The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Opinion Spam Detection with Attention-Based Neural Networks

Zeinab Sedighi,[1,3] **Hossein Ebrahimpour-Komleh,**[1] **Ayoub Bagheri,**[2] **Leila Kosseim**[3]
[1]Dept. of Computer Engineering, Faculty of Computer and Electrical Engineering, University of Kashan, Kashan, I.R.Iran
[2]Dept. of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, Utrecht, Netherlands
[3]Dept. of Computer Science and Software Engineering, Concordia University, Montreal, Canada
zeinab.sedighi@concordia.ca, ebrahimpour@kashanu.ac.ir, a.bagheri@uu.nl, leila.kosseim@concordia.ca

Abstract

Fake opinion attacks, consisting of the malicious dissemination of fake reviews, can be detrimental to both customers as well as organizations. Several methods have been proposed to automatically detect deceptive reviews; however, these rely on manual feature engineering methods to build classifiers based on labeled data. Deep Neural Networks coupled with an Attention mechanism have recently been shown to improve the performance of many classification tasks, as it enables the model to learn and focus automatically on the most the important features. This paper describes our approach to apply an attention based deep neural network for the detection of truthful versus fake reviews. The evaluation of our model shows that its performance is significantly better than traditional models, and requires no hand feature engineering.

Introduction

Today, the ease of sharing comments and experience online has lead consumers as well as companies to rely and monitor the social media for decision making. However, this phenomenon has also led to an increase in fake review attacks by individuals or groups. In fact, it is estimated that as much as one-third of opinion reviews on the Internet constitute spam (Streitfeld 2012).

Manually identifying spam reviews from non-spam ones is both time-consuming and inaccurate (Ott, Cardie, and Hancock 2013); therefore developing automatic approaches to detect spam review has become a necessity. Much research has addressed the problems of opinion mining (e.g. (Bagheri, Saraee, and De Jong 2013; Sun, Luo, and Chen 2017)), however, the automatic detection of spam opinion still remains an open problem.

Most previous work on spam opinion detection have used classic supervised machine learning methods based on handcrafted features and the identification of discriminating linguistic features to increase the performance of the classification.

In this paper we present a deep learning model that uses an attention mechanism to learn representations and features automatically to detect spam reviews. To reduce the number of model inputs, we introduce a preprocessing level which performs basic feature extraction. The proposed model obtains significantly better results compared to traditional approaches.

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.

Related Work

Spam reviews are typically categorized into three types (Dixit and Agrawal 2013):

1. Untruthful reviews which try to deliberately affect user decisions;
2. Reviews whose purpose is to advertise specific brands;
3. Non-reviews which are irrelevant to the topic.

Type 2 and type 3 spam reviews are more easy to detect as the topic of the content differs significantly from that of truthful reviews. However type 1 spam are more difficult to identify. This article focuses on reviews of type 1, which try to mislead users through topic-related deceptive comments.

Previous work on the automatic detection of spam reviews have used a variety of approaches, ranging from unsupervised learning (e.g. (Abbasi et al. 2010)), semi-supervised learning (e.g. (Li et al. 2011; Jindal, Liu, and Lim 2010)) as well as supervised methods (e.g. (Li 2016; Zhang et al. 2016; Jindal and Liu 2008)). Supervised methods that rely on human feature engineering have however constituted the most common models. In particular, (Li et al. 2014) experimented with a Naïve Bayes classifier, logistic regression and a Support Vector Machine (SVM) using features such as part-of-speech (POS) tags and LIWC features. (Lau et al. 2011) experimented with a dataset of reviews of Three Domains to avoid the dependency to a specific domain. They also examined to use of SVM and Sparse Additive Generative Model (SAGE) for the classification. On the other hand, (Ott, Cardie, and Hancock 2013) focused on the generation of a synthetic data set to improve the performance of the classifiers.

Apart from classical machine learning techniques, classification based on Deep Neural Networks have lead to significant improvements in the state of the art in many Natural

Language Processing (NLP) tasks. In particular, Recurrent Neural Networks (RNNs) have been successful (e.g. (Pascanu et al. 2013)) as they address the issue of long term dependencies that are prominent in natural language. (Kuefler 2016) employed an RNN in parallel with a Convolutional Neural Network (CNN) to improve the analysis of sentiment phrases. (Socher 2016) used a recursive neural network to create sentence representations. (Vu et al. 2016) presented a context representation for relation classification using a ranking recurrent neural network.

Attention mechanisms have shown much success in the last few years (Bahdanau, Cho, and Bengio 2014). Using such mechanisms, neural networks are able to better model sequences of information in texts, voices, videos, etc (Xu et al. 2015; Chan et al. 2016). They are particularly useful in NLP applications that require modeling dependencies regardless of their distances (Bahdanau, Cho, and Bengio 2014). (Vaswani et al. 2017) used self-attention mechanisms for learning representations and explored their effects in different NLP tasks.

Attention mechanisms are inspired by the workings of the human brain which can extract pertinent information of different levels of data and ignore information which is not necessary for the task. Computationally, 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.

Scaled Dot-Product Attention (Vaswani et al. 2017) is a specific attention mechanism that calculates the similarity using Scaled Dot-Product. This method has an extra dimension which adjusts the inner product from becoming too large. If the calculation are performed several times instead of once, it enables the model to learn more relevant information concurrently in different sub-spaces. This model is called Multi-Headed Self-Attention.

Given the recent successes of attention-based neural networks, we experimented with the use of such an architecture to learn representations and features automatically to detect spam reviews.

Methodology

In order to evaluate the use of attention mechanisms for review spam detection, we propose an attention based deep structure as a means to improve the state of the art in opinion spam review detection.

Our model is composed of a bidirectional LSTM coupled with a Multi-Headed Self-Attention mechanism. The architecture of the model is composed of four layers: an input embedding layer, a BiLSTM layer, an attention layer and a softmax layer.

The Embedding Layer: The embedding layer uses the pretrained Word2vec word embeddings (Mikolov et al. 2013; Lau and Baldwin 2016) of size 300 to represent each input word.

The LSTM Layer: Each embedding is then passed to the LSTM layer that tries to account for the relations between distant words. Specifically, this layer uses Bidirectional Long Short Term Memory (BiLSTM) cells where the input word embeddings propagate in both directions using parallel layers of forward and backward LSTM. Thus, the model can be trained three ways: in a forward direction, backward direction, and in a combination of both. The model is composed of two BiLSTM layers, each composed of 150 LSTM units. Training is performed after each 32 time steps using Back Propagation Through Time (BPTT) with a learning rate $\eta = 0.001$ and a dropout rate of 30%.

The Attention Layer: The results of the BiLSTM layer is fed into the Multi-Headed Self-Attention mechanism and information is extracted using a weighting mechanism.

The Softmax Layer: Finally, softmax is applied to perform the final classification into either truthful reviews from non-truthful ones.

The model was implemented using Keras and Tensorflow.

Experiments and Results

To compare our proposed model, we also experimented with traditional supervised machine learning approaches coupled with classic linguistic features, as well as with other deep learning models but without an attention mechanism.

Models

For the traditional models, we experimented with Support Vector Machines (SVM), Naive Bayes (NB) and Logistic Regression (LogReg). 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). The CNN uses two Convolutional and Pooling layers connected to one fully connected hidden layers.

Features

To extract features to feed the traditional feature-engineered models (SVM, NB and LogReg), we proceeded as follows. First, we pre-processed the reviews to remove stop words, then stemmed the remaining words using the Porter stemmer (Porter 1980). Then, to distinguish the role of words, we used part-of-speech (POS) tags as provided by the NLTK Toolkit (Bird, Klein, and Loper 2009). Finally, to keep only discriminating features, we used bigrams and TF-IDF to extract more repetitive words in the document.

Dataset

The dataset used is the Three-Domain Dataset proposed by (Li et al. 2014). This data set is a collection of 3032 reviews in three different domains: Hotel, Restaurant and Doctor. The dataset was annotated by three types of annotators: Turker, Expert and Customer. Each review was assigned a binary label: truthful (P) or deceptive (N). Table 1 shows

Table 1: Statistics of Three-Domain Dataset

Data set	Turker	Expert	Customer	Total
Hotel (P/N)	400/400	140/140	400/400	1880
Restaurant (P/N)	200/0	120/0	200/200	720
Doctor (P/N)	200/0	32/0	200/0	432
Total	1200	432	1400	3032

Table 2: Results with the Three-domain Dataset

Classifier	Precision	Recall	F-measure
SVM	72.33	68.50	70.36
NB	61.69	63.32	62.49
LogReg	55.70	57.34	56.50
CNN	79.23	69.34	73.95
RNN	75.33	73.41	74.35
Proposed method	90.68	84.72	87.59

statistics of the data set. From the dataset, we used the combined data that includes 3031 reviews (1880 reviews for the Hotel domain, 720 for the Restaurant domain and 432 for the Doctor domain). The distinction between Turker versus Expert versus Customer annotation was not exploited. For training our models, we used 10-fold cross-validation.

Results

Table 2 shows the results of all 6 models evaluated. As the table shows, classic machine learning methods achieve the lowest performances, with F-measures close or below 70%. In is worth noting that the SVM does perform significantly better that the Naïve Bayes or the Logistic Regression. Deep learning methods, on the other hand, easily break the 70% barrier in F-measure. The CNN and the RNN with no attention mechanism do achieve a higher F-measure than the SVM, but it is our proposed method, the BiLSTM+attention, that yields the most significant F-measure with 87.59%.

Conclusion

In this paper, we evaluated the use of an attention based neural network to learn document representations automatically for the task of opinion spam detection. By using a hierarchy of feature extraction in a deep structure neural network, we construct a semantic model in different resolution to detect fake reviews. Engineering features in document, sentence and word level features produce features with higher quality. The proposed model clearly outperforms classic supervised learning models in terms of F-measure. In addition, the model requires no manual feature engineering.

For future work, it would be interesting to analyze the performance of the model for each discourse domain, and across domains. A cross-domain analysis would allow us to measure how domain-specific the learned representations are, and to what extend the features and model learned from one discourse domain for which much data is available can transferred to another domain for which less training data is available.

References

Abbasi, A.; Zhang, Z.; Zimbra, D.; Chen, H.; and Nunamaker Jr, J. F. 2010. Detecting fake websites: The contribution of statistical learning theory. *Mis Quarterly* 435–461.

Bagheri, A.; Saraee, M.; and De Jong, F. 2013. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems* 52:201–213.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Beijing: O'Reilly.

Chan, W.; Jaitly, N.; Le, Q.; and Vinyals, O. 2016. 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)*, 4960–4964.

Dixit, S., and Agrawal, A. 2013. Survey on review spam detection. *International Journal of Computer & Communication Technology* 4:0975–7449.

Jindal, N., and Liu, B. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM 2008)*, 219–230. Palo Alto, California, USA: ACM.

Jindal, N.; Liu, B.; and Lim, E.-P. 2010. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, 1549–1552.

Kuefler, A. R. 2016. Merging recurrence and inception-like convolution for sentiment analysis. *https://cs224d.stanford.edu/reports/akuefler.pdf*.

Lau, J. H., and Baldwin, T. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*.

Lau, R. Y.; Liao, S.; Kwok, R. C.-W.; Xu, K.; Xia, Y.; and Li, Y. 2011. Text mining and probabilistic language modeling for online review spam detection. *ACM Transactions on Management Information Systems (TMIS)* 2(4):25.

Li, F.; Huang, M.; Yang, Y.; and Zhu, X. 2011. Learning to identify review spam. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI 2011)*, 2488–2493.

Li, J.; Ott, M.; Cardie, C.; and 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 (Volume 1: Long Papers) (ACL-2014)*, volume 1, 1566–1576.

Li, H. 2016. *Detecting Opinion Spam in Commercial Review Websites*. Ph.D. Dissertation, University of Illinois at Chicago.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Ott, M.; Cardie, C.; and Hancock, J. T. 2013. 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)*, 497–501.

Pascanu, R.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2013. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*.

Porter, M. 1980. An algorithm for suffix stripping. *Program* 14(3):130–137.

Socher, R. 2016. Deep learning for sentiment analysis – invited talk. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 36.

Streitfeld, D. 2012. The best book reviews money can buy. *The New York Times* 25.

Sun, S.; Luo, C.; and Chen, J. 2017. A review of natural language processing techniques for opinion mining systems. *Information Fusion* 36:10–25.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 5998–6008.

Vu, N. T.; Adel, H.; Gupta, P.; and Schütze, H. 2016. Combining recurrent and convolutional neural networks for relation classification. *arXiv preprint arXiv:1605.07333*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML 2015)*, 2048–2057.

Zhang, D.; Zhou, L.; Kehoe, J. L.; and Kilic, I. Y. 2016. 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.