# Deceptive Opinion Spam Detection Using Deep Level Linguistic Features

Changge Chen[1,2], Hai Zhao[1,2](✉), and Yang Yang[1,2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China
changge.chen.cc@gmail.com, {zhaohai,yangyang}@cs.sjtu.edu.cn

[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

**Abstract.** This paper focuses on improving a specific opinion spam detection task, deceptive spam. In addition to traditional word form and other shallow syntactic features, we introduce two types of deep level linguistic features. The first type of features are derived from a shallow discourse parser trained on Penn Discourse Treebank (PDTB), which can capture inter-sentence information. The second type is based on the relationship between sentiment analysis and spam detection. The experimental results over the benchmark dataset demonstrate that both of the proposed deep features achieve improved performance over the baseline.

**Keywords:** Spam detection · Shallow discouese parsing · Sentiment analysis

## 1 Introduction

In nowadays, online reviews of products and services have increasingly large impact on consumer purchasing decisions[1]. Accordingly, there comes an increasing potential to gain money through deceptive opinion spam - fraudulent reviews that are deliberately written to deceive readers.

While various studies have been carried out in recent years, most of them focused on extracting scattered syntactic features which capture local information. In this work, we consider more deep features on the basis of syntactic

This work was supported in part by the National Natural Science Foundation of China under Grants 60903119, 61170114, and 61272248, the National Basic Research Program of China under Grant 2013CB329401, the Science and Technology Commission of Shanghai Municipality under Grant 13511500200, the European Union Seventh Framework Program under Grant 247619, the Cai Yuanpei Program (CSC fund 201304490199, 201304490171, and the art and science interdiscipline funds of Shanghai Jiao Tong University (a study on mobilization mechanism and alerting threshold setting for online community, and media image and psychology evaluation: a computational intelligence approach under Grant 14X190040031 (14JCRZ04).

[1] http://www.conecomm.com/2011coneonlineinfluencetrendtracker

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 465–474, 2015.
DOI: 10.1007/978-3-319-25207-0_43

feature set. Specifically, we consider two types of features. Firstly, we explore the possibility of integrating results generated by a shallow discourse parser into opinion spam detection. Instead of treating the whole text as an unordered set of sentences or terms, from a high level of viewing, we inspect the text consisting of sentences that are glued together by discourse relations in a systematic way. Automatic discourse parsing is considered as one of the most challenging Natural Language Processing tasks [1]. In this work, we integrate the most frequent type of discourse relations - *'Explicit'* relation, and its corresponding sense into our module. Secondly, we design an extensive feature set to capture sentiment cues for further performance improvement.

This work will be strictly evaluated on a public golden standard dataset. Ott et al.[2] constructed a negative deceptive opinion spam dataset. Combining with the positive opinion spams, a dataset consists of both negative and positive reviews is available. This dataset includes 1,600 reviews with four categories: positive truthful, positive deceptive, negative truthful, and negative deceptive.

The rest of paper is organized as follows: in Section 2, we discuss and compare recent related works. In Section 3, we give a description of the construction and analyze the dataset. In Section 4, we describe features employed by our detection methodology. In Section 5, we present and discuss our experiment results. Finally, conclusions are given in Section 6.

## 2 Related Work

Opinion spam detection has attracted growing interests from academia and industry in recent years. While enormous researches focus on mining opinions, the source of reviews are seldom concerned. Jindal and Liu [3] were the first to study the trustworthiness of opinions in reviews according to our knowledge. They categorized the spam reviews into three types: a) untruthful opinion spam, b) reviews on brands, and c) non-reviews. While they found that it is easy to distinguish non-reviews, it is difficult to disambiguate the first and second types of spams due to absence of annotated data. By the observation that spams often appear many times, they annotated duplicated reviews as spams. Based on this novel assumption, they got positive results by comparing area under receiver operating characteristic curve (ROC)[2].

Ott et al.[4] constructed the first public DECEPTIVE opinion dataset by hiring experienced online writers through crowd-sourcing service. The paid writers were asked to simulate staff working in 20 hotels located in Chicago. Then the 'staff' were requested by their manager to write positive reviews towards their 'own' hotels in order to promote their reputation. Based on several filtering rules, they solicited 400 deceptive positive reviews. To construct a truthful reviews, they collected reviews from TripAdvisor[3]. After filtering unqualified reviews and balancing the number with deceptive reviews, they got 400 truthful reviews. One concern would be that reviews from TripAdvisor may also contain

[2] http://en.wikipedia.org/wiki/Receiver_operating_characteristic

[3] http://tripadvisor.com

opinion spams. Since we focus on the detection of opinion spam, the impurity of truthful dataset only demonstrates the performance of the classifier. Ott et al.[2] later constructed another dataset in the same manner. But at this time, both truthful reviews and deceptive reviews are in negative attitude.

Various studies have been carried out based on Ott's dataset. Xu et al.[5][6] exploited linguistic features generated from dependency parsing tree. Li et al.[7] proposed a generative LDA-based topic modeling approach for fake review detection. They introduced a semi-supervised manifold ranking algorithm for this task. Feng et al.[8] investigated syntactic stylometry for deception detection. Banerjee et al.[9] developed a linguistic framework to distinguish between genuine and deceptive reviews based on their readability, genre and writing style.

Howerer, all these works treat the text as an unordered set of terms. Beyond that, we exploit discourse relations that hold the text together to extract information from inter-sentence level. Furthermore, on the availability of the dataset with both positive and negative emotion polarities, we also study the relation between sentiment analysis and deceptive detection.

## 3 Construction of Dataset

As this work focuses on deceptive spam that is very hard to be exactly identified even by human observation, corpus or dataset should be carefully selected. We adopt a public dataset with necessary extensions for our evaluation, i.e., the corpus annotated by Ott et al.[2][4]. According to our best knowledge, it is the first public dataset on deceptive spam detection. In this section, a brief description is given to the procedure of data collection and annotation. A simple rule-based deceptive spam detection method is deployed to compare with human judge on this dataset.

### 3.1 Truthful Dataset

The truthful reviews are from TripAdvisor concerning 20 most popular Chicago hotels. Filtering rules are given for better quality control. Reviews thus must be:

- 5-star reviews;
- Only English reviews;
- More than 150 characters (To comply with deceptive reviews);
- Written by non-first-time authors;

To balance truthful and deceptive opinion reviews, 400 truthful reviews are selected randomly.

### 3.2 Deceptive Dataset

While truthful opinions are ubiquitous online, deceptive opinions are difficult to obtain without resorting to heuristic methods [3]. Through the crowd-sourcing service provided by Amazon Mechanical Tuck (AMT)[4], Ott created a pool of 400 Human Intelligence Tasks (HIT) to solicit golden standard positive, deceptive opinion spam toward the 20 chosen hotels. With a reward of 1$ for each review, they restricted their task to Tuckers located in United States with an approval rating of at least 90%. The time duration for each task should be between 1 and 30 minutes. Each Tucker was presented with the name and website of the hotel. They were asked to assume that they were the staff of the hotel's marketing department with the mission to write positive, realistic sounding reviews for their own hotel. At the end, after filtering out unqualified reviews (e.g., unreasonably short, plagiarized and so on), they obtained 400 golden deceptive opinions. Later, they constructed 400 golden negative deceptive opinions using the same way [2].

### 3.3 Human Judge

Ott et al.[4] adopted a skeptical meta judge which labels a review as spam if any of the judges believes so. However, native speakers' performance on detecting deceptive reviews was slightly over random guess. Xu et al.[5] demonstrated that non-native speakers did even worse than natives, which means that opinion spam may do more harm for non-native speakers.

While reading a couple of deceptive reviews, we find that the names of the hotels that the turkers were asked to review on, continuously appear in the first line of the crafted reviews. We investigate this phenomenon in all the corpus. The result is given in Table 1. About 73% of the 800 deceptive reviews have the hotels' name in the first line, which is much higher than that in true reviews (46.8%, 374/800) This phenomenon is not sensitive to sentiment polarity, as there are 486 and 472 reviews fall in positive and negative categories respectively. We can see that this phenomenon is apparent in deceptive reviews. This difference can be ascribed to the difference of motivation between spam writer and true consumers. With reward in minds, spammers need to craft a positive or negative comment on the target as soon as possible, and at the same time, complying with the requirements the HIT asked. Whereas true customers just want to express their own feelings, pleasant or unpleasant, after their personal experiences. Based on this observation, we set up a simple rule-based method which labels a review as deceptive if the hotel name appears in the first line of the review. Otherwise we adopt the rule to label it as truthful. We compare the result with human judges [4] in Table 2.

**Table 1.** Hotel names appearance in the golden standard English Dataset.

| | Total | Positive | Negative |
|---|---|---|---|
| Deceptive | 584 | 284 | 300 |
| Truthful | 374 | 202 | 172 |

[4] http://mturk.com

**Table 2.** Performance of native speaker, non-native speaker, and rule on deceptive detection.

| | Accuracy(%) | Precision(%) | Recall(%) | F-score(%) |
|---|---|---|---|---|
| Rule-based | **66** | **61.0** | **73.0** | **66.4** |
| Native speaker | 60.6 | 60.5 | 61.3 | 60.9 |
| Non-native speaker | 58.1 | 56.3 | 61.5 | 58.8 |

We can see that a simple rule outperforms both native and non-native judge. It recalls 73% of the all opinion spam reviews, which demonstrates our observation. The overall F-score outperforms native speaker by almost 6 percents which shows the vulnerability of human exposed of opinion spam. Therefore, we integrate this syntactic cue into our feature set.

## 4 Deep Features for Deceptive Spam Detection

### 4.1 Shallow Syntactic Features

We view the spam detection task as a text categorization problem with the following features.

**a) Bag-of-words (baseline)** A model only with bag of word features may outperform human judge by achieving a F-score of 88.3% [2] [4]. This serves as our baseline.

**b) POS-*n*-gram** Since the frequency distribution of part-of-speech (POS) tag in a text often depends on the genre of the text [10], and POS tag bigram will not only show frequency information of POS, but also the structure of the sentence, we therefore adopt the POS-unigram and POS-bigram.

**b) Punctuation** This feature indicates the appearance of exclamation and question marks.

**d) Hotel name** This feature indicates whether the hotel name is appear in the first sentence.

### 4.2 Discourse Parsing Features

A typical text consists of sentences that are glued together in a systematic way to form a coherent discourse. Shallow discourse parsing [11] is to parse a piece of text into a set of discourse relations between two adjacent or non-adjacent discourse units. Thus discourse relations convey the infomation of the structure of the article. Combining sentences together enables us to represent a text from an inter-sentence level. PDTB [11] defines five types of discourse relations: *'Explicit'*, *'Implicit'*, *'AltLex'*, *'EntRel'*, *'NoRel'*. *'Explicit'* relation, as its name indicates, is the relation signaled by a connective explicitly. There are 100 connectives annotated by PDTB. Senses have been annotated in the form of sense tags for *'Explicit'* and *'Implicit'* connectives, and *'AltLex'* relations. Sense tags provide a semantic description of the relation between the arguments of connectives.

The tag set of senses is organized hierarchically as in Table 3. The top level has four tags representing four major semantic classes: *'Temporal'*, *'Contigency'*, *'Comparison'* and *'Expansion'*. For each class, a second level of types is defined to further refine the semantics levels.

**Table 3.** Hierarchy of sense tags.

| First Level | Second Level | Third Level |
|---|---|---|
| TEMPORAL | Asynchronous | – |
| | Synchronous | precedence, succession |
| COMPARISON | Constrast | juxtaposition, opposition |
| | Pragmatic Contrast | – |
| | Concession | exception, contra-exception |
| | Pragmatic Concession | – |
| CONTINGENCY | Cause | reason, result |
| | Pragmatic Cause | justification |
| | Condition | hypothetical, general, unreal present, unreal past, factual present, factual past |
| | Pragmatic Condition | relevance, implicit assertion |
| EXPANSION | Conjunction | - |
| | Instantiation | - |
| | Alternative | conjunction, disjunction, chosen alternative |

The following example shows a *'Condition'* relation. And the underlined spans are explicitly signaled by the connective *'if'*.

- *The Treasury said the U.S. will default on Nov. 9* ***if*** *Congress doesn't act by them* .

The shallow discourse parsing can be divided into two parts: explicit and non-explicit. Giving the fact that parsing the non-explicit relation is relatively hard at this time, we only deploy the explicit part here. Noting not all the connectives that appear in the text represent discourse relations, we use a classifier to disambiguate those connectives that convey a relation from those not. In addition, the classifier also determines the sense the connective conveys. Seven features generated from constituent parsing tree are employed to train such a classifier:

**a) Self Category** The highest dominated node which covers the connective.

**b) Parent Caterogy** The category of the parent of the self category.

**c) Left Sibling Category** The syntactic category of immediate left sibling of the self-category. It would be *'NONE'* if the connective is the leftmost node.

**d) Right Sibling Category** The immediate right sibling of the self category. It also would be assigned *'NONE'* if the self-category has been the rightmost node.

**e) VP Existence** The binary feature is to indicate whether the right sibling contains a VP.

**f) Connective** In addition to those features proposed by Pilter et al. [12], we introduce connective feature. The potential connective itself would be a strong sign of its function. Thus we use the POS tag of the candidate connective, its preceding and following word.

The explicit discourse parser is trained and evaluated on the dataset provided by CoNLL 2015 shared task[6]. With an explicit discourse parser, we may adopt related features derived from the parser outputs for the task of deceptive spam detection. In detail, all the 100 connectives defined by PDTB, their frequencies and corresponding senses are used as features. Considering the difficulty of disambiguating all the three levels of senses, in this work we only adopt the first level senses, i.e., *'Temporal'*, *'Contigency'*, *'Comparison'* and *'Expansion'*.

### 4.3 Sentiment Features

Ott et al.[2] gave a brief description between the sentiment and deceptive detection. They claimed that fake negative reviewers over-produced negative emotion terms relative to the truthful reviews in the same way that fake positive reviewers over-produced the positive emotion terms. Based on their assumption, we design a set of sentiment features to express the sentiment information like polarity, intensity, and so on. Similar to what Zhou et al.[13] did, we use term frequency to quantify these features. Note that a negation word that is adjacent to a sentiment word can change its emotional polarity. For any sentiment words within a window following a negation word, we reverse its sentiment polarity from positive to negative, or vice versa. Based on annotation provided by Inquirer Dictionaries[7], we construct a sentimental dictionary consisting of positive, negative, and negation terms. A partial list of the dictionary is given in Table 4.

**Table 4.** A partial list of positive, negative, and negative words.

| Positive terms | Negative terms | Negation terms |
|---|---|---|
| *accurate, agile,*<br>*apt, boost,confident, ...* | *abandon, blame,*<br>*stubborn, torment, ...* | *not, never, none,*<br>*little, few, seldom, ...* |

## 5 Experiments

After combining the positive and negative dataset we have a total of 1,600 reviews. The dataset is divided into four categories: a) positive, deceptive reviews; b) positive, truthful reviews; c) negative, deceptive reviews; d) negative, truthful reviews. For every category, there are 20 reviews toward each of the 20 hotels. All the 400 reviews in each category are divided evenly into 5 folds. Each fold contains 80 reviews toward the corresponding 4 hotels.

We use a 5-fold nested cross validation [14] for evaluation. 4 folds are to do model selection. The selected model is then tested on the left out fold. The division of folds guarantees learned models are always evaluated on reviews from unseen hotels.

[6] http://www.cs.brandeis.edu/~clp/conll15st/index.html

[7] http://www.wjh.harvard.edu/~inquirer/homecat.htm

For basic language processing tools, BasePOS [15][16][17] is used to do POS tagging. We train our explicit discourse parser based on features extracted from constituent tree. The phrase structure parsing tree is predicted by the Berkeley Parser[8]. The parser outputs are converted by Standford dependency converter[9]. All the approaches we describe in Section 3 are used on a maximum entropy classifier[18][19].

## 6 Results and Discussion

As shown in Table 5, our explicit discourse parser obtains a good enough result. This prevents aggravating unnecessary errors into the following steps.

**Table 5.** Performance of explicit discourse parser.

| | Precision(%) | Recall(%) | F-score(%) |
|---|---|---|---|
| Explicit discourse Classifier | 91.2 | 89.1 | 90.1 |

**Table 6.** Performance of our approaches based on 5-fold nested cross validation.

| Features (%) | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|
| Bigram$^+$ + LIWC (baseline) | 88.4 | 89.1 | 87.5 | 88.3 |
| baseline + Sentiment | 88.6 | 89.2 | 88.2 | 88.7 |
| baseline + Syntactic | 89.1 | 89.1 | 89.3 | 89.2 |
| baseline + Syntactic + Discourse Parsing | **89.5** | **89.4** | **89.8** | **89.6** |

The result is given in Table 6. We can see that integrating sentiment features improves the recall by 0.8 percents, whereas the precision stays almost the same as the baseline. Based on annotation provided by Inquirer Dictionaries[7], the distribution of emotion terms in the dataset is shown in Table 7. We have two interesting findings. First, contrary to the analysis by Ott et al.[2], we find that the emotion intensity of deceptive reviews are tend to be milder than truthful reviews: Negative deceptive reviewers use less negative terms than truthful reviewers. And positive deceptive reviewers use less positive terms than truthful reviewers. Second, to our surprise, we find that in negative reviews, both deceptive and truthful reviews over-produce positive terms relative to negative terms. One reason would be that a negation word that is adjacent to a sentiment word can change its emotional polarity. We thus collect a list of negation words. The distribution across different categories is given in Table 8. We can see that there are more negation words in negative reviews.

The syntactic feature continues improving the recall with a slight drop on precision. Finally, combining with discourse parsing features, the model improves the accuracy by 1.1 percents.

[8] http://nlp.cs.berkeley.edu/

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

[7] http://www.wjh.harvard.edu/~inquirer/homecat.htm

**Table 7.** The number of sentiment words per review across different categories.

| Positive reviews | Deceptive reviews | Truthful reviews |
|---|---|---|
| Positive terms | 7.22 | 7.42 |
| Negative terms | 3.26 | 2.77 |

| Negative reviews | Deceptive reviews | Truthful reviews |
|---|---|---|
| Positive terms | 7 | 6.89 |
| Negative terms | 6.08 | 6.3 |

**Table 8.** The number of negation words per review across different categories.

| Negation Terms | Deceptive reviews | Truthful reviews |
|---|---|---|
| Positive reviews | 2.49 | 2.49 |
| Negative reviews | 4.44 | 4.64 |

## 7 Conclusion

In this work, we extend the existing standard deceptive opinion spam dataset. Then, we analyze its sentiment term distribution across different categories. Contrary to previous work, we find deceptive reviews tend to have a milder emotion than truthful reviews. We also receive a useful observation that paid-writers tend to demonstrate their truthfulness by stressing their presence as early as possible. Based on this discovery, a rule-based method can be effectively used for deceptive spam annotation, which demonstrates a better performance than human judge. To build a better model for deceptive spam detection, we consider two types of deep level linguistic features that are respectively given by an explicit discourse parser built from constituent parsing tree and a sentiment polarity classifier. After integrating the proposed features, our model gives a performance improvement by 1.1 percents. This result verifies the effectiveness of the proposed techniques.

## References

1. Ghosh, S., Johansson, R., Tonelli, S.: Shallow discourse parsing with conditional random fields. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, pp. 1071–1079, November 2011
2. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: Proceedings of NAACL-HLT 2013, Atlanta, Georgia, pp. 497–501, June 2013
3. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, California, USA, pp. 219–230, February 2008
4. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the ACL-2011, Stroudsburg, PA, USA, pp. 309–319, June 2011
5. Xu, Q., Zhao, H.: Using deep linguistic features for finding deceptive opinion spam. In: Proceedings of the 23th International Conference on Computational Linguistics, Mumbai, pp. 1341–1350, December 2012

6. Zhang, J., Zhao, H., Lu, B.L.: A comparative study on two large-scale hierarchical text classification tasks' solutions, ICMLC-2010, pp. 3275–3280 (2010)
7. Jiwei, L., Cardie, C., Sujian, L.: Topicspam: a topic-model-based approach for spam detection. In: Proceedings of ACL-2013, Sofia, Bulgaria, pp. 217–221, August 2013
8. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of ACL-2012, Jeju, Republic of Korea, pp. 171–175, July 2012
9. Banerjee, S., Chua, A.Y.: A linguistic framework to distinguish between genuine and deceptive online reviews. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. 1, Hong Kong, pp. 501–506, March 2014
10. Newman, M.L., Pennebaker, J.W., Berry, D.S., Richards, J.M.: Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin **29**(5), 665–675 (2003)
11. Lin, Z., Ng, H.T., Kan, M.Y.: A PDTB-styled end-to-end discourse parser. Natural Language Engineering **20**, 151–184 (2014)
12. Pitler, E., Nenkova, A.: Using syntax to disambiguate explicit discourse connectives in text. In: Proceedings of ACL-IJCNLP 2009, Suntec, Singapore, pp. 13–16, August 2009
13. Xiang, B., Zhou, L.: Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In: Proceedings of ACL-2014, Baltimore, Maryland, USA, pp. 434–439, June 2014
14. Quadrianto, N., Smola, A.J., Caetano, T.S., Le, Q.V.: Estimating labels from label proportions. The Journal of Machine Learning Research **10**, 2349–2374 (2009)
15. Zhao, H., Kit, C.: Parsing syntactic and semantic dependencies with two single-stage maximum entropy models. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning, Manchester, pp. 203–207, August 2008
16. Zhao, H., Chen, W., Kazama, J., Uchimoto, K., Torisawa, K.: Multilingual dependency learning: exploiting rich features for tagging syntactic and semantic dependencies. In: Proceedings of CoNLL-2009, Boulder, Colorado, pp. 61–66, June 2009
17. Zhao, H., Zhang, X., Kit, C.: Integrative semantic dependency parsing via efficient large-scale feature selection. Journal of Artificial Intelligence Research **46**, 203–233 (2013)
18. Jia, Z., Wang, P., Zhao, H.: Grammatical error correction as multiclass classification with single model. In: Proceedings of CoNLL- 2013, Sofia, Bulgaria, pp. 74–81, August 2013
19. Wang, P., Jia, Z., Zhao, H.: Grammatical error detection and correction using a single maximum entropy model. In: Proceedings of CoNLL-2014, Baltimore, Maryland, USA, pp. 74–82, July 2014