EasyChair Preprint

№ 11608

Optimizing Arabic Spam Filtering Through Unsupervised and Ensemble Learning Approaches

Marouane Kihal and Lamia Hamza

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

December 23, 2023

Optimizing Arabic Spam Filtering through Unsupervised and Ensemble Learning Approaches

Marouane KIHAL
Laboratory of Medical Informatics (LIMED), Faculty of Exact Sciences, University of Bejaia, 06000 Bejaia, Algeria
marouane.kihal@univ-bejaia.dz

Lamia HAMZA
Laboratory of Medical Informatics (LIMED), Faculty of Exact Sciences, University of Bejaia, 06000 Bejaia, Algeria
lamia.hamza@univ-bejaia.dz

***Abstract*—**

This paper presents a new Ensemble Learning approach for filtering Arabic spam. The proposed approach utilizes four unsupervised Machine Learning algorithms, including One Class Support Vector Machine (OCSVM), the Histogram-Based Outlier Score (HBOS), Local Outlier Factor (LOF) and Isolation Forest (IF), to construct a robust spam filter. The performance of our proposed approach is evaluated on a textual Arabic dataset. The experimental results show that our model achieves more than 84% of accuracy outperforming other Machine Learning algorithms. The use of Ensemble Learning and multiple unsupervised algorithms in our approach proves to be a promising solution for effective Arabic spam filtering.

Keywords—Spam filtering; Arabic Spam; Unsupervised Learning; Ensemble Learning; Machine Learning

I. INTRODUCTION

Spam is unsolicited messaging or content that is frequently distributed to a large number of recipients. It may be employed for maliciousness, advertising or commercial purposes. To avoid this problem, researchers and companies are proposing filtering software and spam detectors based on Artificial Intelligence (AI). Until today, computer users and especially social network users still suffer from spam. Consumer scams, threatened privacy, security breaches and lost productivity for businesses are the most relevant dangers of spams. To protect against these dangers, researchers have proposed several filtering techniques.

Nowadays with the evolution of Machine Learning algorithms, several filters have been proposed, such as K-Nearest Neighbors (KNN) [1], Support Vector Machine (SVM) [2], ETC. Most of the existing work focuses on supervised learning, but recently there are some new works based on unsupervised learning, which attempts to teach a Machine Learning algorithm information that is neither classified nor labeled, and to allow this algorithm to react to this information independently.

However, the problem of spam remains, especially spam written in languages other than English, such as Arabic, which really causes a problem because of the complexity of this language and the very few resources available to train models in Arabic. That is why we have chosen to address Arabic spam.

In this paper, we propose a new spam filter based on unsupervised learning and ensemble learning, Our architecture is based on a late fusion of four Machine Learning algorithms namely One Class Support Vector Machine (OCSVM), the Histogram-Based Outlier Score (HBOS), Local Outlier Factor (LOF) and Isolation Forest (IF).

To summarize, the primary contributions of this study are:

- Proposal of an effective Arabic anti-spam filter based on four unsupervised learning.
- Utilizing ensemble learning fusion for result optimization.
- We compared our model with supervised and unsupervised learning algorithms.
- Our model surpassed the Machine Learning models, achieving 84,78% accuracy, 99,99% precision, 84,78% recall, and 91,76% F1-score.

The remainder of this paper is structured as follows:

Section 2 entails an overview of related works. Section 3 presents our proposed approach. Section 4 discusses the experimentation results. Section 5 gives the conclusion and perspectives for future research.

II. RELATED WORKS

In this section we present related works in two parts: unsupervised learning spam filteret and arabic spam filter.

A. Unsupervised Learning Spam filter

In [3], the authors presented a text-mining framework for extracting textual signatures from unlabeled documents. The system contains preprocessing stages like term reduction and matrix formation in addition to Latent Semantic Analysis, which boosts semantics. As a supplement to textual signatures in spam detection, it also covers the extraction of HTML and URL signatures. The article assesses the detection accuracy of the various types of signatures as well as the accuracy of the created campaign clusters.

In [4], the authors used M-DBSCAN's unsupervised learning capabilities to identify spam and legitimate emails. Modified Density-Based Spatial Clustering of Applications with Noise (M-DBSCAN) was used in the method's implementation. The online test uses the N-representative points that were retrieved from each cluster. These points are created during the training phase of the distance-based spam email detection process.

In [5], the authors used the digest procedure to cluster emails, and then they used the DBSCAN clustering method. Comparing the results to the regular DBSCAN, the accuracy was found to have improved. This study demonstrates the efficacy of DBSCAN and clustering algorithms for email spam detection.

B. Arabic Spam filter

Najadat et al.[6] proposed a keyword-based technique for identifying fake reviews in Arabic. To extract keywords from Arabic text, the weight of a term, the Term Frequency-Inverse Document Frequency (TF-IDF) matrix, and filter approaches have been utilized. Then, Supervised Machine Learning algorithms including Decision Tree, kNN, SVM, and Nave Bayes have been used for classification.

The Gradient Boosting algorithm and the most efficient hyperparameter selection were used by [7] to provide a novel method for improving Arabic spam filtering effectiveness.

In [8], the authors provided a method that does not rely on hand-crafted features that are frequently time-consuming to get and are designed for a specific type of low-quality information, but instead automatically extracts textual features using deep learning techniques. They additionally set up a rapid system that makes use of a selection of arabic textual features to spot spammy Twitter accounts in real time.

III. OUR METHOD

Our approach is based on two concepts: unsupervised learning to get the primary decision on whether the Arabic content is spam or not, by using four unsupervised algorithms. Moreover, the second notion of our approach is ensemble learning to make the final decision by fusing the results of the four previous algorithms.

A. Unsupervised Learning

In the context of spam filtering, unsupervised learning is a technique for data analysis where a model finds key patterns in a dataset without the use of labels. Instead of depending just on labels, the algorithms used in this method seek to cluster related data or reveal hidden patterns in order to determine if content is Spam or Ham.

a) One class SVM:

One Class Support Vector Machines (OCSVM) [9] is a kind of outlier detection method. This unsupervised learning method used for developing the capacity to distinguish test samples from one class from test samples from other classes.

b) The Histogram-based Outlier Score:

The Histogram Based Outlier Score (HBOS) [10] determines the outlier score for that variable using the histogram of each variable. The multivariate outlier score for an observation can be calculated by summing the outlier scores of all the variables. The HBOS is an effective unsupervised method to identify anomalies since histograms are simple to create.

c) Local Outlier Factor:

The Local Outlier Factor (LOF) [11] algorithm calculates the local density deviation of a particular data point with respect to its neighbors. It is an unsupervised anomaly identification technique. The samples that have a significantly lower density than their neighbors are considered as outliers.

d) Isolation Forest:

The binary tree-based Isolation Forest (IF) algorithm[12] works well with massive quantities of data since it has a linear time complexity and needs little memory. Fundamentally, the system does a quick density estimation approximation, recognizing data points with noticeably low density estimates as anomalies.

B. Ensembel Learning

Ensemble learning improve results by combining various models. This technique permits the production of higher prediction performance as compared to employing a single model.

A number of fusion techniques exist, including decision-level fusion, which combines feature classification decisions, model-level fusion, which is based on the relationships between features in various algorithms, and rule-level fusion, which employs techniques like majority voting or weighted fusion [13]. “Fig.1” shows the general architecture of our proposed system.

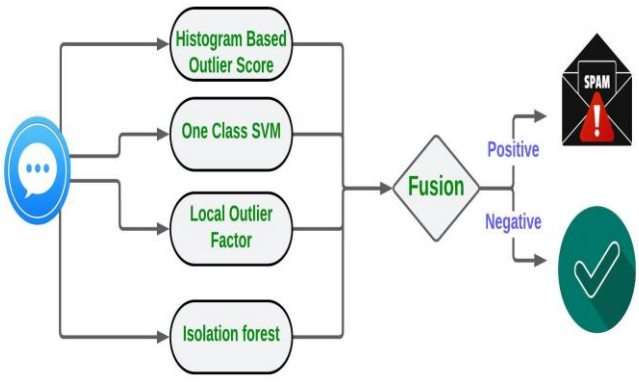

Fig. 1. Our system architecture.

We propose to make the classification with the four algorithms already mentioned (IF, HBOS, LOF and One class SVM). Therefore, we will arrive at four decisions on the content: 1

for spam, 0 for not spam. The final decision is determined by averaging the decisions made by unsupervised Machine Learning classifiers. As shown in equation 1:

$$D(f)=Round((D(OCSVM)+D(HBOS)+D(LOF)+D(IF))/4). \quad (1)$$

Where D(f) is the final decision, and D(OCSVM), D(HBOS), D(LOF) , and D(IF) are the decision of the four unsupervised algorithms.

IV. Experimentations

A. Dataset

Our model was trained and tested using health-related spam campaigns[14]. Between May 2018 and November 2020, this unbalanced dataset was collected from prominent Arabic hashtags using Twitter's standard search application programming interface (API). The dataset composed of 3000 tweets; 2500 training and 500 testing dataset.

B. Results

For the results we used four evaluation metrics most commonly used in spam detection [15], namely Accuracy, Precision, Recall and F1-score.

Table 1 shows the experimental results of our model in comparison with seven known Machine Learning algorithms, four unsupervised above-mentioned and three supervised ML such as: Naïve Bayes, Logistic Regression, and Support Vector Machine.

TABLE I. HEALTH-RELATED SPAM CAMPAIGNS RESULTS

Classifier	Justesse	Precesion	Recall	F1-Score
Isolation Forest	79,35%	93,59%	83,91%	88,48%
HBOS	82,61%	97,44%	84,44%	90,48%
Local Outlier Factor	83,70%	98,72%	84,62%	91,12%
One class SVM	80,43%	94,87%	84,09%	89,16%
Naive Bayes	68,85%	33,71%	54,12%	41,54%
Logisitc Regression	79,92%	57,14%	7,33	13,00%
Support Vector Machine	80,86%	81,81%	8,25%	15,00%
Our method	**84,78%**	**99,99%**	**84,78%**	**91,76%**

From the results displayed in the table 1, we conclude that our model has overcome the Machine Learning algorithms under four evaluation metrics with accuracy equal to 84.78%, a precision 99.99%, Recall 84.78% and F1-score equally to 91.76%.

V. Conclusion

In this paper, we have proposed a new approach based on Unsupervised and Ensemble Learning to effectively filter Arabic textual spam, using four powerful unsupervised Machine Learning algorithms, namely One Class Support Vector Machine (OCSVM), The Histogram-Based Outlier Score (HBOS), Local Outlier Factor (LOF) and Isolation Forest (IF). The experiment's findings show that our suggested model surpassed the individual supervised and unsupervised Machine Learning algorithms in terms of four evaluation metrics, achieving an accuracy of 84.78%, a precision of 99.99%, Recall of 84.78% and a F1-score equally to 91.76%.

These findings demonstrate the potency and reliability of our suggested approach, which may be applied as a successful countermeasure to the spreading Arabic textual spam across a variety of internet platforms. Future research can examine how well this method can be applied to different domains and datasets.

Acknowledgment

This work has been sponsored by the General Directorate for Scientific Research and Technological Development, Ministry of Higher Education and Scientific Research (DGRSDT), Algeria.

References

[1] L. FIRTE,C. LEMNARU, and R. POTOLEA. “Spam detection filter using KNN algorithm and resampling”. In : Proceedings of the 2010 IEEE 6th international conference on intelligent computer communication and processing. IEEE, 2010. p. 27-33.

[2] RC. PATIL and DR. PATIL, “Web spam detection using SVM classifier”. In : 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO). IEEE, 2015. p. 1-4.

[3] F. QIAN, A. PATHAK, YC. HU, ZM. Mao and Y. Xie, “A case for unsupervised-learning-based spam filtering”. ACM SIGMETRICS performance evaluation review, 2010, vol. 38, no 1, p. 367-368.

[4] M. MANAA, A. OBAID, M. DOSH. “Unsupervised approach for email spam filtering using data mining”. EAI Endorsed Transactions on Energy Web, 2021, vol. 8, no 36.

[5] A. HARISINGHANEY, A. DIXIT, S. GUPTA,A Arora. “Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm”. In : 2014 International Conference on Reliability Optimization and Information Technology (ICROIT). IEEE, 2014. p. 153-155.

[6] H. NAJADAT, MA. ALZUBAIDI and I. QARQAZ. “Detecting Arabic spam reviews in social networks based on classification algorithms”. Transactions on Asian and Low-Resource Language Information Processing, 2021, vol. 21, no 1, p. 1-13.

[7] M. KIHAL and L. HAMZA. “Enhancing Efficiency of Arabic Spam Filtering Based on Gradient Boosting Algorithm and Manual Hyperparameters Tuning”. In : International Conference on Applied CyberSecurity. Cham : Springer Nature Switzerland, 2023. p. 49-56.

[8] R. ALHARTHI, A. ALHOTHALI and K. MORIA. “A real-time deep-learning approach for filtering Arabic low-quality content and accounts on Twitter”. Information Systems, 2021, vol. 99, p. 101740.

[9] M. MANEVITZ, and M. YOUSEF, “One-class SVMs for document classification”. Journal of machine Learning research, 2001, vol. 2, no Dec, p. 139-154.

[10] M. GOLDSTEIN, and A. DENGEL, “Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm”. KI-2012: poster and demo track, 2012, vol. 1, p. 59-63.

[11] O. ALGHUSHAIRY, R. ALSINI, T. SOULE, “A review of local outlier factor algorithms for outlier detection in big data streams”. Big Data and Cognitive Computing, 2020, vol. 5, no 1, p. 1.

[12] FT. LIU,KM. TING, Z. ZHOU, “Isolation forest“. In : 2008 eighth ieee international conference on data mining. IEEE, 2008. p. 413-422.

[13] X. DONG, Z. YU, W. CAO. “A survey on ensemble learning”. Frontiers of Computer Science, 2020, vol. 14, p. 241-258.

[14] M. KIHAL and L. HAMZA. “Robust multimedia spam filtering based on visual, textual, and audio deep features and random forest”. Multimedia Tools and Applications, 2023, p. 1-19.

[15] Imam N (2020), Health-related Spam Campaigns, Mendeley Data, V1, https://doi.org/10.17632/rgrvt5x4tk.1