Deep Semantic Frame-based Deceptive Opinion Spam Analysis

Seongsoon Kim
Korea University
Seoul, Korea
seongkim@korea.ac.kr

Hyeokyoon Chang
Korea University
Seoul, Korea
hyeokyoon@korea.ac.kr

Seongwoon Lee
Korea University
Seoul, Korea
seongwoon@korea.ac.kr

Minhwan Yu
Korea University
Seoul, Korea
minhwan90@korea.ac.kr

Jaewoo Kang*
Korea University
Seoul, Korea
kangj@korea.ac.kr

ABSTRACT

User-generated content is becoming increasingly valuable to both individuals and businesses due to its usefulness and influence in e-commerce markets. As consumers rely more on such information, posting deceptive opinions, which can be deliberately used for potential profit, is becoming more of an issue. Existing work on opinion spam detection focuses mainly on linguistic features such as n-grams, syntactic patterns, or LIWC. However, deep semantic analysis remains largely unstudied. In this paper, we propose a frame-based deep semantic analysis method for understanding rich characteristics of deceptive and truthful opinions written by various types of individuals including crowdsourcing workers, employees who have expert-level domain knowledge about local businesses, and online users who post on Yelp and TripAdvisor. Using our proposed semantic frame feature, we developed a classification model that outperforms the baseline model and achieves an accuracy of nearly 91%. Also, we performed qualitative analysis of deceptive and truthful review datasets and considered their semantic differences. Finally, we successfully found some interesting features that existing methods were unable to identify.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

General Terms

Algorithms, Experimentation

*Corresponding author.

CIKM'15, October 19–23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806551.

Keywords

Deceptive Opinion Spam, FrameNet, Semantic Analysis

1. INTRODUCTION

With the rapid growth of social network services and internet media, online review communities receive a significant amount of attention from both individuals and businesses. Online users increasingly rely on such reviews when making decisions. Recent surveys report that over 79% of users trust reviews written by others and consult them before making a purchase decision [1]. Naturally, there is also a growing concern about the potential for deceptive opinion spam or false reviews—fictitious opinions that are deliberately written to sound authentic to deceive readers for potential monetary gain [15]. Human readers usually cannot detect this kind of opinion spam because most opinion spam is carefully designed to avoid being identified by users or review content providers.

Since Jindal et al. introduced the first deceptive opinion spam problem to the research community [9], much significant work has been done on identifying deceptive opinion spam contents [4, 11, 15], individual spammers [10, 12] and spamming groups [13]. Recently, the research community as well as businesses have been concerned about the opinion spam issue. Yelp[1] may be one of the few top local business review hosting sites in the U.S. that takes this opinion spam issue seriously. It has been working on its own spam filtering method for over 10 years, and its spam filtering algorithm is still trade secret. Mukherjee et al. investigated the characteristics of real-life dataset collected from Yelp [14]. Even though the simple linguistic feature-based (such as n-gram) classification model is highly accurate in identifying crowdsourced deceptive opinion spam datasets [15], their study showed that it was no longer useful for the Yelp dataset. They used behavioral features of the users along with n-grams to classify Yelp's "filtered" and "non-filtered" reviews. The model that used behavioral features outperformed the baseline model that used only linguistic features. Mukherjee et al. assume that Yelp uses primarily the behavioral patterns of users with many internal metrics that are unavailable to the public.

[1]http://www.yelp.com

However, even though behavioral features can be used to identify deceptive opinions, they are ineffective at times. Suppose we used spammers' activity patterns to successfully find and block them from a site. If the spammers change their account name or make a new account (cold start), blocking spammers' access is no longer a useful filtering method. This scenario illustrates that content level review analysis still needs to be considered with individuals' behavioral patterns. Hence, the scope of this work also focuses on content level review analysis.

The majority of existing work focuses on review (or content) level classification tasks that are based on supervised learning approaches that use linguistic features. Linguistic features can be used as simple, yet powerful features in most classification tasks. However, we hardly know why specific features are highly weighted in trained models. Linguistic features do not convey semantic information that can help humans intuitively understand data. To gain insight into texts using linguistic features such as part-of-speech (POS) tags or n-grams, we need further analysis to interpret the characteristics of those features.

For example, Li et al. found that truthful reviews tend to include more nouns, adjectives, prepositions, and determiners (N, JJ, IN, and DT, respectively), while deceptive reviews tend to include more verbs, adverbs, and personal pronouns (V, RB, and PRP, respectively) [11]. The authors argue that domain experts tend to include more details about a local business using nouns, adjectives, preposition or determiners. However, the presence of particular linguistic features such as part-of-speech tags, LIWC categories, or n-gram tokens still cannot be solid evidence of truthful review contents. Also, Feng et al. showed that deep syntactic features driven from context free grammar (CFG) can improve opinion spam detection performance [4]. Their production rule is based on the CFG parse tree (e.g., NPˆS $\rightarrow$ NN, NPˆPP $\rightarrow$ DT NNP) which also does not carry meaningful semantic information.

In this paper, we view deceptive opinion analysis as a problem of understanding natural language. To overcome the limitations of the existing linguistic features mentioned above, we introduce a novel frame-based semantic feature based on FrameNet [5], a knowledge base in both human and machine-readable format. While there have been many studies on linguistic features such as n-gram, part-of-speech and syntactic structure, semantic analysis that uses frame features remains unstudied. Our contributions in this paper are four-fold:

- To the best of our knowledge, this is the first study to analyze deceptive opinion spam from a deep semantic level perspective.
- We propose two statistical analysis methods which we call Normalized Frame Rate (NFR) and Normalized Bi-frame Rate (NBFR) for the study of semantic frames.
- Using semantic frame features, we successfully improved the classification accuracy by 4.3% compared to the baseline.
- Last, we qualitatively analyzed the characteristics of all the truthful reviews and deceptive reviews in terms of their semantic differences.

We have found that truthful reviewers are more likely to focus on spatial details, which is consistent with previous studies [11, 15]. Moreover, truthful reviewers usually include exact dates or dimensions of objects to describe their real experience. Another interesting observation is that even though employees had expert-level knowledge of their specific business location and are able to include details in their review, the employees failed to mention details about their travel experience, itinerary, or travel activities. We also found different sentiment expression patterns in deceptive reviews and truthful reviews in terms of linguistic theory-based sentiment word categorization.

The rest of this paper is organized as follows. In Section 2, we briefly discuss previous opinion spam research trends. In Section 3, we describe our proposed Normalized Frame Rate (NFR) and Normalized Bi-frame Rate (NBFR) methods and their statistical validation. In Section 4, we qualitatively analyze semantic frame features. In Section 5, we report our own implementation of the baseline model in [15] and our classification results using semantic frame features with and without n-gram features. We conclude the paper and discuss future work in Section 6.

2. RELATED WORK AND BACKGROUND

Previous studies on spam detection. The opinion spam issue was first introduced by Jindal et al. [9] and was investigated after studies on web or email spam were published [2, 3, 8, 17]. The authors categorized review spam into several different types (untruthful opinions, reviews on brands only, and non-reviews) by characteristics, and proposed detection techniques using review centric, reviewer centric, and product centric features. In [12], target and deviation based spamming behavioral patterns were used to detect spammers. To support human evaluators' decision making, they also developed review spammer evaluation software that provides a visualization interface for reviews. The interface contains information about review duplication, review ratings, other recent ratings, and multiple reviews on the same products. Wang et al. proposed a review graph-based approach for finding fake reviewers [18]. They introduced three concepts regarding the trustiness of reviewers, the honesty of reviews, and the reliability of stores and their interrelationships. The experimental results and human participants confirmed that the proposed method can identify subtle spamming activities.

An in-depth study about deceptive opinion spam was conducted by Ott et al. [15]. They used Amazon Mechanical Turk[2] (AMT) to solicit fake reviews (which will be referred to as AMT dataset in this paper) about hotels in the Chicago area. Their model achieved high accuracy (89.8%) using simple n-gram-based features. Feng et al. introduced syntactic structures of review sentences and thus further boosted classification accuracy [4]. Li et al. developed multi-domain deceptive datasets (which will be referred to as Employee dataset) that include reviews by domain experts to establish general rules for identifying deceptive opinions in reviews [11]. Their Employee dataset compensates for the lack of experience of Turkers who also do not put much effort into faking reviews. Li et al. found that a lack of spatial details may not be a universal cue for deception. They also proposed general linguistic cues of deceptive opinion spam as different usage of part-of-speech tags, LIWC categories [16], and first-person singular pronouns.

[2]https://www.mturk.com/mturk/welcome

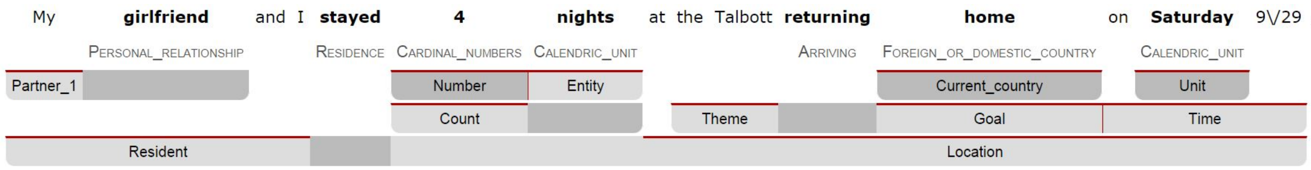

Figure 1: FrameNet parsing example from CMU's ARK syntactic & semantic parsing demo page (http://demo.ark.cs.cmu.edu/parse)

Although some studies focused mainly on identifying opinion spam content, some existing studies have used consumers' behavioral patterns to detect individual fake reviewers and fake review groups. For instance, Jindal et al. captured unusual review patterns using domain-independent rules [10]. Mukherjee et al. proposed several spamming behavioral indicators of spamming activities [13].

Semantic Frame Theory and FrameNet. FrameNet [5] is a lexical resource for English with semantic representations in combination of word sense disambiguation and semantic role labeling and is based on frame semantics [6]. Frames are cognitively founded and formally explored devices that represent knowledge about objects and categories by means of their attributes and values [7]. A word (or a lexical unit (LU)) in a sentence evokes a frame of semantic knowledge and the frame describes prototypical situations spoken in natural language. The frame contains a set of semantic roles that correspond to the participants (Frame Elements (FEs)) of a described event or situation. Many NLP researches use FrameNet especially for semantic role labeling, word sense disambiguation and question-answering.

In this work, we use FrameNet as our core analysis tool. In FrameNet, the machine processes all review content using semantic frame units. Unlike linguistic features, output results that are based on frames are also in human-readable and understandable format, which helps us understand the data more efficiently. We use a simple example to explain frame-based analysis.

Suppose we have the following sentence: *"My girlfriend and I stayed 4 nights at the Talbott returning home on Saturday 9/29."*[3] The FrameNet parsing result of the above sentence is described in Figure 1. In this example, a total of seven frames are identified. The darker gray box represents a frame that was evoked by a lexical unit and the successive lighter gray boxes represent their frame units (and frame range). Note that the verbs *stayed* and *returning* evoke their dedicated Residence and Arriving frames respectively, and the nouns *girlfriend* and *home* evoke Personal_relationship and Foreign_or_domestic_country frames, respectively. By analyzing these extracted frames, we can figure out that the reviewer would like to convey some sense of his or her relationships, travel schedule, and duration of stay.

We hypothesize that the frame occurrence may be different because of the different mindsets or experiences of truthful users and spammers. On this basis, we propose metrics for measuring the prevalence of frames for each dataset in the following subsection.

[3]This sentence is borrowed from the opinion spam dataset of [15].

3. METHODOLOGY

In this section, we present our methodology of using semantic frames for analyzing user-generated content. We first define measures of the difference between frame proportions for each dataset. We then obtain frames that have high discriminative power and explore their interesting characteristics.

3.1 Discriminative Frame Selection

Although there are many, not every frame that appears in a sentence is useful. To select the most effective frames for an analysis, we propose Normalized Frame Rate (NFR) and Normalized Bi-frame Rate (NBFR), both of which quantitatively measure the discriminative power of a frame and frame pairs, respectively. We will explain each metric using examples in Table 1.

3.1.1 Normalized Frame Rate (NFR)

We define Normalized Frame Rate (NFR) to compute the frame distribution differences between truthful reviews and deceptive opinion spam. The NFR value of a specific frame indicates the number of times the frame occurred compared with the total number of occurrences of frames in a dataset. Before we count each frame occurrence from the dataset, we first investigate whether there is a statistically significant difference of frame proportion between the deceptive opinion spam dataset and the truthful review dataset using a two-proportion z-test. The test result successfully rejects the null hypothesis. ($H0$: sample proportions of each frame are equal on both datasets in significant level $p < 0.01$)

Now, let us explain how to calculate the NFR value for each class. The *Personal_relationship* frame (f_1) occurred twice in the deceptive class but did not occur in the truthful class. Since the total sum of frame occurrence of the deceptive class is 11, the NFR value of f_1 is 2/11 or 0.18 for the deceptive class and 0 for the truthful class. Similarly, the *Calendric_unit* frame (f_4) occurred three times in the truthful class and twice in the deceptive class. The NFR value of f_4 for the truthful class is 0.33 and 0.18 for the deceptive class. Then we calculate the difference between two NFR values of a frame by subtracting the NFR value for the truthful class from the NFR value for the deceptive class. The difference between two NFR values for a specific frame f_m is calculated as below:

$$\Delta NFR_{f_m} = NFR_{D_{deceptive}} f_m - NFR_{D_{truthful}} f_m \quad (1)$$

For example, the ΔNFR_{f_4} value is -0.15 (0.18 - 0.3); it is negative because the *Calendric_unit* frame (f_4) is more likely to appear in the truthful class.

Table 1: Example with frame and bi-frame in the sentences

Class	Example Sentences	Bi-frame(s)
Deceptive	s_1: My **girlfriend**(f_1) and I **stayed**(f_2) **4**(f_3) **nights**(f_4) at the Talbott on **Saturday**(f_4) 9/29.	(f_1, f_2), (f_2, f_3), (f_3, f_4), (f_4, f_4)
	s_2: The Talbott **Hotel**(f_7) is **fantastic**(f_8).	(f_7, f_8)
	s_3: Me and my **husband**(f_1) **stayed**(f_2) at the Hyatt and I **satisfied**(f_9) with this **hotel**(f_7).	(f_1, f_2), (f_2, f_9), (f_9, f_7)
Truthful	s_4: We **stayed**(f_2) **2**(f_3) **nights**(f_4) in a cosy **bedroom**(f_{10}).	(f_2, f_3), (f_3, f_4), (f_4, f_{10})
	s_5: We **visited**(f_5) Chicago on the 4th of **July**(f_4) **weekend**(f_4) .	(f_5, f_4), (f_4, f_4)
	s_6: Breakfast **starts**(f_{11}) at 7 am and it was only **20**(f_3) $.	(f_{11}, f_3)
Frame Index	f_1: Personal_relationship f_2: Residence f_3: Cardinal_numbers f_4: Calendric_unit f_5: Arriving f_6: Foreign_or_domestic_country f_7: Buildings f_8: Desirability f_9: Experiencer_focus f_{10}: Building_subparts f_{11}: Process_start f_{12}: Sole_instance	

3.1.2 Normalized Bi-frame Rate ($NBFR$)

A bi-frame refers to a pair of successively occurring frames in a sentence. While bigram catches context of a text in n-gram-based approach, a bi-frame also conveys semantic context of a given sentence. Furthermore, we consider the frame sequence as the latent writing style of a reviewer.

We use Table 1 again to explain NBFR. In a sentence s_3, f_1, f_2, f_9 and f_7 frames are identified. We obtain bi-frame pairs by grouping successive frames into two such as (f_1, f_2), (f_2, f_9), and (f_9, f_7). We calculate NBFR value in the exact same way as we calculate NFR, but now we count two frames as one counting unit. For example, the bi-frame *Personal_relationship* (f_1) $\rightarrow$ *Residence* (f_2) occurred twice in the deceptive class but did not occur in the truthful class. Therefore, the NBFR value of the bi-frame *Personal_relationship* (f_1) $\rightarrow$ *Residence* (f_2) in the deceptive class is 2/7 (0.29). In general, ΔNBFR for frame f_i and f_j is calculated as below:

$$\Delta NBFR_{f_if_j} = NBFR_{D_{deceptive}f_if_j} - NBFR_{D_{truthful}f_if_j} \tag{2}$$

Hence, $\Delta NBFR_{f_1f_2}$ is a positive value of 0.29 which means that this bi-frame is more likely to appear in the deceptive class.

4. QUALITATIVE EVALUATION OF FRAME FEATURES

This section presents a qualitative analysis of frame features. We first introduce the dataset that was used in this paper in Section 4.1. Then, we investigate the individual frame characteristics that appeared in each dataset and look into bi-frames in Section 4.2.

4.1 Dataset

Table 2 shows the statistics of the datasets used in this paper. For the truthful opinion reviews, we used the TripAdvisor dataset from Ott et al. [15]. Details about the deceptive opinion spam datasets solicited from crowdsource, employees, and real-life screened reviews are described in what follows.

Table 2: Statistics of datasets

Class	Dataset	#of docs	#of unique words	size (in kB)
Deceptive	AMT	400	4,180	254
	Employee	140	1,856	62
	"not-recommended" (Yelp)	3,361	18,466	2,177
Truthful	TripAdvisor	400	4,934	265
	"recommended" (Yelp)	3,869	20,259	2,747

4.1.1 AMT and Employee Dataset

Gathering quality and reliable gold-standard deceptive opinion datasets is a challenging task in opinion spam analysis research. Ott et al. employed a crowdsourcing framework provided by Amazon to generate gold-standard deceptive opinion spam datasets [15]. However, Li et al. pointed out that the AMT dataset is only representative of a specific type of deception that is generated by people who do not have particular experience with a target local business [11]. The reviewers' lack of experience is complemented by hiring real employees and allowing them to write deceptive opinion reviews based on their higher-level domain knowledge.

For the experiment, we explore the gold-standard deceptive opinion datasets from [15], [11] and our own real-life dataset collected from Yelp. Since the AMT dataset contains only hotel reviews, we also only use the hotel employee dataset in [11]. We will investigate the characteristics of each dataset in relation to semantic frames, and not just linguistic differences which were already studied in previous work [4, 11, 15].

4.1.2 Dataset from Yelp

We gathered 3,869 "recommended" reviews and 3,361 "not currently recommended"[4] reviews for 265 hotels in the Chicago area from Yelp. We collected only 5-star reviews to constantly maintain the sentiment polarity. We mixed popular and unpopular hotels based on the number of reviews but we used only reviews of hotels whose overall ratings were over 3.5. The number of unique unigrams in the Yelp dataset was almost 4 times larger than that in the AMT dataset due to the different size of the dataset (~20,000 compared with ~5,000).

4.2 Frame-based Semantic Analysis

4.2.1 Frame-by-frame Analysis

The goal of our analysis is to understand how frames are distributed in different datasets. Figure 2 and Figure 3 illustrate the top 50 frames that were sorted by their ΔNFR value, which were differentially expressed in the deceptive class and the truthful class from each of the two datasets. The NFR value of each frame is ranged from -1.0 to 1.25 in the AMT dataset, and from -2.0 to 1.6 in the Employee dataset. By the definition of ΔNFR, if a frame's ΔNFR

[4] Yelp recently changed its review filtering service name from "filtered" to "not currently recommended". The word "currently" would give the sense that its review recommendation policy continues to change, so that blocked reviews are temporary.

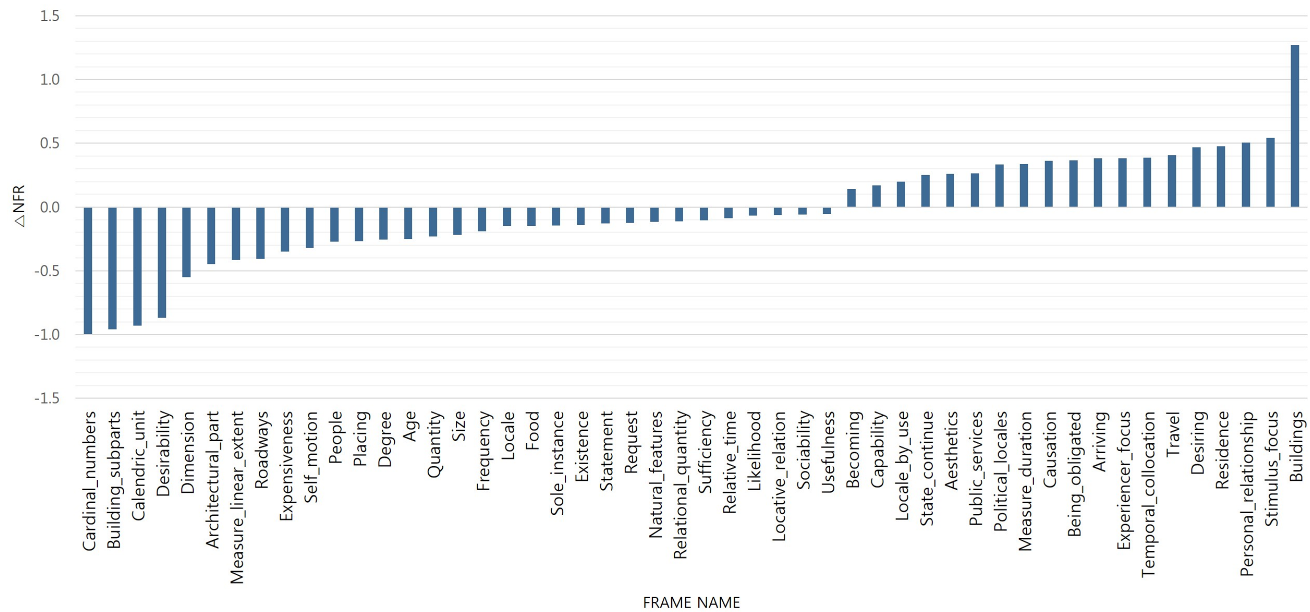

Figure 2: Top 50 differentially expressed frames sorted by ΔNFR in AMT dataset

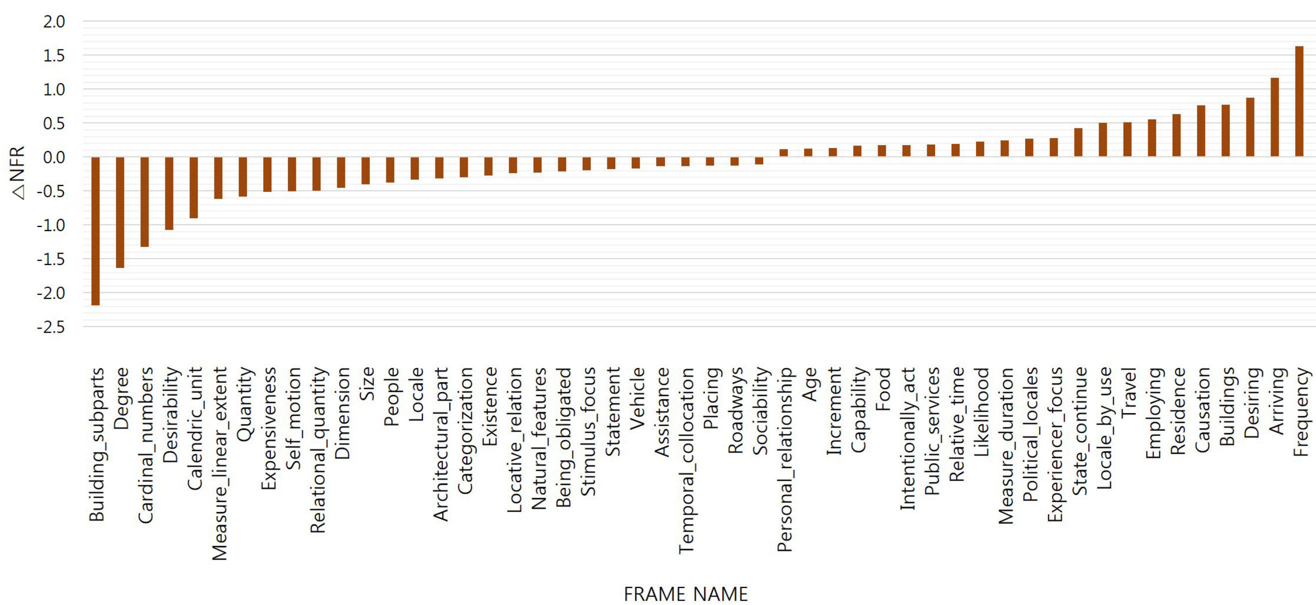

Figure 3: Top 50 differentially expressed frames sorted by ΔNFR in Employee dataset

value is negative, the frame is more likely to occur in the truthful dataset than in the deceptive class and vice-versa.

In this analysis, we found that the nature of the frames detected by our method is also largely consistent with previous deceptive opinion spam studies.

Spatial and business detail-related frames. We observed that frames such as Cardinal_numbers, Building_subparts, Calendric_unit, Dimension, Expensiveness, which are used to described the spatial details about a business are conspicuous in the truthful dataset. For example, the Cardinal_numbers and Calendric_unit frames express a specific date of travel or arrival, duration of stay, and so on. The Building_subparts frame evoked by lexical units such as *room, floor, bathroom*, and *lobby* is frequently used in both truthful datasets showing strong negative ΔNFR value for describing spatial details. These building parts are explained using specific measurements (Cardinal_number and Dimension). Furthermore, we can see that the truthful reviewers frequently commented on the price of a hotel (negative ΔNFR value of Expensiveness frame in both datasets).

This observation is consistent with previous studies that truthful reviewers are more likely to include spatial details mentioned above [11, 13, 15]. On the other hand, frames that have positive ΔNFR values (e.g., Buildings, Residence, Travel, etc.) are rather general and do not contain much detailed information.

Personal relationship-related frames. We found that the spammers deliberately emphasize relationship keywords to convince the reader that their review is based on real experience. For example, the Personal_relationship frame that is evoked by the lexical units *wife, husband, friend* is highly ranked in Table 4 which is also confirms the previous studies. In Ott et al., The highest weighted truthful and deceptive lexical features learned by the supervised model contain the word *husband* [15]. Similarly, the highly weighted LIWC feature list contains the word *family* in the study of Li et al. [11]. Unlike the previous studies, instead of capturing each individual word as a feature, our approach captures a semantic group of a word as a feature.

Comparison of our top-ranked lexical units and top-weighted terms in previous work. We observed that some of our top-k lexical units in Table 4 were also reported in previous studies [11, 14, 15]. For example, in the study of Li et al., the top weighted LIWC feature list contains the category word *number* in the truthful class. This corresponds to our ΔNFR negative Cardinal_numbers frame [11]. Furthermore, the lexical units *small, hotel*, and *husband* of the Dimension, Buildings, and Personal_relationship frames, respectively, are in the top-weighted lexical features list in [15]. The verb *felt* as well as nouns, which were captured by Mukherjee et al., would correspond to the lexical unit *felt like* of the Desiring frame [14].

Sentiment expression pattern analysis. Interestingly, we discovered that there are different sentiment expression patterns between deceptive Turkers and truthful reviewers. Table 3 shows that the Stimulus_focus frame is used more in the deceptive reviews while the Desirability frame is used more in the truthful review dataset even though both frames are related to sentiment expression. According to the FrameNet, the Stimulus evokes a particular emotion or experience in the Experiencer. For example, in the following sentence, *"The movie was quite fascinating."*[5] the stimulus *the movie* brings out positive emotions in the reviewer. On the other hand, the Desirability frame describes the event of judging the gradable attribute Evaluee for its quality. For the given sentence, *"The view was astonishing."*[6] the reviewer judges the goodness of the *view* aspect. We assume that Turkers tend to fabricate their emotions, and that truthful reviewers tend to evaluate a target using their real experience. Human readers may not be able to differentiate two sentiment words in spoken language. However, when machines read and process sentences, they capture subtle differences using the linguistic knowledge provided by the FrameNet lexical database.

4.2.2 Bi-frame Co-occurrence Matrix Similarity

Table 5 and Table 6 are the lists of the top differentially expressed bi-frames from the AMT dataset and the Employee dataset, respectively. Each table illustrates the top 15 and the bottom 15 bi-frames, all of which were ranked according to their ΔNBFR values. To investigate how bi-frame co-occurrence patterns of two datasets are different, we use their co-occurrence pattern similarity.

[5]Example sentence of Stimulus_focus frame in FrameNet.

[6]Example sentence of Desirability frame in FrameNet.

Table 3: Top 5 frame list of two datasets

Dataset	Deceptive	Truthful
AMT	Buildings	Cardinal_numbers
	Stimulus_focus	Building_subparts
	Personal_relationship	Calendric_unit
	Residence	Desirability
	Desiring	Dimension
Employee	Frequency	Building_subparts
	Arriving	Degree
	Desiring	Cardinal_numbers
	Buildings	Desirability
	Causation	Calendric_unit

Let L be the union of the top bi-frames in Table 5 and Table 6 where $|L|$=42. Then, we define $n \times n$ matrices M_d and M_t for the deceptive and the truthful dataset, respectively, where n is the number of frames in set L. Each cell of matrix M represents the NBFR value of bi-frame. By comparing the co-occurrence matrices of the two datasets, we can see how spammers and non-spammers write reviews differently in terms of semantic flow. The similarity between the two matrices M_d and M_t is calculated as follows.

$$sim(M_d, M_t) = \frac{vec(M_d) \cdot vec(M_t)}{\|vec(M_d)\|\|vec(M_t)\|} \quad (3)$$

Table 7 shows the similarity values computed between dataset pairs. To compare the similarity values, we needed a threshold value to make a decision whether two datasets show different bi-frame co-occurrence patterns. So we randomly divided a single dataset into two subsets and calculated their within-set similarity. We repeated this procedure 10 times and reported the averages in Table 7.

As we can see in Table 7, the similarity value of AMT (deceptive) and TripAdvisor (truthful) is less than the within-set similarity value ($0.86 < 0.92$). This means that the bi-frame patterns of deceptive and truthful reviews are different. There are also differences between AMT and Employee even though they are in the deceptive category. Li et al. also mentioned these characteristics and reported that their two-class classifier achieves an accuracy over 0.76 in distinguishing between Turker and Employee reviews [11]. Note that the similarity between Employee and TripAdvisor is almost equal to that of the Employee-within set. This implies that employees put more effort into writing fake reviews to try to sound authentic.

4.2.3 Qualitative Analysis of Bi-frame Occurring Patterns

In Table 5 and Table 6, we also observe a similar pattern in the top ΔNFR list. We see that in the truthful group (the negative value of ΔNBFR), the reviewers include details about the business. For instance, a bi-frame Building_subparts→Dimension appeared in both datasets and was most likely used to describe the interior of the hotel. The Self_motion frame evoked by the lexical unit *walk* or

Table 4: The most used frame lexical units of the top 5 frames

Class	Deceptive (AMT)					Truthful				
Frame Name	Buildings	Stimulus_focus	Personal_relationship	Residence	Desiring	Cardinal_numbers	Building_subparts	Calendric_unit	Desirability	Dimension
Lexical Unit	hotel	nice	husband	stay	wanted	one	room	night	great	large
	hotels	amazing	wife	stayed	want	2	rooms	nights	excellent	small
	home	relaxing	friends	staying	felt like	two	bathroom	day	wonderful	larger
	bar	relax	friend	living	feel like	3	lobby	weekend	good	high
	house	pleasant	couple	live	wish	4	bathrooms	morning	best	dark
Class	Deceptive (Employee)					Truthful				
Frame Name	Frequency	Arriving	Desiring	Buildings	Causation	Building_subparts	Degree	Cardinal_numbers	Desirability	Calendric_unit
Lexical Unit	always	get	feel like	hotel	made	room	very	one	great	night
	every	return	want	home	make	rooms	totally	2	excellent	nights
	never	visit	felt like	bar	making	bathroom	somewhat	two	good	day
	frequently	arrived	wanted	living	makes	lobby	fairly	3	wonderful	weekend
	rate	come	hope	pavilion	brought	bathrooms		4	best	morning

Table 5: The top and the bottom 15 most differentially expressed bi-frames in the AMT dataset

Top 15 Bi-frames	$\Delta NBFR$	Bottom 15 Bi-frames	$\Delta NBFR$
Residence Buildings	0.00354	Cardinal_numbers Calendric_unit	-0.004083
Building_subparts Public_services	0.00186	Desirability Locale	-0.002952
Locative_relation Measure_duration	0.00177	Building_subparts Dimension	-0.002282
Businesses Travel	0.00170	Calendric_unit Calendric_unit	-0.002097
Direction Buildings	0.00153	Dimension Building_subparts	-0.001859
Buildings Buildings	0.00140	Cardinal_numbers Measure_linear_extent	-0.001818
Buildings Inclusion	0.00136	Residence Locative_relation	-0.001796
Experiencer_focus Residence	0.00134	Building_subparts Desirability	-0.001756
Locale State_continue	0.00124	Locale Desirability	-0.001441
Buildings Residence	0.00120	Building_subparts Degree	-0.001294
Temporal_collocation Residence	0.00118	Self_motion Range	-0.001292
Arriving Buildings	0.00104	Desirability Frequency	-0.001191
Measure_duration Buildings	0.00103	Relational_quantity Cardinal_numbers	-0.001057
Desirability Expertise	0.00096	Age Buildings	-0.001057
Aesthetics Buildings	0.00093	Cardinal_numbers Measure_duration	-0.001049

Table 6: The top and the bottom 15 most differentially expressed bi-frames in the Employee dataset

Top 15 Bi-frames	$\Delta NBFR$	Bottom 15 Bi-frames	$\Delta NBFR$
Buildings Buildings	0.00357	Cardinal_numbers Calendric_unit	-0.00551
Capability Waiting	0.00250	Building_subparts Degree	-0.00476
Measure_duration Arriving	0.00248	Residence Locative_relation	-0.00449
Waiting Arriving	0.00248	Degree Stimulus_focus	-0.00303
Frequency Residence	0.00217	Calendric_unit Calendric_unit	-0.00263
Food Desirability	0.00204	Building_subparts Dimension	-0.00258
Capability Statement	0.00198	Building_subparts Desirability	-0.00247
Locative_relation Desirability	0.00191	Desirability Locale	-0.00237
Causation Residence	0.00188	Locale Desirability	-0.00219
Physical_artworks Building_subparts	0.00188	Desirability Measure_linear_extent	-0.00209
Locale State_continue	0.00172	Cardinal_numbers Measure_linear_extent	-0.00203
Ordinal_numbers Roadways	0.00167	Self_motion Range	-0.00192
Memory Being_named	0.00167	Degree Desirability	-0.00190
Activity_ongoing Locative_relation	0.00161	Degree Sociability	-0.00184
Personal_relationship Arriving	0.00161	Buildings Building_subparts	-0.00170

hike explains that the Self_mover[7] moves under its own direction along a path. Accordingly, the Self_motion→Range bi-frame was categorized in the truthful group and thus can be interpreted as a situation where the reviewer describes his or her activity around the hotel.

However, bi-frames in the deceptive group are mostly not notable and unspecific, compared with those in the truthful group. The Businesses→Travel bi-frame in the AMT dataset may indicate the reason of travel that fake reviewers fabricate for their reviews. This may be because paid Turkers need to quickly fabricate plausible reasons to try to sound authentic and pick a general topic such as business. On the other hand, employees are more likely to comment on the quality of a food (Food→Desirability).

[7]This is a core frame element of Self_motion frame in FrameNet.

Physical_artworks→Building_subparts well represent employees' detailed knowledge about the business. Personal_relationship→Arriving is also a conspicuous bi-frame which is more like a cliché in deception. Note that the Cardinal_numbers→Calendric_unit bi-frame is top ranked in both Tables 5 and 6 and has lowest ΔNBFR values, which means this bi-frame is more likely to appear in truthful reviews. Although employees can write details about the business (such as Physical_artwork), we can assume that they cannot fabricate details about an activity-based experience such as a specific date (Calendric_unit, Self_motion, etc.).

5. FRAME FEATURE VALIDATION ON CLASSIFICATION TASKS

We perform machine learning-based classification tasks to distinguish truthful reviews from deceptive reviews. In Section 5.1, we explain the pre-processing procedure used in the experiment. In Section 5.2, we report the classification results of varying features on different classification models.

5.1 Data Pre-processing and Frame Extraction

We set a default analysis unit as a single sentence for frame analysis. Since both datasets do not annotate sentence boundaries, we divide each review document into multiple sentences using OpenNLP Sentence Detector[8]. Then, each sentence is analyzed for dependency parsing and n-gram tokenization using Stanford Parser[9]. The parsed results are passed to SEMAPHORE V2.1, an automatic frame-semantic annotation[10] system for frame extraction.

Table 8 shows the statistics of the frame extraction results. The truthful dataset contains more frames than the deceptive dataset, but both datasets contain almost same number of unique frames. Due to the difficulties of gathering datasets from the real employees, the size of the Employee dataset is smaller than that of the AMT dataset.

Table 7: Bi-frame co-occurrence matrix similarity

Settings	Dataset	Similarity
Cross set	AMT&TripAdvisor	0.86
	AMT&Employee	0.78
	Employee&TripAdvisor	0.68
Within set	TripAdvisor	0.92
	AMT	0.91
	Employee	0.66

Table 8: Frame extraction statistics obtained from two datasets

Doc / Frame	Deceptive		Truthful
	AMT	Employee	
#of docs	400	140	400
#of frames	17131	3641	19153
#of unique frames	467	322	463

Table 9: Our implementation of the baseline

System	Feature	Acc	Deceptive			Truthful		
			Prec	Rec	F1	Prec	Rec	F1
Ott'11 [15]	Uni	0.884	0.870	0.903	0.886	0.899	0.865	0.882
	Bi^+	0.896	0.891	0.903	0.897	0.901	0.890	0.896
Our Impl.	Uni	0.870	0.868	0.873	0.870	0.872	0.868	0.870
	Bi^+	0.876	0.873	0.880	0.877	0.879	0.873	0.876

5.2 Classification Task

For a reliable comparison, we also report our own implementation[11] of [15] for the AMT dataset (Table 9). We use SVM (SVM^{light}, Joachim, 1999) and Naive Bayes classifiers as classification models. To find the separating hyperplane for SVM, a linear kernel is used. Each model is validated based on a nested 5-fold cross validation scheme which was also done by Ott et al. [15].

Nested cross-validation works in the following way. There is an inner CV loop where we search for optimal parameters using the parameter search algorithm (e.g., grid search). At the end of this process, we end up with k models (k being the number of folds in the outer loop) that gives the best accuracy within the inner CV. Note that the performance results of our implementation are slightly different from those of the original results reported in [15]. It is due to the different experimental settings such as model parameter, n-gram tokenization, and so on, which were not provided in the original paper.

[8]https://opennlp.apache.org/

[9]http://nlp.stanford.edu/software/stanford-dependencies.shtml

[10]http://www.ark.cs.cmu.edu/SEMAFOR/

[11]We were not able to obtain the original implementation from the authors.

Table 10: Classification result using frame feature-only

Dataset	Feature	Acc	Deceptive			Truthful		
			Prec	Rec	F1	Prec	Rec	F1
AMT vs. Truthful	Frm3	0.660	0.639	0.735	0.684	0.688	0.585	0.632
	Frm5	0.710	0.688	0.768	0.726	0.737	0.653	0.692
	Frm10	0.737	0.716	0.788	0.750	0.764	0.688	0.724
	Frm_{full}	0.774	0.753	0.815	0.783	0.798	0.733	0.764
Employee vs. Truthful	Frm3	0.824	0.727	0.514	0.603	0.846	0.933	0.887
	Frm5	0.856	0.763	0.643	0.698	0.882	0.930	0.905
	Frm10	0.878	0.803	0.700	0.748	0.900	0.940	0.919
	Frm_{full}	0.870	0.773	0.707	0.739	0.900	0.928	0.914

Table 11: Classification result varying the number of bi-frames with a fixed number of frame features

Feature	Bi-frm50	Bi-frm100	Bi-frm200	Bi-frm500	Bi-frm1000
AMT vs. Truthful	0.785	0.791	0.791	0.796	0.796
Employee vs. Truthful	0.876	0.878	0.876	0.870	0.872

5.2.1 Frame-only Feature

We report the classification result of using only frame feature and investigate how frame features are effective for differentiating spam and non-spam. Frm-k, (k = 1, 2 ...n) in Table 10 represents the number of k frames from the top and bottom in the frame list sorted by ΔNFR values.

First, we trained a SVM model using only frame features. Interestingly, we obtained a much higher classification accuracy when we used only six frame features (Frm3 setting in Table 10) than random guess of 50% (0.660 and 0.824) for both the AMT dataset and the Employee dataset. Adding more frame features gradually increases the accuracy. As we can see in Table 10, the model that uses the Frm5 feature yields almost the same accuracy as the model using the Frm_{full} feature. This proves that the top-k frames can be used as discriminative features for training supervised classification models. The top 5 frames from both datasets are listed in Table 3.

5.2.2 Frame and Bi-frame Features

We also use bi-frames along with frame features. To compare the results of using only frame features reported in Table 10, we mixed the bi-frame feature onto the frame feature. Instead of using every combination of frames, we take only the top 1000 and the bottom 1000 bi-frame features in the bi-frame list which were sorted by their ΔNBFR value.

Figure 4 shows that the classification accuracy is gradually improved by adding bi-frame features to the model

Table 12: Classification result using frame features with n-grams

Dataset	Features	Acc	Deceptive			Truthful		
			Prec	Rec	F1	Prec	Rec	F1
AMT	Frm5+Uni_500$_{svm}$	0.877	0.866	0.892	0.879	0.889	0.862	0.876
	Frm15+Bi-Frm160+Uni_150$_{svm}$	0.881	0.877	0.888	0.882	0.886	0.875	0.881
	Frm30+Bi-Frm10+Bi$^+$_100$_{svm}$	0.887	0.880	0.897	0.889	0.895	0.877	0.886
	Frm5+Bi-Frm20+Bi$^+$_full$_{NB}$	0.914	0.903	0.928	0.915	0.925	0.900	0.913
	Our impl. of [15] Uni	0.870	0.868	0.873	0.870	0.872	0.868	0.870
	Our impl. of [15] Bi$^+$	0.876	0.873	0.880	0.877	0.879	0.873	0.876
Employee	Frm5+Uni_100$_{svm}$	0.900	0.830	0.771	0.800	0.921	0.945	0.933
	Frm5+Bi-Frm15+Uni_full$_{svm}$	0.924	0.872	0.829	0.850	0.941	0.958	0.949
	Frm5+Bi-Frm40+Bi$^+$_100$_{svm}$	0.917	0.852	0.824	0.836	0.938	0.950	0.944
	Frm5+Bi-Frm10+Bi$^+$_full$_{NB}$	0.867	0.746	0.736	0.741	0.908	0.913	0.910
	Same setting in [15] Uni$_{svm}$	0.916	0.857	0.814	0.835	0.936	0.953	0.944
	Same setting in [15] Bi$^+{}_{svm}$	0.894	0.837	0.736	0.783	0.911	0.950	0.930

compared to the model that used frame-only feature in both datasets. The semantic context delivered by bi-frames can be useful for detecting deceptive opinions. Next, we investigate how the number of bi-frames affects classification performance when the number of bi-frame features varies and when the number of frame features is fixed to "full". In this setting, the classification result did not change significantly.

In the AMT vs. Truthful setting in Table 11, the classification accuracy improved from 0.785 to 0.796 as a result of increasing the number of bi-frames. However, in the Employee vs. Truthful setting, the performance improvement was not as large as that in the AMT vs. Truthful setting. The accuracy did not improve even though we increased the number of bi-frames. From this result, we can assume that the more frame and bi-frame features we used, the more likely the performance will improve.

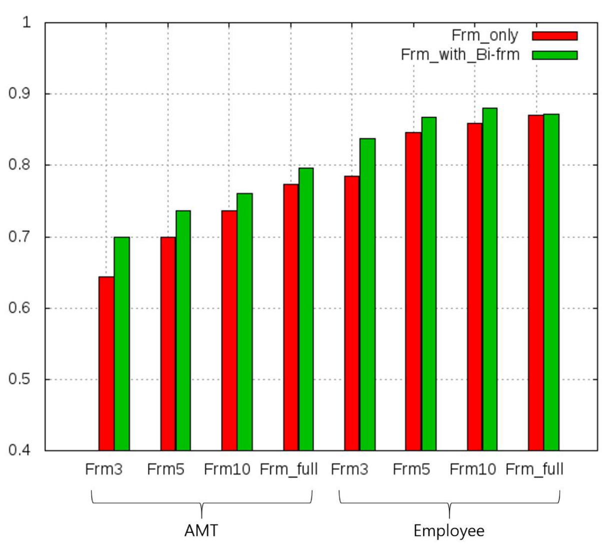

Figure 4: Classification accuracy using frame and bi-frame feature together

5.2.3 Frame with Lexico-syntactic Features

In this experiment, we train Naive Bayes and SVM classifiers using individual frame and bi-frame features in combination with lexical n-grams and the result is reported in Table 12. To address the question about the power of lexical n-grams, we vary the number of n-grams. We have evaluated every combination of features with various numbers of n-grams. However, we report only the meaningful results due to the limit of space. Likewise Frm-k, Uni-k indicates the top-k unigrams that were selected using the same ΔNFR method that we used to obtain the top-k frames. The superscript$^+$ of Bi$^+$ feature indicates that the bigram feature subsumes the unigram feature set.

In Table 12, the feature combination of Frm5 and Uni500 achieved an accuracy of 0.877 whereas the baseline result achieved an accuracy of 0.870 (using unigram-only features in our implementation of Ott et al. [15]). After various tries, we obtained the highest classification accuracy of 0.914 in the AMT dataset using the feature combination of Frm5, Bi-Frm20 and full Bi$^+$ with a Naive Bayes model (a 4.34% increase compared with our baseline implementation). In the Employee dataset, Frm5, Bi-Frm15 and Uni_full features obtained an accuracy of 0.924 using an SVM model (a 0.87% increase compared with the baseline).

5.2.4 Classification on Yelp

In Table 13, we also report an SVM classification result on the Yelp dataset. First, the classification experiments that use an n-gram feature on the Yelp data yielded an accuracy of 0.625. This result is consistent with [14] that simple linguistic features are hardly effective for real-life datasets (an accuracy of 0.676 was obtained from their Yelp dataset). Next, we added frame and bi-frame features (Frm5 and Bi-Frm20) to n-grams, which improved the performance of the previous classification task. However, the semantic frame feature did not improve classification performance. We assume that Yelp's review recommendation system does not consider the semantic features of review contents but utilizes other non-content features such as users' behavioral patterns, which were investigated by Mukherjee et al. [14].

Table 13: SVM 5-fold CV result on our Yelp dataset

Feature	Acc	Not recommended			Recommended		
		Prec	Rec	F1	Prec	Rec	F1
Uni	0.625	0.594	0.608	0.601	0.653	0.640	0.646
Bi$^+$	0.624	0.589	0.628	0.608	0.657	0.620	0.638
Frm+Uni_full	0.620	0.590	0.597	0.593	0.646	0.640	0.643
Frm+Bi$^+$_full	0.619	0.585	0.618	0.601	0.651	0.620	0.635

6. CONCLUSIONS

In this paper, we investigate deceptive opinion spam and truthful reviews using semantic frames. We also propose an analysis method for measuring the proportion of frame in a sentence. To the best of our knowledge, this is the first study to perform semantic frame analysis of customer reviews.

Our experimental results show that the classification model that used frame features outperformed the baseline model by 4.34% for the AMT dataset. Furthermore, using new semantic frame features, we successively captured some interpretable differences between fake and real reviews in terms of their unique semantic features, which cannot be done by other existing methods. For example, even though employees who have expert-level domain knowledge attempt to write a seemingly truthful review, they cannot fabricate specific details about an activity-based experience. We also captured subtle differences of sentiment expression patterns in deceptive and truthful review contents using a linguistic theoretical method. Looking at other domains such as restaurants or medical services using our proposed semantic analysis method remains for the future work in this area.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] 2013 study: 79% of consumers trust online reviews as much as personal recommendations, "http://searchengineland.com/2013-study-79-of-consumers-trust-online-reviews-as-much-as-personal-recommendations-164565". Accessed: 2015-04-05.

[2] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. Spamrank–fully automatic link spam detection work in progress. In *Proceedings of the first international workshop on adversarial information retrieval on the web*, AIRWeb '05, Chiba, Japan, 2005.

[3] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. In *Proceedings of SIGIR*, Amsterdam, Netherlands, July 2007. ACM.

[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*, 2012.

[5] C. Fillmore, C. Johnson, and M. Petruck. Background to framenet. *International journal of lexicography*, 16(3):235, 2003.

[6] C. J. Fillmore. Frame semantics and the nature of language. In *Origins and Evolution of Language and Speech, 280*, 1976.

[7] T. Gamerschlag, D. Gerland, R. Osswald, and W. Petersen. *Frames and Concept Types: Applications in Language and Philosophy*, volume 94 of *0924-4662*. Springer International Publishing, 1 edition, 2014.

[8] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, VLDB '04, pages 576–587, 2004.

[9] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, 2008.

[10] N. Jindal, B. Liu, and E.-P. Lim. Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1549–1552, New York, NY, USA, 2010. ACM.

[11] J. Li, M. Ott, C. Cardie, and E. Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics*, 2014.

[12] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, 2010.

[13] A. Mukherjee, B. Liu, and N. S. Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012*, pages 191–200.

[14] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance. What yelp fake review filter might be doing? In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.*

[15] 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*, 2011.

[16] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The Development and Psychometric Properties of LIWC2007. Austin, TX, USA LIWC. Net.

[17] N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter*, 13(2):50–64, 2012.

[18] G. Wang, S. Xie, B. Liu, and P. S. Yu. Review graph based online store review spammer detection. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 1242–1247, Washington, DC, USA, 2011. IEEE Computer Society.