Opinion Spam Detection with Attention-based LSTM Networks

Zeinab Sedighi[1], Hossein Ebrahimpour-Komleh[2], Ayoub Bagheri[3], and Leila Kosseim[4]

[1] Dept. of Computer Science and Software Engineering, Concordia University, Montreal, Canada
zeinab.sedighi@concordia.ca
[2] Dept. of Computer Engineering, University of Kashan, Kashan, I.R.Iran
ebrahimpour@kashanu.ac.ir
[3] Dept. of Methodology and Statistics, Utrecht University, Utrecht, Netherlands
a.bagheri@uu.nl
[4] Dept. of Computer Science and Software Engineering, Concordia University, Montreal, Canada
leila.kosseim@concordia.ca

Abstract. Today, online reviews have a great influence on consumers' purchasing decisions. As a result, spam attacks, consisting of the malicious inclusion of fake online reviews, can be detrimental to both customers as well as organizations. Several methods have been proposed to automatically detect fake opinions; however, the majority of these methods focus on feature learning techniques based on a large number of handcrafted features. Deep learning and attention mechanisms have recently been shown to improve the performance of many classification tasks as they enable the model to focus on the most the important features. This paper describes our approach to apply LSTM and attention-based mechanisms for detecting deceptive reviews. Experiments with the Three-domain data set [15] show that a BiLSTM model coupled with Multi-Headed Self Attention improves the F-measure from 81.49% to 87.59% in detecting fake reviews.

Keywords: Deep Learning, Attention Mechanisms, Natural Language Processing, Opinion Spam Detection, Machine Learning.

1 Introduction

Due to the increasing public reliance on social media for decision making, companies and organizations regularly monitor online comments from their users in order to improve their business. To assist in this task, much research has addressed the problems of opinion mining [2, 22]. However, the ease of sharing comments and experience on specific topics has also led to an increase in fake review attacks (or spam) by individuals or groups. In fact, it is estimated that as much as one-third of opinion reviews on the web are spam [21]. These, in turn,

[112] Zeinab Sedighi, Hossein Ebrahimpour-Komleh, Ayoub Bagheri and Leila Kosseim (2019). Learning to Identify Fake Opinion Reviews Using a Neural Network with Attention. In Proceedings of the 20th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2019), April 7–13, La Rochelle, France.

decrease the trustworthiness of all online reviews for both users and organizations.

Manually discerning spam reviews from non-spam ones has been shown to be both difficult and inaccurate [17]; therefore developing automatic approaches to detect review spam has become a necessity. Although automatic opinion spam detection has been addressed by the research community, it still remains an open problem. Most previous work on opinion spam detection have used classic supervised machine learning methods to distinct spam from non-spam reviews. Consequently, much attention has been paid to learning appropriate features to increase the performance of the classification.

In this paper we explore the use of an LSTM based model that uses an attention mechanism to learn representations and features automatically to detect spam reviews. All the deep learning models tested obtained significantly better performance than traditional supervised approaches and the BiLSTM+Multi-Headed Self Attention yielded a best F-measure of 87.59%, a significant improvement over the current state of the art.

This article is organized as follows. Section 2 surveys related work in opinion spam review detection. Our attention-based model is then described in Section 3. Results are presented and discussed in Section 4. Finally, Section 5 proposes future work to improve our model.

2 Related Work

According to [7], spam reviews can be divided into three categories 1) untruthful reviews which deliberately affect user decisions, 2) reviews whose purpose is to advertise specific brands and 3) non-reviews which are irrelevant. Types 2 and 3 are more easy to detect as the topic of the spam differs significantly from truthful reviews; however type 1 spam are more difficult to identify. This article focuses on reviews type 1, which try to mislead users using topic-related deceptive comments.

2.1 Opinion Spam Detection

Much research has been done on the automatic detection of spam reviews. Techniques employed range from unsupervised (e.g. [1]), semi-supervised (e.g. [13, 10]) and supervised methods (e.g. [14, 27, 9]) with a predominance for supervised methods. Most methods rely on human feature engineering and experiment with different classifiers and hyper parameters to enhance the classification quality.

To train from more data, [17] generated an artificial data set and applied supervised learning techniques for text classification. [10] uses spam detection for text summarization and [10] applies Naive Bayes, logistic regression and Support Vector Machine methods after feature extraction using Part-Of-Speech tags and LIWC. To investigate cross domain spam detection, [12] uses a data set consist of three review domains to avoid the dependency to a specific domain. They examine SVM and SAGE as classification models.

2.2 Deep Learning for Sentiment Analysis

Deep learning models have been widely used in natural language processing and text mining [16] such as sentiment analysis [20], co-reference resolution [5], POS tagging [19] and parsing [6] as they are capable to learn relevant syntactic and semantic features automatically as opposed to hand-feature engineering.

Because long term dependencies are prominent in natural language, Recurrent Neural Networks (RNNs) and in particular Long Short Term Memories (LSTMs) have been very successful in many applications (e.g. [18, 20, 25]). [11] employed an RNN in parallel with a Convolutional Neural Network (CNN) to improve the analysis of sentiment phrases. [20] used an RNN to create sentence representations. [25] presented a context representation for relation classification using a ranking recurrent neural network.

2.3 Attention Mechanisms

Attention mechanisms have also shown much success in the last few years [3]. Using such mechanisms, neural networks are able to better model dependencies in sequences of information in texts, voices, videos, etc [26, 4] regardless of their distance. Because they learn which information from the sequence is more useful to predict the current output, attention mechanisms have increased the performance of many Natural Language Processing tasks such as [3, 23].

An attention function maps an input sequence and a set of key-value pairs to an output. The output is calculated as a weighted sum of the values. The weight assigned to each value is obtained using a compatibility function of the sequence and the corresponding key. In a vanilla RNN without attention, the model embodies all the information of the input sequence by means of the last hidden state. However, when applying an attention mechanism, the model is able to glance back at the entire input; not only by accessing the last hidden state but also by accessing a weighted combination of all input states.

Several types of attention mechanisms have been proposed [23]. Self Attention, also known as intra-attention, aims to learn the dependencies between the words in a sentence and uses this information to model the internal structure of the sentence. Scaled Dot-Product Attention [24], on the other hand, calculates the similarity using scaled dot-product. As opposed to Self Attention, Scaled Dot-Product Attention uses an additional dimension to adjust the inner product from becoming too large. If the calculation is performed several times instead of once, it will enable the model to learn more relevant information concurrently in different sub-spaces. This last model is called Multi-Headed Self Attention. In light of the success of these attention mechanisms, this paper evaluates the use of these techniques for the task of opinion spam detection.

3 Methodology

3.1 Attention-Based LSTM Model

Figure 1 shows the general model used in our experiments. The look-up layer maps words to a look-up table by applying word embeddings. In order to better capture the relations between distant words, the next layer uses LSTM units. In our experiments (see Section 4), we investigated with both undirectional and bi-directional LSTMs (BiLSTM) [8]. We considered the use of one LSTM layer, one BiLSTM layer and two BiLSTM layers. In all cases we used 150 LSTM units in each layer and training phase was applied after each 32 time steps using Back Propagation Through Time (BPTT) with a learning rate of 1e-3 and a dropout rate of 30%.

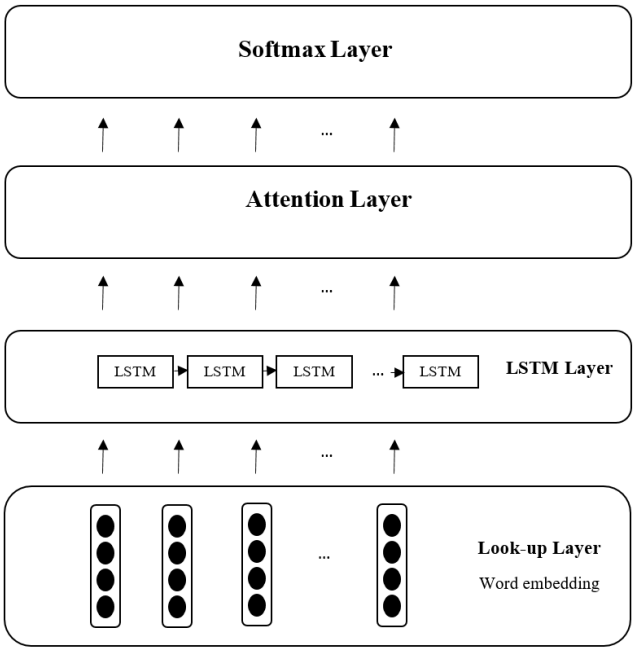

Fig. 1. General Architecture of the Attention-based Models

The results from the LSTM layer is fed into the attention layer. Here, we experimented with two mechanisms: Self Attention and Multi-Headed Self Attention mechanisms [24]. Finally, a Softmax layer is applied for the final binary classification.

4 Experiments

To evaluate our models, we use the Three-Domain data set [15] which constitutes the standard benchmark in this field. [15] introduced the Three-domain review spam data set: a collection of 3032 reviews in three different domains (Hotel, Restaurant and Doctor) annotated by three types of annotators (Turker, Expert and Customer). Each review is annotated with a binary label: truthful or deceptive. Table 1 shows statistics of the data set.

Table 1. Statistics of the Three-domain Dataset

Data Set	**Turker**	**Expert**	**Customer**	**Total**
Hotel	800	280	800	1880
Restaurant	200	120	400	720
Doctor	200	32	200	432
Total	1200	432	1400	3032

To compare our proposed models with traditional supervised machine learning approaches we also experimented with SVM, Naive Bayes and Log Regression methods. For these traditional feature-engineered models, we pre-processed the text to remove stop words, and stemmed the removing words. Then, to distinguish the role of words, POS tagging was applied. To extract helpful features for classifying reviews, feature engineering techniques are required. Bigrams and TF-IDF are applied to extract more repetitive words in the document. For other deep learning models, we used both a CNN and an RNN. The CNN and the RNN use the same word embeddings as our model (see Section 3). The CNN uses two Convolutional and Pooling layers connected to one fully connected hidden layers.

5 Results and Analysis

Our various models were compared using all three domains, in-domain and cross domains.

5.1 All Three-domain Results

In our first set of experiments we used the combination of all three domains (Hotel, Restaurant and Doctor) for a total of 3032 reviews. Table 2 shows the results of our deep learning models compared classic machine learning methods where using cross-validation. As Table 2 shows, in general all deep learning models yield a higher F-measure than SVM, Naive Bayes and Log Regression. It is interesting to note that both precision and recall benefit from the deep models and those attention mechanisms yield a significant improvement in performance. The Multi-Headed Self Attention performs better than the Self Attention; and the bidirectional LSTM does provide a significant improvement compared to the unidirectional LSTM.

Table 2. Results in All Three-Domain Classification

Methods	Precision	Recall	F-measure
SVM	72.33	68.50	70.36
Naive Bayes	61.69	63.32	62.49
Log Regression	55.70	57.34	56.50
CNN	79.23	69.34	73.95
RNN	75.33	73.41	74.35
LSTM	80.85	68.74	74.30
BiLSTM	82.65	80.36	81.49
BiLSTM + Self Attention	85.12	83.96	84.53
BiLSTM + Multi-Headed Self Attention	90.68	84.72	87.59

5.2 In-Domain Results

Table 3 shows the results of same models for each domain. As Table 3 shows, the same general conclusion can be drawn for each specific domain: deep learning methods show significant improvements compared to classical ML methods, and attention mechanisms increase the performance even more.

These results seem to shows that neural models are more suitable for deceptive opinion spam detection. The results on the Restaurant data are similar to those on the Hotel domain. However, the models yield lower results on the Doctor domain. A possible reason is that the number of reviews in this domain are relatively lower, which leads to relatively lower performance.

5.3 Cross-domain Results

Finally, to evaluate the generality of our model, we performed an experiment across domains, where we trained models on one domain and evaluated them on another. Specifically, we trained the classifiers on Hotel reviews (for which we had more data), and evaluated their the performance on the other domains. Table 4 shows the result of these experiments. Again the same general trend appears. One notices however, that the performance of the models does drop compared to Table 3 where the training was done on the same domain.

5.4 Comparison with Previous Work

In order to compare the proposed model with the state of the art, we performed a last experiment in line with the experimental set up of [15]. As indicated in Section 2, [15] used an SVM and SAGE based on unigram + LIWS + POS tags. To our knowledge, their approach constitutes the sate of the art approach on the Three Domain dataset. Although [15] performed a variety of experiments, the set up used applied the classifiers on the turker and the customer sections of the dataset only (see Table 1).

The model was trained on the entire data set (Customer+Expert+Turker), and tested individually on the Turker, Customer, and Expert sections using

Table 3. Results in In-Domain Classification

Data Sets	**Methods**	**Precision**	**Recall**	**F-measure**
Hotel	SVM	69.97	67.36	68.64
	Naive Bayes	58.71	62.13	60.37
	Log Regression	55.32	56.45	55.87
	RNN	72.25	70.31	71.27
	CNN	78.76	74.30	76.47
	LSTM	84.65	81.11	82.84
	BiLSTM	86.43	83.05	84.70
	BiLSTM + Self Attention	90.21	85.73	87.91
	BiLSTM + Multi-Headed Self Attention	89.33	92.59	90.93
Restaurant	SVM	73.76	69.43	71.52
	Naive Bayes	63.18	66.32	64.71
	Log Regression	59.96	63.87	61.85
	RNN	77.12	74.96	76.02
	CNN	78.11	76.92	77.51
	LSTM	80.23	78.74	79.47
	BiLSTM	86.85	87.35	87.10
	BiLSTM + Self Attention	88.68	86.47	87.56
	BiLSTM + Multi-Headed Self Attention	89.66	91.00	90.32
Doctor	SVM	72.17	74.39	73.26
	Naive Bayes	63.18	67.98	65.49
	Log Regression	65.83	69.27	67.51
	RNN	75.28	67.98	71.44
	CNN	77.63	70.03	73.63
	LSTM	79.92	74.21	77.49
	BiLSTM	79.85	78.74	79.29
	BiLSTM + Self Attention	82.54	80.61	81.56
	BiLSTM + Multi-Headed Self Attention	84.76	81.10	82.88

Table 4. Results in Cross Domain Classification

Datasets	**Methods**	**Precision**	**Recall**	**F-measure**
Hotel vs. Doctor	SVM	67.28	64.91	66.07
	Naive Bayes	59.94	63.13	61.49
	Log Regression	55.47	51.93	53.64
	RNN	72.33	68.50	70.36
	CNN	61.69	63.32	62.49
	LSTM	74.65	70.89	72.72
	BiLSTM	76.82	71.63	74.13
	BiLSTM+Self Attention	78.10	73.79	75.88
	BiLSTM+Multi-Headed Self Attention	81.90	77.34	79.55
Hotel Vs. Restaurant	SVM	70.75	68.94	69.83
	Naive Bayes	64.87	67.11	65.97
	Log Regression	60.72	57.90	59.27
	RNN	79.23	69.34	73.95
	CNN	75.33	73.41	74.35
	LSTM	80.15	73.94	76.91
	BiLSTM	80.11	79.92	80.01
	BiLSTM+Self Attention	87.73	82.27	84.91
	BiLSTM+Multi-Headed Self Attention	90.68	84.72	87.59

Table 5. Comparison of the proposed model with [15] on the Customer data

Data	Customer					
	Precision		Recall		F-Measure	
	Our Model	[15]	Our Model	[15]	Our Model	[15]
Hotel	85	67	93	66	89	66
Restaurant	90	69	92	72	91	70

Table 6. Comparison of the proposed model with [15] on the Expert data

Data	Expert					
	Precision		Recall		F-Measure	
	Our Model	[15]	Our Model	[15]	Our Model	[15]
Hotel	80	58	85	61	82	59
Restaurant	79	62	84	64	81	70

Table 7. Comparison of the proposed model with [15] on the Turker data

Data	Turker					
	Precision		Recall		F-Measure	
	Our Model	[15]	Our Model	[15]	Our Model	[15]
Hotel	87	64	92	58	89	61
Restaurant	88	68	89	70	88	68

cross-validation. To compare our approach, we reproduced their experimental set up, and as shown in Table 5 to Table 7, our BiLSTM+Multi-Headed Self Attention model outperforms this state of the art.

6 Conclusion and Future Work

In this paper we showed that an attention mechanism can learn document representation automatically for opinion spam detection and clearly outperform non-attention based models as well as classic models. Experimental results show that the Multi-Headed Self Attention performs better than the Self Attention; and the bidirectional LSTM does provide a significant improvement compared to the unidirectional LSTM. Utilizing a model with no need for manual feature extraction from documents with high performance is effective to improve the detection of spam comments. Utilizing an attention mechanism and an LSTM model enable us to have a comprehensive model for distinguishing different reviews in different domains. This shows the generality power of our model.

One challenge left for the future is to improve the performance of cross domains spam detection. This would enable the model to be used widely to reach the performance of in domain results in all domains.

Acknowledgments

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., Nunamaker Jr, J.F.: Detecting fake websites: The contribution of statistical learning theory. Mis Quarterly pp. 435–461 (2010)
2. Bagheri, A., Saraee, M., De Jong, F.: Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowledge-Based Systems **52**, 201–213 (2013)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016). pp. 4960–4964 (2016)
5. Clark, K., Manning, C.D.: Improving coreference resolution by learning entity-level distributed representations **1**, 643–653 (2016)
6. Collobert, R.: Deep learning for efficient discriminative parsing. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. pp. 224–232. FL, USA (2011)
7. Dixit, S., Agrawal, A.: Survey on review spam detection. International Journal of Computer & Communication Technology **4**, 0975–7449 (2013)

8. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with lstm recurrent networks. Journal of machine learning research **3**(Aug), 115–143 (2002)
9. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM 2008). pp. 219–230. ACM, Palo Alto, California, USA (2008)
10. Jindal, N., Liu, B., Lim, E.P.: Finding unusual review patterns using unexpected rules. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010). pp. 1549–1552. Toronto, Canada (Oct 2010)
11. Kuefler, A.R.: Merging recurrence and inception-like convolution for sentiment analysis. https://cs224d.stanford.edu/reports/akuefler.pdf (2016)
12. Lau, R.Y., Liao, S., Kwok, R.C.W., Xu, K., Xia, Y., Li, Y.: Text mining and probabilistic language modeling for online review spam detection. ACM Transactions on Management Information Systems (TMIS) **2**(4), 25 (2011)
13. Li, F., Huang, M., Yang, Y., Zhu, X.: Learning to identify review spam. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2011). pp. 2488–2493 (2011). https://doi.org/https://dl.acm.org/citation.cfm?id=2283811
14. Li, H.: Detecting Opinion Spam in Commercial Review Websites. Ph.D. thesis, University of Illinois at Chicago (2016)
15. Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a general rule for identifying deceptive opinion spam. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL-2014). vol. 1, pp. 1566–1576 (2014)
16. Manning, C.D.: Computational linguistics and deep learning. Computational Linguistics **41**(4), 701–707 (2015)
17. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2013). pp. 497–501 (2013)
18. Pascanu, R., Gulcehre, C., Cho, K., Bengio, Y.: How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026 (2013)
19. Santos, C.D., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014). pp. 1818–1826 (2014)
20. Socher, R.: Deep learning for sentiment analysis – invited talk. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. p. 36. San Diego, California (2016)
21. Streitfeld, D.: The best book reviews money can buy. The New York Times **25** (2012)
22. Sun, S., Luo, C., Chen, J.: A review of natural language processing techniques for opinion mining systems. Information Fusion **36**, 10–25 (2017)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA (2017)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
25. Vu, N.T., Adel, H., Gupta, P., Schütze, H.: Combining recurrent and convolutional neural networks for relation classification. arXiv preprint arXiv:1605.07333 (2016)

26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML 2015). pp. 2048–2057. Lille, France. (2015)
27. Zhang, D., Zhou, L., Kehoe, J.L., Kilic, I.Y.: What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. Journal of Management Information Systems **33**(2), 456–481 (2016)