ANALYSIS OF TF-IDF TEXT MINING BASED MACHINE LEARNING METHODS FOR TEXTUAL SPAM

İrfan KILIÇ*[1], Aytaç KAŞOĞLU[2], Orhan YAMAN[3]
[1]Fırat University Rectorate, Information Technologies Department, Elazığ/Turkey, irfankilic@firat.edu.tr
[2] Fırat University, Digital Forensic Engineering Department, Elazığ/Turkey, 190509026@firat.edu.tr
[3] Fırat University, Digital Forensic Engineering Department, Elazığ/Turkey, orhanyaman@firat.edu.tr
[*] Corresponding Author: irfankilic@firat.edu.tr

Abstract: *The most popular messaging applications that form the basis of Internet technology are E-mail and SMS applications. The biggest problem with these applications is spam. When comparing audio, video and text spam, text spam is known to be common. It is very important to detect textual spam with high accuracy and to know which category the spam content belongs to. This study focuses on the development of TF-IDF (Term Frequency-Inverse Document Frequency) based machine learning methods in the field of textual spam filtering. The study discussed the prominence of TF-IDF as a technique that plays an important role in the field of text mining. This technique uses the ratio between the frequency of the term in the document and its frequency in all documents to determine the importance of a term in a document. Within the scope of the study, data sets consisting of SMS and E-mail texts were collected and the effect of using Support Vector Machines (SVM), Decision Trees, Random Forests, Naïve Bayes, k-NN and TF-IDF together on spam filtering performance was examined. It has been demonstrated that the developed TF-IDF-based machine learning methods have the potential to obtain more effective results in spam filtering systems.*

Keywords*: Spam Filtering, Text Mining, Machine Learning, TF-IDF, Support Vector Machine*

1. Introduction

It lies in the imperative demand for ingenious identification and blocking measures against undue demands in text-based communication spaces. Annoying attacks in the form of spam comments and messages not only disrupt smooth interactions but also pose potential security hazards to end users. Therefore, formulating precise and proficient methodologies to screen and eliminate spam-laden content becomes crucial to preserve both user experience and data integrity. Worldwide spam volume between 2011 and 2023 is shown in Fig. 1 [1].

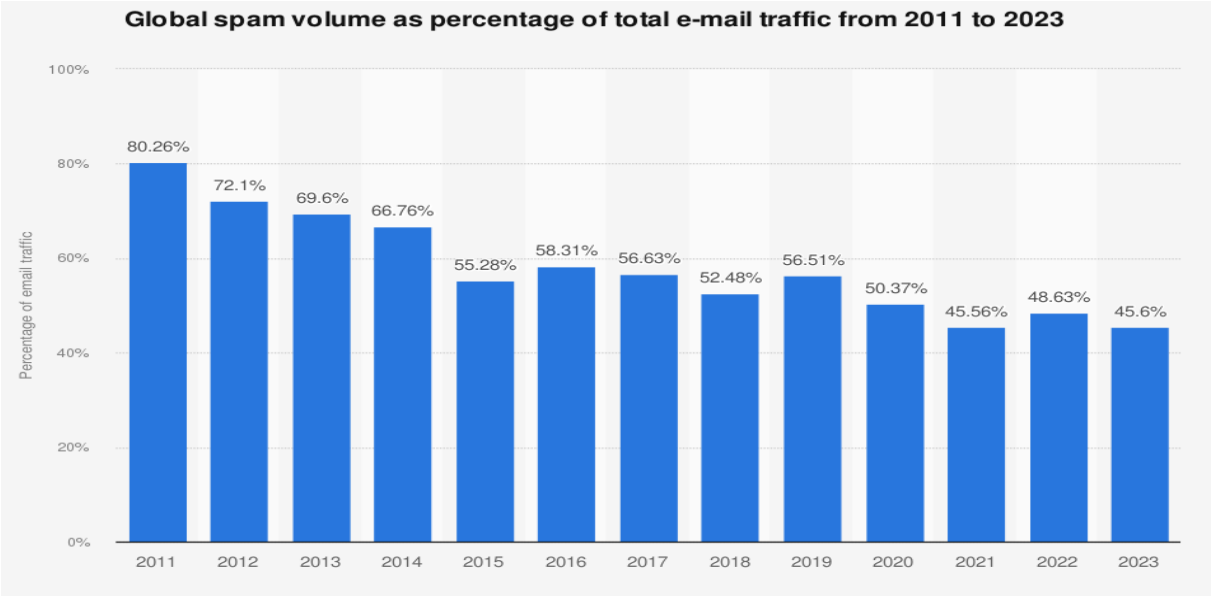

Fig. 11. *Total spam traffic worldwide (2011-2023)*

The study tries to examine the use of machine learning algorithms and TF-IDF (Term Frequency - Inverse Document Frequency) in the field of text-based spam filtering applications. Common machine learning algorithms (SVM, k-NN, Random Forest, Decision Tree, Naive Bayes) have been used synergistically for classification attempts and TF-IDF, an analytical framework for detecting term importance in text mining.

1.1. Literature

Text-based feature extraction methods are: [2];

- Frequency Based Word Representation Methods:

1. One-Hot Encoding
2. N-gram
3. Counting Vectorizer (Bag of Words)
4. TF-IDF

- Predictive Word Representation Methods:

1. Word2Vec
2. FastText

- Pre-Trained Word Representation Methods:

1. GloVe
2. Bert
3. GPT

There are many methods based on text mining and machine learning in the literature [2]–[13]. Some of the textual spam detection-based methods used in the literature are listed in Table I.

Tablo I. *Some methods used in the literature*

Reference	**Methods Used**	**Explanation**
Feng vd. (2020) [6]	TF-IDF	Keyword extraction and text mining
Wang vd. (2021) [13]	TF-IDF	Keyword extraction from news, social media and medical texts
Alan Chavez [5]	Multinomial Naïve Bayes, TF-IDF	Email spam filtering
Mohammed & Omar (2020) [11]	TF-IDF, Word2Vec	Word comparison of TF-IDF and Word2Vec

Joachims (1998) [9]	SVM	Text classification and spam filtering
Alias vd. (2019) [3]	SVM	Video spam comment filtering
Jáñez-Martino vd. (2020) [10]	TF-IDF, SVM	Spam email classification
Haqimi vd. (2019) [7]	TF-IDF, SVM	Instagram spam comment detection
Jehad vd. (2023) [8]	Naïve Bayes	Spam detect
Pham & Le-Hong (2016) [12]	Content Based Approach	Spam SMS filtering

As seen in Table 1, there are TF-IDF and machine learning based studies in the literature.

1.2. Motivation and Purpose of the Study

Fighting spam has been a huge issue in today's digital world, and it is important to effectively filter, prevent and detect such spam messages.

One of the main reasons for choosing TF-IDF, one of the frequency-based word representation algorithms of the study, is that it can effectively solve document classification, word suggestion and similar text mining problems thanks to its ability to highlight important terms. It is also encouraging that it is language independent, increasing its usability on text data in different languages. Although it uses a simple and effective calculation method, it can handle complexities such as measuring and ranking similarities between documents. For this reason, the TF-IDF algorithm is widely used in text mining and information retrieval applications by measuring the importance of terms in documents.

This method was chosen because it was thought that the TF-IDF algorithm would help highlight important terms in documents by mining text. It is thought that combining these two methods will help TF-IDF, with the classification capabilities of machine learning algorithms, achieve successful results by extracting the meaning of text-based data and identifying the highlights.

The overall goal is to develop a flexible strategy to distinguish between documents tainted by spam and documents that are free of such intrusions.

By combining machine learning algorithms and TF-IDF, the study attempts to create a reliable framework to screen out spam-infested content and measure its strength through practical verification. The study aims to increase the applicability and usability of machine learning methods in the field of text-based spam filtering. With this study;

- To demonstrate methods that can be used in the field of text-based spam detection and prevention.,

- Achieving at least 90% accuracy with the help of TF-IDF and machine learning algorithms,
- It is intended to detect types of spam as well as spam and non-spam.

2. Method

TF-IDF (Term Frequency-Inverse Document Frequency) is a term weighting method used in text mining. This metric is used to determine the importance of a particular term in a document and to calculate the prevalence of that term in the overall collection.

2.1. Term Frequency (TF)

TF measures the frequency of a particular term in a document. That is, the more frequently a term is used, the higher its TF value will be. For example, TF for term 't' in a document is calculated as in Equation 1:

$$\boldsymbol{TF\ (t,d)} = \frac{\boldsymbol{Number\ of\ term}\ \text{"t" in document}}{\boldsymbol{Totam\ number\ in\ document}} \quad (1)$$

This determines the overall frequency of the term in the document.

2.2. Inverse Document Frequency (IDF)

IDF is used to determine the overall importance of a term. If a term is rarely used in general, its IDF value will be high. This shows how unique the term is within the document collection. IDF is calculated as in Equation 2.

$$\boldsymbol{IDF\ (t,D)} = \boldsymbol{log}\left(\frac{\boldsymbol{Total\ number\ of\ documents}}{\boldsymbol{Number\ of\ documents\ containing\ term\ "t"+1}}\right) \quad (2)$$

This equation measures the overall importance of a term within the document collection.

2.3. TF-IDF Score

The TF-IDF score, obtained by multiplying the TF and IDF values, balances the frequency of a term in a document but with its overall importance across all documents. This score is calculated as in Equation 3.

$$\mathrm{TF-IDF(t,d,D)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t,D)} \quad (3)$$

While this score determines the importance of a term in a particular document, it also takes into account the importance of that term within the overall collection.As a result, TF-IDF is a powerful method that is frequently used in text mining to determine term importance between documents and classify documents. This metric helps us better understand the meaning in text data by focusing on the frequency of terms and their importance within the overall collection. Therefore, TF-IDF is considered an important tool in the field of text mining and has been successfully used in various application areas. Fig. 2 shows the implementation architecture of the method proposed in our study.

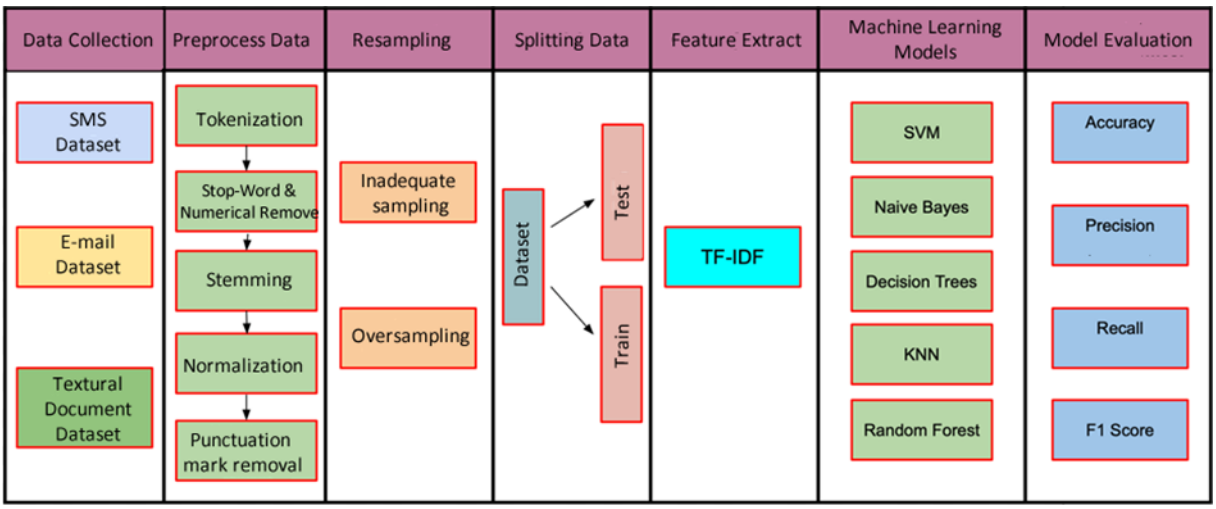

Fig. 12. *Implementation architecture of the proposed method*

2.4. Data Collection

The data of the study consists of SMS, E-mail and textual documents. Methods and resources such as Kaggle, Google Scholar, and UCI Machine Learning Repository were used to find datasets. The majority of the collected data set consists of SMS data sets. Fig. 3 shows sample SMS data collected.

v1	v2
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your frie
spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE
ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 in
spam	URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LC
ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You ha
ham	I HAVE A DATE ON SUNDAY WITH WILL!!
spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJ
ham	Oh k...i'm watching here:)
ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham	Fine if thatåÕs the way u feel. thatåÕs the way its gota b
spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ì¼1.2
ham	Is that seriously how you spell his name?
ham	I‰Û÷m going to try for 2 months ha ha only joking
ham	So Ì_ pay first lar... Then when is da stock comin...

Fig. 13. *Sample SMS datas*

2.5. Preprocessing Datas

The data was preprocessed using the following steps;

- Tokenization has been done to divide text data into smaller pieces.
- Stop words have been removed.
- Words have been stemmed or lemmatized.

- Data has been converted to lowercase notation.
- After the normalization process, any punctuation marks are removed.

2.6. Resampling

It is the process of changing the number of observations in the data set. There are two basic data resampling techniques.

a. Oversampling: It is used to balance the number of observations between classes by multiplying classes with a small number of samples. This is often done by replicating observations from the minority class or by generating synthetic examples.

b. Reduced Sampling (Undersampling): It is mostly used to balance the number of observations between classes by reducing the classes that have a large number of samples. This is done by randomly selecting observations from the majority class or removing some samples.

2.7. Splitting Datasets

Datasets will be divided into training and testing sets, and 80% of the data will be used for training and 20% for testing.

2.8. Feature Extraction

The TF-IDF score will be used to extract features from text data.

2.9. Model Determination

The model was trained using machine learning algorithms and TF-IDF features. The machine learning algorithms used are given below.

a. SVM (Support Vector Machines): It is a learning algorithm used for classification and regression tasks. It tries to best separate data points with a hyperplane between two classes.

b. Decision Trees: Decision trees are an algorithm that creates a model in a tree structure used to classify or regression a data set. Each internal node makes a decision by splitting the data with a feature.

c. Random Forest: It is an ensemble learning algorithm that creates a more powerful and generalizable model by combining many decision trees.

d. kNN (k-Nearest Neighbors): k-NN is an algorithm that determines the class or value of a particular point based on the labels of its k nearest neighbors.

e. Naive Bayes: Naive Bayes is a probability-based classification algorithm that calculates the probability of a class based on the assumption of independence between features.

3. Results and Discussion

An application has been developed with Python language to classify with SVM, kNN, Naïve Bayes, Decision Tree and Random Forest algorithms, which are Machine Learning algorithms. Using the TF-IDF scores obtained with this application, the data is marked as Spam or Clean (Not Spam) by classifying it with each algorithm. Performance results are given in Table II.

Table II. *TF-IDF based spam classification performance results*

Models / **Class**	**Naïve Bayes**	**Support Vector Machine (SVM)**	**k-NN**	**Decision Tree**	**Random Forest**
Spam	Accuracy: 95.70% Precision: **100.00%** Recall: 67.79% F1-score: 80.80%	Accuracy: **98.30%** Precision: 98.51% Recall: **88.59%** F1-score: **93.29%**	Accuracy: 91.66% Precision: **100.00%** Recall: 37.58% F1-score: 54.63%	Accuracy: 96.05% Precision: 85.23% Recall: 85.23% F1-score: 85.23%	Accuracy: 97.67% Precision: **100.00%** Recall: 82.55% F1-score: 90.44%
Not Spam	Accuracy: 95.70% Precision: 95.27% Recall: **100.00%** F1-score: 97.58%	Accuracy: **98.30%** Precision: **98.27%** Recall: 99.79% F1-score: **99.02%**	Accuracy: 91.66% Precision: 91.22% Recall: **100.00%** F1-score: 95.41%	Accuracy: 96.05% Precision: 97.72% Recall: 97.72% F1-score: 97.72%	Accuracy: 97.67% Precision: 97.38% Recall: **100.00%** F1-score: 98.67%

When Table 2 is examined, the results are classified as spam and non-spam by running 5 different machine learning models (Naïve Bayes, Support Vector Machine (SVM), k-NN, Decision Tree, Random Forest) on the extracted data. The best classification performance was given by the SVM classifier (**98.3%/98.3%**) for both classes. In terms of F1 score, the best results (**93.29%/99.02%**) were obtained with the SVM classifier. When evaluated in terms of all classifiers, it is seen that the accuracy results are over **91.0%**.

4. Conclusion

In this study, it was seen that TF-IDF feature extractor offers an effective solution in spam classification. Correct classification of data significantly affects the performance of TF-IDF feature extractor. The study showed that the SVM classifier provides state-of-the-art results. It is planned to conduct similar studies with deep learning in future studies.

5. Acknowledgement

This study was supported by TÜBİTAK 2209A project number 1919B012321026.

6. References

[1] "Global spam volume as percentage of total e-mail traffic from 2011 to 2023," 2024. Global spam volume as percentage of total e-mail traffic from 2011 to 2023 (accessed Apr. 15, 2024).

[2] Rukiye Özdemir Tekir, "Text classification using graph convolutional networks," İstanbul Üniversitesi-Cerrahpaşa, 2023. [Online]. Available: https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=i5n1F6nW3nLKanZgrGP9VQ

[3] N. Alias, C. F. M. Foozy, and S. N. Ramli, "Video spam comment features selection using machine learning techniques," *Indones. J. Electr. Eng. Comput. Sci.*, 2019, doi: 10.11591/ijeecs.v15.i2.pp1046-1053.

[4] G. Forman, "BNS feature scaling: An improved representation over TF-IDF for SVM text

classification," 2008. doi: 10.1145/1458082.1458119.

[5] A. Chavez, "TF-IDF classification based Multinomial Naïve Bayes model for spam filtering," National College of Ireland, 2020. [Online]. Available: https://norma.ncirl.ie/4490/1/alanchavez.pdf

[6] Y. Feng, M. A. Naeem, F. Mirza, and A. Tahir, "Reply using past replies—a deep learning-based e-mail client," *Electron.*, 2020, doi: 10.3390/electronics9091353.

[7] N. A. Haqimi, N. Rokhman, and S. Priyanta, "Detection Of Spam Comments On Instagram Using Complementary Naïve Bayes," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, 2019, doi: 10.22146/ijccs.47046.

[8] M. Jehad, A. Bou Nassif, and M. A. AlShabi, "Spam SMS detection based on long short-term memory and recurrent neural network," 2023. doi: 10.1117/12.2674147.

[9] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," 1998. doi: 10.1007/bfb0026683.

[10] F. Jáñez-Martino, E. Fidalgo, S. González-Martínez, and J. Velasco-Mata, "Classification of Spam Emails through Hierarchical Clustering and Supervised Learning." 2020.

[11] M. Mohammedid and N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS One*, 2020, doi: 10.1371/journal.pone.0230442.

[12] T. H. Pham and P. Le-Hong, "Content-based approach for Vietnamese spam SMS filtering," 2017. doi: 10.1109/IALP.2016.7875930.

[13] W. A. N. G. Zhuohao, W. A. N. G. Dong, and L. I. Qing, "Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF," *Chinese J. Electron.*, 2021, doi: 10.1049/cje.2021.05.007.