Using Deep Linguistic Features for Finding Deceptive Opinion Spam*

Qiongkai Xu[1,2] Hai Zhao[1,2†]
(1) MOE-Microsoft Key Laboratory of Intelligent Computing and Intelligent System;
(2) Department of Computer Science and Engineering, Shanghai Jiao Tong University,
#800 Dongchuan Road, Shanghai, China, 200240
xuqiongkai@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

ABSTRACT
While most recent work has focused on instances of opinion spam which are manually identifiable or deceptive opinion spam which are written by paid writers separately, in this work we study both of these interesting topics and propose an effective framework which has good performance on both datasets. Based on the golden-standard opinion spam dataset, we propose a novel model which integrates some deep linguistic features derived from a syntactic dependency parsing tree to discriminate deceptive opinions from normal ones. On a background of multiple language tasks, our model is evaluated on both English (gold-standard) and Chinese (non-gold) datasets. The experimental results show that our model produces state-of-the-art results on both of the topics.

TITLE AND ABSTRACT IN ANOTHER LANGUAGE (MANDARIN)

欺骗性垃圾信息中的高级语言特征探索

最近，许多研究工作分别着重研究了可手动识别的意见垃圾和由收费写手编写的欺骗性意见垃圾。本文中，我们提出了一个行之有效的框架，并在这两个不同的主题的数据集上都获得了良好的处理效果。基于标准的意见垃圾数据集，本文提出了一个集成了依存句法树等高级语言特征的新模型，用于欺骗性垃圾意见的识别。在多语言任务背景下，所提出的模型分别在英文数据集（职业写手编写的）和中文（手工标注的）数据集上进行了评估。实验结果表明，所提出的模型均能获得目前为止的最优结果。

KEYWORDS: Opinion Spam, Multi-Language, Deep Linguistic Features.

KEYWORDS IN MANDARIN: 意见垃圾, 多语言, 深度语言特征.

†Corresponding author

*This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119 and Grant No. 61170114), the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 20110073120022, the National Basic Research Program of China (Grant No. 2009CB320901), and the European Union Seventh Framework Program (Grant No. 247619).

Proceedings of COLING 2012: Posters, pages 1341–1350,
COLING 2012, Mumbai, December 2012.

1 Introduction

With the growing number of review websites where users can express opinions (e.g.,TripAdvisor [1]), there is an increasing potential to gain money through opinion spam–inappropriate or fraudulent reviews. The large benefits of this result in the occurrence of a group of writers who only write articles that deceive the users. Their comments always mislead readers to buy or use their products. Our goal is to find the hidden features of the deceptive opinions written by these writers. In Chinese, paid writers who flood websites are called "water army" and recognizing them is called "Water Army detection".

Till now, considerable attentions have been paid to other kinds of spam, such as web spam, e-mail spam and so on. Research focused on opinion spam is rarely reported till now. Jindal and Liu (2007) was one of the earliest work about Internet review spam. Further more, most previous work in this area focused on finding methods of detecting opinion spam which can be identified by a human reader. Only detecting this kind of opinion spam is not enough, because users can easily recognize the useless information by themselves and will not be misled. A much more challenging task, detection of deceptive opinion has been proposed by (Yoo and Gretzel, 2009; Ott et al., 2011) which is based on a gold-standard dataset. They have done the data selection and initial analysis work on this interesting topic.

Our work uses gold-standard dataset collected by (Ott et al., 2011) and non-gold standard Chinese dataset collected by ourselves. We give a machine learning model which is about 2 percent better than previous works on gold standard dataset and is also very effective on non-gold standard dataset. Later, we analyze the close relationship between sentence structure and deceptive opinion. Finally, statistical methodologies have been used to analyze all the feature sets and some theoretical contributions are summarized.

Our work mainly focuses on deceptive opinion spam. They are fictitious opinions that have been deliberately written by paid writers to sound authentic, in order to deceive the reader. If a deceptive opinion is mixed with a huge amount of truthful opinions, it is very hard for users to ignore or even identify it. The existing study proved that even a native speaker cannot identify most deceptive opinions. However, automatic classifiers have really good performance on these disturbing texts.

To obtain better performance on this task, we optimize it in two parts. Firstly, we test some other machine learning models. We use a support vector machine (SVM) as baseline to compare with the maximum entropy model (MEM). The result is that MEM performs better. Secondly, we try other approaches on this dataset and find that sentence structure features give really good performance. We combine all these improvements and train a final model that outputs state-of-the-art performance on gold-standard dataset. Later, this model is used on Chinese dataset collected by our group and also obtain good performance. Finally, we make some theoretical contributions to this topic.

2 Related Work

Most Internet spam detection work can be divided into two stages of development. The earliest work tried to detect spam which contained little useful information. They focused on the media of e-mail spam (Sahami et al., 1998; Drucker et al., 1999), web spam (Fetterly et al., 2004; Ntoulas et al., 2006),blog spam (Bhattarai et al., 2009), Twitter spam (Grier et al., 2010) and

[1] http://tripadvisor.com

review spam (Jindal and Liu, 2008). They used statistics and machine learning methodologies to analyze and investigate this topic extensively. In recent years, some researchers have begun to pay attention to the detection of spam which is deceptive. They analyzed review spam (Yoo and Gretzel, 2009; Wu et al., 2010; Ott et al., 2011; Li et al., 2011; Lau et al., 2012) and rumors on microblogs (Qazvinian et al., 2011). Following these works, our work deals with the second problem, deceptive opinion detection.

Although, opinion spam is widely spread on the Internet(Jindal and Liu, 2008). It is quite difficult to obtain a first-hand deceptive opinion dataset. Jindal and Liu (2008) used duplicate reviews as positive data and other views as negative examples[2]. Wu et al. (2010) tried to detect deceptive opinion spam by comparing popularity rankings. Qazvinian et al. (2011) annotated rumors (a similar concept to deceptive opinion) by experts manually which is a huge project. Our work may save such human judgement as we use gold-standard deceptive opinions.

Yoo and Gretzel (2009) first tried to collect a small gold-standard dataset from a group of tourism marketing students and statistical methods were used to analyze the difference between them from a psychological viewpoint. Ott et al. (2011) extended their work and collected a gold-standard dataset of 400 truthful and 400 deceptive opinions to develop automated deception classifiers. Following them, we also collect two datasets, 800 gold-standard opinions in English and 1800 non-gold standard opinions in Chinese (See section 3). We try to find the sentence structure or deep linguistic characteristics of deceptive opinions. By this effort, we improve the automated deceptive classifier by about 2 percent on gold-standard dataset and our model also works well on non-gold standard Chinese dataset collected by our group.

Chen et al. (2011) introduced the spam detection work into Chinese forums. They focus on detecting deceptive writers by using both semantic and non-semantic analysis. Spam writer detection was also investigated by (Lim et al., 2010; Mukherjee et al., 2011). Different from their works, our work only focuses on the content of opinions themselves with no additional information.

3 Dataset Construction

It is pointed that most of the opinions online are truthful(Jindal and Liu, 2008). Insidious deceptive opinions are very difficult to obtain. In this part, we describe where our English gold-standard deceptive opinions and Chinese manually annotated dataset come from.

3.1 Gold-standard English Dataset

Since Ott et al. (2011) have already provided a dataset which contains deceptive and truthful opinions, we use their dataset as our English gold standard dataset for our research. Bellow, we describe the detailed methods of collection.

3.1.1 Deceptive Opinions

To obtain a credible deceptive dataset, data collection procedure imitates the real way how these deceptive opinions are collected by asking those true deceptive opinion authors to do their jobs again. They created 400 Human Intelligence Tasks (HITs) using the Amazon Mechanical Turk[3](AMT)[4] with a one dollar award and allocated them to Turkers located in the United

[2]They suppose duplicate opinions are likely to be deceptive opinions

[3]http://mturk.com

[4]20 HITs for each of the 20 hotels they selected.

		TRUTHFUL			DECEPTIVE		
	Accuracy	P	R	F	P	R	F
META-old(native)	60.6%	60.8%	60.0%	60.4%	60.5%	61.3%	60.9%
META-new(non-native)	54.5%	52.8%	54.7%	53.7%	56.3%	61.5%	58.8%

Table 1: Performance of meta judgement of three college students, corresponding to the cross-validation experiment in Section 5.

States. They imposed a restriction that all the opinions should be written by unique authors to avoid that classifiers are over-tuned by different author styles, and all the tasks should be finished in 30 minutes.

They told the Turkers the name and website of a hotel. The Turkers were asked to assume that they worked for the hotel and write a deceptive, realistic-sounding and positive review for the hotel. Finally, they filtered out all the insufficient quality reviews(e.g., unreasonably short, plagiarized and so on) and obtained 400 golden deceptive opinions. These opinions were used as the deceptive part of the dataset.

3.1.2 Truthful Opinions

For truthful part, they first got all 6,977 opinions of the 20 most popular hotels (Same as the 20 hotels chosen for HIT) form TripAdvisor. To balance the number of truthful opinions and deceptive opinions, 20 opinions for each of the 20 hotels that meet the following conditions were selected[5]:

- 5-star[6] review;
- Only English reviews;
- More than 150 characters, because most deceptive opinions have at least 150 characters;
- Not written by first-time authors (new users who have not previously posted an opinion);[7]

3.1.3 Human Performance

Ott et al. (2011) have proved the human performance on their dataset is low and made this a baseline for further discussion. The highest result is from meta-judge[8] of the three students, as presented in Table 1. We also ask three Chinese college students who have passed CET6 (College English Test Level 6) for help to make the judgement on a subset of this data. We label the review deceptive when any of the students believe that the review is deceptive. The result is shown in the second line of Table 1.

The result in Table 1 shows that non-native speakers perform even worse than native speakers. Both these meta judges will be used as baselines to compare with the automatic approach of detecting deceptive opinions.

[5]Same as the hotels selected for deceptive opinion dataset
[6]Score given by user from 1-star to 5-star shows the support of the user for the hotel.
[7]First-time authors are more likely to give opinion spam(Wu et al., 2010)
[8]Meta judge labels a review deceptive when any human judge believes the review to be deceptive.

total	DECEPTIVE	TRUTHFUL
23397	22337	1060

Table 2: Number of each class of instances.

3.2 Chinese Dataset

Chinese dataset is collected from a famous Chinese online forum [9]. Since it is very hard to find qualified deceptive opinion authors, we use a collection-and-annotation method. We collected over 20000 reviews and asked two experts who are very familiar with photography to go through all these reviews and marked each review a spam or not, see Table 2. Finally, we get a dataset of 1800 reviews (900 positive and 900 negative opinions) for the balance of data.

To evaluate the accuracy of our Chinese non-gold dataset, we annotated 800 reviews randomly selected instances twice and Kappa coefficient (κ) was calculated to compare the result of each annotator by the following formula:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{1}$$

where $Pr(a)$ is the proportion of times that the two annotators agree and $Pr(e)$ is the proportion of times that they would are expected to agree by chance(Carletta, 1996). The Kappa coefficient result is 0.97. This result shows that our annotators reach a high agreement in our deceptive opinion annotation task.

4 Deceptive Opinion Detection

To train an effective classifier, Ott et al. (2011) mainly focus on the following three approaches:

- Genre identification;
- Psycholinguist deceptive detection;
- Text categorization;

Another approach to identify deceptive opinion spam is using structural features of sentences. We are inspired by the idea that the genre of the text can be used to detect deceptive opinion spam. The structure of the sentence shows the genre of a writer, in some sense. We use the following two approaches to describe the structure of the sentence :

- BIPOS($POS_{-1} + POS$)
- DEP($DEP\_label + form + head\_form$)

Since the frequency distribution of part-of-speech (POS) tags in a text is often dependent on the genre of the text (Newman et al., 2003) and POS tag bigrams (BIPOS) will not only show frequency information of POS, but also show the structure of the sentence, we suppose that POS tag bigrams will give a better performance than pure POS features. By counting the frequency

[9] http://bbs.fengniao.com/forum/. Most topics on this forum are about photography.

of different structural features of the sentence, we get the hidden genre information of the text. These features also provide a baseline with which to compare our other sentence structure features.

The structure of the sentence is usually represented by a parsing tree. Among various existing syntactic parsers, the dependency parser is chosen for this work due to its simplicity. We extract the corresponding feature from the output of the Stanford parser (De Marneffe et al., 2006). Three dependency parsing features are integrated, dependency label (*DEP_label*), word forms (*form*) of the current word and the head word (*head_form*). These features are used as part of a sentence structure feature set in our classifier.

4.1 Classifiers

Features from the approaches just introduced are used to train support vector machine and maximum entropy classifiers, both of which are well known machine learning models which have performed well in related work (Zhang and Yao, 2003; Ott et al., 2011).

We train a support vector machine (SVM) classifier, which finds a high-dimensional separating hyperplane between two groups of data. This method is proved useful in (Ott et al., 2011). Instances can be classified by the following formula:

$$y(x) = sign\left[\sum_{k=1}^{N} \alpha_k y_k \Phi(x, x_k) + b\right] \tag{2}$$

Where N is the number of instance, x_k is the kth input pattern and y_k is the kth output pattern. Φ is the kernel function : $\Phi(x, x_k) = x_k^T x$ (linear SVM).

We use LIBSVM(Chang and Lin, 2011) to train our linear support vector machine (SVM) models on all the approaches mentioned in Section 4.

We also train a maximum entropy model (MEM), which finds the probability distribution that satisfies the constraints and minimizes the relative entropy. In general, a conditional Maximum Entropy model is an exponential (log-linear) model which has the form:

$$p(a|b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)} \tag{3}$$

where $p(a|b)$ denotes the probability of predicting an outcome a in the given context b with constraint or "feature" functions $f_j(a, b)$. Here k is the number of features and $Z(b) = \sum_a \prod_{j=1}^{k} \alpha_j^{f_j(a,b)}$.

We use openNLP MAXENT[10] (Berger et al., 1996) and iterate 200 times to train our models on all the approaches mentioned in Section 4, the same as the approaches used for the support vector machine (SVM) model.

5 Experiment and Discussion

5.1 Experiment

We use a five-fold nested cross validation (CV) (Quadrianto et al., 2009) procedure to evaluate the performance of each feature set. The result is given in Table 3. SVM-linear line is the result

[10] http://incubator.apache.org/opennlp/

			TRUTHFUL			DECEPTIVE		
Model	Feature	Accuracy	P	R	F	P	R	F
SVM-linear (baseline)	UNIGRAM	88.4%	89.9	86.5	88.2	87.0	90.3	88.6
	BIGRAM$^+$	89.6%	**90.1**	89.0	89.6	89.1	**90.3**	89.7
	BIGRAM$^+$ + LIWC	**89.8%**	89.8	**89.8**	**89.8**	**89.8**	89.8	**89.8**
MEM	POS	74.0%	72.0	75.0	73.5	76.0	73.1	74.5
	BIPOS	76.9%	76.3	77.2	76.7	77.5	76.5	77.0
	DEP	86.3%	86.3	86.3	86.3	86.3	86.3	86.3
	UNIGRAM	90.6%	89.0	92.0	90.5	92.2	89.3	90.8
	UNIGRAM + DEP	**91.6%**	**90.8**	92.4	**91.6**	92.5	**90.9**	**91.7**
	UNIGRAM + LIWC	91.4%	89.8	**92.8**	91.2	**93.0**	90.1	91.5
	BIGRAM$^+$	90.5%	89.3	91.5	90.4	91.8	89.5	90.6
META_JUDGEMENT_old		60.6%	60.8	60.0	60.4	60.5	61.3	60.9
META_JUDGEMENT_new		58.1%	53.1	56.3	55.5	56.3	54.4	55.3

Table 3: Performance of our approaches based on 5-fold cross-validation (CV) experiments with accuracy, precision, recall and F-score. Baseline is the performance of the approaches according to (Ott et al., 2011) on our dataset.

		TRUTHFUL			DECEPTIVE		
Feature	Accuracy	P	R	F	P	R	F
UNIGRAM+DET	79.1%	89.8	74.0	81.1	68.5	87.0	76.6
BIGRAM$^+$	78.3%	89.8	73.0	80.5	66.8	86.7	75.4
BIGRAM$^+$+DET	78.8%	89.8	73.6	80.9	67.8	86.9	76.1

Table 4: Performance of 5-fold cross-validation (CV) experiments with accuracy, precision, recall and F-score on Chinese dataset

POS		UNIGRAM		DEPENDENCY label only	
DECEPTIVE	TRUTHFUL	DECEPTIVE	TRUTHFUL	DECEPTIVE	TRUTHFUL
-LRB-	JJ	prime_JJ	why_WRB	punct	mwe
,	WP	home_NN	etc._NN	prt	purpcl
PRP	NN	well_NN	commented_VBD	possessive	advmod
-RRB-	MD	convention_NN	extras_NNS	iobj	aux
CC	.	round_NN	downstairs_NNS	nsubj	amod
BIPOS		**BIGRAM**		**DEPENDENCY detail**	
DECEPTIVE	TRUTHFUL	DECEPTIVE	TRUTHFUL	DECEPTIVE	TRUTHFUL
WRB_FW	RB_,	checking_VBG_out_PRP	was_VBD_worth_IN	pobj(for&members)	pobj(with&amenity)
VBG_PDT	IN_$	want_VB_to_TO	Just_RB_returned_VBN	mark(visiting&while)	nn(Michigan&Lake)
NNP_DT	VBD_RP	the_DT_wine_NN	feeling_NN_._.	cop(spacious&is)	prep(surrounded&by)
VBN_VBP	WP_PRP	next_JJ_year_NN	level_NN_._.	root(ROOT&leave)	xcomp(in&check)
IN_VBZ	CC_FW	a_DT_breakfast_NN	and_CC_take_VB	amod(staff&excellent)	prep(reminded&of)

Table 5: 15 most frequently occurring features of each feature set. Ranks of deceptive and truthful opinion are separated.

of the approach in (Ott et al., 2011). MEM line is our new model which outperforms previous work.

We also give the results of non-native speaker performance on gold-standard dataset in Table 1. We find that they do worse work than native speaker. We attribute this to the reason that the students that we asked for help are Chinese college students and English is not their native language. That suggests that deceptive text can mislead foreigners more easily.

On a background of multiple language tasks, our model is also tested on our Chinese dataset. All the approaches described in Section 4 were used on the MEM. We use BaseSeg (Zhao et al., 2006) as word segmenter, BasePos[11] as POS (part of speech) tagger and FudanNLP tools [12] as dependency parser. The highest three results are shown in Table 4. BIGRAM$^+$+DEP outperforms BIGRAM$^+$ shows that sentence structure features also give good performance on Chinese and our model keeps effective.

5.2 Discussion

Comparing POS feature alone with POS tag bigram (BIPOS) features, we find BIPOS always performs better. That means POS feature alone cannot fully represent the genre of an opinion. On the other hand, the BIPOS feature set has much richer features and can classify the genre more easily. Since sentence structure also indicates the genre of a text, we will use this feature set as an optimization feature set.

We have tested different combinations of feature sets and listed the representative results, see Table 3. We find that UNIGRAM+DEP works best on maximum entropy model(MEM), about 2% higher than best result of SVM-linear model(BIGRAM$^+$+LIWC). This proves that sentence structure can decide the genre of a text and detect deceptive opinions.

To make the following analysis clear, the 5 highest weighted features (learned by MEM) for each feature set for deceptive opinion and truthful opinion are listed in Table 5. Observation results are shown below: (1) PRP has a high weight in deceptive opinion spam, which means that deceptive opinions are more likely to use personal words. (2) Such forms, nn(Michigan&Lake) weights high showing that truthful opinion always provided concrete information like a location. (3) Words like home, well, wine, breakfast, excellent which are normally used in daily life get higher weight in deceptive opinion, while words like feeling, downstairs which can reflect self feeling are more likely to occur in truthful opinion. (4) etc. obtain high weight in truthful opinion detection meaning that truth authors can sometimes give concrete examples to elaborate their views.

6 Conclusion

In this work, we made an extensive annotation based on the existing dataset for deceptive opinion spam detection. We tried a new approach, using deep linguistic features, for this task and proved it useful. We also tested some other classifiers and improved the classification models for the task. The proposed model outperforms the baseline system by about 2%. On the background of multiple language tasks, our new model was tested on both English and Chinese datasets and proved to be useful.

[11] http://bcmi.sjtu.edu.cn/ zhaohai/index.html
[12] http://code.google.com/p/fudannlp/

References

Berger, A., Pietra, V., and Pietra, S. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Bhattarai, A., Rus, V., and Dasgupta, D. (2009). Characterizing comment spam in the blogosphere through content analysis. In *Computational Intelligence in Cyber Security, 2009. CICS'09. IEEE Symposium on*, pages 37–44. IEEE.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Chang, C. and Lin, C. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Chen, C., Wu, K., Srinivasan, V., and Zhang, X. (2011). Battling the internet water army: Detection of hidden paid posters. *CoRR*, abs/1111.4297.

De Marneffe, M., MacCartney, B., and Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Drucker, H., Wu, D., and Vapnik, V. (1999). Support vector machines for spam categorization. *Neural Networks, IEEE Transactions on*, 10(5):1048–1054.

Fetterly, D., Manasse, M., and Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*, pages 1–6. ACM.

Grier, C., Thomas, K., Paxson, V., and Zhang, M. (2010). @ spam: the underground on 140 characters or less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37. ACM.

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

Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230. ACM.

Lau, R. Y. K., Liao, S. Y., Kwok, R. C.-W., Xu, K., Xia, Y., and Li, Y. (2012). Text mining and probabilistic language modeling for online review spam detection. *ACM Trans. Manage. Inf. Syst.*, 2:25:1–25:30.

Li, F., Huang, M., Yang, Y., and Zhu, X. (2011). Learning to identify review spam. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Lim, E., Nguyen, V., Jindal, N., Liu, B., and Lauw, H. (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 939–948. ACM.

Mukherjee, A., Liu, B., Wang, J., Glance, N., and Jindal, N. (2011). Detecting group review spam. In *Proceedings of the 20th international conference companion on World wide web*, pages 93–94. ACM.

Newman, M., Pennebaker, J., Berry, D., and Richards, J. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665.

Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. (2006). Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). 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*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Qazvinian, V., Rosengren, E., Radev, D. R., and Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Quadrianto, N., Smola, A., Caetano, T., and Le, Q. (2009). Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10:2349–2374.

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*, volume 62, pages 98–105. Madison, Wisconsin: AAAI Technical Report WS-98-05.

Wu, G., Greene, D., Smyth, B., and Cunningham, P. (2010). Distortion as a validation criterion in the identification of suspicious reviews. In *Proceedings of the First Workshop on Social Media Analytics*, pages 10–13. ACM.

Yoo, K. and Gretzel, U. (2009). Comparison of deceptive and truthful travel reviews. *Information and Communication Technologies in Tourism 2009*, pages 37–47.

Zhang, L. and Yao, T. (2003). Filtering junk mail with a maximum entropy model. In *Proceeding of 20th international conference on computer processing of oriental languages (ICCPOL03)*, pages 446–453.

Zhao, H., Huang, C., and Li, M. (2006). An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, volume 1082117. Sydney: July.