammad Adeel, Saleem Ullah, Muhammad Abubakar Siddique, Muhammad Faheem Mushtaq, Wajdi Aljedaani, and Furqan Rustam. 2022. "Spam SMS Filtering Based on Text Features and Supervised Machine Learning Techniques." *Multimedia Tools and Applications* 81(28):39853–71.

Artama, M., I. N. Sukajaya, and G. Indrawan. 2020. "Classification of Official Letters Using TF-IDF Method." P. 12001 in *Journal of Physics: Conference Series*. Vol. 1516. IOP Publishing.

Cahyani, Denis Eka, and Irene Patasik. 2021. "Performance Comparison of Tf-Idf and Word2vec Models for Emotion Text Classification." *Bulletin of Electrical Engineering and Informatics* 10(5):2780–88.

Chen, Chao, Jun Zhang, Xiao Chen, Yang Xiang, and Wanlei Zhou. 2015. "6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection." Pp. 7065–70 in *2015 IEEE international conference on communications (ICC)*. IEEE.

Hosseinpour, Shaghayegh, and Hadi Shakibian. 2023. "An Ensemble Learning Approach for Sms Spam Detection." Pp. 125–28 in *2023 9th International Conference on Web Research (ICWR)*. IEEE.

Julis, M. Rubin, and S. Alagesan. 2020. "Spam Detection in SMS Using Machine Learning through Textmining." *International Journal Of Scientific & Technology Research* 9(02).

Junior, Antonio P. Castro, Gabriel A. Wainer, and Wesley P. Calixto. 2022. "Weighting Construction by Bag-of-Words with Similarity-Learning and Supervised Training for Classification Models in Court Text Documents." *Applied Soft Computing* 124:108987.

Lubis, A. Ridho, Mahyuddin K. M. Nasution, O. Salim Sitompul, and E. Muisa Zamzami. 2021. "The Effect of the TF-IDF Algorithm in Times Series in Forecasting Word on Social Media." *Indones. J. Electr. Eng. Comput. Sci* 22(2):976.

Nahari, Galit, Tzachi Ashkenazi, Ronald P. Fisher, Pär-Anders Granhag, Irit Hershkowitz, Jaume Masip, Ewout H. Meijer, Zvi Nisin, Nadav Sarid, and Paul J. Taylor. 2019. "'Language of Lies': Urgent Issues and Prospects in Verbal Lie Detection Research." *Legal and Criminological Psychology* 24(1):1–23.

Pimpalkar, AMIT PURUSHOTTAM, and R. Jeberson Retna Raj. 2020. "Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features." *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* 9(2):49.

Sykora, Martin, Suzanne Elayan, and Thomas W. Jackson. 2020. "A Qualitative Analysis of Sarcasm, Irony and Related# Hashtags on Twitter." *Big Data & Society* 7(2):2053951720972735.

Teja Nallamothu, Phani, and Mohd Shais Khan. 2023. "Machine Learning for SPAM Detection." *Asian Journal of Advances in Research* 6(1):167–79.