Research Article

Context-dependent model for spam detection on social networks

Razan Ghanem[1] · Hasan Erbay[2]

Received: 2 November 2019 / Accepted: 19 August 2020 / Published online: 29 August 2020

Abstract
Social media platforms are getting an important communication medium in our daily life, and their increasing popularity makes them an ideal platform for spammers to spread spam messages, known as spam problems. Moreover, messages on social media are vague and messy, so a good representation of the text may be the first step to address spam problem. While traditional weighting methods suffer from both high dimensionality and high sparsity problems, traditional word embedding methods suffer from context independence and out of vocabulary problems. To overcome these problems, in this study, we propose a novel architecture based on a context-dependent representation of text using the BERT model. The model was tested using the Twitter dataset, and experimental results show that the proposed method outperforms traditional weighting methods, traditional word embedding based methods as well as the existing state of the art methods used to detect spam on the twitter platform.

Keywords Spam detection · Word embedding · Bidirectional encoder representations from transformers

1 Introduction

Social media are interactive computer-mediated technologies that facilitate the creation or sharing of information, ideas, career interests, and other forms of expression via virtual communities and networks. Twitter is one of the most popular social media nowadays. Twitter reported that its worldwide monetizable daily active users (mDAUs) grew by 24% to 166 million in Q1 2020. Each twitter user has, on average, 208 followers, and they post 140 million tweets daily. This popularity of the Twitter platform has made it a suitable environment for spreading spam messages, which have become a challenging problem due to the messy and ambiguity of short text messages on social media.

Social spam messages might be defined as irrelevant or unsolicited messages sent over social media such as malicious links, advertisements, or any low-quality content. Unlike long messages like e-mails, social spam messages are more sparse and ambiguous, and thus spam classification problem in social networks becomes a more challenging problem. One of the important tasks that could be utilized to handle short text on social media is word representation.

The traditional word representation methods are based on the Bag of Word (BoW) model in which each word or n-gram is linked to a vector index and marked as 0 or 1 depending on whether it occurs in a given document. Although it produces acceptable results, it suffers from some problems like high dimensionality and high sparsity. Word Embedding methods solve these problems by representing the words as dense vectors, where a vector represents the projection of the word into a continuous vector space. Word2vec is the first-word embedding model introduced by Tomas Mikolov in 2013 at Google.

There are two main training algorithms for Word2vec: Skip Gram and Continous Bag Of Words (CBOW). Fast Text, an extension of the word2vec model, is another common

✉ Razan Ghanem, razan@kku.edu.tr; Hasan Erbay, herbay@thk.edu.tr; hasan_erbay@yahoo.com | [1]Department of Computer Engineering, Kırıkkale University, Kirikkale, Turkey. [2]Department of Computer Engineering, University of Turkish Aeronautical Association, Ankara, Turkey.

SN Applied Sciences (2020) 2:1587 | https://doi.org/10.1007/s42452-020-03374-x

word embedding method developed by Facebook in 2016. In Fast Text, each word in the corpus is represented as an n-gram of characters, which helps to capture the meaning of shorter words and allows the embeddings to understand suffixes and prefixes. The other benefit of using the Fast Text model is the ability to work with the rare words that haven't seen them before during the training data. Another common model for word representation is Global Vectors (GloVe) which was developed by Pennington et al. at Stanford University. Unlike word2vec, a predictive model, GloVe is a count-based model and learns its vectors by dimensionality reduction on the co-occurrence counts matrix. The topic model is another method that might be used for representing the texts. The topic model is a statistical model and discovers the abstract "topics" that occur in a collection of documents. The most common form of the topic modeling is Latent Dirichlet Allocation (LDA), LDA aims to find topics in which a document belongs to, based on the words in it. Its main limitation is the inability to work with short text like tweets.

On the other hand, traditional word embedding models suffer from so-called the context-independent problem in which they generate only one vector for each word, regardless of its meaning and its position in the sentence, and also suffer from the out of vocabulary (OOV) problem that the models cannot process new data not previously present in the models' dictionary. To overcome these problems, we propose to use Bidirectional Encoder Representations from Transformers (BERT) to represent the tweet text. BERT is a context-dependent model that generates different word embeddings for a word depending on its meaning and its location in a sentence. Any word that does not occur in the vocabulary is broken down into sub-words to deal with the out of vocabulary problem. In short, in this study, we implement various types of deep learning-based models with the help of different word embedding methods to address the problem of spam messages on social networks and test the models on twitter and SMS datasets. Then, we propose a novel architecture based on the BERT model to detect spam messages on twitter. Finally, we compare traditional weighting methods, traditional word embedding-based methods, and some state of the art methods with our proposed method in terms of their performances. The rest of the paper is organized as follows. Section 2 presents some existing methods in the literature to detect spam on social networks. Section 3 describes the used methodologies. Section 4 is devoted to the proposed models. It presents the architecture of the models and describes models' parameters. Section 5 presents experimental results, Section 6 includes discussion, and Section 7 includes conclusion and future works.

2 Related works

Spam problem on social networks has gained attention in the last decade. Researchers in various areas made a great effort to address spam on social media. In this study, we will focus on spam detection on the Twitter network. To detect spam on Twitter, most of the studies were oriented on two main directions, detecting spammer accounts as in [1–6], and detecting spammed tweet messages as in [7–11].

The majority of the researches treated spam as messages that contain malicious links [1] and others like [11] generalized the spams' definition by including advertisements, automatically generated content, and any low-quality content.

Due to the convincing performances provided by machine learning algorithms, many researchers have used machine learning methods to detect spam in social networks. For example in [7], the authors used direct features which were extracted from crawled JSON tweet to detect content polluter using random forest classifier, and in [10] a framework based on tweet-based and user-based features along with text-based features was proposed to detect spam on the tweet text and evaluated using some machine learning classifiers and neural network. Table 1 summarizes some studies based on machine learning algorithms to detect spam on social networks.

Along with the remarkable development of deep learning algorithms in the field of natural language processing, various types of text classifications based on deep neural networks emerged. Since neural networks are designed to learn from numerical data, word embedding methods improve the ability of networks to learn from a text

Table 1 Baseline machine learning methods used to detect spam on social networks

Paper	Classification method	Dataset	Remarks
[4]	Support vector machine (SVM)	Sina Weibo	Detect spammer accounts on Sina Weibo social media based on message content and users' social behavior
[5]	Naïve Bayes (NB) and SVM	Twitter	The integrated approach was more accurate than machine learning used alone
[20]	J48, Decorate and NB	Twitter	J48 outperformed all other classifiers
[21]	NB and decision tree (DT)	YouTube	the performance of DT classifier was better than NB classifier in most cases

by representing the words as lower-dimensional vectors. Besides, embedding methods enable us to measure the similarity between different words by computing the distance between embedding vectors. With fixed-length embedding vectors as input, deep learning-based models with word embedding methods achieve competitive results in text classification problems which prompted researchers to integrate deep learning methods to their detection architecture to achieve better performances, for example In [12], a recurrent convolutional neural network was used in document classification, and in [13], the model called "C-LSTM" which combines Recurrent Neural Network (RNN) with Convolutional Neural Network (CNN) was proposed for sentence representation and text classification. The model utilized CNN to extract a sequence of higher-level phrase representations and used a long short-term neural network (LSTM) to obtain the sentence representation. In [14], the authors proposed two combination models, named BLSTM-2DPooling and BLSTM- 2DCNN, where the 2D max-pooling operation was applied to obtain a fixed-length representation of the text, and 2D convolution was utilized to sample more meaningful information of the matrix. In [9], the authors proposed semantic Convolutional Neural Network (SCNN) for spam classification on Twitter with the help of WordNet and ConceptNet knowledge bases, and in [15, 16] the syntax of each tweet was learned through Word2Vector model and trained by deep learning, then a binary classifier was used to differentiate between spam and non-spam tweet.

Table 2 summarizes some deep learning-based methods to detect spam on social networks.

Furthermore, the topic model, especially LDA, was used significantly in text classification problems. For example in [17] a new algorithm based on the LDA and the Support Vector Machine was used in the Arabic texts classification, and in [18] an improved method based on the LDA topic model and K-Nearest Neighbor algorithm was proposed to handle the problem of short text classification. On the other hand in [19], a so-called Labeled-LDA to enhance the traditional LDA is proposed to integrate the class information, and based on it a new algorithm was introduced to figure out the latent topics' quantities of each class synergistically.

Herein this study, we propose to use a novel method based on BERT encoder to digitize tweets in detecting spam on Twitter. To the best of our knowledge, it is the first time that this method has been used on the Twitter spam detection domain. Also, the most common word representation methods with different architectures were implemented to analyze the effect of word representation methods on spam detection tasks.

3 Methodology

3.1 Used dataset

We used the dataset published by Chen et al. in their study [11] in which 100,000 tweets were collected through public streams provided by Twitter Streaming API. The original dataset contains only Tweet IDs with labels without tweets information, so we built a crawler that collects tweet information from Twitter based on Tweet ID using a Python program with the help of the Tweepy library. We couldn't access all tweets in the dataset, where only 58,159 tweets were reached, perhaps due to the deletion of these tweets by the owners or by Twitter. Since the extracted dataset was imbalanced, data augmentation was applied to the text to balance the dataset to improve the performance of classification tasks. We also used the SMS spam dataset, which was available in the UCI repository. Table 3 presents the details of the datasets. The datasets were split into 90% training and 10% testing in all models, except for the

Table 3 Used dataset

Dataset	No. spam	No.ham	Total
Twitter	38,205	27,822	66,027
SMS	747	4827	5574

Table 2 Baseline deep learning based methods used for spam detection

Paper	Method	Dataset	Remarks
[9]	CNN,word2vec WordNet and ConceptNet	-Twitter -SMS spam	The authors' approach outperforms the-state-of-the-art results
[15]	Word2vec and binary classifier	Twitter	Proposed model's performance outperforms other text-based and non-text-based methods
[22]	cost-sensitive ensemble learning techniques with regularized deep neural networks	-Twitter-Hayves	The proposed approach outperform other common used classifiers such as random forest, Naive Bayes or support vector machines
[23]	CNN, word embedding (Glove, word2vec)	Twitter	performance of the proposed approach better than the baseline methods

proposed model the splitting was 80% training and 20% testing. The table presents the number of spam messages, the number of non-spam messages for both twitter and SMS spam datasets, and the total for each of them.

3.2 Data preparation

The first step in any text classification task is the preprocessing step. The preprocessing step is very important because it helps to clean unnecessary data to improve the performance of the classification task.

The following preprocessing steps were applied to the text: removing punctuations, removing extra spaces, removing URLs, removing special characters and emojis, removing stop words, and converting the text to lower case. After that, the text was converted to a sequence of tokens and padded with zeroes to ensure equal length.

3.3 Traditional weighting–based method

In this study, we experimented with the most popular BoW based weighting methods such as count vectorizer and TF-IDF with one of the most common machine learning classifiers used in text classification problems such as Naïve Bayes (NB) classifier.

3.4 Traditional word embedding based methods

Three different deep neural network architectures were implemented and experimented with the most common word embedding methods such as word2vec, Glove, and fast Text. We chose simple architectures from each of the following types of neural networks such as a convolutional neural network (CNN), Long short term memory neural network (LSTM), and bidirectional long short term neural network (bi-LSTM). Table 4 shows the hyperparameters used in each model. Where Adam is used as an optimizer in each model, binary cross-entropy is used as a loss function, and a constant value of dropout is used in all models.

3.5 Bidirectional encoder representations from transformers

BERT stands for Bidirectional Encoder Representations from Transformers, which is Google's neural network-based technique for natural language processing (NLP). It was created and published in 2018 by Devlin et al. in Google [24].

Traditional word embedding models are context-independent models output just one vector for each word, regardless of the position of the word in the sentence and regardless of the meaning of that word. In contrast, the BERT model can generate different vectors (embeddings) for the word depending on the meaning and the order of that word in the sentence. BERT is designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers [24]. Figure 1 shows the block diagram of the BERT model where E1 is the embedding representation, T1 is the final output, and Trm are the intermediate representations of the same token.

4 The proposed model

Figure 2 shows the proposed architecture. The model consists of an input layer, a BERT encoder layer with sequence length equal to 20 tokens, then 768 nodes in a hidden layer with Relu as an activation function, and finally one output layer with sigmoid activation function applied to classify the text to spam or ham.

Table 5 displays the hyperparameters used in the model where Adam is used as an optimizer, binary cross-entropy is used as a loss function.

Table 4 Parameters used in word embedding-based models

Parameter	CNN	LSTM	BiLSTM
Optimizer	Adam	Adam	Adam
No. units	–	300	300
Dropout	0.2	0.2	0.2
Recurrent dropout	–	–	0.2
No. filters	32	–	–
kernel size	3	–	–
Pooling	Global max		
Loss function	Binary cross entropy	Binary cross entropy	Binary cross entropy
Batch size	32	64	64
Epoch	30	30	30
Optimization metric	Accuracy	Accuracy	Accuracy

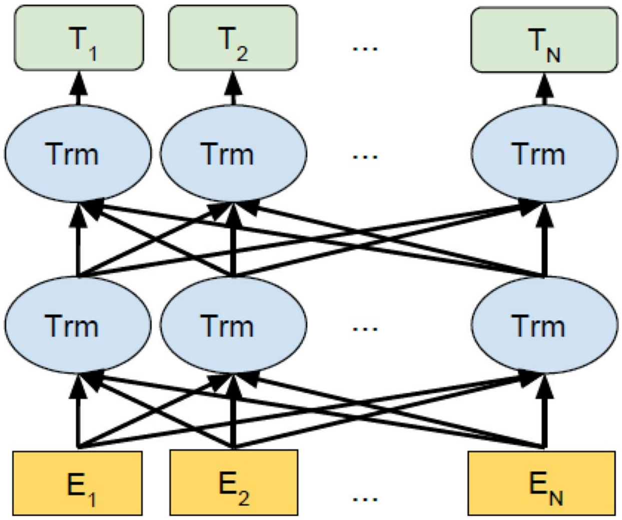

Fig. 1 The block diagram of BERT model adopted from [24]

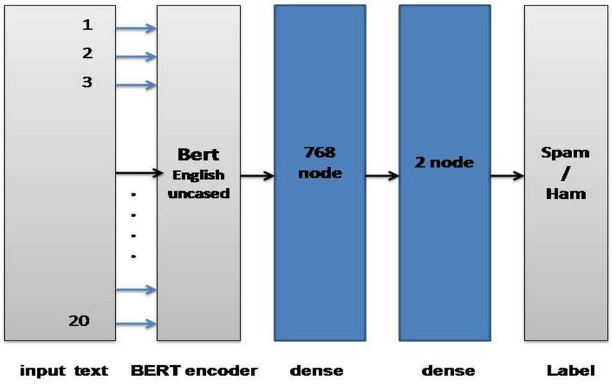

Fig. 2 The block diagram of the proposed model

Table 5 The used parameters in the proposed model

Parameter	Value
Optimizer	Adam
Activation function	RELU, Sigmoid
Sequence length	20
Loss function	Binary cross entropy
Batch size	64
No. epochs	3
Learning rate	0.000001
Optimization metrics	Accuracy, F1-score, precision, recall

Table 6 Confusion matrix

		Predicted class	
True class		**Spam**	**Ham**
	Spam	**98.3%**	1.7%
	Ham	2.1%	**97.9%**

5 Experimental results

Herein, we proposed a novel architecture based on the BERT model for text representation and tested on twitter dataset, then compared the results with traditional feature weighting methods and traditional most common word embedding methods. All experiments carried out using Intel Xeon E5-2680 8 cores and 64 GB of RAM.

5.1 Evaluation metrics

Standard evaluation metrics were used to measure our proposed method, such as accuracy, F1-score, recall, and precision. Table 6 presents the confusion matrix for the proposed model. Here, True Positive (TP) indicates the number of tweets that classified correctly into the spam class. The meanings of other quantities like True negative (TN), False positive (FP), and False negative (FN) follows from TP. Accuracy, precision, recall, and F1-score can be computed by equations below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision}(P) = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall}(R) = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} - \text{score} = \frac{2 * P * R}{P + R} \tag{4}$$

6 Discussion

In this section, we discuss the results obtained in both of the traditional weighing-based methods and neural networks models based on word embedding methods then we will compare the results with the results obtained by the proposed method.

Figure 3 shows the comparisons between the traditional word embedding methods for spam detection on

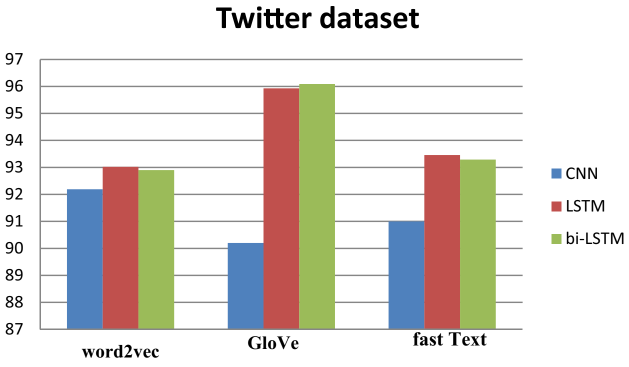

Fig. 3 Accuracy comparison of word embedding methods with different neural network architectures on the Twitter dataset

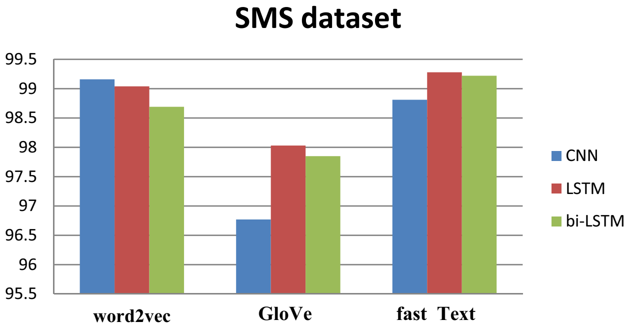

Fig. 4 Accuracy comparison of word embedding methods with different neural network architectures on the SMS dataset

Table 7 The test results of the proposed method on twitter dataset

Accuracy	Precision	Recall	F1-score
98.1	98	98.1	98.05

the Twitter dataset. The structures based on the representation of GloVe gave the best results compared to the rest of the models. As for the neural network structure, it appears that the LSTM-based structure is the best structure among all models. As for the SMS dataset, the LSTM structure based on FastText embedding achieved the highest accuracy compared to the rest of the structures, see Fig. 4.

As for the traditional weighing methods, as shown in Table 9, the performance of the Count vectorizer with Naive Bayes (NB) classifier outperforms the other experimented methods on the Twitter data set, while word-level TF-IDF achieved the best results in the SMS data set. On the other hand, by moving to the proposed method as shown in Table 7, the proposed method outperformed all previous methods with a remarkable performance reached 98.1% of accuracy and 98.05% of F1-score (Tables 8 and 9).

6.1 Comparison with other works

Table 10 shows the performance comparison between the proposed method and studies using similar methods available in the literature. Jain et al. [9] proposed an LSTM structure to detect spam messages on twitter and used word2vec for text representation with the help of knowledge bases to enhance the representation of text by adding more semantic. Madisetty et al. [23] used different word embedding methods such as word2vec and GloVe to represent text and fed into the CNN-based architecture to detect spam messages on twitter (Figure 5). On the other side, Ameen et al. [16] preferred the word2vec model to represent text and linear classification methods to distinguish spam from non-spam tweets.

7 Conclusions and future works

With the noticeable increase in the popularity of social networks in recent years, short texts classification problem has become quite common. One of the most common problems experienced by social networks' users is spam messages. An appropriate word representation could be a good solution to handle these messages.

Table 8 The accuracy results of word embedding models

Word embedding method		Word2vec			GloVe			Fast text		
Neural network architecture		CNN	LSTM	bLSTM	CNN	LSTM	bLSTM	CNN	LSTM	bLSTM
Accuracy (%)	Twitter dataset	92.19	93.02	92.9	90.2	95.93	96.09	91	93.46	93.29
	SMS spam dataset	99.16	99.04	98.69	96.77	98.03	97.85	98.81	99.28	99.22

CNN convolutional neural network, *LSTM* long short term memory neural network, *bLSTM* bi-directional long short term memory neural network

Table 9 performance evaluation of traditional weighting methods with Naive Bayes classifier

Dataset	Method	Accuracy	Precision	Recall	F1-score
Twitter dataset	Count vectors	0.927	0.95	0.92	0.94
	Word level TF-IDF	0.866	0.92	0.84	0.88
	N-gram level TF-IDF	0.818	0.80	0.92	0.86
	Character level TF-IDF	0.811	0.92	0.74	0.82
SMS spam dataset	Count vectors	0.974	0.87	0.93	0.90
	Word level TF-IDF	0.980	1.00	0.87	0.93
	N-gram level TF-IDF	0.976	1.00	0.81	0.89
	Character level TF-IDF	0.980	0.98	0.85	0.91

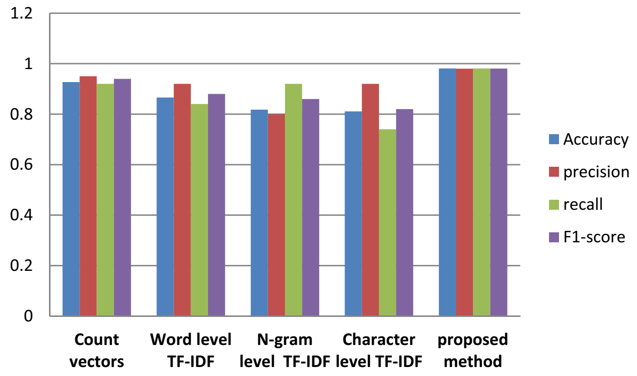

Fig. 5 Performance comparison of proposed approach with other traditional weighting methods on Twitter dataset

Traditional weighting methods and word embedding methods cannot handle the problem of short text classification effectively because of some previously mentioned limitations. Therefore, to overcome these limitations, novel context-dependent word representation methods were developed like CoVe, ELMO, and BERT.

In this study we proposed to use the BERT encoder to represent the text in the Twitter dataset, then we compared the performance with both traditional weighting methods and traditional word embedding methods. Finally, we compared the results of the proposed model with the results of some state of the art methods in the literature. We can conclude that the proposed method outperforms all other methods. to represent the text in the Twitter dataset, then we compared the performance with both traditional

7.1 Future works

As future work, we intend to improve the architecture of the proposed model and conduct an empirical study to tune the hyperparameters. Combining traditional weighting methods with contextual word representation could be a new open trend. Furthermore, other modern word representation methods could be integrated with deep learning algorithms.

Table 10 performance comparison of the proposed method with other famous existing methods

Work	Method	Dataset	Accuracy	precision	Recall	F1-score
[9]	CNN,word2vec, wordNet and conceptNet	twitter	0.944	–	–	–
[23]	CNN,word2vec and gloVe	HSpam14	0.957	0.880	0.909	0.894
		1KS10KN	–	0.922	0.876	0.893
[16]	Word2vec with linear classification methods	twitter	–	0.92	0.88	0.89
Our method	BERT encoder	twitter	**0.981**	0.98	0.981	0.9805

Acknowledgement The authors would like to thank Chen et al. [11] and UCI repository for sharing their datasets.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Yang C, Harkreader R, Gu G (2013) Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Trans Inf Forensics Secur 8(8):1280–1293
2. Lee S, Kim J (2014) Early filtering of ephemeral malicious accounts on Twitter. Comput Commun 54:48–57
3. El-Mawass, N. and S. Alaboodi, Hunting for spammers: Detecting evolved spammers on twitter. arXiv preprint arXiv:1512.02573, 2015.
4. Zheng X et al (2015) Detecting spammers on social networks. Neurocomputing 159:27–34
5. Kandasamy, K. and P. Koroth. An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques. In: 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science. 2014. IEEE
6. Miller Z et al (2014) Twitter spammer detection using data stream clustering. Inf Sci 260:64–73
7. Chen, W., et al. Real-time Twitter Content Polluter Detection Based on Direct Features. In: 2015 2nd International Conference on Information Science and Security (ICISS). 2015. IEEE
8. Wu, T., et al. Twitter spam detection based on deep learning. In: Proceedings of the australasian computer science week multi-conference. 2017
9. Jain G, Sharma M, Agarwal B (2018) Spam detection on social media using semantic convolutional neural network. Int J Knowl Discov Bioinf (IJKDB) 8(1):12–26
10. Gupta, H., et al. A framework for real-time spam detection in Twitter. In: 2018 10th International Conference on Communication Systems & Networks (COMSNETS). 2018. IEEE
11. Chen W et al (2017) A study on real-time low-quality content detection on Twitter from the users' perspective. PLoS ONE 12(8):e0182487
12. Lai, S., et al. Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence. 2015
13. Zhou, C., et al., A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630, 2015.
14. Zhou, P., et al., Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. arXiv preprint arXiv:1611.06639, 2016.
15. Wu T et al (2017) Detecting spamming activities in twitter based on deep-learning technique. Concurr Comput Pract Exp 29(19):e4209
16. Ameen, A.K. and B. Kaya Spam detection in online social networks by deep learning. In: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). 2018. IEEE.
17. Zrigui M et al (2012) Arabic text classification framework based on latent dirichlet allocation. J Comput Inf Technol 20(2):125–140
18. Chen Q. L. Yao, and J. Yang. Short text classification based on LDA topic model. In: 2016 International Conference on Audio, Language and Image Processing (ICALIP). 2016. IEEE
19. Li W, Sun L, Zhang D-K (2008) Text classification based on labeled-LDA model. Chin J Comp Chin Ed 31(4):620
20. Mateen M. et al. A hybrid approach for spam detection for Twitter. In: 2017 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST). 2017. IEEE
21. Uysal AK (2018) Feature selection for comment spam filtering on YouTube. Data Sci Appl 1(1):4–8
22. Barushka A, Hajek P (2019) Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks. Neural Comput Appl 32:4239–4257. https://doi.org/10.1007/s00521-019-04331-5
23. Madisetty S, Desarkar MS (2018) A neural network-based ensemble approach for spam detection in Twitter. IEEE Trans Comput Soc Syst 5(4):973–984
24. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018