Detecting Deceptive Opinion Spam using Linguistics, Behavioral and Statistical Modeling

Arjun Mukherjee
Department of Computer Science
University of Houston
501 PGH, 4800 Calhoun Rd. Houston, TX
arjun@cs.uh.edu

1 Introduction

With the advent of Web 2.0, consumer reviews have become an important resource for public opinion that influence our decisions over an extremely wide spectrum of daily and professional activities: e.g., where to eat, where to stay, which products to purchase, which doctors to see, which books to read, which universities to attend, and so on. Positive/negative reviews directly translate to financial gains/losses for companies. This unfortunately gives strong incentives for *opinion spamming* which refers to illegal human activities (e.g., writing fake reviews and giving false ratings) that try to mislead customers by promoting/demoting certain entities (e.g., products and businesses). The problem has been widely reported in the news. Despite the recent research efforts on detection, the problem is far from solved. What is worse is that opinion spamming is widespread. While credit card fraud is as rare as 0.2%, based on our research we estimated that up to 30% of the reviews on many Web sites could be fake. Thus, detecting fake reviews and opinions is a pressing and also profound issue as it is critical to ensure the trustworthiness of the information on the web. Without detecting them, the social media could become a place full of lies, fakes, and deceptions, and completely useless.

Major review hosting sites and e-commerce vendors have already made some progress in detecting fake reviews. However, the task is still extremely challenging because it is very difficult to obtain large-scale ground truth samples of deceptive opinions for algorithm development and for evaluation, or to conduct large-scale domain expert evaluations. Further, in contrast to other kinds of spamming (e.g., Web and link spam, social/blog spam, email spam, etc.) opinion spam has a very unique flavor as it involves fluid sentiments of users and their evaluations. Thus, they require a very different treatment. Since our first paper in 2007 (Jindal and Liu, 2007) on the topic, our group and many other researchers have proposed several algorithms and bridged algorithmic methodologies from various scientific disciplines including computational linguistics (Ott et al., 2011), social and behavioral sciences (Jindal and Liu, 2008; Mukherjee et al., 2013a, b), machine learning, data mining and Bayesian statistics (Mukherjee et al., 2012; Fei et al., 2013; Mukherjee et al., 2013c; Li et al., 2014b; Li et al., 2014a) to solve the problem. The field of deceptive opinion spam has gained a lot of interest in communications (Hancock et al., 2008), psycholinguistics communities (Gokhman et al., 2012), and economic analysis (Wang, 2010) apart from mainstream NLP and Web mining as attested by publications in top tier venues in their respective communities. The problem has far reaching implications in various allied NLP topics including Lie Detection, Forensic Linguistics, Opinion Trust and Veracity Verification and Plagiarism Detection. However, owing to the inherent nature of the problem, a unique blend of NLP, data mining, machine learning, social, behavioral, and statistical techniques are required which many NLP researchers may not be familiar with.

In this tutorial, we aim to cover the problem in its full depth and width, covering diverse algorithms that have been developed over the past 7 years. The most attractive quality of these techniques is that many of them can be adapted for cross-domain and unsupervised settings. Some of the methods are even in use by startups and established companies. Our focus is on insight and understanding, using illustrations and intuitive deductions. The goal of the tutorial is to make the inner workings of these techniques transparent, intuitive and their results interpretable.

2 Content Overview

The first part of the tutorial presents the problem in its various flavors, the NLP techniques, and the algorithms motivated from social and behavioral sciences. It also presents a detailed insight into commercial vs. crowdsourced deceptive opinions using information theory and linguistics. The second section includes detailed math and algorithms for training supervised, unsupervised, semi-supervised, and partially supervised machine learning and statistical models for deceptive opinion spam

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 21–22,
Beijing, China, July 26-31, 2015.

detection. These algorithms allow us to work on unlabeled data which is a key aspect of the problem as generating high quality labels of fake reviews in large scale is hard if not impossible. We also discuss some new evaluation methods. Additionally, we draw connections to Authorship Attribution to discover fake reviewers with multiple accounts based on their writing styles, which is a new frontier in deceptive opinion spamming. The last part of the tutorial gives a general overview of the different applications of the methods in allied NLP problems and domains, data sources, and the limitations of the existing methods.

3 Tutorial Outline

I. Introduction
- a. The socio-economic value of opinions
- b. Deceptive Opinion Spam and Fraud
- c. Opinion Spam Types: Individual, Group, Singular, and Campaigns

II. Leveraging Linguistic Signals
- a. N-gram language models
- b. Psycholinguistics
- c. Stylometry

III. Leveraging Behavioral Signals
- a. Rating, Reviewing, & Collusion Behaviors
- b. Distributional and Time-Series Analysis
- c. Graph Based Methods
- d. Linguistic vs. Behavioral Features: A case study on Commercial vs. Crowdsourced Fake Reviews

IV. Machine Leaning & Statistical Modeling
- a. Supervised vs. Unsupervised Methods
- b. Positive and Unlabeled (PU) and Semi-Supervised Learning
- c. Latent Variable Models

V. The Next Frontier: Sockpuppets
- a. Authorship Attribution and Beyond
- b. Modeling Latent Spaces of Language
- c. Learning in Similarity Spaces

VI. Discussion and Resources
- a. Applications
- b. Data sources
- c. Evaluation
- d. Discussion

4 Instructor Biography

Arjun Mukherjee is an Assistant Professor in the Department of Computer Science at the University of Houston. He is an active researcher in the area of opinion spam, sentiment analysis and Web mining. He is the lead author behind several influential works on opinion spam research. These include group opinion spam, commercial fake review filters (e.g., Yelp), and various statistical models for detecting singular opinion spammers, burstiness patterns, and campaign. His work on opinion mining including deception detection have also received significant media attention (e.g., ACM Tech News, NYTimes, LATimes, Business Week, CNet, etc[1]). Mukherjee has also served as program committee members of WWW, ACL, EMNLP, and IJCNLP.

References

G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. 2013. Exploiting Burstiness in Reviews for Review Spammer Detection. *ICWSM*.

S. Gokhman, J. Hancock, P. Prabhu, M. Ott, and C. Cardie. 2012. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*.

J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth. 2008. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*.

N. Jindal and B. Liu. 2007. Review spam detection. *WWW*.

N. Jindal and B. Liu. 2008 Opinion Spam and Analysis. *WSDM*.

H. Li, B. Liu, A. Mukherjee, and J. Shao. 2014a. Spotting Fake Reviews using Positive-Unlabeled Learning. *Computaci{ó}n y Sistemas*, 18(3).

H. Li, A. Mukherjee, B. Liu, R. Kornfield, and S. Emery. 2014b. Detecting Campaign Promoters on Twitter using Markov Random Field. *ICDM*.

A. Mukherjee, V. Venkataraman. 2014. Opinion Spam Detection: An Unsupervised Approach using Generative Models. *UH-CS-TR-2014-07*.

A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance. 2013a. What Yelp Fake Review Filter might be Doing? *AAAI ICWSM*.

A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance. 2013b. Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews. *UIC-CS-2013-03*

A. Mukherjee, A. Kumar, B. Liu, J. Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013c. Spotting Opinion Spammers using Behavioral Footprints. *KDD*.

A. Mukherjee, B. Liu, and N. Glance. 2012. Spotting Fake Reviewer Groups in Consumer Reviews. *WWW*.

M. Ott, Y. Choi, C. Cardie, and J. T Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. *ACL*.

Z. Wang. 2010. Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews. *The B.E. Journal of Economic Analysis & Policy*, 10(1):1–34, January.

[1] http://www.cs.uic.edu /~liub/FBS/media-coverage.html