E6893 Big Data Analytics: Yelp Fake Review Detection

Mo Zhou, Chen Wen, Dhruv Kuchhal, Duo Chen

Columbia University in the City of New York

December 11th, 2014

Overview

1. Problem Summary
2. Technical Approach
3. Result and Discussion
4. Future Work and Conclusion

Introduction

- The goal of this project is to flag opinion spammers on rating websites such as Yelp.com. These users are usually receiving incentives to post positive reviews for certain businesses.
- Opinion spam creates unfair competitions, provides deceptive information for users and is detrimental to the credibility of rating websites.
- We use dataset acquired from Yelp Dataset Challenge, which provides >1 million ratings and reviews of >40,000 businesses.

Background

- Content Based Detection
  - Analyze the reviews and detect spams using computational linguistics analysis
  - Characteristics like not having been to the place and still being able to write the review differentiates it from an original review, with an accuracy of approximately 68%
- Behavioral Detection
  - Maximum number of reviews submitted by a user based on the fact that an average user does not submit more than a few reviews in a day. Percentage of positive review used to identify false accounts that submit only false reviews

Related Work

- Lin et al. (2014) Finding Valuable Yelp Comments by Personality, Content, Geo, and Anomaly Analysis
  - Used to reorder the reviews to improve visibility to the user
  - Utilized both the content and the behavior of the reviewers for the detection
  - Used Natural Language processing and Personality Analyser tool, with an accuracy of approximately 80%
- Mukherjee et al. (2012) Spotting Fake Reviewer Groups in Consumer Reviews
  - Used behavioral model among fake reviewers and relational model between different groups, individual reviewers and products they reviewed
  - Utilized GS Rank (Group Spam Rank) algorithm, a relation based algorithm to rank the spamming groups

Strategy Overview

- Opinion spam signals
  - Large deviation in ratings
  - Large deviation in text reviews
  - Interdependent feature extraction
- Our approach to detect opinion spam
  - Compute average rating similarity score between one user and all others
  - Compute average review similarity score between one user and all others
  - Cluster users with both low rating similarity score and review similarity score
  - Generate a list of opinion spam candidates

Algorithm - Compute Rating Similarity

- Let $r_1, r_2, \ldots, r_n$ be the rating of user n of the same business, with a maximum rating of 5, and let N be the number of users, then

$$rSim_n = 1 - \frac{\sum_{i=0,k=0,i \neq k}^{N} (r_i - r_k)/5}{N} \in [0, 1]$$

denotes the average rating similarity score between user n and all other users, limited for the same business

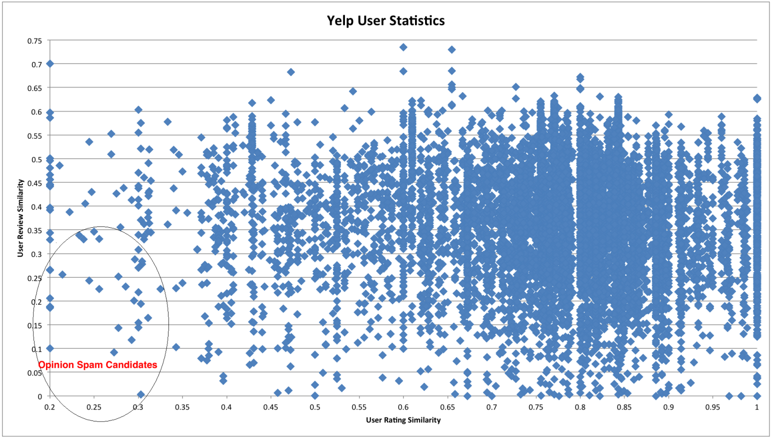
Algorithm - Compute Review Similarity

- Let $c_1, c_2, \ldots, c_n$ be the cosine similarity between review posted by user n and all other users of the same business, and let N be the number of users, then

$$cSim_n = \frac{\sum_{i=0,k=0,i\neq k}^{N} c_i \times c_k}{\sqrt{\sum_{i=0}^{N} c_i^2} \times \sqrt{\sum_{k=0}^{N} c_k^2}} \in [0,1]$$

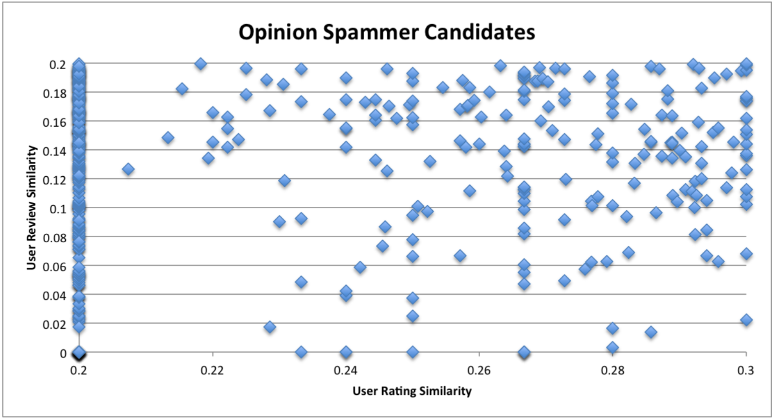
denotes the average review similarity score between user n and all other users of the same business

Result Overview



- The majority of users have high rating similarity score and decent review similarity score.
- The deviants at bottom left corner are users of interest.

Opinion Spam Candidates



- Cut-off score
  - $rSim = 0.2$
  - $cSim = 0.3$

Discussion

- The location of data points sustantiates our hypothesis that most ratings and reviews are genuine.
- Yelp.com has its own filter which may result in the lack of the true negative. Nevertheless, we still obtain several hundred potential fake review candidates.
- The threshold setting for *rSim* and *cSim* is worth further studies.
    - *cSim* seems to have larger impact on the process of determining opinion spam.
    - *rSim* is served as simply a filtering process.

Future Work

- Gather ground truth for review data
    - Domain expert
    - MTurk
- Incorporate more features into spam signals
    - Time window
    - User relationships
- Compare our approach with other learning Algorithm
    - GSRank
    - Supervised Classification

References

- Lin et al. (2014) Finding Valuable Yelp Comments by Personality, Content, Geo, and Anomaly Analysis
- Mukherjee et al. (2012) Spotting Fake Reviewer Groups in Consumer Reviews
- Dot products - the Stanford NLP, http://nlp.stanford.edu/IR-book/html/htmledition/dot-products-1.html
- Yelp Dataset Challenge, http://www.yelp.com/dataset_challenge