Detection of fake online hotel reviews

Anna V. Sandifer
Department of Mathematics &
Computer Science
The Citadel
Charleston, SC
asandife@citadel.edu

Casey Wilson
Department of Computer Science
College of Charleston
Charleston, SC
cawilson1@g.cofc.edu

Aspen Olmsted
Department of Computer Science
College of Charleston
Charleston, SC 29401
olmsteda@cofc.edu

Abstract— **Individuals use online reviews to make decisions about available products and services. In recent years, businesses and the research community have shown a great amount of interest in the identification of fake online reviews. Applying accurate algorithms to detect fake online reviews can protect individuals from spam and misinformation. We gathered filtered and unfiltered online reviews for several hotels in the Charleston area from yelp.com. We extracted part-of-speech features from the data set, applied three classification models, and compared accuracy results to related works.**

Keywords: fake review detection, machine learning, part-of-speech tags, Multinomial Naïve Bayes classifier, Bernoulli Naïve Bayes classifier, logistic regression classifier.

I. INTRODUCTION

There are thousands of reviews online, which makes it convenient for people to make decisions, but the amount of data makes it difficult to sort through [1]. The real value of online reviews is in its content and the certainty that reviewer indeed received products or services prior to writing the review. Promotion or demotion of the products and services is one of the main reasons for deceptive reviews. At times, to create better ratings for the venue, hotel owners pay employees to fabricate false reviews [2]. Alternatively, some reviewers write negative reviews for malicious reasons, like to distort the reputation of the business reviewed [3].

Yelp.com is one of the biggest online review sites. It uses a filtering algorithm to detect fake reviews. However, the algorithm is a trade secret. In this work, we collected reviews from yelp.com for 100 random hotels in the Charleston area. We labeled filtered reviews as real and unfiltered reviews as fake. We extracted part-of-speech features, trained and tested the data set, built a model and compared results to related work.

The rest of the paper is organized as follows: Section II reviews related work. Section III describes the motivation for our work. Section IV describes the data collection mechanism and analysis. Finally, Section V will provide a conclusion and future work.

II. RELATED WORK

Ott et al. [4] define deceptive opinion spam as “fictitious opinions that have been deliberately written to sound authentic in order to deceive the reader.” To analyze deceptive opinion spam, authors performed a hotel review analysis, using a data set of 800 reviews. Truthful reviews were gathered from tripadvisor.com, and fake reviews were written by Amazon Mechanical Turk online workers (known as Turkers). Each Turker was tasked to write a deceptive review on one of the 20 Chicago hotels. Reviews were written from the perspective of the hotel’s marketing department employees. They satisfied certain length and complexity criteria and portrayed the hotel in a positive light. The Turker got paid $1 per accepted review. There were 6977 truthful reviews for the same Chicago hotels that were mined from tripadvisor.com, but only 400 reviews were selected for research.

Two approaches were chosen to detect deceptive opinion spam: human performance, and automated approach. Three graduate students were selected to analyze the reviews. For an automated approach, authors used n-gram features as well as a combination of the n-gram features and psycholinguistically-motivated features. Accuracy results using human judges was reported at 61.9%, using bigram features 89.6%, and using psycholinguistic deception detection approach 76.8%.

Mucherjee et al. [5] used filtered and unfiltered data from yelp.com to analyze real and pseudo reviews. Yelp’s filter claimed to be very accurate with a few false positives. Authors used reviews from 85 hotels and 130 restaurants in the Chicago area. While analyzing reviews, authors noticed that the quantity of fake reviews on yelp.com is much smaller than real reviews. They used balanced data (50% fake and 50% real reviews) to train the data set and used natural distribution to test the set. Mucherjee et al. used POS (part-of-speech) based features to build classifiers. All experiments were based on 5-fold Cross Validation and yielded 67.8% accuracy.

McCallum and Nigam [6] performed an evaluation of two event models for the Naïve Bayes Text Classification. The first model was a multi-variate Bernoulli event model. This model calculates the probability of the document, by multiplying the probability of all attribute values, including the probability of non-occurrence for words that were not present in the document. The second model was the multinomial event model. This model calculates the probability of the document, by multiplying the probability of words that occurred in that document. Authors completed the empirical study with different data sets, such as the Newsgroups and WebKB. Results showed that the second model outperformed the first one in terms of large vocabulary sizes. The multinomial model reduced accuracy error by 27%.

978-1-908320/93/3/$31.00©2017 IEEE

III. MOTIVATING EXAMPLE

Product online reviews provide an information base for consumer experience. Online reviews appear on websites that sell products and offer services. Consumers are likely to look for information online before making purchase decisions. According to Fang et al. [1], 65% of leisure service consumers will read online reviews, before deciding on the vacation destination. The quality of individual hotels is measured through different means, such as stars or ratings on characteristics of cleanliness, view, and staff friendliness. However, online reviews provide support for those means. It is critical to be able to identify fake online reviews with great accuracy.

IV. IMPLEMENTATION

Our first step was to collect the data. We used Chrome Web Scraper to gather reviews from yelp.com for 100 random hotels in the Charleston area. Non–filtered reviews were tagged as real and filtered reviews were tagged as fake. Table I provides dataset statistics.

TABLE I. DATASET STATISTICS

	Hot els	Fake	Non-fake	% of fake	Total reviews
Our work	100	640	3310	16.2%	3950
Mucherjee et al.	85	4876	4876	14.1%	5678
Ott et al.	20	400	400	50%	800

Next, we tokenized both datasets, extracted adjectives, adverbs, and verbs and tagged words as real or fake. Most informative features and their probabilities are shown in Table II, where w is the word, P(w|R) is the probability of the word occurring in the real review, and P(w|R) is the probability of the word occurring in the fake review.

TABLE II. MOST INFORMATIVE FEATURES

w	P(w|R)	P(w|F)
assume	1.0	14.4
suggested	1.0	11.1
packing	1.0	10.3
tells	1.0	10.3
easier	1.0	8.7

We calculated a frequency distribution of extracted features. To train the data, we used half of the features, and the other half was used for testing. For model building, we used NLTK [7] and Scikit-Learn [8] modules. To identify fake online reviews, we used Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and logistic regression models. The best accuracy result was achieved by applying the Multinomial Naïve Bayes model and yielded 85.9% accuracy. Figure 1 compares accuracy results for our work and related work.

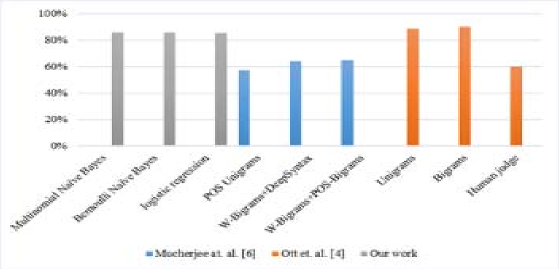

Figure 1. *Accuracy Result*

V. CONCLUSION AND FUTURE WORK

In this paper, we collected data, extracted part-of-speech features, and applied three different classification models to identify fake online reviews. We found that highest accuracy was achieved by applying the Multinomial Naïve Bayes classification model to our dataset.

In our future work, we plan to improve on our methods for extracting features from datasets. We also intend to use more classification models, develop voting algorithms, and introduce a reliability score.

REFERENCES

[1] B. Fang, Q. Ye, D. Kucukusta and R. Law, "Analysis of the perceived Value of online tourism reviews: Influence of readability and reviewer characteristics," *Tourism Management,* pp. 498-506, 2016.

[2] K. L. Short, "Buy My Vote: Online Reviews for Sale," *Vanderbilt Journal of Entertainment and Technology Law,* 2013.

[3] S. Banerjee and A. Y. Chua, "Understanding the process of writing fake online reviews," in *Ninth International Conference On Digital Information Management (ICDIM 2014)*, 2014.

[4] M. Ott, Y. Choi, C. Cardie and J. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination.," in *49th Annual Meeting of the Association for Computational Linguistics*, Portland, 2011.

[5] A. Mukherjee, V. Venkataraman, B. Liu and N. Glance, "Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews," Department of Computer Science (UIC-CS-2013-03), Chicago, 2013.

[6] A. Mccallum and K. Nigam, "A comparison of event models for Naive Bayes text classification.," 2010.

[7] "Natural Language Toolkit," [Online]. Available: http://www.nltk.org/. [Accessed 26 March 2017].

[8] "Scikit-learn Machine Learning in Python," [Online]. Available: http://scikit-learn.org/stable/. [Accessed 26 March 2017].