Constructing and Evaluating a Novel Crowdsourcing-based Paraphrased Opinion Spam Dataset

Seongsoon Kim[†], Seongwoon Lee[†], Donghyeon Park, Jaewoo Kang[*]
{seongkim, seongwoon, parkdh, kangj}@korea.ac.kr
Department of Computer Science and Engineering
Korea University, Seoul, Korea

ABSTRACT

Opinion spam, intentionally written by spammers who do not have actual experience with services or products, has recently become a factor that undermines the credibility of information online. In recent years, studies have attempted to detect opinion spam using machine learning algorithms. However, limitations of gold-standard spam datasets still prove to be a major obstacle in opinion spam research. In this paper, we introduce a novel dataset called Paraphrased OPinion Spam (POPS), which contains a new type of review spam that imitates real human opinions using crowdsourcing. To create such a seemingly truthful review spam dataset, we asked task participants to paraphrase truthful reviews, and include factual information and domain knowledge in their reviews. The classification experiments and semantic analysis results show that our POPS dataset most linguistically and semantically resembles truthful reviews. We believe that our new deceptive opinion spam dataset[1] will help advance opinion spam research.

Keywords

Deceptive Opinion Spam, Paraphrased Opinion Spam; Crowdsourcing

1. INTRODUCTION

Today, opinion reviews on the Web influence the decision making of ordinary people [10]. As the importance of opinion reviews continues to grow, the number of fictitious reviews written for commercial gain also increases. By way of example, one may ask a third person who has not used his/her company to leave positive opinions for marketing his/her company or to leave malicious opinions about a rival company. As such, an opinion written with intent regardless of the experience of using a service or a product is called opinion spam. Most opinion spam has been written too skillfully to be recognized by human readers and has become increasingly more difficult to detect. Accordingly, in recent years, studies have attempted to systematically distinguish different types of opinion spam using classification algorithms [1, 4, 5, 18, 19, 21, 22, 23, 24, 25]. Even though many research studies on detecting spam reviews have been conducted, limitations of gold-standard datasets still prove to be a major hindrance to opinion spam research. One of the representative gold-standard deceptive opinion spam datasets was made by Ott et al. [25]. The deceptive reviews in their dataset were written by Amazon Mechanical Turk (AMT) task respondents (also known as Turkers). However, since the Turkers did not actually visit the target hotels, their reviews about the hotels lacked factual information and details [14].

Recently, Li et al. [19] have pointed out that Turkers are not a representative of the general population since they do not have any experience or information about hotels that they were asked to review. To address such a problem, the authors hired real hotel employees qualified to write reviews about certain hotels, and created a dataset of all the reviews. Because the reviews in this dataset were written by the hotel employees who knew the target hotels, it does not lack factual information. However, their dataset lacks details about the subjective experiences of travelers [14]. Furthermore, since employing domain experts is a high-cost method, the amount of opinion spam it generates may be insufficient for representing real-world spam.

While there are many types of opinion spam, there exist few gold-standard spam datasets. To construct a robust spam classifier that can be used in a real-world setting, one needs to train their model using spam from as many different spam domains as possible because spam in the real world is diverse. In this regard, we believe that a new type of review spam will help to construct a better classification model and ultimately advance opinion spam research.

In this paper, we propose a new method for creating a deceptive review dataset, based on the observation that spammers reference truthful reviews to write a deceptive review [9]. For our Paraphrased OPinion Spam (POPS) dataset, we asked AMT Turkers to paraphrase actual reviews of hotels so that they can convey factual information and include personal experiences of original truthful reviewers in their deceptive reviews. Thus, our dataset overcomes the limitations of existing datasets and contains reviews that are similar to truthful reviews. The major contributions of this research can be summarized as follows:

- We introduce a novel deceptive opinion spam dataset, Paraphrased OPinion Spam (POPS) dataset, which contains high-quality fictitious reviews.
- We provide a detailed explanation of the dataset generation process including the task set preparation, crowdsourcing task assignment, quality control, and response filtering.

[†] These authors contributed equally to this work. S. Lee is now in Kiwiple Inc., Seoul, Korea.
[*] Corresponding author.
[1] https://github.com/jacobis/paraphrased-opinion-spam

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052607

- Through various experiments and analyses, we found that the deceptive reviews in our novel POPS dataset most resemble truthful reviews, and that our dataset is difficult to classify using state-of-the-art classification models.

- To the best of our knowledge, this is the first study to create a deceptive opinion spam dataset using the concept of paraphrasing given truthful reviews. As our POPS dataset provides a new perspective on review spam, it will raise several new research questions and will help further spam research.

2. BACKGROUND AND RELATED WORK

Research on review spam has advanced since the studies by Jindal and Liu [11, 12, 13]. The following opinion spam datasets and methods have been recently proposed.

Existing deceptive spam datasets. Ott et al. [25] utilized Amazon Mechanical Turk (AMT) to create a deceptive spam dataset. A pool of 400 HITs (Human Intelligence Tasks) were created to obtain deceptive reviews on 20 hotels in the Chicago area. The Turkers were paid one US dollar for an accepted submission. 400 truthful reviews in their dataset were collected from TripAdvisor[2]. Ott et al. [23, 24], Feng et al. [4, 5], Mukerjee et al. [21], Li et al. [18], and Barnerjee et al. [1] conducted follow-up research studies. For instance, Mukerjee et al. [22] used real-life Yelp.[3] data in their study. Also, Li et al. [19] developed multi-domain (i.e., Hotel, Restaurant, and Doctor) deceptive datasets that include reviews by domain experts to establish general rules for identifying deceptive opinions in reviews. They asked two hotel employees (experts in the hotel domain) from each of seven hotels to each writes 10 deceptive positive and 10 deceptive negative sentiment reviews. As a result, they obtained a total of 280 deceptive reviews. In our study, we used only 140 positive sentiment spam reviews for the sake of experiment.

Deceptive spam generation using truthful reviews. Sun et al. [26] developed a review synthesis method that generates review spam using truthful reviews. To create a spam review, a review synthesizer replaces sentences in a real review with sentences that are the most similar to the replaced sentences. The sentences that replace the original sentences are taken from a pool of truthful reviews that are obtained from the review website. Although this method can effectively mass-produce fictitious reviews, the generated reviews may read awkwardly to humans as sentences are drawn from multiple reviews written by different people.

Automatic text generation using a deep neural network. Recently, researchers who work in the natural language processing domain have been actively conducting studies on automatically generating text using deep learning techniques. Such automatic text generation methods can be used to generate a large volume of opinion spam. Li et al. [17] proposed a hierarchical neural autoencoder to reconstruct input paragraphs while preserving the syntactic and semantic properties and discourse coherence of an original text. In their experiments, the authors used hotel review data as one type of training data and provided output results that were obtained after paraphrasing input review text. The results were interesting, but contain awkward expressions that are insufficient to be used as sophisticated review spam.

Semantic frame-based review analysis. Kim et al. [14] proposed a frame-based semantic analysis method to better understand review characteristics. To analyze the difference between truthful and deceptive review datasets at the semantic level, they proposed two statistical analysis methods: Normalized Frame Rate (NFR) and Normalized Bi-frame Rate (NBFR). This method is suitable for analyzing how similar a real review is to the reviews in our proposed dataset at the semantic level. Therefore, in this work, we applied the NFR and NBFR method to analyze the semantic differences between our POPS dataset and the other datasets.

3. POPS DATASET GENERATION PROCESS

In this section, we explain the philosophy of creating our POPS dataset. In the following subsections, we include an overview of generating the POPS dataset using crowdsourcing, and provide details of the task description and response filtering.

3.1 POPS Dataset Philosophy

Why we need a new type of spam. As interest in deceptive opinion spam has been increased recently, a large body of works on identifying and analyzing opinion spam have been conducted. However, existing opinion spam datasets are limited in representing various types of spams that exist in real-life. Of course, there is no possible way to know how real-world spam is composed, how many different spam types exist, which spam type is prevalent, and so on. Therefore, it is important to acquire and analyze various types of spam datasets as much as possible in order to understand the nature of opinion spam. This motivated us to develop a novel method of paraphrasing fictitious reviews to create deceptive opinion spam.

Spammers often reference written truthful reviews. We have noticed the tendency of spammers to often reference already written truthful reviews to imitate opinions of a real situation [9]. To model this, we aimed to create a dataset of paraphrased deceptive positive sentiment reviews using the AMT crowdsourcing service. For the task, we gave Turkers reviews that contain factual information and real opinions and feelings of actual visitors of specific hotels [7], and then provided guidelines for paraphrasing the reviews. Even though paraphrased reviews are based on truthful reviews, they should be considered as spam because they were written by Turkers who do not have actual experience of target hotels.

3.2 Dataset Preparation and Task Description

Truthful review grouping for task set generation. After we created a HIT, we provided Turkers with four truthful reviews of a specific hotel taken from the TripAdvisor dataset of Ott et al. [25]. The four reviews were prepared in the following way. The TripAdvisor dataset contains 20 reviews of each hotel. We made a task review set by randomly sampling four out of the 20 reviews. We repeated this process until five task sets per hotel were created (Figure 1-(a)). Then, we assigned four different Turkers to each task set and created a total of 400 paraphrased opinion spam reviews (20 hotels × 5 task sets × 4 Turkers).

Paraphrasing task description. To retain the contents of real reviews, we asked Turkers to select five aspects that are emphasized the most in the provided reviews (Figure 1-(b)). Suppose that the following sentence is given to Turker: *"This hotel may be a bit difficult for your cab driver to find, but it is worth getting lost for. It truly is two steps from the main plaza, but is hidden among the azaleas and honeysuckle behind a privacy gate that opens up on a paradise of blooming plants and trees laden with flowers and citrus fruit."* In this review, features such as "view" and "location" are the most emphasized. This process enables Turkers to understand aspects described in reviews, and can be used to see how many different aspects are included. The provided list of aspects[4] is as follows:

[2] https://www.tripadvisor.com

[3] https://www.yelp.com

[4] The aspects were selected from hotel reservation sites such as Hotels.com.

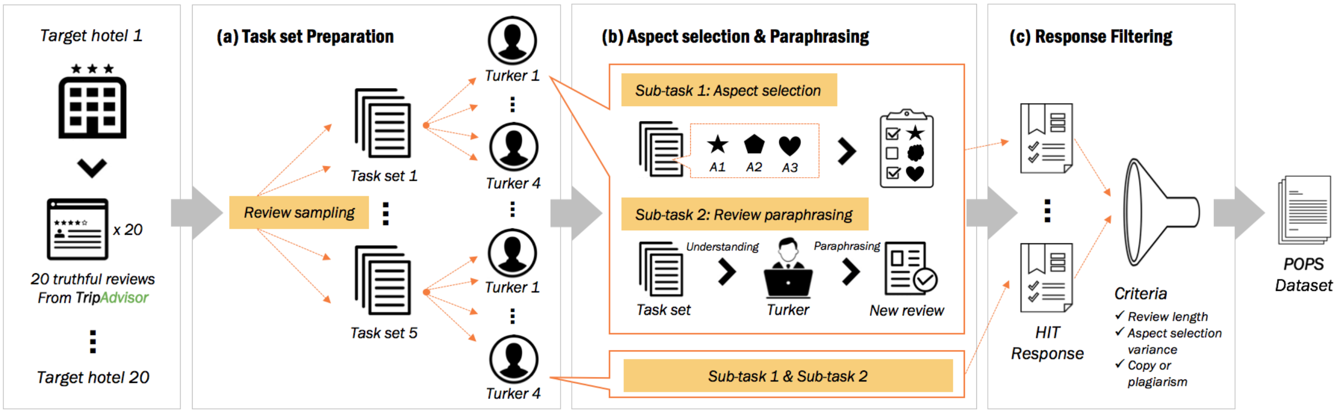

Figure 1: Overview of the POPS dataset generation process.

Building (exterior), Room (condition or size), Bathroom, Bedding, Electronics, Amenities (bar, business facilities, gym, parking area, pool, restaurant, smoking areas, spa, etc.), Additional service (airport transfers, free Wi-Fi, pet friendly, etc.), Service (concierge, hotel), Price, Sights (tourist spots), Location, Hotel food, Plan (reviewer schedule), View, and 'Input of your own'. For the next step, we asked Turkers to write a new paraphrased hotel review based on the selected aspects using the following instructions:

- Write a new paraphrased review for the hotel based on the selected features. Assume that you have indirect experiences of the hotel through the provided reviews.
- Try to use your own style of writing and different expressions when you state factual information from the provided reviews (e.g., beds and pillows ⇒ bedding).
- The length and emotional intensity of the new review should be similar to those of the provided reviews.

The key feature of our dataset is newly created deceptive spam that contains factual information from given truthful reviews. The deceptive reviews have almost the same characteristics as the given truthful reviews. Furthermore, when writing deceptive reviews, Turkers were instructed to use the same or similar review length and emotional intensity of given truthful reviews. Note that emotional intensity was very difficult to quantify. Hence, we instructed Turkers to use a similar intensity level and their own judgment, instead of strict instructions.

3.3 Response Filtering

We rejected some of the submitted reviews written by Turkers to control the quality of our dataset (Figure 1-(c)). First, we filtered reviews if their length was too short (less than 150 words) or differed too much from the length of the given truthful reviews. Second, we compared the selected aspects of responses of Turkers who received the same test set to check whether Turkers satisfactorily understood the given reviews and they are consistent in aspect selection. A response was considered as an outlier and rejected if its selected aspects differed too much from those of the responses submitted from same group of Turkers. Lastly, we also discarded reviews that Turkers copied or plagiarized from the Web[5].

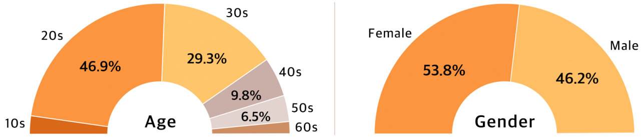

Figure 2: Age and gender distribution of Turkers.

Table 1: Comparison of statistics between datasets.

	TripAdvisor	AMT	Employee	POPS
Num. of reviews	400	400	140	400
Num. of unique words	4171	4657	1645	3112
Size (in kB)	265	254	62	255
Avg. length	677.7	636.0	399.7	636.6
Time spent (min)	-	8.06	-	13.4

4. DATASET STATISTICS IN DETAIL

4.1 Basic Statistics

Turkers' age and gender distribution. Figure 2 shows the age and gender distribution of Turkers. Over 76% of Turkers were in their 20s and 30s. 53.8% and of the Turkers who participated in the paraphrasing task were female and 46.2% were male.

Review length and processing time. In Table 1, TripAdvisor and AMT refer to the truthful review and deceptive review datasets, respectively, from the study of Ott et al [25]. Employee refers to the deceptive review dataset from the study of Li et al. [19]. The POPS dataset and the dataset of Ott et al. are about the same in size and length. However, as we have our novel dataset generation pipeline in place, we can easily expand POPS if requested by the community. The average length of the POPS reviews did not differ much from that of the TripAdvisor reviews that were provided for Turkers. The processing times for our POPS dataset were 5 minutes longer on average than those for the AMT dataset. This difference can be attributed to the guideline that instructs to carefully read provided truthful reviews.

The most selected aspects in POPS dataset. We asked Turkers to select the five aspects that are emphasized the most in the given reviews. Table 2 shows the list of the top five and bottom five aspects in our POPS dataset. Turkers mainly selected aspects such as Location, Service, Price, and Room; however, they did not select aspects such as Plan, Electronics, and Building in the provided reviews written by truthful reviewers. This result resembles

[5] http://plagiarisma.net;http://copyscape.com

Table 2: List of the top and bottom five aspects of the POPS dataset.

	Aspect Name	# of selected aspects
Top 5	Room	336
	Location	332
	Service	323
	Price	179
	Sights	149
Bottom 5	Additional Information	78
	Addition	77
	Building	36
	Electronics	20
	Plan	4

the frame analysis result presented in a later section, which indicates that aspects selected by Turkers largely affects the contents of paraphrased reviews.

5. DATASET EVALUATION

We are mainly focused on i) verifying how well our dataset is constructed using task instructions explained in the previous section and ii) evaluating the existing classification model performance on our dataset. First, we classified our POPS dataset using Review Skeptic[6] to find out how the existing spam classification service handles our POPS dataset. Second, we classified our POPS dataset using our own implementation of the classification model proposed by Ott et al. [25]. Then, we cross-trained/tested existing opinion spam datasets to justify the importance of creating a new type of spam dataset. Finally, we performed a classification task using an up-to-date deep neural model (Bidirectional LSTM).

5.1 Datasets

5.1.1 Truthful Review Dataset

We used the truthful positive reviews on 5-star rated popular hotels in the Chicago area from the dataset of Ott et al. [25]. The reviews were gathered from TripAdvisor.

Is it okay to treat all TripAdvisor reviews as truthful? It is nearly impossible to know how real-world spam is composed, how many different spam types exist, which spam type is prevalent, and so on. Thus, we cannot guarantee that the truthful reviews from TripAdvisor are in fact truthful. However, by careful filtering, we can increase the integrity of a truthful dataset. To control the quality of their truthful dataset, Ott et al. [25] selectively filtered reviews gathered from TripAdvisor using filtering strategies (e.g., checking review rating consistency, filtering by review length, eliminating first-time author, etc.) that were previously studied in spam research. The truthful dataset has been used not only in our research study but also in many other previous spam research studies (e.g., the study by Li et al. [19]).

5.1.2 Deceptive Opinion Spam Datasets

For the comparative experiments, we use the following three datasets: AMT, Employee, and POPS. Each dataset is briefly described below.

a) AMT dataset. This dataset by Ott et al. [25] contains 400 deceptive hotel reviews. Note that no prior information about target hotels was provided to Turkers when writing a review.

b) Employee dataset. Li et al. [19] created a total of 280 positive and negative sentiment hotel reviews. Only 140 positive reviews were used for our experiment.

Table 3: The classification results on different datasets by Review Skeptic.

Dataset	TripAdvisor	AMT	Employee	POPS
# of reviews	400	400	140	400
# of correct answers	348	400	58	236
Accuracy (%)	0.87	1	0.41	0.59

Table 4: Classification results using SVM on different datasets. AMT(Ott'11) is the result published in [25] and the remaining rows contain the classification performance results of our implementation of the model.

Dataset	Feature	Acc	Truthful			Deceptive		
			Prec.	Rec.	F1	Prec.	Rec.	F1
AMT (Ott'11)	UNI	0.884	0.899	0.865	0.882	0.870	0.903	0.886
	BI+	0.896	0.901	0.890	0.896	0.891	0.903	0.897
AMT	UNI	0.870	0.872	0.868	0.870	0.868	0.873	0.870
	BI+	0.876	0.879	0.873	0.876	0.873	0.880	0.877
Employee	UNI	0.916	0.936	0.953	0.944	0.857	0.814	0.835
	BI+	0.894	0.911	0.950	0.930	0.837	0.736	0.783
POPS	UNI	0.807	0.808	0.808	0.808	0.808	0.808	0.808
	BI+	0.818	0.824	0.810	0.817	0.813	0.828	0.820

c) POPS dataset. The POPS dataset is a paraphrased dataset of 400 deceptive positive sentiment reviews, which is described in detail in the previous section.

5.2 Classification Experiments using Review Skeptic

Current classification model is data dependent. Review Skeptic is an online tool that implements a classification model from Ott et al. [25]. We tried four different datasets using Review Skeptic (Table 3). Review Skeptic classified the AMT deceptive dataset with a perfect accuracy of 100% and classified the TripAdvisor truthful dataset with a high accuracy of 87%. From this result, we assume that the classification model of Review Skeptic is trained on the AMT dataset. It is important to note that Review Skeptic performs additional training using the feedback of users.

We also found that Review Skeptic's classification model is data dependent. Review Skeptic classified the POPS and Employee datasets with far lower accuracies of 59% and 41%, respectively. More importantly, these results show that current state-of-the-art spam detection tools cannot accurately identify specific types of spam that are based on factual information but still deceptive such as spam in the POPS or Employee datasets.

5.3 Classification Experiments using Model of Ott et al.

We implemented a linear support vector machine (SVM) based on the spam classification model of Ott et al. [25] to discover how well their model classifies our POPS dataset. (Table 4[7]) The best feature combination of their model was LIWC[8] and n-grams (Uni+Bi), which achieved an accuracy of 89.8%. Note that even though LIWC helped achieve their best performance, this feature does not significantly contribute to the model. Hence, to simplify experiments, we used UNIGRAM and BIGRAM as main features for training our model. The difference in performance (2.0% in accuracy) between our implementation of the model by Ott et al. (87.6%) and their implementation (89.6%) resulted from several reasons such as tokenization, model parameter values, and so on.

[6] http://reviewskeptic.com

[7] Superscript $^+$ subsumes the previous feature set.

[8] http://liwc.wpengine.com

Table 5: Cross-set classification experiment result.

		Training set													
		AMT		POPS		Employee		AMT & POPS		POPS & Employee		AMT & Employee		ALL	
		Bi+	Frame+	Bi+	Frame+	Bi+	Frame+	Bi+	Frame+	Bi+	Frame+	Bi+	Frame+	Bi+	Frame+
Test set	**AMT**	**0.876**	**0.878**	0.733	0.733	0.53	0.533	0.82	0.818	0.764	0.773	0.844	0.854	0.808	0.805
	POPS	0.68	0.686	**0.799**	**0.811**	0.509	0.506	0.791	0.789	0.793	0.795	0.675	0.69	0.779	0.78
	Employee	0.767	0.757	0.698	0.689	**0.916**	**0.913**	0.72	0.709	0.807	0.806	0.831	0.841	0.746	0.744
	AMT & POPS	0.753	0.761	0.758	0.768	0.378	0.382	0.808	0.809	0.788	0.798	0.749	0.764	**0.835**	**0.838**
	POPS & Employee	0.652	0.657	0.746	0.749	0.564	0.562	0.786	0.784	**0.817**	**0.825**	0.716	0.732	0.807	0.811
	AMT & Employee	0.822	0.822	0.687	0.694	0.582	0.585	0.811	0.81	0.795	0.803	**0.862**	**0.857**	0.835	0.838
	ALL	0.725	0.729	0.724	0.735	0.431	0.432	0.83	0.83	0.807	0.816	0.77	0.781	**0.849**	**0.851**

To compare the model's classification performance on the POPS dataset with that on the AMT and Employee deceptive datasets, we trained our model on the TripAdvisor (truthful) dataset and the POPS (deceptive) dataset (similarly AMT and Employee). Each model was evaluated using 5-fold nested cross-validation. The classification accuracy of the POPS dataset (81.8%) is about 6-7% lower than that of the AMT (87.6%) and Employee (89.4%) datasets. This result shows that the POPS dataset is more difficult to classify than the AMT and Employee deceptive datasets because there are fewer linguistic differences between the deceptive reviews of the POPS dataset and the truthful reviews.

5.4 Cross-train & Cross-test Datasets

We cross-trained and cross-tested the datasets where each dataset was used alone or shuffled with other datasets (Table 5). We built an SVM classifier using BIGRAM and FRAME[9] features with a default parameter setting utilized in the study by Kim et al. [14]. We sampled each dataset using the same proportion due to the different sizes of the datasets. For example, we randomly sampled 200 reviews from the AMT dataset and 200 reviews from the POPS dataset to create a pool of 400 deceptive reviews. To minimize the deviation of the results, we reported an average accuracy of a total of 10 repeated experiments.

The result in Table 5 shows that when training the same target dataset for testing, the classification performance was the highest in most cases (highlighted in bold). In other words, the classification model that trained solely on the POPS dataset can much better classify a POPS type of dataset than a model trained on the other dataset or combined datasets (AMT, Employee). In addition, the model trained on datasets without Paraphrased OPinion Spam (POPS) could not classify the POPS dataset as well as the models trained on datasets with POPS (10 to 20% decrease in classification accuracy).

In reality, since many types of spam with various characteristics exist, it is extremely difficult to classify all spam. To resolve such a problem, we presumed that spam, in reality, is *all* the datasets in Table 5 which contains all types of existing deceptive review datasets, and then proceeded with our experiment. When we trained a model on all the datasets, we obtained the best classification performance with an accuracy of 85.1%. As a result, the most practical way to better classify all types of spam is to find as many various spam types as possible and train models on all the types that are found.

5.5 Classification Task using a Deep Neural Model

Deep neural models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are popular because they achieve state-of-the-art results in many natural language processing (NLP) tasks [15]. Due to their capability of processing variable-length text, RNNs are widely used in many NLP tasks such as text classification. While there exists a large body of work on text classification using a deep learning approach [16], opinion spam is rarely considered. In this experiment, we investigate how well a deep neural model identifies opinion spam using different datasets.

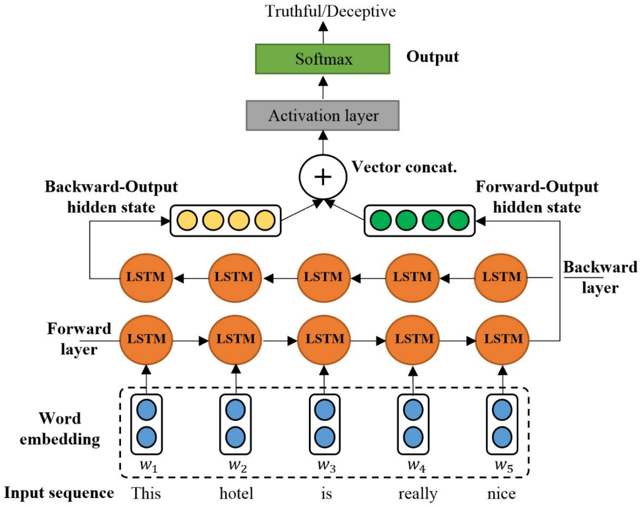

Figure 3: Bi-directional LSTM architecture of our classification experiment. $w_1, w_2, \cdots, w_5$ is an input sequence. The final hidden states of the forward and the backward layer are concatenated and fed to an activation layer. The class label is determined in the final softmax layer.

Table 6: Classification results on different datasets using deep neural model.

Dataset	Acc.	Truthful			Deceptive		
		Prec.	Rec.	F1	Prec.	Rec.	F1
AMT	0.864	0.880	0.852	0.866	0.848	0.876	0.865
POPS	0.813	0.835	0.799	0.817	0.790	0.827	0.808

Hyper-parameters and training details. We used bidirectional LSTM [8], a commonly used variant of the basic LSTM architecture, for our experiment. Bidirectional LSTM consists of two LSTM layers that run parallel so that the past and future context of an input are captured (Figure 3). We initialized the input word representations of our LSTM model using publicly available 300-dimensional GloVe[10] vectors (Pennington et al., 2014), which are trained on 840 billion tokens of Common Crawl data. Word representations that are not present in the GloVe vectors are randomly initialized in a uniform distribution of (-0.25, 0.25). Our models

[9] From the frame list, we select the top and bottom k frames sorted by ΔNFR value.

[10] http://nlp.stanford.edu/projects/glove/

Table 7: List of top 10 words that have caused a given review to be truthful analyzed by Review Skeptic ($p(w)$ in 1.0$\times 10^{-3}$).

Word (w)	TripAdvisor		AMT		POPS	
	Count	p(w)	Count	p(w)	Count	p(w)
<number>	171	3.08	19	0.37	23	0.46
location	94	1.69	21	0.41	88	1.76
)	83	1.50	8	0.16	19	0.38
(	70	1.26	7	0.14	19	0.38
<money>	55	0.99	2	0.04	10	0.20
floor	39	0.70	1	0.02	7	0.14
large	34	0.61	4	0.08	30	0.60
small	30	0.54	0	0.00	14	0.28
reviews	20	0.36	1	0.02	4	0.08
door	17	0.31	3	0.06	7	0.14
Sum	613	11.05	66	1.29	221	4.42

Table 8: Part-Of-Speech sequence similarity between different review sets.

Set Category	Similarity Value
Within-Set	0.639
Cross-Set	0.636

were trained using AdaGrad [3], with a learning rate of 0.1, a mini-batch size of 10, a dropout rate of 0.5, and a total epoch of 30.

Classification result using Bidirectional LSTM. Using bidirectional LSTM, we perform a classification task to compare our POPS deceptive dataset and the AMT deceptive dataset (here, the Employee dataset was excluded due to its limited number of training examples), and report the results in Table 6. Each model was evaluated using 5-fold nested cross validation and the same experiment setting used for the SVM model described above. Interestingly, the classification accuracy of the POPS dataset (81.3%) is lower than that of the AMT dataset (86.4%). Even with the up-to-date deep neural model, the result shows again that the POPS dataset is more difficult to classify than the other datasets because its reviews are linguistically similar to the truthful reviews.

6. LINGUISTIC SIMILARITY ANALYSIS

Through the classification experiment in the previous section, we have shown that the POPS dataset is more difficult to classify than other existing datasets since its reviews resemble truthful reviews. In this section, we will analyze how the POPS and truthful reviews are similar from a linguistic and informatics point of view.

Did Turkers use their own style of writing? The major advantage of the paraphrasing method is that the review contents submitted by four Turkers can be diverse due to the different writing styles of the Turkers, even though the given task set is the same. Lipka et al. [20] stated that an author's intrinsic text synthesis traits (writing style) can be captured by Part-of-Speech (POS) n-gram vectors. On this basis, we compared POS sequences of sentences in reviews submitted by Turkers who were given the same task set to see how writing style affects the diversity of reviews. To do this, we computed the cosine similarity of POS vectors between reviews submitted by four Turkers who shared the same task review set in a pairwise manner, which we call a within-set. Next, we randomly selected review pairs from different task review sets (cross-set). We repeated this procedure 10 times and reported the averages in Table 8. Interestingly, the similarity values of two different sets were almost the same. From this result, we conclude that even though four Turkers share the same task set, their review contents and expressions are as diverse as the reviews generated from different task review sets, due to the various writing styles of Turkers.

Table 9: Comparison of KL-distance between truthful and two deceptive datasets.

$P(x)$	$Q(x)$	$D_{KL}(P \parallel Q)$	$D_{KL}(Q \parallel P)$	$D_{KLD}(P \parallel Q)$
TripAdvisor	AMT	0.994	0.807	1.801
	POPS	0.779	0.524	1.303

How many truthful words are included in each dataset? Review Skeptic highlights truthful or deceptive words as its analysis result. As shown in Table 4, the performance of the Review Skeptic's truthful review classification is very high (87%). Therefore, the words highlighted as truthful by Review Skeptic are reliable. We used this analysis function of Review Skeptic to measure how many specific words were included in each data set. Table 7 shows a list of the top 10 words that caused a given review to be truthful, and were derived from Review Skeptic. We observed that our POPS dataset contains words that are used more often in the truthful reviews (TripAdvisor), compared with the AMT dataset. We also calculated the probability of the top 10 words that appear in each dataset. The sum of the probability is $\sum_{i=1}^{k} p(w_i)$ where k is the top-k words in the list, and $p(w_i)$ is the probability of word w_i in the dataset. Although both the POPS and AMT datasets were created using the same crowdsourcing service, the POPS dataset contains 3.4 times more truthful words than the AMT deceptive dataset (4.42 vs. 1.29).

Information-Theoretic analysis. To compare the probability distributions of all words including specific words, we calculated the Kullback-Leibler (KL) divergence between each dataset. The KL-divergence of probability distribution Q with respect to distribution P over words w in a vocabulary W is defined as:

$$D_{KL}(P \parallel Q) = \sum_{w \in W} P(w) \log \frac{P(w)}{Q(w)} \quad (1)$$

It is known that the KL-divergence is not symmetric, i.e., $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$. Since in this research we care only about the distance between truthful and deceptive reviews that are generated by different methods, we use the following symmetrized Kullback-Leibler distance measure, which is defined based on the Kullback-Leibler divergence:

$$D_{KLD}(P \parallel Q) = D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P) \quad (2)$$

Table 9 shows the KL-divergence and KL-distance values between the truthful review dataset and the other two deceptive datasets. As the values in Table 9 show, the KL-distance value between the truthful dataset and the POPS dataset is lower than that between the truthful dataset and the AMT dataset, implying that the POPS dataset is more linguistically similar to the truthful dataset.

7. FRAME-BASED SEMANTIC ANALYSIS

In this section, using the semantic frame-based analysis method, we will investigate how the internal composition of our POPS dataset may differ from that of other datasets and how the POPS reviews are similar to the truthful reviews. According to Kim et al. [14], there is a meaningful difference between deceptive reviews and truthful reviews and it is possible to capture some interpretable and unique semantic features using semantic frames[11] [6]. Their methodology was created to find the difference of frames between

[11] We use SEMAPHORE V2.1 [2], an automatic frame semantic annotation system used for frame extraction.

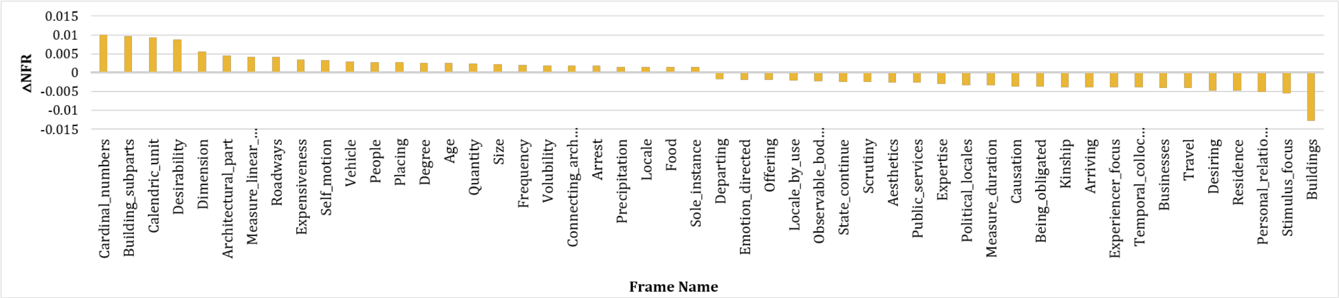

Figure 4: Top 50 differentially expressed frames sorted by ΔNFR in the AMT dataset.

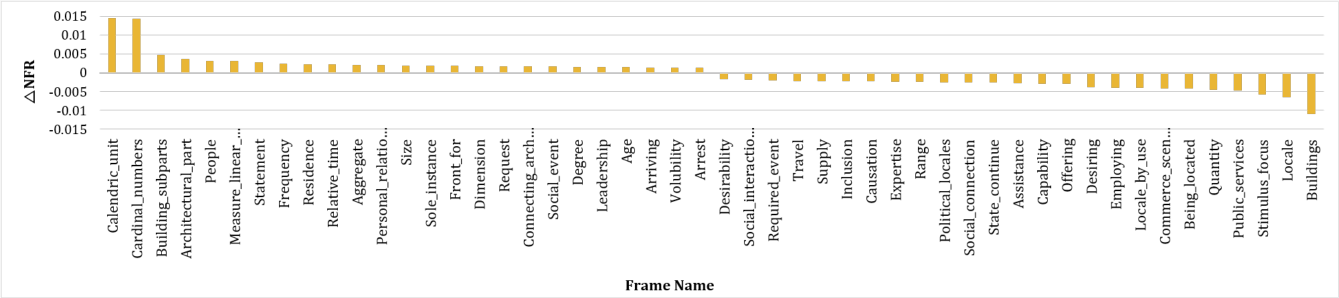

Figure 5: Top 50 differentially expressed frames sorted by ΔNFR in the POPS dataset.

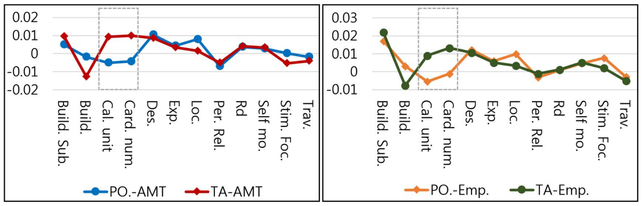

Figure 6: Graphs of ΔNFR value of PO.-AMT and TA-AMT; PO.-Emp and TA-Emp. The two lines of each graph show a similar trend except for Cardinal_numbers and Calendric_Unit frames. (PO: POPS; TA: truthful reviews from TripAdvisor; Emp.: Employee).

Table 10: NFR differences of key frames between POPS dataset and the other datasets.

Frame Name	PO.-AMT	TA-AMT	PO.-Emp.	TA-Emp.
Building_subparts	0.00487	0.00958	0.01717	0.02188
Buildings	-0.00182	-0.01272	0.00320	-0.00769
Calendric_unit	-0.00517	0.00933	-0.00547	0.00903
Cardinal_numbers	-0.00431	0.00996	-0.00107	0.01320
Desirability	0.01035	0.00871	0.01235	0.01071
Expensiveness	0.00431	0.00348	0.00601	0.00518
Locale	0.00796	0.00150	0.00982	0.00336
Personal_relationship	-0.00706	-0.00504	-0.00316	-0.00114
Roadways	0.00365	0.00405	0.00084	0.00124
Self_motion	0.00281	0.00322	0.00466	0.00507
Stimulus_focus	0.00028	-0.00541	0.00765	0.00196
Travel	-0.00193	-0.00408	-0.00294	-0.00509

truthful reviews and deceptive reviews. Frames differences are measured by the Normalized Frame Rate (NFR) which represents how often a specific frame appears in a dataset. The NFR value is the ratio of how many times a particular frame appears in a dataset relative to other frames. The difference between two NFR values, ΔNFR for a specific frame f_i appeared in each dataset D_m and D_n is calculated as below:

$$\Delta NFR_{f_i} = NFR_{D_m} f_i - NFR_{D_n} f_i \tag{3}$$

Suppose D_m is a truthful dataset and D_n is a deceptive dataset. According to the definition of ΔNFR, when the value of ΔNFR is positive, the frame frequently appears in the truthful dataset and vice versa.

7.1 Qualitative Analysis of Frames

Figure 4 and 5 are graphs of the top 50 frames that are sorted by ΔNFR value between the truthful dataset (TripAdvisor) and each of the deceptive datasets (AMT, POPS). Table 10 contains the specific frames' NFR difference between the POPS dataset and the other deceptive datasets. Similarly, if the ΔNFR value is positive, the frame (blue in Table 10) frequently appears in the POPS dataset.

We observe that the ΔNFR values between the POPS dataset and other datasets pairs show a trend similar to ΔNFR values between the TripAdvisor and other datasets (Figure 6 and Table 10). The Pearson product-moment correlation coefficients of PO.-AMT/TA-AMT and PO.Emp./TA-Emp., except Cardinal_numbers and Calendric _unit frames, are 0.76 and 0.84, respectively, which indicates a strong correlation. As shown in Figure 6, the difference between the NFRs of the frames, except Cardinal_numbers, Calendric_unit, is very small (a dashed line denotes a large gap between two ΔNFR values). The difference in the frequencies of the two frames is due to the reproduction of the information. This issue will be explained in more detail in the Discussion section.

1. Spatial and figure-related frames. In the study by Kim et al. [14], frames such as Building_subparts, Cardinal_numbers, Calendric_unit, Dimension, and Expensiveness were used to describe spatial and figure-related frames. For example, the Building_subparts frame represents lexical units such as room, bathroom, elevator, and lobby. In the POPS dataset, the NFR differences of the above five frames were larger than those of the five frames in the other deceptive datasets because the POPS dataset contains detailed information about target hotels as intended.

In Table 10, the ΔNFR of the Building_subparts frame in the POPS dataset is +0.005[12] and +0.017 higher than the ΔNFRs of the Building_subparts frames in the AMT dataset and the Employee dataset, respectively. This result proves that the lexical units related to the Building_subparts frame, which generally has strong positive ΔNFR values in the truthful dataset were used more frequently in the POPS dataset. The Expensiveness frame represents lexical units related to the prices of a hotel. Lexical units such as "affordable" and "pricey" that actual visitors of hotels frequently used were well described in the POPS dataset. On the other hand, the Cardinal_numbers and Calendric_unit frames representing specific dates of travel and duration of stay appeared less frequently in the POPS dataset.

2. Location-related frames. We also focused on the Locale, Roadways, and Self_motion frames that represent a hotel's specific location and position. For the POPS dataset, rather than general expressions, Turkers often used specific words such as "walking distance," "Avenue," "across the street," and "intersection" that describe specific locations and directions. In other words, POPS better describes expressions in truthful reviews written by truthful reviewers.

3. Personal relationship-related frames. Many previous studies [14, 19, 25] mentioned that spammers tend to use personal relationship terms to resemble a truthful reviewer. For example, Personal_relationship lexical units such as "husband," "wife," "friend," and "family" are used intentionally so a review sounds convincing. However, the Personal_relationship frame was used 0.007 and 0.003 less in the POPS dataset than in the AMT and Employee datasets, respectively. This result suggests that although POPS, like the AMT deceptive reviews, is generated using the crowdsourcing method, there are few relational expressions that Turkers use habitually because Turkers are more focused on restating factual information contained in provided truthful reviews rather than trying to make it sound real.

4. Sentiment expression-related frames. Kim et al. [14] focused on the frame difference between the Stimulus_focus and Desirability frames to study emotional expressions in truthful reviews and deceptive reviews. According to Kim et al. and FrameNet, the Stimulus focus frame brings about a particular emotion or experience in the Experiencer.

On the other hand, the Desirability frame concerns an Evaluee being judged for its quality. Truthful users tend to objectively evaluate their actual experience (Desirability) whereas Turkers tend to overexpress emotions about an imaginary circumstance (Stimulus_focus). Interestingly, the Desirability frame, which generally appeared more often in the truthful dataset than in the deceptive datasets, appeared more frequently (ΔNFR value of -0.001 in Figure 5) in the POPS deceptive dataset than in the TripAdvisor truthful dataset and the other deceptive datasets (Table 10).

8. DISCUSSION

In this section, we point out the following general limitations of our POPS dataset. First, the POPS dataset is dependent on the characteristics of original truthful review sets since it is basically a paraphrased version of the review sets. Thus, the quality of fabricated deceptive reviews is lower than that of the original reviews. Second, even though truthful reviews were imitated to create the POPS dataset, we found that not all the factual information was mentioned. For example, Cardinal_numbers and Calendric_unit frames (known to frequently appear in truthful reviews) representing specific dates of travel and duration of stay appeared less frequently in the POPS dataset.

• **Review 1.** My husband and I first stayed at the Amalfi in 8/06 totally based on reviews I read on this website. ··· I returned to Chicago with 3 girlfriends in 9/07 and booked us in a suite at the Amalfi. It was great ···
• **Review 2.** ··· The 'Taste of Chicago' was on during the 4th of July. Let me tell you, this was the best 'taste' of a city that I have ever been to. ··· I took the train in from the airport (it was around $2.00 but it was a solid 45 mins but for $2.00, how can you complain?)···
• **Review 3.** My wife and I just returned from a 3 night stay at the Amalfi Hotel in Chicago and enjoyed everything about the Hotel. ···
• **Review 4.** We stayed in a "Corner King" room. The room was amazing. Two of the four walls are floor to ceiling windows. Even just being on the 5th floor, we had great views of the surrounding area. ···

• **Resulting Paraphrased Review.**
The Amalfi hotel in Chicago is a wonderful hotel. It is close by many great locations to shop and dine out. There is a great atmosphere around the hotel and it is a very inexpensive ride by train to the hotel. The amenities are fantastic and the staff is wonderful! All the food I ate at the Amalfi was top notch. Even though there are great restaurants very close by, I would recommend just eating at the hotel because it's such a great environment and the food is fantastic. I loved the rooms, very spacious and lovely. There were great bath products (TOP OF THE LINE!) and the staff even left me champagne. I was thoroughly impressed by the entire stay and would recommend it highly to my friends and family.

Figure 7: An example of paraphrased review.

As Figure 7 shows, there are four given truthful reviews and a resulting paraphrased review. In this example, we do not give the full text of the truthful reviews due to the space limitations. Instead, we present review snippets which contain information about the date or numbers related to specific travel schedule. We observed that the resulting paraphrased review does not contain specific information from the given truthful reviews, which we expected to be included. We came up with the following two possible explanations for this:

1. Instruction bias. The HIT instruction to select five aspects for writing a new paraphrased spam review caused Turkers to miss some aspects of the given truthful reviews. For example, from their response reports, Turkers selected the Plan aspect (equivalent of Cardinal_numbers and Calendric_unit frames) only five times because the Plan aspect did not comprise a large portion of truthful reviews. This instruction bias may have hindered Turkers' ability to imitate factual information.

2. Too specific to imitate. The Turkers, who did not actually visit the hotels that they were asked to review, were unable to imitate actual visitors' itineraries with specific numbers or dates. Although the Turkers could refer to the truthful reviews, it was too difficult for them to fabricate specific itineraries similar to those of actual visitors.

[12] Previous study on frame-based opinion spam analysis [14] proved that a ΔNFR value near 0.01 is statistically significant ($p < 0.01$).

9. CONCLUSION AND FUTURE WORK

In this work, we proposed a new opinion spam generation method that paraphrases truthful reviews using crowdsourcing, based on the observation that spammers refer to truthful reviews when writing fictitious reviews. To model this observation, we carefully designed the human intelligence tasks (HITs) and instructed the AMT workers to use information from truthful reviews to make their reviews seem truthful.

To the best of our knowledge, our study is the first to use a paraphrased opinion spam dataset to address the deceptive opinion spam detection problem. The experimental results show that it is difficult for existing models (including up-to-date deep neural models), which are not trained on our POPS dataset, to classify our dataset because it contains deceptive reviews that resemble truthful reviews. The information theoretic frame-based semantic analysis results also confirm that our new dataset is similar to the truthful dataset.

However, there are some limitations in our dataset. For example, detailed information such as dates and itineraries were not well retained in paraphrased reviews. In future work, we will explain the HIT instruction in more detail so that Turkers can better understand the task. As our POPS dataset provides a new perspective on review spam, we hope that it will lead to several new research questions and will be a valuable resource for those who work in deceptive opinion spam research.

10. ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2014R1A2A1A10051238, 2012M3C4A7033341).

11. REFERENCES

[1] S. Banerjee and A. Y. Chua. Applauses in hotel reviews: Genuine or deceptive? In *Science and Information Conference (SAI)*, pages 938–942. IEEE, 2014.

[2] D. Das, N. Schneider, D. Chen, and N. A. Smith. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 948–956, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[3] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[4] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics, 2012.

[5] V. W. Feng and G. Hirst. Detecting deceptive opinions with profile compatibility. In *The 6th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 338–346. Association for Computational Linguistics, 2013.

[6] C. J. Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976.

[7] S. Gokhman, J. Hancock, P. Prabhu, M. Ott, and C. Cardie. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 23–30. Association for Computational Linguistics, 2012.

[8] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 273–278, 2013.

[9] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari. Detection of review spam: A survey. *Expert Systems with Applications*, 42(7):3634–3642, 2015.

[10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[11] N. Jindal and B. Liu. Analyzing and detecting review spam. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 547–552. IEEE, 2007.

[12] N. Jindal and B. Liu. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM, 2007.

[13] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM, 2008.

[14] S. Kim, H. Chang, S. Lee, M. Yu, and J. Kang. Deep semantic frame-based deceptive opinion spam analysis. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1131–1140. ACM, 2015.

[15] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[16] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2267–2273. AAAI Press, 2015.

[17] J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015.

[18] J. Li, M. Ott, and C. Cardie. Identifying manipulated offerings on review portals. In *EMNLP*, pages 1933–1942. ACM, 2013.

[19] J. Li, M. Ott, C. Cardie, and E. Hovy. 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)*, pages 1566–1576. Association for Computational Linguistics, 2014.

[20] N. Lipka and B. Stein. Identifying featured articles in wikipedia: writing style matters. In *Proceedings of the 19th international conference on World wide web*, pages 1147–1148. ACM, 2010.

[21] A. Mukherjee, B. Liu, and N. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200. ACM, 2012.

[22] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance. What yelp fake review filter might be doing. In *The 7th International AAAI Conference on Weblogs and Social Media*. AAAI, 2013.

[23] M. Ott, C. Cardie, and J. Hancock. Estimating the prevalence of deception in online review communities. In *Proceedings*

of the 21st international conference on World Wide Web, pages 201–210. ACM, 2012.

[24] M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In *HLT-NAACL*, pages 497–501. Association for Computational Linguistics, 2013.

[25] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics, 2011.

[26] H. Sun, A. Morales, and X. Yan. Synthetic review spamming and defense. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1088–1096. ACM, 2013.