Fake Reviews Detection Based on LDA

Shaohua Jia
College of Computer Science
Inner Mongolia University
Hohhot,China
e-mail: 1424221609@qq.com

Xianguo Zhang, Xinyue Wang, Yang Liu
College of Computer Science
Inner Mongolia University
Hohhot,China
e-mail: 2595083628@qq.com
2365435187@qq.com
136564850@qq.com

***Abstract*—It is necessary for potential consume to make decision based on online reviews. However, its usefulness brings forth a curse – deceptive opinion spam. The deceptive opinion spam mislead potential customers and organizations reshaping their businesses and prevent opinion-mining techniques from reaching accurate conclusions. Thus, the detection of fake reviews has become more and more fervent. In this work, we attempt to find out how to distinguish between fake reviews and non-fake reviews by using linguistic features in terms of Yelp Filter Dataset. To our surprise, the linguistic features performed well. Further, we proposed a method to extract features based on Latent Dirichlet Allocation. The result of experiment proved that the method is effective.**

Keywords-Review detection; Linguistic features; Latent Dirichlet Allocation

I. INTRODUCTION

With the dramatically increasing of online reviews, the review spam come along due to the fact that there is no control, anyone can write anything on the web [1]. The review that describe authentic post-purchase experience can help potential consume get a satisfactory commodity, business have its own accurate positioning. Instead, review spam misleads consume and business. Thus, detection of review spam has become increasingly urgent and important.

There are generally three types of spam reviews: Type 1: untruthful opinions (also known as fake reviews). Type 2: reviews on brands only，Type 3: Non-reviews [1]. In this paper, we aim to detect deceptive fake reviews by looking into deep-level semantics of reviews. Our goal is then to cast the deceptive fake review detection problem into binary classification task and build classification model. Using term frequency, LDA, word2vec to extract features, then we fed those kinds of features extracted from each review of our dataset into several Machine Learning models for classification and finally compare the performances of features in those Machine Learning models.

We perform our experiment in Yelp dataset. As Yelp.com is a well-known large-scale online review site that filters fake or suspicious reviews which can be used as fake reviews in our experiment. In the end, we demonstrate the validity of our method through experiments.

The rest of the paper is organized as follows: in Section 2 we summarize related work; in Section 3 we discuss our dataset, features and classifiers; we show the results and discussion in Section 4; finally, conclusion and future work are given in Section 5.

II. RELATED WORKS

There are many significant study on how to classify authentic and fictitious reviews. Ott Collected 800 deceptive opinions via Mechanical Turk and 800 truthful opinions from TripAdvisor, then Integrating work from psychology and computational linguistics, they develop and compare three approaches to detecting deceptive opinion spam, and ultimately develop an admirable classifier that is nearly 90% accurate on their gold-standard opinion spam dataset [2, 3]. It is admirable for Ott to open their gold-standard opinion spam dataset, which have make a great impact on the field of fake reviews detection. Nitin Jindal and Bing Liu deals with a restricted problem, identifying unusual review patterns which can represent suspicious behaviors of reviewers [4]. Snehasish Banerjee and Alton YK Chua extract linguistic features to distinguish fake reviews by word n-gram, psycholinguistic deception words, part-of-speech distributions, readability of reviews and review writing style [5]. Heydari focuses on systematically analyzing and categorizing models that detect review spam [6]. Michael Crawford mainly provide a strong and comprehensive comparative study of current research on detecting review spam using various machine learning techniques [7]. Interestingly, Lim, P and Liu, B proposed ranking and supervised methods to discover spammers and outperform other baseline method based on helpfulness votes along [8].

In terms of Yelp Filter Dataset, Mengqi Yu found sentiment features are very useful for rating prediction [9]. Dao Runa treated the fake review detection problem as binary classification task and built classification models by extracting semantic based features and relational based features with several data mining techniques [10]. Boya Yu use a Support Vector Machine model to decipher the sentiment tendency of each review from word frequency. Word scores generated from the SVM models are further processed into a polarity index indicating the significance of each word for special types of restaurant [11].

978-1-5386-6147-5/18/$31.00 ©2018 IEEE

Arjun Mukherjee used linguistic features and behavioral features to train classification model, the result of linguistic features are shown in Table I, which consists of hotel and restaurant. Because the low accuracy of linguistic features, the author paid attention to behavioral features, and yield a respectable accuracy in [12]. In this paper, we will further research linguistic features on Yelp Filter Dataset based on the experiment conducted by Arjun Mukherjee.

We will use linguistic features to train classification model based on Yelp Filter Dataset by conducting three classification model, using term frequency, LDA, word2vec to extract features. Then making contrastive study with experiments conducted by Arjun Mukherjee in linguistic features [12].

III. DESIGN

A. Data

Obtain gold standard dataset for detecting fake review is always a challenging problem. With deceptive and disguised characteristics, the fake reviews are hard to be identified just by looking at individual review text. In addition to large scale number of reviews online, manual labeling is hard for ground truth reviews.

We use the subset of Yelp dataset in [12]. Yelp.com is a well-known large-scale online review site that filters fake or suspicious reviews which can be used as fake reviews in our work without manual labeling. Our dataset contains 64195 reviews across 85 hotels and 130 restaurants in the Chicago area. For each review there are information about <date, reviewID, reviewerID, reviewContent, rating, usefulCount, coolCount, funnyCount, flagged, restaurantID> (in which flagged are shown as yes or no that represent the reviews are fake or non-fake). Dataset statistics are shown in Table II.

But the class distribution is extremely imbalanced. So the review processed basically will be abandoned that review's length below 25. Then selected all fake reviews and Non-fake reviews is twice as much as the fake reviews. New Dataset statistics are shown in Table III.

B. Feature Extraction

In order to classify the reviews either fake or non-fake, we needed a set of features that can distinguish them. In our work, we mainly use linguistic feature, which respectively aims to term frequency, Latent Dirichlet Allocation and word2vec, then merged into one model to conduct experiment, then selecting the best method.

Followings are the details of extracted features:

1) Term frequency

Term frequency can be used to measure the importance of a word in a paper. In this paper, we extract term frequency by using Scikit-learn, and adopt 5000 words of the highest term frequency.

2) Word2vec

Mikolov proposed word2vec, a model generates word embedding for semantic modeling [13].We train a skip-gram model on our corpus by gensim.models.word2vec.Word2Vec and represent every review as a vector by calculating the average of the embedding vectors of each word in the review. The vector size is set to 300, the window size is 5, the min-count is 2, and the iteration is 5.

3) Latent Topic Distribution [14]

Topic modeling is a technique in natural language processing and tries to extract hidden topics from a collection of documents. In our work, we treated fake and non-fake reviews as two document and used LDA model to extract topic-words. We decided the number of topics by minimizing the model's perplexity on held-out data, the result showed that the best number of topics chosen from 100, 150, 200, 250 and 300 for fake reviews was 150, non-fake reviews was 200, each topic contain 30 words, there will show respectively 5 topics and each topic contains 8 words in Table IV and Table V. Each topic-words is viewed as a new review, the LDA was implemented by Liu Yang [15], then integrating new reviews with old reviews. Finally we have a distorted data shown in Table VI.

TABLE I. THE RESULTS OF EXPERIMENT [12]

Features	P	R	F1	A		P	R	F1	A
	hotel					***restaurant***			
Word unigrams (WU)	62.9	76.6	68.9	65.6		64.3	76.3	69.7	66.9
WU + IG (top 1%)	61.7	76.4	68.4	64.4		64	75.9	69.4	66.2
WU + IG (top 2%)	62.4	76.7	68.8	64.9		64.1	76.1	69.5	66.5
Word-Bigrams (WB)	61.1	79.9	69.2	64.4		64.5	79.3	71.1	67.8
WB + LIWC	61.6	69.1	69.1	64.4		64.6	79.4	71	67.8
POS Unigrams	56	69.8	62.1	57.2		59.5	70.3	64.5	55.6
WB + POS Bigrams	63.2	73.4	67.9	64.6		65.1	72.4	68.6	68.1
WB + Deep Syntax	62.3	74.1	67.7	64.1		65.8	73.8	69.6	67.6
WB + POS Seq. Pat	63.4	74.5	68.5	64.5		66.2	74.2	69.9	67.7

TABLE II. DATASET STATISTIC

Domain	fake	Non-fake	%fake	total
Hotel and Restaurant	802	4876	14.1%	5678
	8368	50114	14.3%	58517

TABLE III. NEW DATASET STATISTIC

Domain	fake	Non-fake	%fake	total
Hotel and restaurant	4017	8034	1/3	12051

C. Classification and Evaluation

Features from the two approaches just introduced are used to train Support Vector Machine and Logistic Regression and Multi-layer Perceptron classifiers。

The classification results of above mentioned techniques are evaluated by accuracy, precision, recall and F-measure.

IV. RESULTS AND DISCUSSION

A. Results

In our experiment, we train SVM, Logistic Regression, and Multi-layer Perceptron models in Python 3.6.

We choose 80% of the dataset as training data and 20% as testing data. As we can see from the experimental results in Table VII.

Then compared the experimental results with the result in [12]. The result of compare will be showed in Fig. 1.

The difference between with LDA processing data and without LDA will be showed in Fig. 2.

Table VII shows that the results by using SVM yielded accuracy of 65.7%, LDA+ SVM yielded a maximum accuracy of 67.9%, which slightly lower than the 68.1% accuracy reported by Arjun Mukherjee on the Yelp Filter Dataset [12]. But LDA+ Logistic Regression yielded a maximum accuracy of 81.3%, which obviously higher than the 68.1% accuracy, and LDA+ Multi-layer Perceptron also yielded a maximum accuracy of 81.3%. The accuracy of LDA+ Logistic Regression keep up to LDA + Multi-layer Perceptron's, but the F1-score of LDA+ Logistic Regression slightly higher than the 71.1% F1-score of LDA+ Multi-layer Perceptron in fake reviews.

In terms of Fig. 1, the green line represent LDA+ Logistic Regression results in this paper, red and gray line respective represent the hotel and restaurant's best results in [12]. We can obviously notice the green line far higher than red and gray line, which indicates that method of this paper has a good effectiveness in binary classification task.

In terms of Fig. 2, red line represent the accuracy by using LDA, green line represent the accuracy without LDA. We can notice that the accuracy with LDA slightly higher than the accuracy without LDA, which indicate the effectiveness of LDA in this experiment.

B. Discussion

LDA can extract topic-words from one document, and to some extent, topic-words can represent whole document. Thus, we use LDA to respectively extract topic-words from fake reviews and non-fake reviews, it is more reflected the features of fake or non-fake reviews. Then when we counts the term frequency of each word, the import words to reflect the features of fake or non-fake reviews will have a higher term frequency, and then increase the accuracy of classification models. But due to the quantity of data is enormous, the quantity of topic-words is far less than that. Therefore, the accuracy with LDA slightly higher than the accuracy without LDA.

V. CONCLUSION AND FUTURE WORK

This paper performed a linguistic investigation of the nature fake reviews in the commercial setting of Yelp.com. Our study shows that linguistic features yielded a respectable 81.3% accuracy, which obviously higher than the 68.1% accuracy reported by Arjun Mukherjee on the Yelp Filter Dataset as far as linguistic features linguistic features [12]. Meanwhile, the study proved the effectiveness of features extracted based on LDA.

Possible directions for future work is to explore why Logistic Regression and Multi-layer Perceptron have a high accuracy, SVM in not. There is a hypothesis that sigmoid make a decisive influence, which will be testified in future work.

TABLE IV. TOPIC-WORDS OF FAKE REVIEWS

Topic1	Topic2	Topic3	Topic4	Topic5
promise	park	stones	writing	cube
quality	comments	discarded	reserve	parings
pushy	ramps	split	injure	shined
rationalize	edge	eavesdrop	damn	pomp
podium	cliff	strict	autographed	bamboo
decorated	spray	breadth	hate	heroin
peeled	shots	settle	zealand	absurd
gulped	care	swirling	olfactory	unsalted

TABLE V. TOPIC-WORDS OF NON-FAKE REVIEWS

Topic1	Topic2	Topic3	Topic4	Topic5
extremely	decadent	confirmed	prospect	collective
burnt	entertain	duke	eaten	smiled
vaguely	hiccup	warm	previous	cultural
arrives	successor	pour	night	mystery
content	troubles	laugh	dish	smothering
unstuck	mustards	transmogrify	completely	observing
twists	brighter	care	recognizable	kindle
redefining	responds	school	notable	tire

TABLE VI. EXPERIMENT DATA

Domain	fake	Non-fake	%fake	total
Hotel and restaurant	4167	8234	33.6%	12401

TABLE VII. EXPERIMENTAL RESULTS

Classifier	Approach	Accuracy (%)	Non-fake			Fake		
			P(%)	*R(%)*	*F(%)*	*P(%)*	*R(%)*	*F(%)*
SVM	SVM	65.7	65.7	100	79.3	1	2.6	5.1
	LDA+ SVM	67.9	67.6	99.9	81.6	96.7	3.5	6.8
	Word2Vec+SVM	61.3	65.5	100	79.2	\	0	\
	LDA+Word2Vec+SVM	61.3	65.7	100	79.3	\	0	\
Logistic Regression	Logistic Regression	80.5	83.9	86.8	85.3	73.6	70	71.8
	LDA+ Logistic Regression	**81.3**	85.2	87.2	**86.2**	73	69.5	**71.2**
	Word2Vec+Logistic Regression	65.1	65.1	1	78.9	\	0	\
	LDA+Word2Vec+Logistic Regression	65.8	66.8	1	80.1	\	0	\
Multi-layer Perceptron	Multi-layer Perceptron	80.3	84	86.2	85.1	72.9	69.4	71.1
	LDA+ Multi-layer Perceptron	**81.3**	85	87.4	**86.2**	73.2	69.1	**71.1**
	Word2Vec+Multi-layer Perceptron	65.5	65.5	1	79.2	\	0	\
	LDA+Word2Vec+Multi-layer Perceptron	65.7	65.7	1	79.3	\	0	\

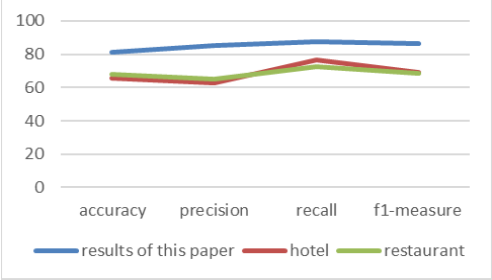

Figure 1. Results of this paper and A. Mukherjee's [12]

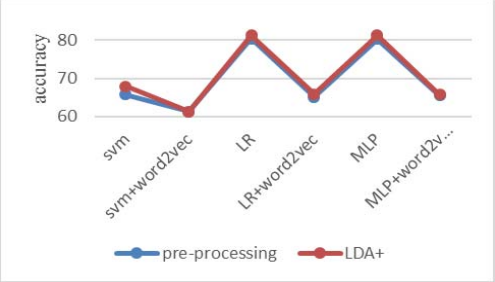

Figure 2. Results of with LDA and without LDA

REFERENCES

[1] Jindal, N., & Liu, B, "Opinion spam and analysis," In Proceedings of the international conference on web search and web data mining, ACM, 2008, pp.219-230.

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

[3] Myle Ott, Claire Cardie, Jeffrey T. Hancock "Negative Deceptive Opinion Spam," North American Chapter of the Association for Computational Linguistics: human language technologies (NAACL HLT 2013), June 9-14, 2013, pp. 497-501.

[4] Nitin Jindal , Bing Liu , Ee-Peng Lim, "Finding unusual review patterns using unexpected rules", Proceedings of the 19th ACM international conference on Information and knowledge management, October 26-30, 2010, pp. 1549-1552.

[5] Banerjee, Snehasish, and Alton YK Chua. "Applauses in hotel reviews: Genuine or deceptive?." Science and Information Conference (SAI), 2014. IEEE, 2014, pp. 938-942.

[6] Heydari, Atefeh & Tavakoli, Mohammadali & Salim, Naomie & Heydari, Zahra. "Detection of review spam: A survey." Expert Systems with Applications, 2015, pp. 3634-3642.

[7] Crawford, M.; Khoshgoftaar, T. M.; Prusa, J. D.; Richter, A. N. & Al Najada, H. Survey of review spam detection using machine learning techniques, J. Journal of Big Data, February 23, 2015.

[8] Lim, P., Nguyen, V., Jindal, N., Liu, B., & Lauw, H, "Detecting product review spammers using rating behaviors." In Proceedings of the 19th ACM international conference on Information and knowledge management. ACM. 2010, pp. 939-948

[9] Yu, Mengqi, Meng Xue and Wenjia Ouyang, "Restaurants Review Star Prediction for Yelp Dataset," 2015.

[10] DaoRuna; XianguoZhang; YongxinZhai.Try to Find Fake Reviews with Semantic and Relational Discovery, 2017 13th International Conference on Semantics, Knowledge and Grids (SKG), 2017, pp. 234 – 239.

[11] Yu, Boya; Zhou, Jiaxu; Zhang, Yi; Cao, Yunong, "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews," ARXIV, 2017.

[12] A. Mukherjee, V. Venkataraman, B. Liu, and N. S. Glance, "What yelp fake review filter might be doing?" Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013, pp. 409-418.

[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, "Distributed representations of words and phrases and their compositionality," J. In Advances in Neural Information Processing Systems, 2013, pp. 3111-3119.

[14] David M. Blei, Andrew Y. Ng, Michael I. Jordan "Latent Dirichlet Allocation," [J].Journal of machine learning research, 2003, pp. 993-1022.

[15] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun and Zhong Chen. "CQARank: Jointly Model Topics and Expertise in Community Question Answering." In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, 2013. pp. 99-108.