Expert Systems with Applications 157 (2020) 113517

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

PV-DAE: A hybrid model for deceptive opinion spam based on neural network architectures

Anass Fahfouh, Jamal Riffi, Mohamed Adnane Mahraz, Ali Yahyaouy, Hamid Tairi

LIIAN Laboratory, Faculty of Sciences Dhar El Mahraz, University Sidi Mohamed Ben Abdelah, Morocco

ARTICLE INFO

Article history:
Received 16 March 2019
Revised 20 February 2020
Accepted 2 May 2020
Available online 7 May 2020

Keywords:
Deceptive opinion spam
Neural networks
Machine learning
Deep learning
Paragraph vector model
Denoising autoencoder model

ABSTRACT

Opinion review is of great importance for both customers and organizations. Indeed, it helps customers in buying decisions and represents a valuable feedback for the companies, allowing them to improve their productions. However, numerous greedy companies resort to fake reviews in order to influence the customer and brighten the brand image, or to defame the one of their competitors. Various models are proposed in order to detect deceptive opinion reviews. Most of these models adopt traditional methods focusing on feature extraction and traditional classifiers. Unfortunately, these models do not capture the semantic aspect while ignoring the opinion's context. In order to tackle this issue, we propose a new approach based on Paragraph Vector Distributed Bag of Words (PV-DBOW) and the Denoising Autoencoder (DAE). The proposed customized model provides a strong representation which is based on a global representation of the opinions while preserving their semantics. Indeed, the embedding vectors capture the semantic meaning of all words in the context of each opinion. The generated review representations are fed into a fully connected neural network in order to detect deceptive opinion spam. The obtained results concerning the deception dataset show that our model is effective and outperforms the existing state-of-the-art methodologies.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

The revolution of web technologies has drastically changed global communication. Indeed, many technologies and systems have swept the daily life of humans and improved their living standard. For instance, the different E-commerce platforms and the social networks have influenced people's lifestyle which has motivated the majority of companies to invest in the digital marketing of all sorts and kinds. Unfortunately, numerous enterprises hire malicious workers in order to change the people's views about the existing products either to promote some products or to relegate competitors' ones. These workers try to manipulate people's opinions by generating a huge amount of deceptive reviews.

Concretely, the purchase decision is influenced by the feedback of the people about a product. Indeed, buyers usually refer to the existing reviews to take the right decisions. These deceptive opinions represent a real threat to both companies and consumers, which makes the detection of fake opinions an urgent task.

Several types of opinion spam are generated by anonymous people for different aims and in different manners. Specifically, they are established in short text such as spam reviews, malicious comments, malicious blogs, malicious social network postings and deceptive texts. Deceptive opinion spams are in some sort more deceitful and mendacious than the above cited spams; they are written in different ways to appear genuine in order to mislead the facts and to delude the purchasers (Fusilier, Montes-y-Gómez, Rosso, & Cabrera, 2015). Due to the nature of deceptive opinion spams and their increasingly huge numbers, their detection remains a hard task for both humans and machines. Ott, Choi, Cardie, and Hancock (2011) stated that it is almost impossible to distinguish deceptive from truthful ones in a satisfactory manner, and only 57.33% is the average accuracy of three human judges. Consequently, additional efforts are required to detect misleading opinion spam to ensure the Web's credibility.

Several studies have been suggested to detect misleading opinion spam, and the famous ones are based on feature extractions which describe the involved reviews. These characteristics are used in order to train a classifier that will be used to predict unlabeled deceptive opinion as spam or ham. Ott et al. (2011) provided three approaches to detect deceptive opinion spam while using three feature representations: ngrams, part of speech (POS) and Linguistic Inquiry and Word Count (LIWC). Using the same features Li, Ott, Cardie, and Hovy (2014) built an additive model to recognize the general rule for deceptive opinion spam detection.

E-mail address: anassfahfouh@gmail.com (A. Fahfouh)

https://doi.org/10.1016/j.eswa.2020.113517
0957-4174/© 2020 Elsevier Ltd. All rights reserved.

In addition, Feng, Banerjee, and Choi (2012) investigated on syntactic stylometry for opinion deception detection in which a large number of researches consider the detection of opinion spam as a stylistic classification, deceptive and truthful opinions have similar content but differ in the way opinions are written. Despite the good performance of these types of models, Ren and Ji (2017) stated that the linguistic features are very sparse, which makes it very hard to obtain the semantic information. Also, Dong et al. (2018) confirmed that these models did not consider the implicit information from the reviews and mining the implicit information is the key to the detection of opinion spam. As to Zhang, Du, Yoshida, and Wang (2018) emphasized that the context information is the important clue to classify a review as deceitful or truthful. Hence, several probabilistic models and hybrid models are provided in order to discover the implicit information in the reviews, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are the main techniques which are used in several models and by several authors. Ren, Ji, and Zhang (2014) proposed a semi-supervised model, mixing population and individual property PU learning (MPIPUL) technique to identify deceptive reviews using LDA. As to Jia, Zhang, Wang, and Liu (2018) proposed a model based on LDA to extract the features. LDA, LSA and Probabilistic LSA (PLSA) models are a kind of topic modeling, they inspect the relationships between words and topics, also documents and topics. Dong et al. (2018) proposed an unsupervised topic-sentiment joint probabilistic model (UTSJ) based on LDA model in order to detect deceptive reviews and mine the topics and sentiments from the reviews. Hernández-Castañeda, Calvo, Gelbukh, and Flores (2016) proposed a continuous semantic space model source represented by the LDA topics, a word-space model, and dictionary based features. A performance comparison between semantic information and behavioral information is made testing several combinations of these features on different datasets designed to identify deception. Zhang, Xu, Yang, Chen, and Ye (2018) stated that these models face several challenges as the requirement of a predefined number of topics, also it is difficult to explain the decomposed matrices and the extracted distributions. In addition, they declared that these models fail to detect the neighboring information of words in a sentence.

To address the above limitations, we propose a new approach based on DAE and PV-DBOW considered as two types of neural network architectures which are characterized by their potential to capture the complex global information and ensuring a highly semantic representation.

The main steps of our approach are: first, we extract the embedding of the involved reviews using two models PV-DBOW and the DAE. Second, we concatenate the embedding feature representation of the two models. Third, we feed them to a fully connected layer.

The rest of the paper is arranged as follows: the next section explores the related work. Section 3 provides a description of the proposed approach. Experimental results are presented in Section 4. And finally, in Section 5 we conclude the paper.

2. Related work

A bunch of models learn discriminant features from the reviews using traditional classification models such as Support Vector Machine (SVM), Naïve Bayes (NB), ETC. These models are illustrated by several works and by different authors. Cagnina and Rosso (2015) provided a model based on three different kinds of features a character n-gram in token, Emotion based features and Linguistic Inquiry Word Count (LIWC) based features which are fed to an SVM classifier. Xu and Zhao (2012) proposed a novel model which the features extracted are deep linguistic derived from a syntactic dependency parsing tree, the Features are used to train SVM and maximum entropy (ME) classifiers. Feng et al. (2012) investigated on syntactic stylometry for opinion deception detection and proved that the features driven from Context Free Grammar (CFG) parse trees consistently improve the detection performance over several baselines that are based only on shallow lexico-syntactic features. Feng and Hirst (2013) extended the proposed model by Feng et al. (2012) in which incorporated profile compatibility features. Also, Shojaee et al. (2013) applied stylometric feature, lexical and syntactic using SVM, Sequential Minimal Optimization (SMO) and NB. Fusilier et al. (2015) obtained the lexical content as well as stylistic information using the character ngrams along with a NB classifier. Hafiz, Siagian, and Aritsugi (2017) combined character ngrams and word ngrams as a feature for detecting positive and negative deceptive opinions using a regression analysis classifier. Patel and Thakkar (2014) proposed an opinion spam detection method extending unigram, bigram and bigram + sequence of words approaches. For each of these approaches, opinions are modeled as Boolean, bag-of-words and term frequency-inverse document frequency (TF-IDF) vectors, NB and Least Squares LS-SVM which are used for the classification task. Mani, Kumari, Jain, and Kumar (2018) proposed a spam review detection algorithm using n-gram (unigram + bigram) features using two ensemble techniques Simple Majority Voting ensemble (Voting) and a Stacked Ensemble to provide a better classification. Perez-Rosas et al. (2015) elaborated a multimodal deception detection on the basis of a combined sets of verbal and nonverbal features. These features are fused from the linguistic and visual modalities for the detection of deceptive opinion spam, which is based on three classifiers SVM, Random Forest (RF) and Decision tree (DT). Saini, Verma, and Sharan (2018) designed a set of features, psychological, linguistic, and other textual features from text reviews, then built a Rough Set Based Optimal Feature Set Partitioning (RS-OFSP) algorithm to construct views for Multi-view Ensemble Learning (MEL). Molla, Biadgie, and Sohn (2018) aimed at the detection of negative deceptive opinions from tweets by combining Syntactic, lexical, content specific and structural features which are the common features used for general text analysis, personal profile and behavioral features of the writer using the NB classifier. Based on three textual characteristics: readability of a review, review genre and review writing Banerjee and Chua (2014) proposed a framework for the identification of authentic online reviews. The above work has been extended by Banerjee, Chua, and Kim (2015) where a supervised learning based on four linguistic clues, namely, understandability, level of details, writing style, and cognition indicators, then ten supervised learning algorithms are used for the analysis.

The number of review sites and reviews has been increased drastically in the recent years. The manipulation and exploration of such amount of data cannot be guaranteed by traditional machine learning algorithms based on limited feature representation (Crawford, Khoshgoftaar, & Prusa, 2016). Currently, Deep learning methods learn hierarchical representations through several processing layers, and have been adopted in many domains. Recently, various models and methods have blossomed in the context of natural language processing (NLP) (Young, Hazarika, Poria, & Cambria, 2018). These models manipulate high dimensional data and handle various types of features in an automatic manner and ensure the semantic representation.

One of the most famous techniques used in deep learning for NLP tasks is word embedding. The aim of word embedding is language modeling and feature learning, which maps words in a vocabulary to vectors of numerical values (Zhang, Wang, & Liu, 2018). Among the most influential models, the word2vec model which preserves the syntactic and the semantic relationships between words (Mikolov, Corrado, Chen, & Dean, 2013). The majority of the proposed models based on deep learning try to provide a representation of the opinion using word representation. However, word representations face an inherent limitation which is their inability to represent phrases. In addition, The use of a small window to generate word embedding may lead to the same embedding for different words (Socher, Pennington, Huang, Ng, & Manning, 2011; Young et al., 2018), Which is likely to bring about a serious issue as far as the sentiment analysis tasks are concerned (Xin Wang, Liu, Sun, Wang, & Wang, 2015; Young et al., 2018). Sometimes, these embeddings cluster semantically similar words which have opposite sentimental polarities. As a result, the downstream model applied for the sentiment analysis task can hardly identify this contrasting polarities, which leads to imperfect performance (Young et al., 2018). The same as deceptive opinion spam, a deceptive opinion is either positively deceitful or negatively deceitful, a poor representation of opinion polarities may mislead the detection of deceptive opinion spam. In order to address these issues several models for sentiment encoded word embedding have been provided. Tang et al. (2014) proposed to learn sentiment specific word embedding (SSWE) by integrating the sentiment information into the loss functions of three neural networks. As to Ren, Zhang, Zhang, and Ji (2016) proposed Topic and Sentiment enriched Word Embedding model for learning topic-enriched multi-prototype word embedding (TMWE) by using two neural networks the first one learns word embeddings from tweets by integrating topic information and the second one learns topic-enriched multiple prototype embeddings for each word. Several studies assumed the words contained in a tweet have the same sentiment polarity as that of the entire tweet, which ignores the sentimental polarity of the word (Xiong, Lv, Zhao, & Ji, 2017). In order to solve this problem, Xiong et al. (2017) developed a multi-level sentiment-enriched word embedding learning method, which utilizes a parallel asymmetric neural network to model n-gram, word-level sentiment, and tweet-level sentiment. Lauren, Qu, Yang, Watta, and Amaury (2018) proposed a new model to generate word representation using an autoencoder architecture based on extreme learning machine (ELM) that operates on a word context matrix. Yu, Wang, Lai, and Zhang (2017) proposed a word vector refinement model that can be applied to any pre-trained word vectors. Where a set of semantically similar nearest neighbors are selected, then classifying similar neighbors higher and dissimilar neighbors lower based on a sentiment lexicon.

Only a few models have been adopted using deep learning and neural network architectures for detecting deceptive opinion spam.

Zhang et al. (2018) proposed a novel approach for the identification of deceptive review based on a recurrent convolutional neural network called DRI-RCNN. This model obtains the word vectors from the skip-gram model, then in order to capture the contextual local information of each word, a recurrent convolutional neural network is adopted. Wang and Chen (2018) provide a model to detect spam reviews using long short-term memory (LSTM) using a variety of parameters included activation function, loss function, optimizers and dropout to test the detection performance. Arguing that deep learning methods LSTM got better performance than traditional machine learning method. Zhao, Xu, Liu, Guo, & Yun, 2018 optimized the convolutional neural network (CNN) model by embedding the word order characteristics in its convolution layer and pooling layer, which makes CNN more suitable for short text classification and deceptive opinion detection. Li, Qin, Ren, and Liu (2017) provided a new model called sentence weighted neural network (SWNN) model based on Sentence CNN (SCNN), the document level representation of each review is learned based on the importance weights of each sentence. Ren and Ji (2017) proposed a new model based on neural network in order to learn the representation of a document for identifying review spam. First, a representation of sentences is provided using CNN. Second, the document vector is constructed by combining the representation of sentences based on a gated recurrent neural network respecting their semantic and the discourse information. Wang, Huang, Zheng, and Wu (2016) proposed a semi-supervised recursive autoencoder neural network model in order to identify social media opinion spams. This model tries to learn the feature vectors form the sentences and their hierarchical structures from the given opinions. The detection of opinion spam is not just an NLP problem while it requires the analysis of people's behavior (Liu, 2012). Wang, Liu, and Zhao (2018) proposed an attention-based neural networks to detect deceptive review spam based on linguistic and behavioral features. Firstly, the behavioral feature vectors are obtained from a Multilayer Perceptron (MLP). Secondly, the linguistic features are acquired by a CNN. After that, a feature attention module is built in order to learn how a review is behaviorally or linguistically suspicious. As to Cambria (2016) stated that affective computing and sentiment analysis have a massive capacity as a subcomponent technology for various systems. They are able to optimize the functionalities of customer relationship management and recommendation systems. Likewise, they might be used for a more or less affective tutoring as well as an affective entertainment or for troll filtering and spam detection in online social communication. Several models have been proposed for affective computing and sentiment analysis. Cambria, Poria, Hazarika, and Kwok (2018) proposed senticNet5 that generates conceptual primitives from texts and links them to common sense concepts and named entities in a new representation of knowledge at three levels for the analysis of sentiments. The supremacy of this model lies in the fact that it is no longer necessary to presume the polarity based on emotional relationships (direct or indirect) in the semantic network of commonsense knowledge; affective reasoning is carried out at the primitive level. Bandhakavi, Wiratunga, Massie, Gordon, and Padmanabhan (2017) considered the document as a mixture of emotional and neutral words, and provided a generative unigram mixture model (UMM) for learning a domain-specific word-emotion lexicon from an input document corpus. Poria, Peng, Hussain, Howard, and Cambria (2016) suggested a multimodal affective data analysis in order to obtain user opinion and emotion from video content. It comprises the extraction of salient features, the development of unimodal classifiers, building feature- and decision-level fusion frameworks. A multiple kernel learning is used to combine visual, audio and textual modalities.

After being proposed by Sabour, Frosst, and Hinton (2017), capsule network has been widely adopted in many NLP tasks, specifically in sentiment analysis. Wang, Sun, Han, Liu, and Zhu (2018) stated that the existing neural network models face some limitation as they focus on the quality of the instance e.g. sentence, paragraph. The use of a vector to represent sentiment is very restricted as opinions are delicate and complex. For this reason, an RNN-Capsule that is a capsule model based on Recurrent Neural Network (RNN) for sentiment analysis is proposed. Zhang et al. (2018) provided a Capsule network for sentiments in the Domain Adaptation scenario with Semantic Rules (CapsuleDAR). In addition, they proposed a network of rules to integrate semantic rules into the network of capsules in order to improve the comprehensive sentence representation learning. Du, Zhao, He, and Guo (2019) proposed a capsule-based hybrid neural network model for sentiment classification of short text that effectively capture the implicit semantic information. A Bidirectional gated recurrent unit (BGRU) is applied in this model to reach the interdependent features with long distance. Despite the importance of capsule network, it is still not applied in the detection of deceptive opinion spam.

Another subject that attracts particular attention in sentiment analysis concerns the aspect extraction which aims to identify opinion targets in opinionated text, i.e. to discover the particular aspects of a product or service that the opinion holder praises or complains about (Poria, Cambria, & Gelbukh, 2016). In this kind of task, Deep learning can be helpful due to its ability in the extraction of complex features. Which can model the correlation or the semantic between the aspect and its context (Zhang et al., 2018). Poria et al. (2016) elaborated a model based CNN architecture for aspect extraction. It comprises seven layers: the first one, the word sentences are converted into word embedding which are trained on two corpora; two convolution layers, each followed by a max-pooling layer; a fully connected layer; and finally the output layer, which contains one neuron per each word. In addition, a bunch of heuristic linguistic patterns are developed and incorporated with the deep learning classifier. Wu, Wu, Wu, Yuan, and Huang (2018) proposed a hybrid unsupervised method which is based on rules and machine learning techniques for aspect term extraction and opinion target extraction tasks. Firstly, a chunk-level linguistic rules are used to retrieve nominal phrase chunks and consider them as candidate opinion targets and aspects. Secondly, they filtered irrelevant candidates based on domain correlation. Thirdly, they used these texts with extracted chunks as pseudo labeled data to train a deep gated recurrent unit (GRU) network. Yang, Zhang, Jiang, and Li (2019) proposed a coattention-LSTM network, which models both target-level and context-level attention alternatively. This model learns the nonlinear representation of both the context and the target which provide better feature representation. The aspect level is widely used in sentiment analysis, in contrast to deceptive opinion spam where the word level, sentence level or the document level is used.

The main contribution of this paper can be outlined as follows: in the first step, we adopt two models, paragraph vector model to learn the representation of the reviews while focusing on the document level and the denoising autoencoder (DAE) model to learn the latent features. In the second step, we concatenate both feature representations generated from the two models, and we feed them to fully connected layers in order to predict opinion spam. Finally, to prove the performance of our model we perform experiments on a deceptive opinion spam corpus from (Ott et al., 2011) and we compare the results with the existing ones.

3. Methodology

Our proposed approach is based on two neural network models involving the Paragraph vector distributed bag of words (PV-DBOW) and the denoising autoencoder. The following section provides a description of these two models.

3.1. The PV-DBOW model

Paragraph Vector, known as Doc2vec, is an unsupervised neural network model for modeling pieces of texts such as sentences, paragraphs or documents into a numerical format. Each document is converted to a low dimensional vector trained to predict words in the document. The Paragraph Vector model is inspired from the famous word vector model (Mikolov & Le, 2014). The word vector representation model learns words embedding in order to capture the closeness between them. It comes in two flavors, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model (Mikolov et al., 2013). There is a slight difference between these two algorithms: CBOW predicts the target word from the context word, while Skip-gram predicts the context words from the target word. Since Paragraph Vector is based on word vector model it comes likewise in two algorithms Paragraph-vector distributed memory (PV-DM) and PV-DBOW. In this paper, we will focus only on the PV-DBOW model as it is a principal component in the proposed model.

The PV-DBOW model which provides document features is similar to the skip-gram model that generate word vectors. The skip-gram model can be trained in order to minimize a well-designed loss function. In a set of training words $w_1, w_2, w_3, \ldots\ldots, w_T$, the aim is to maximize the following function which is the average log probability.

$$P = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{-c\leqslant j\leqslant c, j\neq 0}\log p(w_{n+j}/w_n)\right) \tag{1}$$

The first summation represents the words from the training corpus. The second summation computes the log probability of predicting the word context w_{n+j}, knowing the input word w_n, −c and c are the left and right context. In order to compute $p(w_{n+j}/w_n)$ the skip-gram model adopts several methods which depend on the computation complexity, such as the hierarchical softmax, negative sampling or the softmax function which is defined as follows:

$$p(w_{n+j}/w_n) = \frac{\exp(u_{w_{n+j}}^{T} v_{w_n})}{\sum_{v=1}^{V}\exp(u_v^{T} v_{w_n})} \tag{2}$$

The input vector takes various words in order to predict the context word. Where u_w and v_w are the input and output vector representations of the word w. And v is the number of words in the vocabulary. For the optimization method the skip-gram model uses gradient descent or stochastic gradient descent. First the weights of the network are generated randomly after that the model is trained by the back propagation which calculate the derivative of the loss function and the weights. Finally, for each word in the vocabulary is represented by a low dimension vector called word embedding.

The PV-DBOW uses only paragraph ids as inputs to identify the documents and ignore the contextual words in the input. Then the model is imposed to predict words which are randomly sampled from the paragraph in the output (Mikolov & Le, 2014). The distributional probability of the PV-DBOW is likewise the distributional probability of the skip-gram with a small change where w_n is replaced by the document D_n and $p(w_{n+j}/D_n)$.

3.2. Denoising autoencoder

A basic autoencoder is an unsupervised neural network, which aims to represent the inputs in a more condensed way and reconstruct its inputs. It consists of three layers, the first half represents the encoding layers and the second half represents the decoding layers. In this model, the number of input and output nodes is equal (Bengio, 2009).

An autoencoder can be defined as follows: take an input $x \in [0,1]^{d}$ the encoding layers convert it to a hidden format $y \in [0,1]^{d'}$ called latent representations where $y = \varphi(Wx + b)$ and φ is an activation function (e.g. the sigmoid function). The decoding layers convert the latent representations to a new format ϕ which is a reconstruction of the first inputs $x' = \phi(W'y + b')$ the parameters of the decoding layers $\phi, W'\ and\ b'$ could be chosen as the parameters of the encoding layers. The autoencoder is trained in order to minimize the reconstruction error which can be measured in several techniques for instance the squared error or the cross entropy, which can be defined as follows:

$$H(x,x') = -\sum_{k=1}^{n}\left[x_k \log x'_k + (1-x_k) log(1-x'_k)\right] \tag{3}$$

In order to prevent autoencoders from learning the identity function the number of units in the hidden layers should be less than the input layer. Furthermore, to ameliorate the discovery of rich features, various techniques are proposed such as the sparsity and the randomness. The latter is the basis of the denoising autoencoders (DAE) (Vincent, Larochelle, Bengio, & Manzagol, 2008). The inputs x are corrupted using stochastic mapping to inputs $\tilde{x}$ and the DAE try to predict the corrupted values from these transformed values. The DAE model can better preserve the semantic of the documents than the autoencoder.

3.3. Proposed approach

The following figure Figure 1 describes the proposed model encompassing three steps:

The first step concerns the dataset preprocessing. First, the tokenization process which is the task of chopping up the whole text in opinions into pieces in order to obtain words which produce a set of tokens. Second, for more accurate classification of the opinions the stop words in spam and non-spam opinions will be removed such as "the", "in", etc. Third, words are transformed into their base forms which is called lemmatization. Finally, all the opinions will be tagged by a unique id to differentiate between them, which is required in the PV-DBOW process.

The second step aims to generate the feature embedding while adopting the PV-DBOW and the DAE models. The PV-DBOW model transforms each tagged opinion into a numerical format in a low dimensional space. Let O be a representation of a set of opinions where $O = \{O_1, O_2, ...O_n\}$ a corpus of opinions, every $O_i \in O$ where $i = \{1, 2...., K\}$ is the vector representation of an opinion which is a one hot encoded vector and let D be the weight matrix of the network where D is $E \times N$ matrix, E is the dimension of the middle layer and N is the number of documents (opinions). The activation function is just the weighted sum of the input layer. The document embedding of each opinion O_i can be represented as $Ed_i = D * O_i$. The PV-DBOW model try to predict a subset of words from the words given in the document. P represents the number of unique words in the vocabulary, the output vectors K correspond to the words which can be represented as $K = M * Ed$ where M is $E \times P$ the weight matrix between the middle and the output layer. After the break of the training the documents embedding will be retrieved from the middle layer. On the other hand, the denoising autoencoder is performed to provide a distributed representation for each opinion. Let O_i be the original input and $\tilde{O}_i$ the corrupted input of the DAE model. The corrupted opinion $\tilde{O}_i$ is transmitted to a hidden layer which provides the latent representation $Y_i = \tanh(W\tilde{O}_i + b)$, the weight matrix W and b the bias are initialized randomly. The obtained representation is passed to the decode layer reconstructing the vector $Z \in [0, 1]^d$ representing the input opinions where $Z_i = \tanh(W'y + b')$, after the training the embedding of the documents is obtained from the middle layer.

In the third step, in order to benefit from both representations, we concatenate the two representations issued from both models. After that, the resulted representations are fed to a fully connected layers aiming at predicting opinions either as truthful or deceptive. The fully connected layers include four hidden layers. The RELU function is used as an activation function for the hidden layers. As for the output layer a sigmoid function is adopted.

4. Experiment results and comparative study

Deceptive opinion spam datasets are categorized into four categories based on the different construction methods: rule-based method, human-based method, filtering algorithm based method and AMT (Amazon Mechanical Turk) based method (Ren & Ji, 2019). However, the construction of these datasets faces one of the research challenges of deceptive opinion detection, which is the deficiency of labeled dataset. There exists only one dataset created by Ott et al. (2011) with true gold standard for product opinion domain (Saumya & Singh, 2018). In our experiments, we use the famous dataset called the deceptive opinion spam or the gold standard dataset created by Ott et al. (2011). It contains 1600 reviews divided in four parts: 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. In this experiment, we divide this dataset into two parts deceptive reviews and truthful ones. For each review, we conduct the preprocessing task: firstly, we remove stop words from each review. Secondly, we tokenize the reviews. Thirdly, we lemmatize the reviews, and finally we tag the reviews for the PV-DBOW model.

A set of experiments are conducted according to Zhang et al. (2018) settings using the preprocessed opinions. Firstly, we split

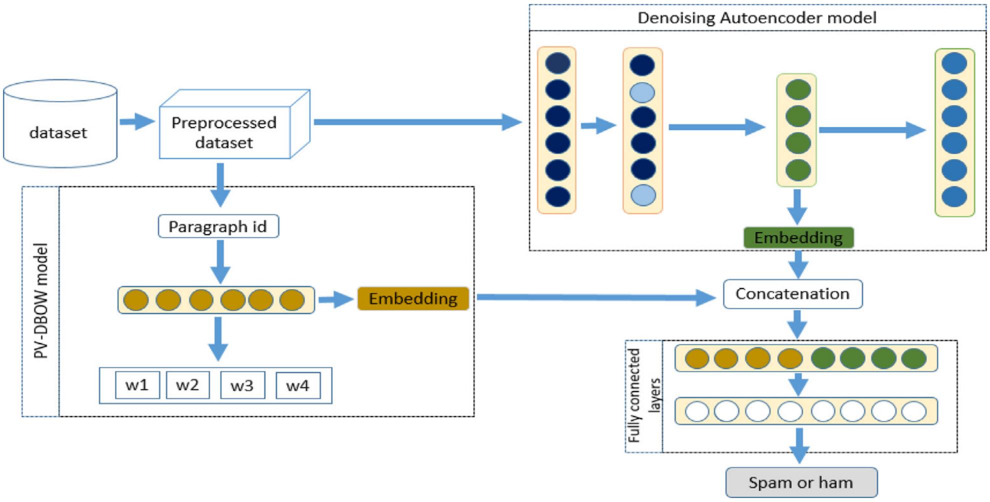

Fig. 1. The architecture of the PV-DAE model.

the dataset into training set and test set. Each time, we vary the training set from 50% to 90% and the remaining data is used for the test. Then each training set is divided into five folds. Four folds are used for the training and one fold for the validation in order to choose the hyper parameters of our model. Finally, the hyper-parameters are obtained from the best average accuracy and F1-score through repeating the cross validation 5 times at each training set. The optimal hyper-parameters of our model are as follows: the embedding features of PV-DBOW involving each review is a vector of 300 dimensions, the context window that represents the distance between the current word and the predicted one is set to 30. On the other hand, the DAE is fed with statistical representation using TF-IDF obtained from the preprocessed opinions. Then, we use the middle layer of the DAE, which contains 300 nodes, to generate the embedding features. After that, we concatenate the two vectors in the purpose of building richer and relevant features. Once we get the representation for each opinion, we pass these vectors to a fully connected layers (dense layers). The input layer contains 600 neurons, four hidden layers are used with 300, 100, 50 and 10 neurons having RELU as an activation function. The output layer is a sigmoid layer.

Table 1 shows the performance metrics of our model in terms of accuracy, F1-score, recall and precision concerning 5 training sets from 50% to 90%. The measures indicate that the number of true predicted opinions divided by the number of all the predictions are being increasing significantly from 85% to 92%. Indeed, the model shows a better accuracy when we have more training data. Therefore, the accuracy is not the only metric to make a decision regarding the model performance. Specifically, when we have unbalanced datasets. The recall metric, representing the number of true positives (ham opinions detected as hams) divided by the number of true positives and false negatives (deceptive opinions detected as hams), also increased during the growth of training set and reaches 91%. On the other hand, the precision is decreased by 3% from the first training set to the second one, and augmented to reach about 93%. The F1 score or F-measure, which reflects the balance between the recall and the precision, increases as the size of the training set increases. The conclusion is that the model performs well when we have more training data.

In order to evaluate the performance of our model, we compare it to the state-of-the-art methodologies, including the DRI-RCNN model introduced by Zhang et al. (2018), a gated recurrent neural network (GRNN) model for deceptive opinion spam detection proposed by Ren and Ji (2017), the (SEL-feature) model based on stylistic feature (n-gram), emotion-based feature and a linguistic feature based on LIWC variables proposed by Cagnina and Rosso (2015), and finally the syntactic stylometry for deception detection (SSDD) model by Feng et al. (2012). Tables 2 and 3 show the accuracies and F1-scores related to the above models which have been addressed in the paper of Zhang et al. (2018) and our model.

The accuracy and the F1-score performances are influenced by the size of the training sets. As shown in Tables 2 and 3, the three models (Our model, DRI-RCNN, GRNN—CNN) achieve better accuracies and F1-scores when we devote more data for the training. In contrast, SSDD and SEL-feature models decrease in terms of accuracy and F1-score.

Table 2 shows that our model and SSDD model reach the best accuracy when the training set is about 50%, which shows the performance of our model regarding small datasets. On the other hand, the accuracy measures of our model, DRI-RCNN and SSDD models are very close when the percentage of the training set is 80%.

The performance of our model is more important when the training set is equal to 90% compared to the other models. Moreover, as shown in Table 3, the F1-score of our model is better than F1-scores of the different state-of-the-art models in the five training sets. In addition, our model's F1-score in the training set of 60% is higher than the ones of the state- of-the-art models concerning greater training sets. We conclude that our model is efficient at both small and big datasets.

Figs. 2 and 3 show, respectively, the accuracy and F1-measure of our model and the state-of-the-art models.

As shown in the Fig. 2 and the Fig. 3 the accuracy and F1-score curves of our model is above all the other ones, which proves the performance and efficiency of our model in comparison with the state-of-the-art methodologies in both accuracy and F1-score. We notice also that the variation of our model's curves related to the first four training sets remains approximately steady then it increases considerably.

The state-of-the-art models of deceptive opinion spam are either based on traditional machine learning or deep learning based word embedding for the representation of opinions. These models face several challenges as shown in the introduction and the related works. In contrast, our model provides a strong representation which is based on a global representation of the opinions while preserving their semantics. In addition, the adoption of a

Table 1
The measures of our model by the four metrics.

Metrics and training sets	50%	60%	70%	80%	90%
Accuracy	0.8575	0.8750	0.8833	0.8875	0.9250
Recall	0.8025	0.8687	0.8791	0.8750	0.9125
Precision	0.9016	0.8797	0.8865	0.8974	0.9358
F1-score	0.8492	0.8742	0.8828	0.8860	0.9240

Table 2
The accuracy of the state-of-the-art and our model.

Models and training sets	50%	60%	70%	80%	90%
Our model	**0.8575**	**0.875**	**0.8833**	**0.8875**	**0.9250**
DRI-RCNN	0.8293	0.8415	0.8589	0.8724	0.8815
GRNN	0.8015	0.8189	0.8324	0.8415	0.8582
SSDD	0.8487	0.8512	0.8558	0.8609	0.8571
SEL-feature	0.8353	0.8481	0.8510	0.8434	0.8314

Table 3
The F1-score of the state-of-the-art and our model.

Models and training sets	50%	60%	70%	80%	90%
Our model	**0.8492**	**0.8742**	**0.8828**	**0.8860**	**0.9240**
DRI-RCNN	0.8123	0.8301	0.8458	0.8536	0.8659
GRNN	0.8037	0.8281	0.8332	0.8417	0.8513
SSDD	0.8203	0.8311	0.8438	0.8312	0.8210
SEL-feature	0.8203	0.8492	0.8432	0.8321	0.8230

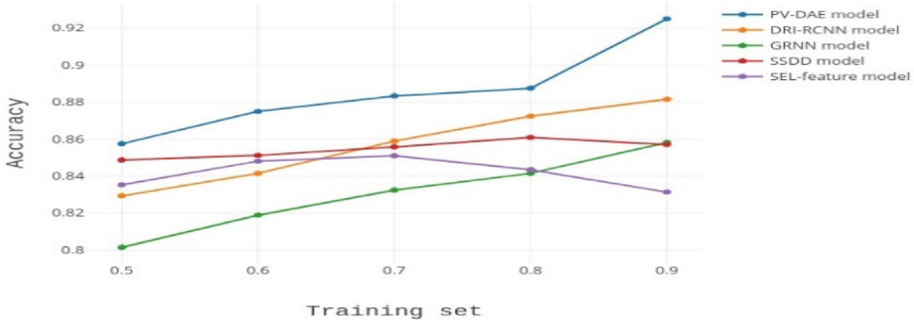

Fig. 2. The accuracy measures of the models.

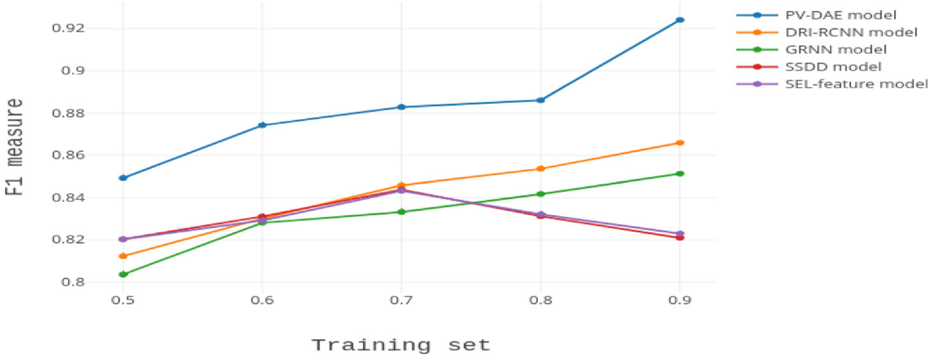

Fig. 3. The F1-score measures of the models.

hybrid model that have shown their performance in the field of sentiment analysis in general. This can explain the performance of our model in terms of accuracy and F1-score.

5. Conclusion and future work

This paper provides a new hybrid model for the detection of deceptive opinion spam based on two architectures the PV-DBOW and DAE models. This model provides a representation of opinions while preserving their semantic. This representation will be fed to a fully connected layers for predicting the deceptive opinion spams. The experiments show that our model is more efficient and outperforms the state-of-the-art models in terms of accuracy and F1-score. The detection of deceptive opinion spam still faces several challenges as the existent models do not show a great improvement in this task such as the advancement of sentiment analysis. To our knowledge, no model is provided for the detection of deceptive opinion spam from an aspect level or from an affective computing point of view. For all these reasons, deceptive opinion spam still needs further investigation.

In future works we plan to combine the semantic and emotional aspects to improve the detection performance of deceptive opinion spam as these aspects show their importance in sentiment analysis in general. While believing that the incorporation of the characteristics of emotional aspect will enrich the representation of opinions that will help in the detection of deceptive opinion spam.

CRediT authorship contribution statement

Anass Fahfouh: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Jamal Riffi:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Supervision, Writing - review & editing. **Mohamed Adnane Mahraz:** Validation, Visualization, Data curation, Visualization. **Ali Yahyaouy:** Validation, Visualization, Data curation, Writing - review & editing. **Hamid Tairi:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bandhakavi, A., Wiratunga, N., Massie, S., Gordon, R., & Padmanabhan, D. (2017). Lexicon generation for emotion detection from text. *IEEE Intelligent Systems, 32*, 102–108.

Banerjee, S., & Chua, A. Y. K. (2014). A theoretical framework to identify authentic online reviews. *Online Information Review. Emerald Group Publishing Limited, 38*, 634–649. https://doi.org/10.1108/OIR-02-2014-0047.

Banerjee, S., Chua, A. Y. K., & Kim, J.-J. (2015). Using Supervised Learning to Classify Authentic and Fake Online Reviews. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication - IMCOM* (pp. 1–7).

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning, 2*, 1–127. https://doi.org/10.1561/2200000006.

Cagnina, L. C., & Rosso, P. (2015). Classification of deceptive opinions using a low dimensionality representation. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015), Lisboa, Portugal, Association for Computational Linguistics* (pp. 58–66).

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 102–107.

Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018). SenticNet 5 : Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). In *Association for the Advancement of Artificial Intelligence (www.aaai.org)* (pp. 1795–1802).

Crawford, M., Khoshgoftaar, T. M., & Prusa, J. D. (2016). Reducing Feature Set Explosion to Facilitate Real-World Review Spam Detection. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference* (pp. 304–309).

Dong, L., Ji, S., Zhang, C., Zhang, Q., Qiu, L., Dong, L., ... Qiu, L. (2018). An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews. *Expert Systems With Applications*, 210–223. https://doi.org/10.1016/j.eswa.2018.07.005.

Du, Y., Zhao, X., He, M., & Guo, W. (2019). A novel capsule based hybrid neural network for sentiment classification. *IEEE Access, 7*, 39321–39328. https://doi.org/10.1109/ACCESS.2019.2906398.

Feng, S., Banerjee, R., & Choi, Y. (2012). Syntactic Stylometry for Deception Detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, Republic of Korea, 8-14 July 2012* (pp. 171–175).

Feng, V. W., & Hirst, G. (2013). Detecting deceptive opinions with profile compatibility. In *International Joint Conference on Natural Language Processing*, October 2013 (pp. 338–346).

Fusilier, D. H., Montes-y-Gómez, M., Rosso, P., & Cabrera, R. G. (2015). Detecting positive and negative deceptive opinions using PU-learning. In *Information Processing and Management* (pp. 1–11).

Hafiz, A., Siagian, M. A., & Aritsugi, M. (2017). Combining Word and Character N-grams for Detecting Deceptive Opinions. In *IEEE 41st Annual Computer Software and Applications Conference* (pp. 828–833). https://doi.org/10.1109/COMPSAC.2017.90.

Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., & Flores, J. J. G. (2016). Cross-domain deception detection using support vector networks. *Soft Computing, UMR, 7030*, 585–595. https://doi.org/10.1007/s00500-016-2409-2.

Jia, S., Zhang, X., Wang, X., & Liu, Y. (2018). Fake Reviews Detection Based on LDA. In *4th International Conference on Information Management (ICIM)* (pp. 280–283).

Lauren, P., Qu, G., Yang, J., Watta, P., & Amaury, G. H. (2018). Generating Word Embeddings from an Extreme Learning Machine for Sentiment Analysis and Sequence Labeling Tasks. *Cognitive Computation, Springer*, 625–638.

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a General Rule for Identifying Deceptive Opinion Spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, 2014 Association for Computational Linguistics* (pp. 1566–1576).

Li, L., Qin, B., Ren, W., & Liu, T. (2017). Document representation and feature combination for deceptive spam review detection. *Neurocomputing* (pp. 1–9). https://doi.org/10.1016/j.neucom.2016.10.080.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Mani, S., Kumari, S., Jain, A., & Kumar, P. (2018). Spam Review Detection Using Ensemble Machine Learning. In *Machine Learning and Data Mining in Pattern Recognition. MLDM 2018. Lecture Notes in Computer Science. Springer, Cham* (pp. 198–209). https://doi.org/10.1007/978-3-319-96133-0.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of International Conference on Learning Representations ICLR 2013* (pp. 1–12).

Mikolov, T., & Le, Q. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of the 31 St International Conference on Machine Learning* (pp. 1188–1196).

Molla, A., Biadgie, Y., & Sohn, K. (2018). Detecting Negative Deceptive Opinion from Tweets. In *Mobile and Wireless Technologies 2017* (pp. 329–339). Springer. https://doi.org/10.1007/978-981-10-5281-1.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 309–319).

Patel, R., & Thakkar, P. (2014). Opinion Spam Detection Using Feature Selection. In *2014 Sixth International Conference on Computational Intelligence and Communication Networks* (pp. 560–564). https://doi.org/10.1109/CICN.2014.127.

Perez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C., & Burzo, M. (2015). Verbal and Nonverbal Clues for Real-life Deception Detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2336–2346).

Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect Extraction for Opinion Miningwith a Deep Convolutional Neural Network. In *Knowledge-Based Systems* (pp. 42–49). https://doi.org/10.1016/j.knosys.2016.06.009.

Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, 217–230. https://doi.org/10.1016/j.neucom.2016.09.117.

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: an empirical study. *Information Sciences*, 213–224. https://doi.org/10.1016/j.ins.2017.01.015.

Ren, Y., & Ji, D. (2019). Learning to detect deceptive opinion spam: A survey. *IEEE Transactions and Journals*, 42934–42945. https://doi.org/10.1109/ACCESS.2019.2908495.

Ren, Y., Ji, D., & Zhang, H. (2014). Positive Unlabeled Learning for Deceptive Reviews Detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 488–498).

Ren, Y., Zhang, Y., Zhang, M., & Ji, D. (2016). Improving Twitter Sentiment Classification Using Topic-Enriched Multi-Prototype Word Embeddings. In *Proceedings of the Thirtieth Aaai Conference on Artificial Intelligence* (pp. 3038–3044).

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic Routing Between Capsules. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA (pp. 1–11).

Saini, M., Verma, S., & Sharan, A. (2018). Multi-view ensemble learning using rough set based feature ranking for opinion spam detection. *Advances in Computer Communication and Computational Sciences, Advances in Intelligent Systems and Computing*, 3–12. https://doi.org/10.1007/978-981-13-0341-8.

Saumya, S., & Singh, J. P. (2018). Detection of spam reviews: A sentiment analysis approach. *CSI Transactions on, ICT*, 137–148. https://doi.org/10.1007/s40012-018-0193-0.

Shojaee, S., Azrifah, M., Muradt, A., Azman, A. Bin, Sharefl, N. M., & Nadali, S. (2013). Detecting Deceptive Reviews Using Lexical and Syntactic Features. In *2013 13th International Conference on Intelligent Systems Design and Applications* (pp. 53–58).

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. In *Proc. Conf. Empirical Methods Natural Language Processing* (pp. 151–161).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning Sentiment-Specific Word Embedding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 1555–1565).

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. (2008). Extracting and Composing Robust Features with Denoising Autoencoders. In *In Proceedings of the 25th International Conference on Machine Learning* (pp. 1–16).

Wang, B., Huang, J., Zheng, H., & Wu, H. (2016). Semi-Supervised Recursive Autoencoders for Social Review Spam Detection. In *12th International Conference on Computational Intelligence and Security* (pp. 116–119). https://doi.org/10.1109/CIS.2016.34.

Wang, C., & Chen, C. (2018). Detecting Spamming Reviews Using Long Short-term Memory Recurrent Neural Network Framework. In *Proceedings of the 2nd International Conference on E-commerce, E-Business and E-Government*, June 13-15, 2018, Hong Kong, 2018 Association for Computing Machinery (pp. 6–10).

Wang, X., Liu, K., & Zhao, J. (2018). Detecting Deceptive Review Spam via Attention-Based Neural Networks. In *The 8th CCF International Conference on Natural Language Processing and Chinese Computing2017* (pp. 866–876).

Wang, X., Liu, Y., Sun, C., Wang, B., & Wang, X. (2015). Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 1343–1353).

Wang, Y., Sun, A., Han, J., Liu, Y., & Zhu, X. (2018). Sentiment Analysis by Capsules. *International World Wide Web Conference, Committee*, 1165–1174.

Wu, C., Wu, F., Wu, S., Yuan, Z., & Huang, Y. (2018). A Hybrid Unsupervised Method for Aspect Term and Opinion Target Extraction. *Knowledge-Based Systems, 148*, 66–73. https://doi.org/10.1016/j.knosys.2018.01.019.

Xiong, S., Lv, H., Zhao, W., & Ji, D. (2018). Towards twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing*, 2459–2466. https://doi.org/10.1016/j.neucom.2017.11.023.

Xu, Q., & Hai Zhao. (2012). Using Deep Linguistic Features for Finding Deceptive. Proceedings of COLING, December 2012, Mumbai, (pp. 1341–1350).

Yang, C., Zhang, H., Jiang, B., & Li, K. (2019). Aspect-based sentiment analysis with alternating coattention networks. *Information Processing and Management, 56*, 463–478. https://doi.org/10.1016/j.ipm.2018.12.004.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine, 13*, 55–75. https://doi.org/10.1109/MCI.2018.2840738.

Yu, L., Wang, J., Lai, K. R., & Zhang, X. (2017). Refining Word Embeddings for Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 534–539).

Zhang, B., Xu, X., Yang, M. I. N., Chen, X., & Ye, Y. (2018). Cross-domain Sentiment Classification by Capsule Network with Semantic Rules. *IEEE Transactions and Journals*, 58284–58294. https://doi.org/10.1109/ACCESS.2018.2874623.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. WIREs Data Mining Knowledge Discovery. 2018 Wiley Periodicals, Inc, (pp. 1–25). http://doi.org/10.1002/widm.1253

Zhang, W., Du, Y., Yoshida, T., & Wang, Q. (2018). DRI-RCNN : An approach to deceptive review identi fi cation using recurrent convolutional neural network. Information Processing and Management (2018), 54, (pp. 576–592). http://doi.org/10.1016/j.ipm.2018.03.007

Zhao, S., Xu, Z., Liu, L., Guo, M., & Yun, J. (2018). Towards Accurate Deceptive Opinions Detection Based on Word Order-Preserving CNN. *Mathematical Problems in Engineering,* 1–9.