Information Sciences 385–386 (2017) 213–224

Contents lists available at ScienceDirect

Information Sciences

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

Neural networks for deceptive opinion spam detection: An empirical study

Yafeng Ren [a,*], Donghong Ji [a,b]

[a] Guangdong Collaborative Innovation Center for Language Research & Services, Guangdong University of Foreign Studies, Guangzhou 510420, China
[b] Computer School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Article history:
Received 11 March 2016
Revised 29 December 2016
Accepted 2 January 2017
Available online 3 January 2017

Keywords:
Deceptive opinion spam
Discrete features
Convolutional neural network
Recurrent neural network
Representation learning

ABSTRACT

The products reviews are increasingly used by individuals and organizations for purchase and business decisions. Driven by the desire of profit, spammers produce synthesized reviews to promote some products or demote competitors products. So deceptive opinion spam detection has attracted significant attention from both business and research communities in recent years. Existing approaches mainly focus on traditional discrete features, which are based on linguistic and psychological cues. However, these methods fail to encode the semantic meaning of a document from the discourse perspective, which limits the performance. In this work, we empirically explore a neural network model to learn document-level representation for detecting deceptive opinion spam. First, the model learns sentence representation with convolutional neural network. Then, sentence representations are combined using a gated recurrent neural network, which can model discourse information and yield a document vector. Finally, the document representations are directly used as features to identify deceptive opinion spam. Based on three domains datasets, the results on in-domain and cross-domain experiments show that our proposed method outperforms state-of-the-art methods.

1. Introduction

In recent years, online reviews on products and services contain rich information related to subjective opinions on certain topics. These information have become an important resource for public opinion that influence our decisions over an extremely wide spectrum of daily and professional activities: e.g.,where to eat, where to stay, which products to purchase, which doctors to see, and so on. As a result, sentiment analysis and opinion mining based on product reviews have become a heated topic in natural language processing (NLP) [6,11,41].

Since reviews information can guide people purchase behavior, positive reviews can result in huge economic benefit and fame for organizations or individuals. This gives powerful incentive to promote the generation of deceptive opinion spam [24,28,39,47]. Deceptive opinion spam is a type of review with fictitious opinions, deliberately written to sound authentic [21,34]. Two reviews are shown as follows:

* Corresponding author.
E-mail address: renyafeng@whu.edu.cn (Y. Ren).

http://dx.doi.org/10.1016/j.ins.2017.01.015
0020-0255/

- *I have stayed at many hotels travelling for both business and pleasure and I can honestly stay that the James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all the great sights and restaurants. Highly recommend to both business travellers and couples. (Date of review: Jun 9, 2006)*
- *My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn't ask for more!! We will definatly be back to Chicago and we will for sure be back to the James Chicago. (Date of review: Jun 9, 2006)*

These two reviews are from the firstly public dataset in the domain of opinion spam [34]. The first is non-spam or truthful review, and the second is deceptive opinion spam. Based on two reviews, we can know that it is very difficult for human readers to distinguish them from truthful reviews. In a test by previous work [34], the average accuracy of three human judges is only 57.33%. Deceptive opinion spam detection is a pressing and also profound issue as it is critical to ensure that trustworthiness of the information on the web. Without detecting them, the social media could become a place full of lies, fakes, and deceptions and completely useless. Hence, machine learning methods for automatically detecting deceptive opinion spam can be very necessary.

Generally, deceptive opinion spam detection is deemed to be a classification problem [34,39]. Based on the positive and negative examples annotated by people, supervised learning is utilized to build a classifier, and then an unlabeled review can be predicted as deceptive review or truthful one. So the objective of the task is to identify whether a given document a spam or not. The majority of existing approaches follow the seminal work of Jindal and Liu (2008) [21], employing classifiers with supervised learning. Most studies focus on designing effective features to enhance classification performance. Typical features represent linguistic and psychological cues, but fail to effectively represent a document from the viewpoint of global discourse structures. For example, Ott et al. (2011) and Li et al. (2014) represent documents with Unigram, POS and LIWC (Linguistic Inquiry and Word Count) feature [17,34]. Although such features give the strong performance, their sparsity makes it difficult to capture non-local semantic information over a sentence or discourse.

Recently, neural network models have been used to learn semantic representations for NLP tasks [16,50], achieving highly competitive results. The potential advantages of neural networks for spam detection are three-fold. First, neural models use real-valued hidden layers for automatic feature combinations, which can capture complex global semantic information that is difficult to express using traditional discrete manual features. This can be useful in addressing the limitation of discrete models mentioned above. Second, neural networks take distributed word embeddings as inputs, which can be trained from large-scale raw text, thus alleviating the scarcity of annotated data to some extent. Third, neural network models can learn continuous document representations, leveraging sentence and discourse models simultaneously.

In this paper, we show that significant improvements can be achieved by learning continuous document representations using a neural network model. In particular, we propose a three-stage system for opinion spam detection, as shown in Fig. 1. In the first stage, a convolutional neural network is used to produce sentence representations from word representations. Then a bi-directional gated recurrent neural network is used to construct a document representation from the sentence vectors by modeling their semantic and discourse relations. Finally, the document representation is used as features to identify deceptive opinion spam. Such automatically induced dense document representation is compared with traditional manually-designed features for the task.

We evaluate the proposed model on a standard benchmark [17], which consists of data from three domain (*Hotel, Restaurant*, and *Doctor*). Results on in-domain and cross-domain experiments show that our proposed neural model significantly outperforms the state-of-the-art methods, demonstrating the advantage of neural models in capturing semantic characteristics.

In remaining parts, Section 2 presents related work. Section 3 gives details of our proposed neural model. Section 4 introduces experimental setup, and then reports experimental results of in-domain and cross-domain settings. Section 5 concludes this work.

2. Related work

2.1. Deceptive opinion spam detection

Spam detection has been historically investigated in the Web-page and E-mail domains [8,33,55]. With the rise of e-commerse, spam detection research has recently been extended to the customer review domain [17,31,34]. Various types of indicator features have been studied. Jindal and Liu (2008) first studied deceptive opinion spam problem, training models using features based on the review content, reviewer, and the product itself [21]. Yoo and Gretzel (2009) gathered 40 truthful and 42 deceptive hotel reviews and manually compared the linguistic differences between them [54].

Ott et al. (2011) created a benchmark dataset by employing *Turkers* to write fake reviews [34]. Their data was adopted by a line of subsequent work [9,10,35]. For example, Feng et al. (2012) looked into syntactic features from context free grammar parse trees to improve the classification performance [9]. Feng and Hirst (2013) built profiles of hotels from collections of reviews, measuring the compatibility of customer reviews to the hotel profile, and using it as a feature for opinion spam detection [10]. Newman et al. (2003) looked into some linguistic cues to deception detection, such as increased negative emotion terms and decreased spatial detail [32]. They found certain writing style difference between informative and imag-