Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/2023400

A systematic review and outlook on machine-learning-based methods for spam-filtering

Yufeng Gao

No. 111, Renai Road, Dushu Lake Science and Education Innovation Zone, China-Singapore Industrial Park, Suzhou City, Jiangsu Province, China

Yufeng.Gao20@student.xjtlu.edu.cn

Abstract. This paper reviews the stages of development and common approaches to spam text classification. This was followed by an introduction to the basic knowledge of spam filtering. After that, some research results are presented. The last part will be a discussion of the literature review and some suggestions for the future development of spam filtering techniques. This paper found that the future direction of spam filtering should be a combination of machine learning techniques and deep learning techniques and propose to take into consideration diversified features of the emails other than text such as IP addresses and IP reputation values into the machine learning models.

Keywords: spam filtering, natural language processing (NLP), deep learning.

1. Introduction

Text-based software like Twitter and email have become a part of our social and work life. However, some unfriendly text may lead to impaired user experience and other potential harms. For example, spam emails, which are referred to as the unsolicited emails, bring a bad experience to the user [1]. In 2019[2], Abdullah Sheneamer classified spam emails into 7 categories: Ads, Chain-letter (scare you into forwarding emails), email spoofing, porn, money scam, hoaxes and malware warning. The harm caused by such spam text includes, but is not limited to, causing damage to user property, damaging user experience, and misleading users. Moreover, according to [3], in 2014, the total number of messages of spam emails reached 54 billion every day, which is a huge number. In addition to spam, there are also emails/tweets that need to be handled properly. The same technique of spam classification is applied to other social software other than email, like YouTube comments [4], tweets and so on [5].

Due to the severity of the problem, different approaches to spam classification have been adopted to handle the spam-text problem. Techniques for spam text classification can be divided into three types: traditional techniques, traditional machine-learning-based techniques and deep learning techniques. The three methods are interrelated rather than isolated from each other, but we find that the integration of these three different technologies is still inadequate. Although there have been many ways to classify spam text, there is still a lack of systematic description of these methods, and this paper will review the existing methods of classifying spam text and try to propose potential methods for the future development of the techniques of spam filtering.

DOI: 10.54254/2755-2721/4/2023400

2. Introduction of spam-filtering methods

2.1. The IP-based methods

The IP-based methods are the longest-established method for spam filtering. The most common method of this category is to divide IP addresses into blacklist and whitelist [6]. A method proposed by Liu Yang (2015) which divides all the IP addresses into two lists is based on the integrity of each IP address [7]. However, these methods are limited because the attacker can use the dynamic IP-address or exchange the IP-address to avoid the detection.

2.2. Feature extraction and selection of texts

2.2.1. Feature extraction. Feature extraction is the first step of machine-learning-based methods to spam text filtering. It refers to the process that transforms the text words into the vectors which computers can "understand".

The most primitive method of feature extraction is called one-hot extraction. It represents each word as a vector of only one 0 and 1s. However, this kind of representation leads to highly sparse vectors. Direct use of the one-hot vectors results in a huge waste of computational resources and little to no benefit. However, [8] one-hot vectors are essential for the training of the dense vectors. Represent each word with a unique dense vector that can represent each word's meaning through the values in the vector (like king-man = queen-women).

There are now many ways to convert text to vectors(word embedding), such as FLAIR[9], [10] GloVe(Global Vectors for Word Representation), word2vec[11] and ConceptNet.

2.2.2. TF-IDF. The full name of TF-IDF is Term Frequency and Inverse Document Frequency. The TF-IDF method is based on the Zip's Law, which, according to [10], means that the relevance of a word to the topic of an article is proportional to the number of times it appears in the article and inversely proportional to the total number of times it appears in the corpus.

$$TF - IDF$$

$$W_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right)$$

$$tf_{i,j} means\ the\ number\ of\ occurences\ of\ i\ in\ j$$

$$df_{i,j}\ means\ the\ number\ of\ occurences\ of\ documents\ containing\ i$$

$$N\ means\ the\ total\ number\ of\ documents$$

2.3. Machine learning methods for the spam filtering

2.3.1. SVM (Support Vector Machine). Support Vector Machine is one of the most popular algorithms in the field of machine learning. SVM can be both linear and non-linear. The core idea of a support vector machine (SVM) is to maximize the sum of the Euclidean distances between points and the decision border.

$$\boldsymbol{Euclidean\ distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \cdots + (y_n - y_n)^2}$$

X and y represent the values of each of the different dimensions of the vector, while the Euclidean distance represents the open root of the interpolated sum of squares of each dimension between the two vectors.

2.3.2. Naïve Bayers. The model of Naïve-Bayers is based on the is based on Bayesian theorem, which calculates the probability of an event using the probability relationship between events. It used to be one of the most accurate models for spam filtering. In the field of text classification, it classifies text on a per-word basis. Typically the input of the model will be the tensors that contain only 0 and 1, where 1 denotes the existence of a certain feature and 0 denotes the opposite. The main disadvantage of Naïve-

DOI: 10.54254/2755-2721/4/2023400

Bayers is that it neglects the order and relationships between words and words. For example, the sentences “not good, very bad” and “not bad, very good” are seen as the same by the Naïve Bayers model.

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_j)P(A|B_j)}$$

2.3.3. Decision tree and random forests. Decision trees can be interpreted as a combination of many if-else statements, it is widely applied in the field of classification. But every decision tree has a tendency to overfit, and this is when the random forest solution is proposed. Random forest is a collection of decision trees which is aimed to mitigate the overfitting of decision trees through combine all the decision trees together and generating a new model.

2.4. The deep-learning-based methods for spam filtering

2.4.1. CNN (Convolutional Neuron Network). The full name of CNN is Convolutional Neuron Network. It is first introduced to the field of Text Classification by Yoon Kim [12]. The convolution operation is composed of 2 parts: convolution and pooling. Convolution represents the operation of extracting features from the text using matrix operations while pooling refers to the operation that simplifies the tensors based on simpler rules (like choosing the maximum).

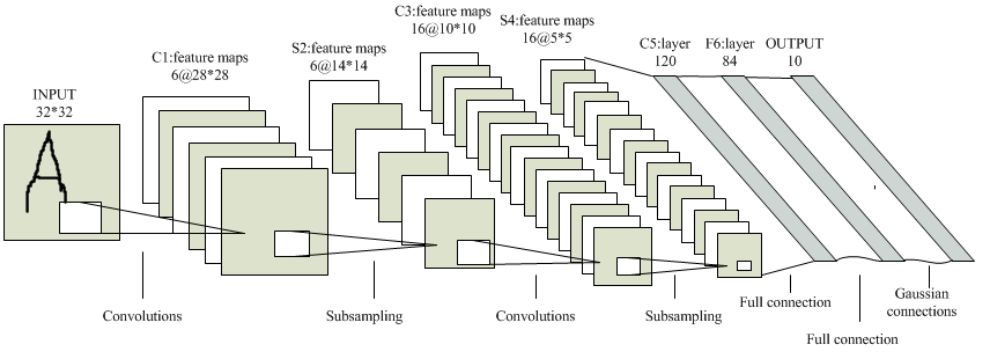

Figure 1. [28] An example of CNN model (LeNet).

2.4.2. RNN (Recurrent Neuron Network). RNN is a neural network layer with a recurrent structure, where the output of the neuron has an impact on the output of the neuron afterward, so in fact, it has a memory function. GRU (Gated Recurrent Unit) and LSTM (Long-Short-Term Memory) are also RNNs in a general sense. Both GRU and LSTM are designed to improve the long-term memory function of traditional RNNs.

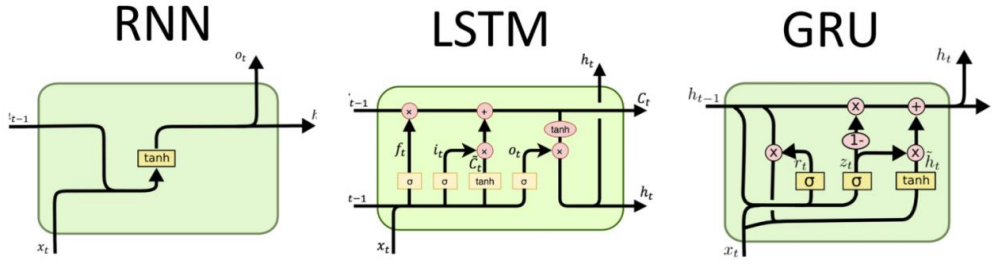

Figure 2. There are three most used RNN models as shown in the figure: GRU, LSTM and) RNN.

2.5. Evaluation metrics

In this paper, I choose accuracy as the evaluation method of the spam filtering model.

DOI: 10.54254/2755-2721/4/2023400

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

TP means the value of true positive which denotes the correct classification of spam. TN means the value of true negative which denotes the correct classification of ham. FP means the false positive which denotes the false classification of spam. FN means false negative which denotes the false classification of ham.

3. Review and outlook of the methods for spam filtering

As we mentioned in the introduction part, the development of techniques of spam filtering has gone through three stages of development: traditional (None-AI) techniques, machine learning-based techniques and deep-learning-based techniques.

Current techniques for classifying spam texts show a tendency towards hybridization and diversity. Part of this diversity is reflected in the classifying models' input data. [13] Alom et al. (2019) proposed a deep learning model for Twitter Spam Filtering. The model takes both the text of tweets and the data of users as the input, reaching the highest accuracy of 99.68% on a dataset. According to [14], Ezpeleta et al. (2016) used Facebook public information to personalize the spam content and create profile-based emails. Another neural-network-based model takes the details (frequency of words in uppercase and the numbers in the message along with the number of different colors, blank lines and the size of the message text) of texts rather than texts themselves as the input and got the accuracy of 97.5% [15]. In 2020, [4] Ezpeleta combine the input of the ham/spam dataset and the results of sentiment analysis and enter them into the given 10 Bayesian spam filtering models, 7 of them gained an increase in accuracy.

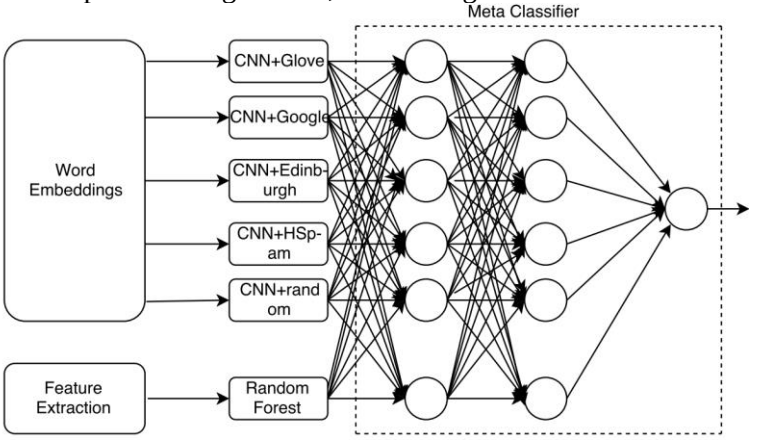

Figure 3. An example of a hybrid model. The author let the results of multiple mixed word embedding pass through an MLP regressor [22].

Another feature of technological diversity is the diversity of the design of the models themselves. Although we mentioned that there are three stages of technological development, the latter two of these three stages are not independent of each other. According to [15], Mai A.Shaaban et al. proposed the DcForest (Deep-Convolutional-Forest) Model, this model consists of alternating random forest layers and CNN layers. This model predicts spam with 98.38% accuracy. Similarly, Alom et al. (citation) designed a model composed of SVM and CNN, the accuracy of model's prediction on Twitter dataset reached 99.32%.

DOI: 10.54254/2755-2721/4/2023400

Another example of the multiplicity of methods used in the model itself is the model proposed by Sreekanth Madisetty [22], she combined 5 models together through put all of their inputs in a meta-classifier which is a Multiple Layers Perceptron composed of 2 layers, each with 6 neurons. This model exceeds the original five sub-models in terms of precision (0.880), accuracy (0.957), Recall (0.909) and F-measure (0.894).

Deep-learning-based methods still play a dominant role in the field of spam filtering. [25] The SSCL (Sequential Stacked CNN-LSTM) model combines the CNN layer and LSTM layers together and reached an accuracy of 99.01%. Another model consisting of a mixture of CNN and LSTM is proposed by P.K. Roy et al. The highest accuracy reached 99.44%.

Traditional machine learning methods still perform well in the field of spam filtering. S. Wang et al. used fuzzy SVM along with the K-Means algorithm and reached an accuracy of 96.5% on the UCI Machine Learning Dataset. B.K. Dedeturk and B. Akay applied logistic regression and Artificial Bee Colony (ABC) algorithm on the email-spam-filtering and get an accuracy of 98.81%. [20]

4. Discussion and conclusion:

As we mentioned before, although the three stages of technology are not mutually exclusive, the integration of their use is still lacking. Most current technologies no longer rely on a single method. [4] and [22] showed that even a simple blend between multiple models with a MLP model can improve the overall performance of prediction. The future direction of this technology will be based on a mixture of existing models. When it comes to the structures of models itself, it is more likely that models will be mixtures of both machine learning and deep learning techniques like DcForest. If machine learning algorithms are separated from deep learning algorithms, the latter generally outperforms the former. A common shortcoming of current spam-text-filtering models is the separation of user data (including the integrity of IP and accounts) from the content of the text itself.

Given the previous research results, the future development direction of spam text filtering may focus on the following points: Combining the mood analysis model with models other than Bayesian models (like SVM and Artificial Neural Network), taking the IP-integrity (and the integrity of accounts) into the input of classifier and taking details such as font and color into account when doing spam filtering. Moreover, when combining the mood analysis model with spam filtering models, emotions should be more finely divided instead of simply making them 1 and 0. (for example, the VADER function in the Python NLTK package can give the score for each emotion that can be calculated).

In conclusion, the current trend in classification models for the text itself is a mixture of existing machine learning and deep learning models. Moreover, the IP-integrity, user profile and details of the texts may play a more important part as the input of deep learning model if we want to continue to improve the deep learning model performance in the future.

Filtering methods	Names	Dataset	categories	Results
[16]	X.Liu et al.	SMS Spam collection	RNN, CNN, Linear, LSTM	98.82% Accuracy
[4]	E. Ezpeleta et al.	YouTube comments dataset	Bayesian spam filtering	Best Accu: 94.38%, the False Positive rate dropped
[17]	Adshish Salunkhe	Deceptive Opinion Spam Corpus	Bidirectional LSTM, CNN, RNN, TF-IDF	Best accuracy: 92.19%

DOI: 10.54254/2755-2721/4/2023400

[15]	M.A. Shaaban et al.	SMS Spam Dataset	Random Forest, CNN	Best Accuracy: 98.9%
[18]	P.K. Roy et al.	SMS Spam dataset	CNN, LSTM	Best accuracy: 99.44%
[19]	S. Wang et al.	UCI ML Repository	Fuzzy-SVM, K-Means	Accuracy: 96.5%
[20]	B.K. Dedeturk, B. Akay	CSDMC2010, Turkish-Email, Enron	Logistic-regression, Artificial Bee colony	Accuracy: 98.81%
[21]	Mohammad Alauthman	Spam-Base	GRU, SVM, RNN	98.65%Accuracy
[22]	Sreekanth Madisetty	HSpam (tweets)	CNN, random forest, SVM	Improve the F1-score to 0.894
[23]	S.E Rahman and S. Ullah	Lingspam and SPMDC	RNN, LSTM, BiLSTM, CNN	Recall, Precision and F-measure are all above 98%
[24]	Z. Alom, B. Carminati, Elena Ferrari	Twitter social honeypot dataset	CNN, SVM	Highest accuracy: 99.32%
[14]	A.S. Katasev et al.	Manually collected e-mails	Neural network	Accuracy: 97.5%
[5]	J. Ma et al.	Twitter and Weibo datasets	RNN(GRU), LSTM	Highest accuracy: 91% on Weibo dataset
[25]	G. Jain, M. Sharma and B. Agarwal	SMS Spam and Twitter datasets	CNN, LSTM (SSCL)	Accuracy: 99.01%
[26]	Qijia Wei	Manually collected	Naïve bayers	----
[27]	S. Kadam et al.	SMS spam and ham messages collected from UCI ML Repository	Naïve Bayers, SVM	Accuracy: 98.28%

References

[1] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," Artif. Intell. Rev., 2008.

[2] Abdullah Sheneamer, “Comparison of Deep and Traditional Learning Methods for Email Spam Filtering” International Journal of Advanced Computer Science and Applications (IJACSA), 12(1), 2021.

DOI: 10.54254/2755-2721/4/2023400

[3] Internet Threats Trend Report. Cyberoam A SOPHOS Campany; 2014
[4] Ezpeleta, E., Iturbe, M., Garitano, I., de Mendizabal, I.V., Zurutuza, U. (2018). A Mood Analysis on Youtube Comments and a Method for Improved Social Spam Detection. In: , et al. Hybrid Artificial Intelligent Systems. HAIS 2018. Lecture Notes in Computer Science, vol 10870. Springer, Cham. https://doi.org/10.1007/978-3-319-92639-1_43
[5] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16). AAAI Press, 3818–3824.
[6] Ezpeleta, Enaitz & Zurutuza, Urko & Gomez Hidalgo, Jose. (2016). A Study of the Personalization of Spam Content using Facebook Public Information. Logic Journal of IGPL. 25. jzw040. 10.1093/jigpal/jzw040.
[7] G. Peng. A review of research based on the naive Bayesian algorithm in spam filtering [J]. 电脑知识与技术 (Knowledge and technologies of computers), 2020,16(14):244-245,247.
[8] Yoav Goldberg and Graeme Hirst. 2017. Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers.
[9] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, June). FLAIR: An easy-to-use framework for state-of-the-art NLP. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) (pp. 54-59).
[10] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[12] Zhang, Ye, and Byron Wallace. "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification." arXiv:1510.03820 (2015).
[13] Ezpeleta, E., Zurutuza, U., & Hidalgo, J. M. G. (2017). A study of the personalization of spam content using facebook public information. Logic Journal of the IGPL, 25(1), 30-41.
[14] Katasev, A. S., Emaletdinova, L. Y., & Kataseva, D. V. (2018, May). Neural network spam filtering technology. In 2018 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM) (pp. 1-5). IEEE.
[15] Shaaban, M. A., Hassan, Y. F., & Guirguis, S. K. (2021). Deep convolutional forest: a dynamic deep ensemble approach for spam detection in text. arXiv preprint arXiv:2110.15718.
[16] Liu, X., Lu, H., & Nayak, A. (2021). A Spam Transformer Model for SMS Spam Detection. IEEE Access, 9, 80253-80263.
[17] Salunkhe, A. (2021). Attention-based Bidirectional LSTM for Deceptive Opinion Spam Classification. arXiv preprint arXiv:2112.14789.
[18] Roy, P. K., Singh, J. P., & Banerjee, S. (2020). Deep learning to filter SMS spam. Future Generation Computer Systems, 102, 524-533.
[19] Wang, S., Zhang, X., Cheng, Y., Jiang, F., Yu, W., & Peng, J. (2018, January). A fast content-based spam filtering algorithm with fuzzy-SVM and K-means. In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 301-307). IEEE.
[20] Dedeturk, B. K., & Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. Applied Soft Computing, 91, 106229.
[21] Alauthman, M. (2020). Botnet spam e-mail detection using deep recurrent neural network. International Journal, 8(5).
[22] Rahman, S. E., & Ullah, S. (2020, June). Email spam detection using bidirectional long short-term memory with convolutional neural network. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 1307-1311). IEEE.
[23] Madisetty, S., & Desarkar, M. S. (2018). A neural network-based ensemble approach for spam

DOI: 10.54254/2755-2721/4/2023400

detection in Twitter. IEEE Transactions on Computational Social Systems, 5(4), 973-984.

[24] Alom, Z., Carminati, B., & Ferrari, E. (2020). A deep learning model for Twitter spam detection. Online Social Networks and Media, 18, 100079.

[25] Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short-term memory neural network. Annals of Mathematics and Artificial Intelligence, 85(1), 21-44.

[26] Kadam, S., Gala, A., Gehlot, P., Kurup, A., & Ghag, K. (2018, August). Word embedding based multinomial naive bayes algorithm for spam filtering. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

[27] Zulfikar Alom, Barbara Carminati, Elena Ferrari, A deep learning model for Twitter spam detection, Online Social Networks and Media, Volume 18,2020, 100079, ISSN 2468-6964, https://doi.org/10.1016/j.osnem.2020.100079.

[28] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324. doi:10.1109/5.726791