TopicSpam: a Topic-Model-Based Approach for Spam Detection

Jiwei Li , Claire Cardie
School of Computer Science
Cornell University
Ithaca, NY, 14853
jl3226@cornell.edu
cardie@cs.cornell.edu

Sujian Li
Laboratory of Computational Linguistics
Peking University
Bejing, P.R.China, 150001
lisujian@pku.edu.cn

Abstract

Product reviews are now widely used by individuals and organizations for decision making (Litvin et al., 2008; Jansen, 2010). And because of the profits at stake, people have been known to try to game the system by writing fake reviews to promote target products. As a result, the task of deceptive review detection has been gaining increasing attention. In this paper, we propose a generative LDA-based topic modeling approach for fake review detection. Our model can aptly detect the subtle differences between deceptive reviews and truthful ones and achieves about 95% accuracy on review spam datasets, outperforming existing baselines by a large margin.

1 Introduction

Consumers rely increasingly on user-generated online reviews to make purchase decisions. Positive opinions can result in significant financial gains. This gives rise to *deceptive opinion spam* (Ott et al., 2011; Jindal et al., 2008), fake reviews written to sound authentic and deliberately mislead readers. Previous research has shown that humans have difficulty distinguishing fake from truthful reviews, operating for the most part at chance (Ott et al., 2011). Consider, for example, the following two hotel reviews. One is truthful and the other is deceptive[1]:

1. *My husband and I stayed for two nights at the Hilton Chicago. We were very pleased with the accommodations and enjoyed the service every minute of it! The bedrooms are immaculate, and the linens are very soft. We also appreciated the free wifi, as we could stay in touch with friends while staying in Chicago. The bathroom was quite spacious, and I loved the smell of the shampoo they provided. Their service was amazing, and we absolutely loved the beautiful indoor pool. I would recommend staying here to anyone.*
2. *We stayed at the Sheraton by Navy Pier the first weekend of November. The view from both rooms was spectacular (as you can tell from the picture attached). They also left a plate of cookies and treats in the kids room upon check-in made us all feel very special. The hotel is central to both Navy Pier and Michigan Ave. so we walked, trolleyed, and cabbed all around the area. We ate the breakfast buffet on both mornings and thought it was pretty good. The eggs were a little runny. Our six year old ate free and our two eleven year old were* $*14 (instead of the adult* $*20). The rooms were clean, the concierge and reception staff were both friendly and helpful...we will definitely visit this Sheraton again when we stay in Chicago next time.*

Because of the difficulty of recognizing deceptive opinions, there has been a widespread and growing interest in developing automatic, usually learning-based methods to help users identify deceptive reviews (Ott et al., 2011; Jindal et al., 2008; Jindal et al., 2010; Li et al., 2011; Lim et al., 2011; Wang et al., 2011).

The state-of-the-art approach treats the task of spam detection as a $text\ categorization$ problem and was first introduced by Jindal and Liu (2009) who trained a supervised classifier to distinguish duplicated reviews (assumed deceptive) from original ones (assumed truthful). Since then, many supervised approaches have been proposed for spam detection. Ott et al. (2011) employed standard word and part-of-speech (POS) n-gram features for supervised learning and built a $gold-standard$ opinion dataset of 800 reviews. Lim et al. (2010) proposed the inclusion of user behavior-based features and found that behavior abnormalities of reviewers could predict spammers, without using any textual features. Li et al. (2011) carefully explored review-related features based on content and sentiment, training a semi-supervised classifier for opinion spam detection. However, the disadvantages of standard supervised learning methods are obvious. First, they do not generally provide readers with a clear probabilistic pre-

[1] The first example is a deceptive review.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 217–221,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

diction of how likely a review is to be deceptive vs. truthful. Furthermore, identifying features that provide direct evidence against deceptive reviews is always a hard problem.

LDA topic models (Blei et al., 2003) have widely been used for their ability to model latent topics in document collection. In LDA, each document is presented as a mixture distribution of topics and each topic is presented as a mixture distribution of words. Researchers also integrated different levels of information into LDA topic models to model the specific knowledge that they are interested in, such as user-specific information (Rosen-zvi et al., 2006), document-specific information (Li et al., 2010) and time-specific information (Diao et al., 2012). Ramage et al. (2009) developed a Labeled LDA model to define a one-to-one correspondence between LDA latent topics and tags. Chemudugunta et al. (2008) illustrated that by considering background information and document-specific information, we can largely improve the performance of topic modeling.

In this paper, we propose a Bayesian approach called TopicSpam for deceptive review detection. Our approach, which is a variation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), aims to detect the subtle differences between the topic-word distributions of deceptive reviews vs. truthful ones. In addition, our model can give a clear probabilistic prediction on how likely a review should be treated as deceptive or truthful. Performance is tested on dataset from Ott et al.(2011) that contains 800 reviews of 20 Chicago hotels. Our model achieves more than 94% accuracy on that dataset.

2 TopicSpam

We are presented with four subsets of hotel reviews, $M = \{M_i\}_{i=1}^{i=4}$, representing $deceptive\ train$, $truthful\ train$, $deceptive\ test$ and $truthful\ test$ data, respectively. Each review r is comprised of a number of words $r = \{w_t\}_{t=1}^{t=n_r}$. Input for the TopicSpam algorithm is the datasets M; output is the label (deceptive, truthful) for each review in M_3 and M_4. V denotes vocabulary size.

2.1 Details of TopicSpam

In TopicSpam, each document is modeled as a bag of words, which are assumed to be generated from a mixture of latent topics. Each word is associated with a latent variable that specifies the topic from which it is generated. Words in a document are assumed to be conditionally independent given the hidden topics. A general background distribution ϕ^B and hotel-specific distributions $\phi^{H_j}(j = 1, ..., 20)$ are first introduced to capture the background information and hotel-specific information. To capture the difference between deceptive reviews and truthful reviews, TopicSpam also learns a deceptive topic distribution ϕ^D and truthful topic distribution ϕ^T. The generative model of TopicSpam is shown as follows:

- For a training review in $r_{1j} \in M_1$, words are originated from one of the three different topics: ϕ^B, ϕ^{H_j} and ϕ^D.
- For a training review in $r_{2j} \in M_2$, words are originated from one of the three different topics: ϕ^B, ϕ^{H_j} and ϕ^T.
- For a test review in $r_{mj} \in M_m, m = 3, 4$, words are originated from one of the four different topics: ϕ^B, ϕ^{H_j} ϕ^D and ϕ^T.

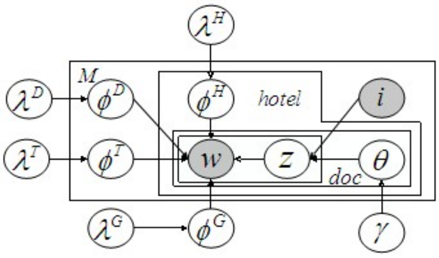

Figure 1: Graphical Model for TopicSpam

The generation process of TopicSpam is shown in Figure 1 and the corresponding graphical model is illustrated in Figure 2. We use $\lambda = (\lambda_G, \lambda_{H_i}, \lambda_D, \lambda_T)$ to represent the asymmetric priors for topic-word distribution generation. In our experiments, we set $\lambda_G = 0.1$, and $\lambda_{H_i} = \lambda_D = \lambda_T = 0.01$. The intuition for the asymmetric priors is that there should be more words assigned to the background topic. $\gamma = [\gamma_B, \gamma_{H_i}, \gamma_D, \gamma_T]$ denotes the priors for the document-level topic distribution in the LDA model. We set $\gamma_B = 2$ and $\gamma_T = \gamma_D = \gamma_{H_i} = 1$, reflecting the intuition that more words in each document should cover the background topic.

2.2 Inference

We adopt the collapsed Gibbs sampling strategy to infer the latent parameters in TopicSpam. In Gibbs

```
1. sample $\phi^G \sim Dir(\lambda^G)$
2. sample $\phi^D \sim Dir(\lambda^D)$
3. sample $\phi^T \sim Dir(\lambda^T)$
4. for each hotel $j \in [1, N]$: sample $\phi^{H_j} \sim \lambda^H$
5. for each review $r$
        if i=1: sample $\theta_r \sim Dir(\gamma_B, \gamma_{H_j}, \gamma_D)$
        if i=2: sample $\theta_r \sim Dir(\gamma_B, \gamma_{H_j}, \gamma_T)$
        if i=3: sample $\theta_r \sim Dir(\gamma_B, \gamma_{H_j}, \gamma_D, \gamma_T)$
        if i=4: sample $\theta_r \sim Dir(\gamma_B, \gamma_{H_j}, \gamma_D, \gamma_T)$
        for each word $w$ in $R$
            sample $z \sim \theta_r$   sample $w \sim \phi^z$
```

Figure 2: Generative Model for TopicSpam

sampling, for each word w in review r, we need to calculate $P(z_w|w, z_{-w}, \gamma, \lambda)$ in each iteration, where z_{-w} denotes the topic assignments for all words except that of the current word z_w.

$$P(z_w = m|z_{-w}, i, j, \gamma, \lambda)$$
$$\frac{N_r^m + \gamma_m}{\sum_{m'}(N_r^{m'} + \gamma'_m)} \cdot \frac{E_m^w + \lambda_m}{\sum_{w'}^V E_m^w + V\lambda_m} \quad (1)$$

where N_r^m denotes the number of times that topic m appears in current review r and E_m^w denotes the number of times that word w is assigned to topic m. After each sampling iteration, the latent parameters can be estimated using the following formulas:

$$\theta_r^m = \frac{N_r^m + \gamma_m}{\sum_{m'}(N_r^{m'} + \gamma_m)} \quad \phi_m^{(w)} = \frac{E_m^w + \lambda_m}{\sum_{w'} E_m^{w'} + V\lambda_m} \quad (2)$$

2.3 Labeling the Test Data

For each review r in the test data, let N_r^D denote the number of words generated from the deceptive topic and N_r^T, the number of words generated from the truthful topic. The decision for whether a review is deceptive or truthful is made as follows:

- if $N_r^D > N_r^T$, r is deceptive.
- if $N_r^D < N_r^T$, r is truthful.
- if $N_r^D = N_r^T$, it is hard to decide.

Let P(D) denote the probability that r is deceptive and P(T) denote the probability that r is truthful.

$$P(D) = \frac{N_r^D}{N_r^D + N_r^T} \quad P(T) = \frac{N_r^T}{N_r^D + N_r^T} \quad (3)$$

3 Experiments

3.1 System Description

Our experiments are conducted on the dataset from Ott et al.(2011), which contains reviews of the 20 most popular hotels on TripAdvisor in the Chicago areas. There are 20 truthful and 20 deceptive reviews for each of the chosen hotels (800 reviews total). Deceptive reviews are gathered using Amazon Mechanical Turk[2]. In our experiments, we adopt the same 5-fold cross-validation strategy as in Ott et al., using the same data partitions. Words are stemmed using PorterStemmer[3].

3.2 Baselines

We employ a number of techniques as baselines:

TopicTD: A topic-modeling approach that only considers two topics: deceptive and truthful. Words in $deceptive\ train$ are all generated from the deceptive topic and words in $truthful\ train$ are generated from the truthful topic. Test documents are presented with a mixture of the deceptive and truthful topics.

TopicTDB: A topic-modeling approach that only considers background, deceptive and truthful information.

SVM-Unigram: Using SVMlight(Joachims, 1999) to train linear SVM models on unigram features.

SVM-Bigram: Using SVMlight(Joachims, 1999) to train linear SVM models on bigram features.

SVM-Unigram-Removal1: In SVM-Unigram-Removal, we first train TopicSpam. Then words generated from hotel-specific topics are removed. We use the remaining words as features in SVMlight.

SVM-Unigram-Removal2: Same as SVM-Unigram-removal-1 but removing all background words and hotel-specific words.

Experimental results are shown in Table 1[4]. As we can see, the accuracy of TopicSpam is 0.948, outperforming TopicTD by 6.4%. This illustrates the effectiveness of modeling background and hotel-specific information for the opinion spam detection problem. We also see that TopicSpam slightly outperforms TopicTDB, which

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

[3] http://tartarus.org/martin/PorterStemmer/

[4] Reviews with $N_r^D = N_r^T$ are regarded as incorrectly classified by TopicSpam.

Approach	Accuracy	T-P	T-R	T-F	D-P	D-R	D-F
TopicSpam	0.948	0.954	0.942	0.944	0.941	0.952	0.946
TopicTD	0.888	0.901	0.878	0.889	0.875	0.897	0.886
TopicTDB	0.931	0.938	0.926	0.932	0.925	0.937	0.930
SVM-Unigram	0.884	0.899	0.865	0.882	0.870	0.903	0.886
SVM-Bigram	0.896	0.901	0.890	0.896	0.891	0.903	0.897
SVM-Unigram-Removal1	0.895	0.906	0.889	0.898	0.887	0.907	0.898
SVM-Unigram-Removal2	0.822	0.852	0.806	0.829	0.793	0.840	0.817

Table 1: Performance for different approaches based on nested 5-fold cross-validation experiments.

neglects hotel-specific information. By checking the results of Gibbs sampling, we find that this is because only a small number of words are generated by the hotel-specific topics. TopicTD and SVM-Unigram get comparative accuracy rates. This can be explained by the fact that both models use unigram frequency as features for the classifier or topic distribution training. SVM-Unigram-Removal1 is also slightly better than SVM-Unigram. In SVM-Unigram-removal1, hotel-specific words are removed for classifier training. So the first-step LDA model can be viewed as a feature selection process for the SVM, giving rise to better results. We can also see that the performance of SVM-Unigram-removal2 is worse than other baselines. This can be explained as follows: for example, word ”my” has large probability to be generated from the background topic. However it can also be generated by deceptive topic occasionaly but can hardly be generated from the truthful topic. So the removal of these words results in the loss of useful information, and leads to low accuracy rate.

Our topic-modeling approach uses word frequency as features and does not involve any feature selection process. Here we present the results of the sample reviews from Section 1. Stop words are labeled in black, background topics (B) in blue, hotel specific topics (H) in orange, deceptive topics (D) in red and truthful topic (T) in green.

1. *My husband and I stayed for two nights at the Hilton Chicago. We were very pleased with the accommodations and enjoyed the service every minute of it! The bedrooms are immaculate,and the linens are very soft. We also appreciated the free wifi, as we could stay in touch with friends while staying in Chicago. The bathroom was quite spacious, and I loved the smell of the shampoo they provided not like most hotel shampoos. Their service was amazing,and we absolutely loved the beautiful indoor pool. I would recommend staying here to anyone.*
   [B,H,D,T]=[41,6,10,1] p(D)=0.909 P(T)=0.091

2. *We stayed at the Sheraton by Navy Pier the first weekend of November. The view from both rooms was spectacular (as you can tell from the picture attached). They also left a plate of cookies and treats in the kids room upon check-in made us all feel very special. The hotel is central to both Navy Pier and Michigan Ave. so we walked, trolleyed, and cabbed all around the area. We ate the breakfast buffet both mornings and thought it was pretty good. The eggs were a little runny. Our six year old ate free and our two eleven year old were $14 (instead of the adult $20) The rooms were clean, the concierge and reception staff were both friendly and helpful...we will definitely visit this Sheraton again when we're in Chicago next time.*
   [B,H,D,T]=[80,15,3,18] p(D)=0.143 P(T)=0.857

background	deceptive	truthful	Hilton
hotel	hotel	room	Hilton
stay	my	)	palmer
we	chicago	(	millennium
room	will	but	lockwood
!	room	$	park
Chicago	very	bathroom	lobby
my	visit	location	line
great	husband	night	valet
I	city	walk	shampoo
very	experience	park	dog
Omni	Amalfi	Sheraton	James
Omni	Amalfi	tower	James
pool	breakfast	Sheraton	service
plasma	view	pool	spa
sundeck	floor	river	bar
chocolate	bathroom	lake	upgrade
indoor	cocktail	navy	primehouse
request	morning	indoor	design
pillow	wine	shower	overlook
suitable	great	kid	romantic
area	room	theater	home

Table 2: Top words in different topics from TopicSpam

4 Conclusion

In this paper, we propose a novel topic model for deceptive opinion spam detection. Our model achieves an accuracy of 94.8%, demonstrating its effectiveness on the task.

5 Acknowledgements

We thank Myle Ott for his insightful comments and suggestions. This work was supported in part by NSF Grant BCS-0904822, a DARPA Deft grant, and by a gift from Google.

References

David Blei, Andrew Ng and Micheal Jordan. Latent Dirichlet allocation. 2003. In *Journal of Machine Learning Research.*

Carlos Castillo, Debora Donato, Luca Becchetti, Paolo Boldi, Stefano Leonardi Massimo Santini, and Sebastiano Vigna. A reference collection for web spam. In *Proceedings of annual international ACM SIGIR conference on Research and development in information retrieval, 2006.*

Chaltanya Chemudugunta, Padhraic Smyth and Mark Steyers. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model.. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference.*

Paul-Alexandru Chirita, Jorg Diederich, and Wolfgang Nejdl. MailRank: using ranking for spam detection. In *Proceedings of ACM international conference on Information and knowledge management. 2005.*

Harris Drucke, Donghui Wu, and Vladimir Vapnik. 2002. Support vector machines for spam categorization. In *Neural Networks.*

Qiming Diao, Jing Jiang, Feida Zhu and Ee-Peng Lim. In *Proceeding of the 50th Annual Meeting of the Association for Computational Linguistics. 2012*

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. In *Advances in kernel methods.*

Jack Jansen. 2010. Online product research. In *Pew Internet and American Life Project Report.*

Nitin Jindal, and Bing Liu. Opinion spam and analysis. 2008. In *Proceedings of the international conference on Web search and web data mining*

Nitin Jindal, Bing Liu, and Ee-Peng Lim. Finding Unusual Review Patterns Using Unexpected Rules. 2010. In *Proceedings of the 19th ACM international conference on Information and knowledge management*

Pranam Kolari, Akshay Java, Tim Finin, Tim Oates and Anupam Joshi. Detecting Spam Blogs: A Machine Learning Approach. In *Proceedings of Association for the Advancement of Artificial Intelligence. 2006.*

Peng Li, Jing Jiang and Yinglin Wang. 2010. Generating templates of entity summaries with an entity-aspect model and pattern mining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.*

Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. Learning to identify review Spam. 2011. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence.*

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting Product Review Spammers Using Rating Behavior. 2010. In *Proceedings of the 19th ACM international conference on Information and knowledge management.*

Stephen Litvina, Ronald Goldsmithb and Bing Pana. 2008. Electronic word-of-mouth in hospitality and tourism management. *Tourism management*, 29(3):458468.

Juan Martinez-Romo and Lourdes Araujo. Web Spam Identification Through Language Model Analysis. In *AIRWeb. 2009.*

Arjun Mukherjee, Bing Liu and Natalie Glance. Spotting Fake Reviewer Groups in Consumer Reviews. In *Proceedings of the 18th international conference on World wide web, 2012.*

Alexandros Ntoulas, Marc Najork, Mark Manasse and Dennis Fetterly. Detecting Spam Web Pages through Content Analysis. In *Proceedings of international conference on World Wide Web 2006*

Myle Ott, Yejin Choi, Claire Cardie and Jeffrey Hancock. Finding deceptive opinion spam by any stretch of the imagination. 2011. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. In *Found. Trends Inf. Retr.*

Daniel Ramage, David Hall, Ramesh Nallapati and Christopher D. Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. 2009. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing 2009.*

Michal Rosen-zvi, Thomas Griffith, Mark Steyvers and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence.*

Guan Wang, Sihong Xie, Bing Liu and Philip Yu. Review Graph based Online Store Review Spammer Detection. 2011. In *Proceedings of 11th International Conference of Data Mining*.

Baoning Wu, Vinay Goel and Brian Davison. Topical TrustRank: using topicality to combat Web spam. In *Proceedings of international conference on World Wide Web 2006 .*

Kyang Yoo and Ulrike Gretzel. 2009. Comparison of Deceptive and Truthful Travel Reviews. In*Information and Communication Technologies in Tourism 2009.*