Deceptive Opinion Spam based On Deep Learning

Fahfouh Anass
dept. LISAC
Faculty of Sciences Dhar El Mahraz
Fes, Morocco
anassfahfouh@gmail.com

Riffi Jamal
dept. LISAC
Faculty of Sciences Dhar El Mahraz
Fes, Morocco
riffi.jamal@gmail.com

Mohamed Adnane Mahraz
dept. LISAC
Faculty of Sciences Dhar El Mahraz
Fes, Morocco
adnane_1@yahoo.fr

Yahyaouy Ali
dept. LISAC
Faculty of Sciences Dhar El Mahraz
Fes, Morocco
ayahyaouy@yahoo.fr

Hamid Tairi
dept. LISAC
Faculty of Sciences Dhar El Mahraz
Fes, Morocco
htairi@yahoo.fr

***Abstract*—The revolution of web technologies and e-commerce platforms such as Amazon and eBay have promoted the businesses and have become essential in our daily life. Despite that these tools have helped the ease of purchases, there is a lot of scams in these kinds of technologies. Unfortunately, several companies use fake opinions to influence the customers about buying a product or to demote the competitors' one. The detection of deceptive opinion spam is a hard task because of the way it is written. The majority of existing deceptive opinion detection models focus on machine learning with hand-engineered feature extraction. Unfortunately, these architectures do not provide the semantic information of the reviews, which is the key to the detection of deceptive opinions. In this paper, we address the comparison between the different neural network architectures and their effectiveness in the detection of deceptive opinion spam. The results show that Convolutional Neural Networks perform better compared to other models.**

Keywords—Deceptive opinion spam; Deep Learning; Machine Learning.

I. INTRODUCTION

Numerous greedy companies enhance their products selling or diminish the competitors' ones by illegal methods. These companies hire malicious workers to generate a huge number of deceptive opinion spams which mimic real people's opinions in order to mislead the buyer's decision. Consequently, the promotion of deceptive opinions constitutes a real threat for both companies and the purchasers. The detection of deceptive opinion spam is considered a hard task because they are written in different ways to appear genuine. Due to the nature of deceptive opinion spams and their increasingly huge numbers, their detection remains a hard task for both humans and machines. Subsequently, powerful models are necessary for the detection of deceptive reviews.

Plenty of studies are suggested to detect deceptive opinion, most of them are based on traditional machine learning and hand-engineered feature extraction. An approach is proposed for the detection of deceptive opinion spams utilizing feature representations such as n-gram, part of speech, and Linguistic Inquiry and Word Count [1]. Several studies suggest that the detection of deceptive opinion spams as a stylistic classification where the truthful and the deceptive opinions have the same content but differ in the way they are written [2]. A novel model is proposed to identify misleading opinions using deep linguistic features derived from a syntactic dependency parsing tree [3]. A model is proposed for the classification of deceptive opinion spam using stylometric features, both lexical and syntactic utilizing supervised machine learning classifiers [4]. A new approach for the detection of deceptive opinion spams using the features extracted from the Context Free Grammar parse trees and the profile compatibility features, which enhances the detection performance and outperforms several baselines [5].

These studies show an improvement of accuracy which attained about 90%. However, there are still some issues that need real solutions. On the one hand, the growing number of opinions cannot be handled by traditional classifiers [6]. On the other hand, the sparsity of the extracted features doesn't keep the meaning of the reviews. According to [7] the clue for the detection of deceptive opinion spam is to mine the implicit information and to obtain the context of this information.

Neural network models likewise Convolutional Neural Networks CNNs, Recurrent Neural Networks RNNs, Long Short Term Memory LSTM, ETC are widely utilized in the text representation. These models have ameliorated several fields, for instance, computer vision, object recognition and natural language processing [8]. Neural network models are able to provide the semantic representation.

The remainder of the paper is organized as follows: in section two related work on deceptive opinion spams is explored. Section three gives insights on the different neural network architectures and methodology description. Section four describes the results of these architectures on the spam dataset. Finally, a conclusion is provided in section five.

II. RELATED WORKS

A. Deceptive opinion spam classification based on deep learning

In deep learning, Neural Networks are made of several layers to obtain representations of data with numerous levels of abstraction. These techniques have drastically enhanced the state-of-the-art in different fields such as speech recognition, natural language processing, visual object recognition and object detection. Deep learning models allow discovering complex structure in huge data sets by utilizing the backpropagation calculation to demonstrate how a machine should change its interior parameters that are utilized to compute the representation in each layer from the representation in the previous layer [9].

According to [10] a new neural network architecture based on the concatenation of a denoising autoencoder and a paragraph vector model is proposed in order to detect deceptive opinion spam. As seen in [11] a new model for the

978-1-7281-8084-7/20/$31.00 ©2020 IEEE

detection of deceptive reviews is provided, which is based on a recurrent convolutional neural network (RCNN) and the information obtained from word context. As demonstrated in [12] a sentence weighted neural network (SWNN) model based on Sentence CNN (SCNN) is efficient in learning the document level, where the representation of opinions is obtained by the score of each sentence. A new neural network model is provided to learn the representation of a document for detecting misleading reviews. First, a CNN model to represent the sentences. Second, a GRU model and discourse information to learn the document [13]. An attention-based on neural networks for the detection of deceitful opinions based on linguistic and behavioral features is proposed. First, the behavioral features are learned from a Multilayer Perceptron (MLP). Second, the linguistic features are obtained using a CNN [14]. A semi-supervised recursive autoencoders model is provided to predict social media spam reviews. Where the model learns the features form the sentences and their hierarchical structures by a given corpus of opinions [15]. A new CNN model based on glove word vectors is provided for opinion spams detection. Plus, some words and characters features extracted from texts are concatenated to the model to enhance the performance [16]. A hierarchical attention neural network model is proposed to learn the discourse information that captures the global and local semantic features in order to classify opinion spams [17]. A new approach is proposed which uses n-grams and skip-gram word embedding to construct a vector model, the latter is fed to a deep feed-forward neural network for spam reviews detection [18]. The majority of models ignore the impact of user's preference on review texts for opinion spams detection. a novel Fusion Convolutional Attention Network (FCAN) to embed the user-level information into a continuous vector space, the representations of which capture essential clues such as user profiles or preferences. Such user representation, in turn, facilitates learning better user-aware textual representation at word and sentence level [19]. An attention based deep neural network is proposed for the detection of truthful versus fake reviews [20]. According to [21] a new spamGAN model is provided which is a generative adversarial network that relies on a limited set of labeled data as well as unlabeled data for opinion spam detection. Saumya et al. [22] proposed an unsupervised learning model combining long short-term memory (LSTM) networks and autoencoder (LSTM-autoencoder) to distinguish spam reviews from other real reviews. The said model is trained to learn the patterns of real review from the review's textual details without any label. According to [23] a new approach is proposed to predict the best helpful online product review, out of the several thousand reviews available for the product using review representation learning. The review texts are embedded into low-dimensional vectors using a pre-trained model. To learn the best features of the review text, three filters are used to learn tri-gram, four-gram, and five-gram features of the text. The prediction is done using a two-layered convolutional neural network model. As seen in [24] a new neural network is proposed based on clustering and attention mechanism to learn the semantic representation of reviews. Specifically, DBSCAN is used to discover the semantic groups in the word embedding space, and then construct the semantics of different semantic groups through the attention mechanism. The model computes the representations of the semantic units and combines them into the sentence representation.

III. PROPOSED METHODOLOGY

A. Methodology description

The classification of deceptive opinion spam still needs further studies. In order to investigate the effectiveness of the different neural network architectures in this field, we follow the steps described in the following sections:

The first step is to pre-process the reviews from the deceptive spam dataset into a suitable format. Firstly, the stop words from each review will be removed such as ''the'', ''on'' as they do not contribute significantly to the classification task. Secondly, the lemmatization process is initiated in order to transform the words into their base forms. Thirdly, each word in the dataset will be transformed into a word embedding representation. Each opinion is a paragraph less than 500 words. For that reason, the reviews are fixed to a length of 500 words where every word is represented as a unique integer. Then, the word embeddings in the dataset are learned via Keras embedding layer. Each word will be embedded into a vector space of 100 dimensions. All the neural network architectures are trained according to binary cross-entropy loss and adam optimizer.

The second step is to examine the performance of the neural network models in the detection of deceptive opinion spam. The parameters and the hyperparameters are chosen based on several experiments since there are no accepted rules to choose the best ones and also they differ on the basis of the dataset characteristics. Four neural networks are described below:

a) Convolutional neural networks (CNNs).

CNNs are deep neural networks inspired by the organization of the Visual Cortex, they are first used in image processing. CNNs are different than a regular neural networks in terms of layers architecture. The neurons of the CNNs are arranged in 3 dimensions which are width, height and depth. In order to build CNNs architecture four types of layers are stacked: the Convolutional Layer, relu layer, Pooling Layer, and Fully-Connected Layer. Convolutional layers extract the feature using filters, these filters slide over the inputs and compute the dot product between the inputs and the entries of the filters producing convolved features. relu layer replaces all the negative feature in the convolved feature with zeros in order to introduce the non-linearity as the most real-world data are nonlinear. The aim of pooling layers is to reduce the dimensionality of each convolved feature in order to reduce the number of parameters and computation in the network, to control overfitting, also to maintain the significant features based on the max, the average, or the sum. The Fully-Connected Layer refers to that the neurons are fully connected to the next layer, classify the inputs from pooling layer to several classes. In this paper, The CNN model is composed of a convolutional layer of 128 filters, a kernel size of 50 with a relu activation and a global max pooling.

b) Recurrent neural networks (RNNs):

Traditional Neural Networks face a major issue which is the persistence of information. RNNs are deep neural networks to deal with this shortcoming. RNNs are powerful models for sequential data. The most important feature of

RNN is to have a memory that keeps the calculated information and provides them to all the layers in order to generate the output. In this paper, a fully-connected RNN with 10 cells and tanh activation function is used where the output to be fed back to the input.

c) Long short-term memory (LSTM):

RNNs suffer from two drawbacks the vanishing gradient and exploding problems, the parameters of the former layers are hard to learn. To deal with these issues, various networks were being proposed such as long short-term memory (LSTM), gated recurrent units (GRUs) and so on.

LSTM is a kind of RNNs, they can handle long term dependencies efficiently, and LSTM is composed of cells which are memory blocks. From one cell to the other, two states are transferred: the cell state and the hidden state. In each cell, the remembering and the handling mechanism are functioning based on three gates: the forget gate aims to remove the unnecessary information from the cell state via a multiplication filter in order to optimize the performance of the network. The input gate analyzes the information in order to keep the important information in order to add them to the cell state and to discard the useless ones. The output gate chooses the right information from the cell state and provides it as output. In this paper, the LSTM and Bidirectional LSTM are used with 100 units for each one of them, which is the dimension of the output with a tanh activation.

d) Gated recurrent units (GRUs):

Same as LSTM, GRU is an improved kind of the RNN. GRU is composed of two gates the updating gate and the reset gate. The first one determines the extent to which the new information is just the old information and by how much the obtained information to be used. The reset gate determines how much of past information to be reduced. In this paper, the GRU and Bidirectional GRU are used with 100 units for each one of them, which is the dimension of the output with a tanh activation.

In the third step, a comparative study between the performances of the different models. After the training, the features of each model are fed to fully connected layers in order to predict the opinion states either as truthful or deceptive. The architecture of the fully connected layers is two hidden layers based on the relu function which is used as an activation function for the two hidden layers. The sigmoid function is adopted in the output layer. Finally, in order to compare the different models. The performance metrics used are the accuracy, F1-score, recall, precision, ROC curve and the AUC metric.

IV. EXPERIMENT RESULTS

Our experiments are based on the dataset called the deceptive opinion spam by [1] in order to compare the state-of-the-art of the different deep learning models. It holds on 1600 reviews divided into four subsets: the first, 400 reviews are truthful positive review from TripAdvisor, the second 400 reviews are deceptive positive reviews from Mechanical Turk, the third 400 reviews are truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp, and the forth 400 reviews are deceptive negative reviews from Mechanical Turk. Before experimenting, we divide the dataset into only two subsets deceptive and truthful reviews.

To evaluate the results of the different models we employ precision, recall, accuracy and F1-score metrics. These metrics are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$\Pr ecision = \frac{TP}{TP + FP} \quad (2)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 * \frac{\Pr ecision * \mathrm{Re}call}{\Pr ecision + \mathrm{Re}call} \quad (4)$$

TP, FN, FP and TN represent the number of true positives, false negatives, false positives and true negatives.

The experiments are performed using the deep learning state-of-the-art models on the deceptive spam dataset. These models are implemented in the Keras 2.0 API with Tensorflow backend using Python 3.0.

In these experiments, the results are obtained using 5 folds cross-validation. The classification results are obtained using the average results of the 5 folds cross-validation on the basis of the metrics described above.

The neural network models utilized for the detection of deceptive opinion spam are RNN, LSTM, Bi-directional LSTM, GRU, Bidirectional GRU and CNN.

TABLE I. THE PERFORMANCE METRICS OF THE MODELS

Model	Accuracy	F1-score	Recall	precision
CNN	0.8729	0.8731	0.8750	0.8713
RNN	0.6958	0.6769	0.6375	0.7216
LSTM	0.7979	0.7999	0.8083	0.7918
Bi-LSTM	0.7917	0.8023	0.8458	0.7631
GRU	0.8000	0.8024	0.8125	0.7926
Bi-GRU	0.7979	0.7983	0.8000	0.7966

Table 1 shows the performance metrics used in the detection of deceptive opinion spam. The metrics show that the recurrent models are close in their performance, which is about 80% in the accuracy only RNN which is under all the models which is about 70% in the accuracy. The CNN model shows a good performance in comparison to all the models, the accuracy and F1-score of this model is about 87%.

The F1-score and accuracy of the CNN model are above the other state of the art models which demonstrate the effectiveness of the CNN model in the detection of the deceptive opinion spam. The results of the LSTM and Bi-LSTM models are very close, the same conclusion for GRU and Bi-GRU.

We can conclude that CNN has the ability to capture more complex features than the other models in this kind of task

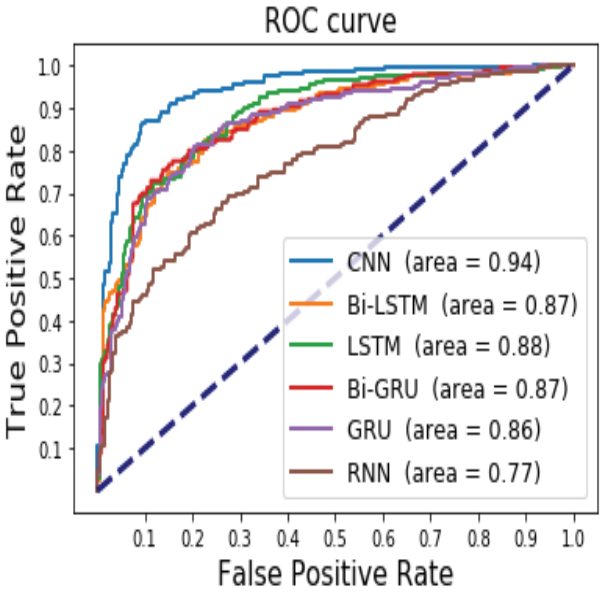

Fig 1: the Roc curve and the AUC metric of the models

Fig. 1 shows the ROC curve and the AUC metric which refers to the area under the Receiver Operating characteristic curve. The ROC curve of CNN model is above the other models also the AUC metric is better than the other models with a value of 0.94. The recurrent models with an AUC value of 0.85 except RNN, which is under 0.68, which can be explained by the vanishing problem.

For both measures the performance metrics and the AUC, we can conclude that the CNN model is more performant than the other models in the detection of deceptive opinion. This can be explained by the ability of CNN in capturing the relationships between the local words.

V. Conclusion

In this paper, we developed and compared several neural network models to identify the deceptive opinion spams. Our experiments compare the different models based on the same dataset in order to have a fair comparison between the models. The results clearly show that the CNN model is superior in comparison to the different models. Which is proved by the high accuracy and F1-score. The performance of CNN relies on its ability to extract high-level features. RNN based models are able to extract contextual information. However, an attention mechanism is necessary in order to enhance the detection accuracy. We believe that deep learning models can find more complex structures and patterns in the detection of deceptive opinions spam. In future studies, we certainly will explore more models based on the deep learning architectures to fit real case studies. In addition, we will focus on the behavior of the spammers to enhance the performance of the detection.

REFERENCES

[1] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding Deceptive Opinion Spam by Any Stretch of the Imagination,” Proc. 49th Annu. Meet. Assoc. Comput. Linguist., p. p (309–319), 2011.

[2] S. Feng, R. Banerjee, and Y. Choi, “Syntactic Stylometry for Deception Detection,” Proc. 50th Annu. Meet. Assoc. Comput. Linguist. pages 171–175, Jeju, Repub. Korea, 8-14 July 2012., no. July, pp. 171–175, 2012.

[3] Q. Xu and Hai Zhao, “Using Deep Linguistic Features for Finding Deceptive,” Proc. COLING December 2012 Mumbai, pp. 1341–1350, 2012.

[4] S. Shojaee, M. Azrifah, A. Muradt, A. Bin Azman, N. M. Sharefl, and S. Nadali, “Detecting Deceptive Reviews Using Lexical and Syntactic Features,” in 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013, pp. 53–58.

[5] V. W. Feng and G. Hirst, “Detecting deceptive opinions with profile compatibility,” Int. Jt. Conf. Nat. Lang. Process., no. October, pp. 338–346, 2013.

[6] M. Crawford, T. M. Khoshgoftaar, and J. D. Prusa, “Reducing Feature Set Explosion to Facilitate Real-World Review Spam Detection,” Proc. Twenty-Ninth Int. Florida Artif. Intell. Res. Soc. Conf., pp. 304–309, 2016.

[7] L. Dong, S. Ji, C. Zhang, Q. Zhang, L. Qiu, L. Dong, S. Ji, C. Zhang, Q. Zhang, and L. Qiu, “An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews,” Expert Syst. Appl., 2018.

[8] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent Trends in Deep Learning Based Natural Language Processing,” arXiv1708.02709v8 [cs.CL] 25 Nov 2018, pp. 1–32, 2018.

[9] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. Macmillan Publishers Limited, 2015.

[10] A. Fahfouh, J. Riffi, M. A. Mahraz, A. Yahyaouy, and H. Tairi, “Expert Systems with Applications PV-DAE : A hybrid model for deceptive opinion spam based on neural network architectures,” Expert Syst. Appl., vol. 157, 2020.

[11] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, “DRI-RCNN : An approach to deceptive review identi fi cation using recurrent convolutional neural network,” Inf. Process. Manag., vol. 54, pp. 576–592, 2018.

[12] L. Li, B. Qin, W. Ren, and T. Liu, “Document representation and feature combination for deceptive spam review detection,” Neurocomputing, vol. 0, pp. 1–9, 2017.

[13] Y. Ren and D. Ji, “Neural networks for deceptive opinion spam detection : An empirical study,” Inf. Sci. (Ny)., vol. 385–386, pp. 213–224, 2017.

[14] X. Wang, K. Liu, and J. Zhao, “Detecting Deceptive Review Spam via Attention-Based Neural Networks,” NLPCC 2017, LNAI 10619, pp. 866–876, 2018.

[15] B. Wang, J. Huang, H. Zheng, and H. Wu, “Semi-Supervised Recursive Autoencoders for Social Review Spam Detection,” 2016 12th Int. Conf. Comput. Intell. Secur., pp. 116–119, 2016.

[16] K. Archchitha and E. Y. A. Charles, “Opinion Spam Detection in Online Reviews Using Neural Networks,” in 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 2019., 2019.

[17] C. Jiang and X. Zhang, “Neural Networks Merging Semantic and Non-semantic Features for Opinion Spam Detection,” in CCF International Conference on Natural Language Processing and Chinese Computing, 2019, pp. 583–595.

[18] A. Barushka and P. Hajek, “Review Spam Detection Using Word Embeddings and Deep Neural Networks,”

in IFIP International Federation for Information Processing 2019 Published by Springer Nature Switzerland AG 2019, 2019, pp. 340–350.
[19] J. Li, Q. Ma, C. Yuan, W. Zhou, J. Han, and S. Hu, "Fusion Convolutional Attention Network for Opinion Spam Detection," in International Conference on Neural Information Processing, 2019, pp. 223–235.
[20] Z. Sedighi, H. Ebrahimpour-komleh, A. Bagheri, and L. Kosseim, "Opinion Spam Detection with Attention-Based Neural Networks," in The Thirty-Second International Florida Artificial Intelligence Research Society Conference (FLAIRS-32), 2019, pp. 245–248.
[21] G. Stanton and A. A. Irissappane, "GANs for Semi-Supervised Opinion Spam Detection," Proc. Twenty-Eighth Int. Jt. Conf. Artif. Intell., pp. 5204–5210, 2019.
[22] S. Saumya and J. Prakash, "Spam review detection using LSTM autoencoder : an unsupervised approach," Electron. Commer. Res., no. 0123456789, 2020.
[23] S. Saumya, J. P. Singh, and Y. K. Dwivedi, "Predicting the helpfulness score of online reviews using convolutional neural network," Soft Comput., no. BrightLocal 2016, 2019.
[24] J. Zhang, X. Du, and B. Wang, "Semantic Representation Based on Clustering and Attention Mechanism to Identify Deceptive Comment Models," J. Comput., vol. 30, no. 4, pp. 130–139, 2019.