Using Decision Tree Algorithms in Detecting Spam Emails Written in Malay: A Comparison Study

Saifuldeen H Abdulrahman [1*], and *Mohammad* Salim [2]

[1] Computer Science Department. College of Science, Knowledge University, 44001 Erbil, Kurdistan Region
[2] IT Department, Faculty of Science, Tishk International University, Erbil, Iraq

Abstract. Emails have become the most economical and fastest communication forms. However, during the past few years, the increment of email users has dramatically increased spam emails. Various anti-spam techniques have been developed to minimize if not eliminate the spam problem. In this paper, we study the disparity in the effectiveness of using different decision tree algorithms in email classification and combat spam problems. For that, we have chosen Universiti Utara Malaysia emails as a case study. To achieve the best possible classification accuracy, we compared all chosen algorithms' performance, which are Random Forest, LMT, Decision Stump, J48, Random Tree, and REP Tree. The experimental results showed that the Decision Stump algorithm is more effective to be used in classifying the emails, and the F-measures, Precision, and recall score for the Decision Stump algorithm are higher than the other comparison algorithms.

1 Introduction

Email is an electronic messaging system that sends messages through networks, allowing users to connect with one another at a low cost while also offering a reliable mail delivery system. Because of its dependability, user-friendliness, and vast range of free email, it is the most popular and preferred communication tool [1]. Email is a repository for messages, and the compose, send, receive, and store methods rely on electronic communication systems. Email management is a significant and growing issue for organizations and individuals because it is prone to mismanagement. Unfortunately, some issues with this service increase dramatically proportional to the development and spread of email services and the increased reliance on email by users [2].

The combination of low-cost, high-bandwidth Internet connections, falling storage costs per megabyte, and increased email users has resulted in an explosion of email data per user. Sorting through all this data and determining what is valuable and what is not is a daunting task. Spam is one of the most destructive problems facing email today. Spam, bulk email, or junk email are all terms that refer to inappropriate or irrelevant messages sent to many recipients via the Internet. Spam also refers to email that was not requested by the user [2, 3]. Typically, spam contains advertisements for dubious products and services such as "Make

* Saifuldeen H Abdulrahman : saifuldeen.abdulrahman@knu.edu.iq

Money Fast" schemes, multilevel marketing, illegally pirated software, and foreign bank scams. Spam may also contain offers to sell real estate, medicine, loans, and investments. Spam, or unsolicited email, is widely regarded as a serious threat to the Internet, as it floods users' inboxes and costs businesses billions of dollars in wasted bandwidth. Spam's global productivity cost increased by $2 billion to $132 billion in 2010 [4].

Additionally, spam causes a number of negative consequences, including an overflow of email storage capacity, which disables the server's ability to receive new emails, as well as a slow response time from the server and insufficient system resources. Additionally, email spam wastes the user's time in containing and deleting unwanted emails, results in the loss and/or delay of critical or emergency email messages and degrades Internet bandwidth and performance. Additionally, spam is one of the most effective methods of spreading malicious programs, worms, and Trojan Horses that corrupt computers and operating systems. Additionally, spam contributes to an increase in the proportion of people exposed to fraud, resulting in the loss of millions of dollars worldwide each year [4].

These and other negative consequences motivate researchers and businesses to work toward the elimination of spam and the abolition of harmful effects. Spam is available in a wide variety of languages, including Malay, Arabic, Korean, Chinese, and other Asian dialects, but is most frequently written in English. As a result, manually classifying too many emails is difficult, and machine learning techniques must be introduced. The machine learning technique entails developing filters to examine and eliminate characteristics from a corpus of spam [5]. Spam detection is not a typical text categorization task since it has some intriguing characteristics. Both spam and legitimate messages can cover a wide variety of topics and genres. In other words, neither class is homogenous. Additionally, email messages range in length from a few text lines to dozens of text lines. Moreover, the message may contain grammatical errors and unusual acronyms (sometimes intentionally used by spammers to fool anti-spam filters). As a result, the learning model should be robust under such circumstances.

This study aims to implement a variety of decision tree algorithm techniques for email classification and determine the most effective algorithms. Due to the growing number of documents, the variety and size of the dataset, and the variety of languages used to write an email. Email classification poses significant challenges due to the difficulty distinguishing spam from legitimate emails [6, 7]. The study's primary objective is to identify the most effective decision tree algorithms for spam email classification. We used the Waikato Environment for Knowledge Learning (WEKA) data mining tool in this study to implement decision tree algorithm techniques on a set of emails obtained from the University Utara Malaysia's Computer Centre. The dataset is collected at various time to ensure that the selected sample contains the required diversity. The remainder of this article will be structured as follows: The following section discusses the context for this study and related works. It is followed by a section describing the methods used to accomplish the study's objectives. Following that, a section is devoted to the study's findings, and finally, there is a conclusion section.

2 Background and Related Works

This section presents the background and related work of the study include spam filtering technique, related work on spam filtering technique, and classifications algorithms related work.

2.1 Spam Filtering Techniques

The growing number of email users has irritated the public over the rise in spam. Email spam is inconvenient for users and costs email server owners and Internet service providers a lot of money to classify. There are numerous spam filtering techniques that fall into several categories. This study will summarize it.

(i) User-defined filters: In this technique, filters automatically determine and remove spam messages based on user-defined rules. These include establishing guidelines for the acceptable subject matter and sources. For example, the user can configure his filter to reject all emails that contain a particular word in the header or emails from senders [8, 9].

(ii) Header filters: This technique relies on inspecting the headers of incoming emails for forgery indicators. Numerous spammers fabricate these headers in order to conceal their identities or locations. Normally, a header contains a wealth of information, including the recipient, subject fields, and sender; it also contains information about the servers that delivered this email, referred to as the relay chain. A good header filter can detect the forged header. Not all spammers, however, forge this information [8, 9].

(iii) Language filters: This technique handles any email that is not written in the language selected by the user. Whether or not the user uses email to communicate with anyone in a foreign language, this technique is beneficial for non-native speakers [8, 10].

(iv) Content filters: This technique is based on a set of rules that analyse the text of emails to determine whether they are spam or not. Unfortunately, this technique has some drawbacks, as these filters may exclude useful emails such as newsletters and other emails that the user specifically requests [5, 8].

(v) Blocklists and allow lists: This technique enables the user to create a blocklist of email addresses from specific users or websites and an allowed list of email addresses that are acceptable. The filter will automatically accept emails from any source on the allow list and will reject emails from any source on the block list. It takes a long time to develop this type of filter. Additionally, it is relatively easy to fool a user by using variations of well-known phrases that are still readable to the user (e.g., p*a*y) or by using bogus email addresses. As a result, new rules must be added on a continuous basis to maintain the filter's efficiency [8, 10].

(vi) Community blacklists: This technique is based on user interaction with spam filtering techniques such as black lists and whitelists. The user of this technology will automatically become a part of the millions of other users fighting spam. Numerous anti-spam techniques make use of this fact to effectively eliminate spam. If several users flag a message as spam, the technique flags it as spam and blocks it from being seen by other users [8, 10].

(vii) Bayesian analysis: Bayesian analysis is a machine learning technique that enables users' systems to learn what constitutes spam and analyses incoming messages using complex algorithms to determine whether they are spam or legitimate. This technique determines the probability that an email is spam based on previously recognized spam versus email considered acceptable. As one of a large number of users subscribed to an anti-spam technique that makes use of Bayesian filtering, the user benefits from community experience and an exact analysis system [11, 12].

2.2 Related Works

Unwanted email messages were first identified as a problem in a 1975 Internet Request for Comments [13]. Once spam or unwanted email becomes prevalent, and the emergence of this problem and the growing circle of spam-affected users. Uncontrolled and manual filtering, on the other hand, is impossible. Numerous businesses and institutions have made sustained efforts to eradicate this phenomenon; numerous techniques have been proposed, and numerous studies have been conducted on email spam filtering techniques [13]. Additionally, these organizations sought to ascertain the economic impact and disadvantages of the email service on its recipients. These efforts resulted in the discovery of numerous studies and numerous strategies, which have been implemented on a variety of datasets to achieve the desired result [3, 6].

In [14], the researcher investigated fourteen classification methods and conducted a cross-experiment to compare them, including Naive Bayesian, linear squares fit, decision tree, neural network, and Rocchio. The researcher discovered that KNN (k-nearest neighbours) is one of the best performers when scaling up to noisy and very large classification problems. A model proposed in [14] for classifying personal emails using Neural Networks and PCA (Principal Component Analysis) as a pre-processor to reduce the data's size and dimensionality.

The researchers in [15] used the minimum distance-to-mean (MDM), maximum likelihood classification (MLC), linear discrimination analysis (LDA), and neural network-based multi-layered perception (MLP) classifiers, as well as the popular C4.5 decision tree. The purpose of this study is to compare the learning ability, generalization ability, and speed of these classifiers. The results of this study demonstrated that MLP is a consistently superior classifier due to its generalization ability. In comparison, the widely used MLC classifier is not noticeably faster. On the one hand, the MDM classifier has a significant speed advantage; on the other hand, [16] compared the prediction accuracy, complexity, and training time of thirty-three classic and novel classification algorithms.

As a conclusion to this section, prior research and related work demonstrate that: there are numerous email spam filtering techniques, each with its own set of advantages and disadvantages. These techniques vary in their approach to remove or reduce spam. Some are dependent on the user, such as the community blacklists technique [8], while others operate independently of the end user, such as the Bayesian analysis technique [11]. Multiple studies indicate that using decision tree algorithms to classify email text is quite effective and efficient at detecting spam. Therefore, we set out to find the best decision tree algorithm for spam email classification.

3 Methodology

This section details the combination of methods used to accomplish the research's stated objective. The proposed methodology consists of six steps: data collection, data pre-processing, feature selection using document frequency, feature weighting using TF-IDF, J48 technique implementation with stratified cross-validation, evaluation of subject vs. body corpora, and comparison to another classification method. These steps are depicted in Fig.3, and the remainder of this section details each one.

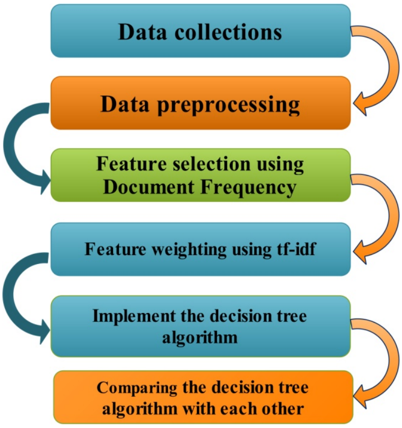
Data collections
Data preprocessing
Feature selection using Document Frequency
Feature weighting using tf-idf
Implement the decision tree algorithm
Comparing the decision tree algorithm with each other

Fig. 1: Methodology Stages

3.1 Data Collections

As previously stated, the required data set was collected over a one-week period from the University Utara Malaysia's Computer Centre to achieve the required diversity in the sample of selected emails. The dataset contains approximately 100 emails, 30 of which are spam and 70 of which are legitimate.

3.2 Data Pre-processing

For each corpus, a bag-of-words representation was used (Spam and Legitimate email). The bag-of-words model is a "simplifying representation used in natural language processing and information retrieval (IR)" [17]. It represents a text (such as a sentence or a document) as a bag (multiset) of its words, disregarding grammar and even word order but retaining multiplicity.

Following that, the data is processed by extracting individual words from the corpus. This step generates a large corpus of individual words divided into two corpora (the first contains 9841 words extracted from spam email, while the second contains 11438 words extracted from legitimate email).

We then eliminate stop-words from the corpus. After removing Malay and English stop words from each corpus, the number of features decreased significantly. We extracted 2316 features from spam emails and 2824 from the body of legitimate emails. Fig. 2 illustrates the dramatic increase in the number of features prior to and following the removal of the stop-words process. The remaining features, however, are still too large to be used in the classification technique. As a result, we must reduce the number of features using a technique described in the following section.

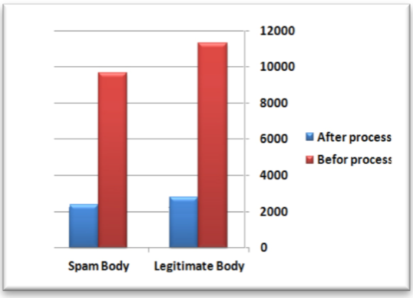

Fig. 2: Features number before and after remove stop-words process

3.3 Feature Selection

Feature selection is often an essential data processing step before applying a learning algorithm. The removal of irrelevant and redundant information often improves the performance of machine learning algorithms [5, 11, 18]. Furthermore, feature selection is an important issue in building classification models, and feature selection is advantageous in that we are thus able to limit the number of input features in a classifier to have a good predictive and less computationally in-tensive model [19].

For feature selection, we used Document Frequency (DF) method, which means for each corpus separately:

- We computed the DF score for each word, document frequency (DF) is defined as "the number of times that term occurs in the document" [20]).
- We selected the top 100 words from each corpus (the ones with the highest DF score) to balance the two corpora in feature selection. We merged the features of two corpora (spam and legitimate) after adding a classification class, which are 1 for spam and 0 for legitimate features.

3.4 Feature Weighting Using TF-IDF

Term frequency-inverse document frequency (TF-IDF), "is a numerical statistic that reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document" [21]. Essentially, TF-IDF determines words' relative frequency in one document compared with the inverse proportion of the same word over the corpus of the entire document. This calculation can determine how a given word is relevant in a particular document.

3.5 Implement Decision Tree Algorithms with Stratified Cross-Validation

To evaluate the performance of the decision tree algorithms in email classification. We use cross validation, or sometimes called rotation estimation, it is an elaborate method, "if we suppose that a number of folds n is specified. The dataset is randomly reordered and then split into n folds of equal size. In each iteration, one-fold is used for testing and the other n-1 folds are used for training the classifier".

The test results are collected and averaged over all folds. This gives the cross-validation estimate of the accuracy. The folds can be purely random or slightly modified to create the same class distributions in each fold as in the complete dataset. In the latter case, the cross-validation is called stratified. For each 10-fold cross validation run, the program prints the accuracy on the test set, and at the end the average accuracy over the 10 runs. In this work, we used 10-fold cross validation to compare the performance of the classifiers on the two corpora (body and subject).

3.6 Comparing the Decision Tree Algorithms with Each other

To discover which is the best decision tree algorithms for the classification of emails between legitimate and spam, we compare the accuracy results of these algorithms with each other. In other words, we applied these techniques to UUM emails corpora. After that, we compare them by using the weighted average for precision, recall, and F-Measure scores. To achieve the goal of this study, which determines the best decision tree algorithms with the best outcome in emails classification.

4 Result and Conclusion

This section presents the comparative result of the 6 decision tree algorithms' performance: Random Forest, LMT, Decision Stump, J48, Random Tree, and REP Tree. The algorithms are compared in terms of their precision, recall, and F-measures.

4.1 Evaluation of Decision Tree Algorithms

After implementing the decision tree algorithms for the email's features corpora, we evaluated the results to choose the best one between them. The evaluation is based on three weighted average scores: precision, recall, and F-measure. Precision "is the proportion of retrieved items that are relevant, measured by the ratio of the number of relevant retrieved items to the total number of retrieved items" [17]. The recall is " the proportion of relevant items retrieved, measured by the ratio of the number of relevant retrieved items to the total number of relevant items in the collection."

F-Measure is "2 * Precision * Recall/ (Precision + Recall), a combined measure of precision and recall" [17], and the third one is recall rate. Table 1 shows the average precision, recall, and F-measure score for each decision tree algorithm. Moreover, Fig. 3 shows the dramatic difference in the weighted average between these decision tree algorithms.

Table 1. The average precision, recall, and F-measure the score for each decision tree algorithm.

	Precision	Recall	F-Measure
Random Forest	58%	41.1%	42%
LMT	11.5%	68.5%	68.5%
Decision Stump	89%	89%	89%
J48	86.5%	86.5%	86.5%
Random Tree	23%	42.5%	29.8%
REP Tree	25%	50%	33.3%

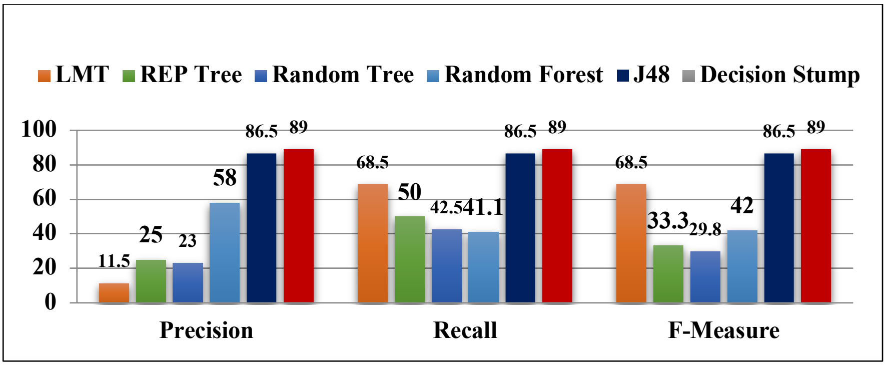

Fig. 3: The Different Accuracy Ratios Between Classifiers When They Applied on Emails Corpora.

As can be seen in figure 3, for the email's corpora, the precision, recall, and F-measure the score for the Decision Stump Algorithm is higher than the other five com-pared algorithms that are 89% for all measures. J48 comes after with 86.5% for all measures. The precision, recall, and F-measure scores for the Random Forest algorithm are 58%, 41.1 %, and 42% respectively. The precision scores for the LMT algorithm are 11.5 % and 68.5% for both recall and F-measure scores. For the REP Tree algorithm, the precision, recall, and F-measure scores are 25 %, 50%, and 33.3% respectively. And finally, the precision, recall, and F-measure scores for Random Tree are 23%, 42.5%, and 29.8% respectively.

5 Conclusion

In conclusion, after examining the performance of 6 decision tree algorithms in the classification of emails and compared its performance by implementing the algorithm on dataset corpora, we have found that the Decision Stump algorithm is the best spam emails classification algorithm for the University Utara Malaysia email service. In addition, the method of applying document frequency on the email features has reduced the number of features on the email dataset without affecting the classification accuracy. The calculation of the tf-IDF weighting technique on the reduced features proved to produce better features to be used in the classification process. This study also discovered that we had proposed the best algorithm to deal with the spam problem in email services. As for recommendations of this study, the proposed methods can also be implemented on the other email system in any part of the world due to the inclusion of language stop word removal in the pre-processing of the email data. Future work in this area includes investigating other pre-processing, feature selection, and feature weighting sections that can better represent email data to increase the accuracy of classification.

6 Acknowledgments

We want to thank the Computer Centre in University Utara Malaysia for their support in providing the data for the study.

References

[1] T. Verma, N. S. J. I. J. o. I. T. Gill, and E. Engineering, "Email Spams via Text Mining using Machine Learning Techniques," **9**, no. 4, pp. 2535-2539, (2020).

[2] N. Saidani, K. Adi, M. S. J. C. Allili, and Security, "A semantic-based classification approach for an enhanced spam detection," **94**, p. 101716, (2020).

[3] H. Taylor, "Making Mass-Spamming Illegal Rises," Harris Interactive (2011).

[4] P. Heymann*, et al.*, "Fighting spam on social web sites: A survey of approaches and future challenges," *Internet Computing, IEEE,* **11**, pp. 36-45, (2007).

[5] Clearbridge. *What is the global cost of spam?* Available: http://www.mailshine.com/2011/06/whats-the-globalcost-of-spam/(2011).

[6] M. Fossi*, et al.*, "Symantec global internet security threat report," *White Paper, Symantec Enterprise Security,* **1**, (2013).

[7] X. Guo and Z. Xia, "Fighting spam," *University of California Berkeley,* (2012).

[8] S. Youn and D. McLeod, "A comparative study for email classification," in *Advances and Innovations in Systems, Computing Sciences and Software Engineering*, ed: Springer, pp. 387-391 (2007).

[9] I. Firdausi*, et al.*, "Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection," in *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*, pp. 201-203, (2010).

[10] S. T. Maller, "Email filtering methods and systems," ed: Google Patents, (2006).

[11] D. Cook*, et al.*, "Catching spam before it arrives: domain specific dynamic blacklists," in *Proceedings of the 2006 Australasian workshops on Grid computing and e-research-* **54**, pp. 193-202**,** (2006).

[12] P. Warkhede*, et al.*, "Fast packet classification for two-dimensional conflict-free filters," in *INFOCOM. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, 2001, pp. 1434-1443 (2001).

[13] P. O. Boykin and V. Roychowdhury, "Personal email networks: An effective anti-spam tool," *arXiv preprint cond-mat/0402143,* (2004).

[14] A. W. Moore and D. Zuev, "Internet text classification using bayesian analysis techniques," in *ACM SIGMETRICS Performance Evaluation Review*, pp. 50-60, (2010).

[15] N. J. Kawale and S. Y. Sait, "A Review on Various Techniques for Spam Detection," in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 1771-1775: IEEE, (2021).

[16] O. El Kouari, H. Benaboud, and S. Lazaar, "Using machine learning to deal with
Phishing and Spam Detection: An overview," in Proceedings of the 3rd International
Conference on Networking, Information Systems & Security, pp. 1-7, (2020).

[17] M. Sahami*, et al.*, "A Bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 2008 workshop*, pp. 98-105, (2008).

[18] B. Cui*, et al.*, "On effective e-mail classification via neural networks," in *Database and Expert Systems Applications*, pp. 85-94, (2005).

[19] G. Leroy and T. C. Rindflesch, "Using symbolic knowledge in the UMLS to disambiguate words in small datasets with a naive Bayes classifier," *Medinfo,* **11**, pp. 381-385, (2004).

[20] E. Byvatov*, et al.*, "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification," *Journal of Chemical Information and Computer Sciences,* **43**, pp. 1882-1889, (2009).

[21] A. Lorenz*, et al.*, "Comparison of different neuro-fuzzy classification systems for the detection of prostate cancer in ultrasonic images," in *Ultrasonics Symposium, 2005. Proceedings., 2005 IEEE*, , pp. 1201-1204 (2005).

[22] N. Widiastuti, "Convolution neural network for text mining and natural language processing," in *IOP Conference Series: Materials Science and Engineering*, **662**, no. 5, p. 052010: IOP Publishing, (2019).