2018

SENTIMENT FEATURES FOR YELP NOT-RECOMMENDED ONLINE REVIEWS STUDY

Na Li
University of Rhode Island, nali.luck@gmail.com

SENTIMENT FEATURES FOR YELP NOT-RECOMMENDED ONLINE REVIEWS STUDY

BY

NA LI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER SCIENCE

UNIVERSITY OF RHODE ISLAND

2018

MASTER OF SCIENCE THESIS

OF

NA LI

APPROVED:

Thesis Committee:

Major Professor Lisa DiPippo

Yan Sun

Joan Peckham

Haibo He

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2018

ABSTRACT

Nowadays, online reviews have become an important source of opinions that people refer to while making decisions. For instance, there are more and more people who refer to Yelp reviews to judge the quality of services that are provided by local businesses. Due to the popularity and guidance of online reviews, many reviews have been imposed for the purpose of either promoting or downgrading target services. Yelp develops its own automatic review recommendation algorithm, which has marked many suspicious reviews as *Not-recommended Reviews*. Yelp has automatically grouped its online reviews in two different categories, and it is a common question "What are the differences between Not-recommended Reviews and Recommended Reviews?". One of the goals in this thesis is to explore the differences. Particularly, it employs the *Text*, one of the most important components of an online review, to develop six different sentiment features, i.e., *Strong Positive*, *Strong Negative*, *Ordinary Positive*, *Ordinary Negative*, *Ordinary*, and *Strong*, and study the differences in terms of sentiment between recommended reviews and not-recommended reviews. It has been found that not-recommended reviews usually contain more polarized (positive or negative) words.

In addition, online reviews are posed for services and products randomly. Generally, the reviews for a service/product are evenly distributed in their lifespan. However, it has been reported in the Amazon system that there are time periods where the reviews for some products are bursty. Put in other words, there are sudden concentrations of reviews in certain time periods. Another goal in this thesis is to investigate review bursts on Yelp. First, it is to explore the *Date* component of a review to develop the Density of Burstiness for the reviews of a business. Second, the normalized burstiness density has been introduced to select *Density Periods*, where

reviews are mostly concentrated. It has been found that Yelp reviews have the following concentration observations, (1) the maximum burstiness density values for density periods vary significantly; (2) the review bursts often occur at the beginning days of the reviews' lifespan; (3) some restaurants have multiple density periods.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my major advisor Lisa DiPippo, from Dept. of Computer Science and Statistics, and co-advisor Yan (Lindsay) Sun, from Dept. of Electrical, Computer and Biomedical Engineering. Their continuous support of my Master Study and related research helped me in all the time of research and writing of this thesis. I really received numerous precious comments and suggestions from both professors. Without helping from my advisors, my research of the thesis would not be possible.

Besides my advisors, I would like to thank the rest of my thesis committees: Prof. Joan Peckham from Dept. of Computer Science and Statistics, Prof. Haibo He from Dept. of Electrical, Computer and Biomedical Engineering, and Prof. Lenore M. Martin from Dept. of Cell & Molecular Biology, for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

Special thanks to a Dr. Yongbo Zeng in Network Security and Trust Laboratory for his support, helping me start my research work.

My sincere thanks also go to many professors and staffs in my department. Especially to Lorraine Berube, Beth Larimer, and Dr. Kevin Bryan. Lorraine gave concrete explanation of the department policy. Beth helped me to set up my office. Kevin provided me department resource for my projects and my research.

Last but not the least, I must express very profound gratitude to my parents and my husband Yihai Zhu for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching. I

also would like to thank to my brother and sister for supporting me spiritually throughout writing this thesis and my life in general. Thanks to my friends for their company and encouragements.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction of Online Reviews

Due to the e-commerce boom in the last decades, reviews posted online have become influential in decision-making. Since online reviews can provide broad and diverse information on target services or products, people rely on online reviews to leverage the experiences of others. A 2014 survey by BrightLocal reports that 88 percent of consumers check online surveys before buying online, and 88 percent also trust reviews as much as personal recommendations (Local Consumer Review Survey 2014). Many opportunities exist on the web today, such as shopping on Amazon.com, looking for a restaurant from Yelp.com and booking a hotel on TripAdvisor.com. However, reviews make it possible to promote or demote products or services, for instance, Amazon sellers can boost their business (Jindal and Liu 2008) by spamming positive reviews, hotels can promote their hotel ratings to receive more financial benefits (Mayzlin and Chevalier 2014), restaurant owners can pay Yelp to remove negative reviews and fraud customers (Schwarz 2014). These activities, creating false, misleading or inauthentic feedback about products or services in an attempt to gain unfair advantages, are called Review Manipulation according to Amazon.com. These reviews are considered fake reviews, and people who manipulate reviews are known as spammers.

1.2 Manipulation of Online Reviews

Businesses are incenting actual reviews with giveaways of cash and prizes in exchange for feedback and social media posts. Unfortunately, these businesses are also

buying fake reviews. If a website has customer feedback functionality, it is nearly certain that it also has online review manipulation. There are more than 20% fake reviews on Yelp (D'Onfro 2013), for example, and the number of fake reviews rose to 20% in 2013 from only 5% in 2006(Luca 2016). The mounting evidence shows that fake reviews have a direct influence on product sales in (Chevalier and Mayzlin 2006, Luca 2011). Business owners might intend to attract more customers by paying someone to write good reviews or to defame their competitors by leaving bad reviews. Fake reviews spread dishonesty, stifle competition, and can cost consumers time, money and trust. Businesses with the money to spend on fake reviews can outspend smaller companies focused on their product, preventing them from achieving market share. Genuine reviews can help to moderate bad business behavior and can improve the quality in the marketplace. For these reasons, it has even been made illegal in some places to post fake reviews (Malbon 2013). Therefore, it is important to develop approaches to distinguish genuine reviews from fake ones. Detecting fake reviews is not a straightforward problem to solve, but rather complex and difficult (Malbon 2013) since major challenge is to obtain large ground truth data to validate the proposed approaches. Because of the difficulty and broad impact, detecting fake reviews has attracted considerable attention from both academy and industry.

1.3 Detections of Fake Reviews

1.3.1 Features of Reviews

In the study of the detection of fake reviews, extracting and analyzing features from reviews, reviewers and products are the most common techniques. There are three types of features (Jindal and Liu 2008):

- review features: the characteristic of reviews that could be text content and metadata known as review's length, time-stamp, rating, review ID.
- reviewer features: the profile information of customer who posted the review for instance location, reviewer ID and others. It also includes reviewers' metadata which could be the percentage of positive reviews written, maximum number of reviews, review length, posted date and time, and so on.
- product features: which are made of the information about a product such as brand name, color and so on.

1.3.2 Text Mining

It is straight forward to obtain metadata features while the challenge is to extract features from the text content of reviews since text mining and Natural Language Processing(NLP) are needed. In the literature, the major approaches used in extracting reviews' text features are 1) Bag of Words, presenting the frequency of individual or groups of words of text. 2)Part of Speech (POS), assigning parts of speech to each word (and other tokens), such as noun, verb, adjective, and counting its frequencies as features[1]. 3) Linguistic Inquiry and Word Count(LIWC), counting the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech and scoring the keywords into 90 psychologically meaningful dimensions 2.

1.3.3 The Basic Assumptions

It is difficult to manually label fake reviews by reading large amounts of reviews and so there is no publicly available, labelled data set of reviews that can be used as

[1] https://nlp.stanford.edu/software/tagger.shtml
[2] https://liwc.wpengine.com/

verified truth. Therefore, several assumptions are made in order to enable detecting fake reviews. The first assumption is that fake reviews will have some textual pattern that is different than authentic reviews since spammers may use a particular type of language or use same review language repeatedly which may cause reviews to have similar words, while a genuine reviewer will leave thoughtful, detailed reviews about the products. For example, (Jindal and Liu 2008) treat duplicate and near-duplicate reviews as fake reviews. The second assumption is abnormal behaviors of reviewers or abnormal rating distribution. Spammers who leave fake reviews have some pattern of behavior or aspect of their profiles that are different than those who leave authentic reviews. Spammers may try to maximize their impact working together as groups to target a specific product. Reviews that arrive on a website in a burst are often considered to be left by spammers, since the impact of the trend will maximize the promotion or demotion of a product in a short time.

Based on these assumptions, the most established literature on the topic of review spam focuses on building models that leverage observable review characteristics, such as textual features, metadata, and reviewers' profile, to identify abnormal reviewing patterns or construct a network about reviews, reviewers or products(stores) to capture correlation among them. For example, textual and metadata features are applied in machine learning methods to train a classifier in (Jindal and Liu 2008, Li F 2011, Li H 2011, Ott 2011, Jindal 2010, Shojaee 2013, Mukherjee 2012, Hammad 2015, Feng 2012). Behavior approach (Mayzlin 2014, Jindal 2007, Mukherjee 2012, Xu 2013, Ye 2015, Li H 2011, Lim 2010, Xie 2012) focuses on the profiles and activity of the reviewers. Abnormal patterns of behavior are defined, and these definitions used to

flag those reviewers who are most likely to leave fake reviews. The network-based dection approach models the fake reviewer problem as a collective classification task on a network (Akoglu 2013, Fei 2013, Li H 2014, Jiang 2014, Ye 2015, Sen 2008, Rayana 2015, Xu 2013). To build the network, the reviews, reviewers, and also product(stores) information is used to create the network graph in the first place. (Fei 2013, Xie 2012, Gu¨nnemann N 2014, Gu¨nnemann S 2014, Hooi 2016) use statistical analysis of time-series and distribution of bursty reviews to identify potentially fake reviews.

1.3.4 Dataset

As mentioned before, another main challenge of this topic is to obtain the dataset, therefore, collecting a reviews dataset from consumer review websites such as Amazon, Yelp and TripAdvisor, and constructing a labeled dataset that can be used as training input for fake review classifiers is also essential related work focused on in the literature. (Jindal and Liu 2008) crawl 5.8 million reviews, 2.14 million and 6.7 million products from amazon.com. They manually label 470 fake reviews as training data. The later work (Jindal and Liu 2010, Lim 2010, Fei 2013) also use the same dataset. However, asking users or domain experts to label deceptive reviews may not reflect the real world online review because of human bias and small dataset constructed. (Ott et al. 2011) work shows that the accuracy of human labelling of fake reviews performs poorly. Therefore, (Ott et al. 2011) manufacture 400 fake reviews about hotels by hiring users on Amazon Mechanical Turk(AMT)- an online labor market – to write fake reviews following a carefully designed procedure. Additionally, 400 truthful reviews are collected from the TripAdvisor by manually labelling at the

same hotels. This dataset is fully balanced with 400 fake positive review and 400 truthful positive reviews. This dataset is also used in (Ott 2012, Li 2014).

In the later work, (Ott 2014) extend the dataset by adding similarly balanced and almost same size but negative reviews from another domain, restaurants and doctors. This method is also hard to scale since it is a costly and slow procedure. Moreover, the labelled data obtained by following a strict procedure may not be considered to have the same features as real online reviews.

(Mukherjee 2013a) study Yelp's filtering reviews across 85 hotels and 130 restaurants in the Chicago area. This dataset is also studied in (Rayana 2015). In this work, they collect two more datasets containing reviews of restaurants located in NYC and NJ, VT, and PA from Yelp. They treat filtered reviews as fake ones. However, the Yelp algorithm is not public information and it filters fake reviews from the Yelp main page as un-recommended. (Feng 2012) crawls 839,442 reviews from 4,000 hotels over 4 years from TripAdvisor. They evaluate 42,766 reviewers as trustworthy members based on historic rating distribution and label fake reviews based on relevant statistics. (Hammad 2013) has similar strategy to flag out fake reviews. They apply text mining methods on reviews collected from TripAdvisor, booking.com and agoda.ae and get text corpus with distinct tokens. With these tokens, they combine data mining and text mining methods to gain spam features. However, these suspicious features could be easily blinded in genuine reviews by spammers with sufficient to domain knowledge generate convincing fake reviews.

To summarize, we have seen features, detection techniques and dataset description and collection from the main contributors to the literatures. More details follow in the formal literature review found in chapter 2.

1.4 Research Objectives

Existing works share a common limitation that is to mainly consider the numeric parts of a review, such as ratings and times to design the system/algorithm to detect the manipulation, and ignore the reviewer opinion of rich textual contents of online reviews. Textual contents are the key components of online reviews, which are rich diverse users' feedback. Review contexts have been adopted to conduct review analysis research in different aspects, such as general online review spam (Jindal and Liu 2008), readability and sentiment analysis of online review manipulation (Hu 2012), and authentic versus fictitious online reviews analysis for different types of hotels (Banerjee 2017). The rich information in review text will help to develop advanced algorithms to detect diverse manipulation behaviors that are hidden in online review systems, which is potentially helpful to build robust and healthy online review systems, and bring benefits to many online users/customers.

There are following study focuses in this thesis:

- Develop different sentiment features from the *Text* component of reviews, and adopt those developed sentiment features to study the differences the terms of sentiment between not-recommended reviews and recommended reviews.
- Develop the Density of Burstiness form the Date component of reviews, and further develop the Density Periods to study the review bursts.

- Adopt the Yelp review dataset for verification studies. It has been found that not-recommended reviews contain more polarized (positive or negative) words, and there are apparent review bursts on Yelp system.

The goals are, therefor:

- To study how sentiment features of text reviews correspond to "recommended" and "not-recommended" Yelp reviews.
- To study how the density of burstiness differs for restaurants.

1.5 Structure of the Thesis

This thesis will be organized as follows. Chapter 2 provides the background of my research and literature review of detection of fake reviews. A detailed research methodology for Yelp review data set is described in Chapter3. Chapter 4 introduces the Yelp dataset adopted and extracted sentiment features in this thesis. Chapter 5 defines the formula of the burstiness reviews from Yelp, followed by experiment discussions in Chapter 6. Chapter 7 discusses the conclusions and future work.

CHAPTER 2

REVIEW OF LITERATURE

In this Chapter, we will present the literature review to fake review detection. As the following indicates, which existing approaches offer a significant contribution to solve this problem, identifying fake reviews is still a challenging task because there is no readily observable way to determine that a review is fake. The following section will present proposed approaches that challenge the spam detection.

2.1 Categories of the detection of fake reviews

To the best of our knowledge, (Jindal and Liu, 2008) were the first to study deceptive opinion spam and they categorized spam reviews into three types:(1) untruthful opinions (2) reviews on brands only and (3) non-reviews. This paper was mostly concerned with the first category. Since then, various approaches were explored and mainly have been categorized into four groups:

- train a classifier using features extracted from reviews.
- find abnormal behaviors from individual reviewer or reviewer group.
- construct a heterogeneous network of reviews, reviewers and products.
- exploit time-series and distributional of "bursty" of reviews.

2.2 Detection from language

Existing approaches in this category have focused on supervised machine learning techniques that classify reviews as two classes: fake and un-fake by leveraging review characteristics such as textual features or meta data. (Jindal and Liu, 2008) started using 2-gram to identify duplicate and near-duplicate reviews as untruthful reviews based on a review dataset crawled from Amazon. However, fake reviews may

not be limited in duplicate reviews, they extracted additional review meta data features and manually labeled spam reviews as training data. Finally, a logistic regression model as a classier was built to detect spam reviews in general. Evaluation was done by using Area Under the reviewer operating characteristic Curve(AUC)[3]. They found that using multiple features yielded a better performance since AUC sore was 0.78 when using all features while the score was 0.6 only when using text features.

(Ott 2011 and Li J 2014) achieved a better detection by using linguistic and psychology features driven from LIWC and POS than text features driven from Bag of Words (unigram, bigram). In the (Li J 2014) study, they drew a conclusion that using multiple features (e.g., LIWC and POS) yields better performance than a single feature. (Ott 2011) obtained fake reviews dataset by paying online users to write fake hotel reviews as mentioned in Chaptter 1 rather than manually label fake reviews. In their work, they used Naïve Bayes and SVM as classifiers and the best model was SVM to get an accuracy of 89.8% by combining bigram and LIWC features.

Another text feature, content similarity, has been a strong indicator were common used to detect spammers. (Jindal 2010, Li 2011, Mukherjee 2012, Fei 2013, Hammad 2013) used Bag of Words (unigram or bigram) to check content similarity. They also combined other features to achieve a better performance. One of most important observations in (Fei 2013) is that the features extracted from synthetic fake reviews using n-gram may not get a good result since synthetic data cannot be represented in the real-world fake reviews.

[3] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

(Hammad 2013) detected spam in Arabic reviews by using methods used in detecting English spam, which demonstrated that those methods can be extended to another language. They believed that reviews gathered online are imbalanced and fake and un-fake reviews have different size, which makes it more difficult to identify spam reviews since classifiers may be biased towards the majority class. A novel approach was proposed by using lexical and syntactic features to detect review spam in (Shojaee 2013). Those features may give an indicative information which reflect the text style of spammer. Lexical features, for example, reflect the types of words and characters such as average word length and syntactic features which represent writing style such as the traditional parts of speech (e.g. noun, verb, preposition, etc.) "he", "the". In this work, the comparison works were done by using either lexical and syntactic features alone or using both features. The dataset gathered in (Ott 2011) was used in SVM and Naïve Bayes classifiers. The SVM with both features achieved the highest F-measure of 84%.

2.3 Detection from Behavior

The approaches in this category have focused on analyzing the reviewer behavior to identify individual spammers or groups of spammers. Identifying spammers or the groups is a more effective method since they may have similar profile characteristics and abnormal behavioral patterns. (Mukherjee 2013b) confirmed that spammers have different behavioral patterns than truthful reviewers based on the study in Amazon dataset. Moreover, it is easier to collect behavioral evidence than detect fake reviews (Lim 2010). Most literatures leveraged suspicious behaviors such as spammers who

manipulate multiple reviewer IDs, and tend to exaggerate sentiment, along with reviewer burstiness that reviewer's and product's bursts happened in the same time.

(Jindal 2010) studied the impact of patterns associated with rating and brand distribution of a user's reviews using Class Association Rules to find unexpected rules and rule groups which tell the identification of spammer activities. They also claim that this technique can be applied to solve a variety of problems because of domain independence. The experiments worked on the same dataset crawled in (Jindal and Liu 2008) with the category of manufactured products. (Li 2011) trained a two-view semi-supervised model by employing a co-training framework to spot fake reviews. This method assigned labels to unlabeled data using a set of labeled data, which made large datasets labeled available for classification. Two behavior features, authority score and brand deviation score in combination with other features, content, sentiment, product and meta data features including review rating, average rating and post time have been studied.

The approaches above are to identify individual reviewers, but a group of spammers can damage a target product severely in the real word since the group can write many reviews in a short time and they are harder to catch than a single reviewer. Detecting the group can be seen in (Mukherjee 2012). They proposed GSRank model to identify unusual review patterns and reviewer behaviors that were correlated with spammer activities based on a (pseudo) ground truth dataset collected in (Jindal and Liu 2008). Group behaviors that may be indicators of spammers, examples of group time windows, group deviation, group content similarity, group member content

similarity, have been defined. They confirmed that spammers have different behaviors than truthful reviewers.

Rating score, an important behavioral feature left by reviewers, has been analyzed in (Lim 2010, Liu 2011, Feng 2012, Zeng 2015). (Liu 2011) detected the manipulated product with inconsistent rating score with time. They argued that the dishonest ratings must cause large enough change in the average rating if spammers want to conduct an effective manipulation. (Zeng 2015) proposed an Equal Rating Opportunity (ERO) Principle with a small dataset to find the manipulated product by arguing that ratings should be primarily because of the quality of the product or service rather than posted time, weekdays or weekend, the number of review content, long or short. (Xie 2012) proposed to detect spammers who only wrote one or few reviews based on the study in the Amazon review dataset (Jindal and Liu 2008) that singleton reviews were from 68% of the reviewers and 90% in their dataset. They spotted those spammers by monitoring their temporal behavior, average rating, review count, and ratio of singleton reviewers.

2.4 Detection from Network

The approach in this category is to construct a review graph to capture the synergy gained from looking at a more complete picture of reviews, reviewers and stores (products). (Wang 2011) built a graph model with three types of nodes representing user, review, and product and introduced trustiness of users, honesty of reviews, and reliability of products assigning scores to find the clue of spam reviews based on the identification of their interrelationships. (Akoglu 2013) proposed a signed bipartite network of users and products based on Markov Random Field (MRF) models. Once

the network was constructed, correlations were used to determine the most likely labelling of each node or edges as Real or Fake, Honest or Fraud and Good or Bad for reviews, reviewers and products respectively. MRFs were also applied in (Fei 2013) to capture spammers in burst based on assumption that spammers write fake reviews for profit and meanwhile can write genuine reviews as a normal customer.

In (Xu 2013) MRF used to label reviewers based on the relation between users and their attributes. (Li 2014a) constructed a reviewer, review and IP address graph inspired by positive and unlabeled examples (PU learning) in (Hernández 2013). They conducted several experiments on the Dianping[4] dataset with label, illustrating that detecting a large number of potential fake reviews hidden can be solved by combining collective classification and PU learning. Finally, (Jiang 2014) and (Ye and Akoglu 2015) have shown promising results that group spammers also can be identified by graph-theory based methods based on their abnormal network footprints. Well-known relational classifiers Loopy Belief Propagation (LBP) (Yedidia 2003) or Iterative Classification Algorithm (ICA) (Sen 2008) are commonly used in fake review or reviewer detection problems in the literatures (Fei et al., 2013; Akoglu et al., 2013; Rayana and Akoglu, 2015, Li et al., 2014a; Xu et al., 2013).

2.5 Detection from Time Series

Approaches in this category exploited the "bursty" nature of reviews by analyzing the pattern of rating distribution to identify review spam. Products that received a larger amount of reviews than usual within a certain time can be due to either the products suddenly becoming popular or they are under attack by spammers or

[4] http://www.dianping.com

spammer groups. (Fei 2013) detected spammers for bursts detection using Kernel Density Estimation (KDE)[5] which is a non-parametric way to estimate the probability density function of a random variable. The properties of smoothness and continuity are desirable for review burst detection. Markov Random Field (MRF) was applied to learn reviewers and their co-occurrence and the Loopy Belief Propagation (LBP) was employed to infer a reviewer is a spammer or not in the graph. They argued that behavioral features in combination with the features of review bursts improve the classification results. (Xie 2012) built up window size to find review burst based on time-series of a single retailer including daily number of reviews, average rating, and ratio of singleton reviews. (Günnemann 2014a; Günnemann 2014b; Hooi 2015) applied Bayesian approaches to detect anomalies in rating time-series. The details of datasets, feature extraction methods are summarized in Table 1.

Table 1. Summary of datasets and feature extraction methods in the literature.

Reference Paper	**Dataset**	**Domain**	**Machine Learning/ Learner**	**Features**	**Category**
Jindal 2008, Fei 2013, Mukherjee 2012,	5.8 million reviews, 2.14reviwers, 6.7million products (Aamzon.com)	Books, Music, DVDs, Manufactured products	Supervised/ Logistic Regression (LB)	Review Reviewers and products	Text, Behavior
Mukherjee 2013a	985,765 reviews, 50,704 reviewers, 112,055products (Aamzon.com)	Manufactured Products	Unsupervised Learning / Author Spamicity Model (ASM) works in Bayesian	Reviews: Duplicate/Near Duplicate Reviews, Extreme Rating, Early Time Frame, etc., Reviewers: content similarity, Maximum	Behavior

[5] https://en.wikipedia.org/wiki/Kernel_density_estimation

				Number of Reviews, etc.,	
Li F 2011	60k reviews, (Epinion.com)	N/A	semi-supervised, co-training/ Naïve Bayes (NB)	Review: unigram, bigrams, cosine similarity Reviewers: authority score, brand deviation score, etc., Product: product description, etc.,	Text, Behavior
Ott 2011	400 truthful reviews, 400 deceptive reviews by Amazon Mechanical Turk (AMT) (TripAdvisor .com)	Hotels	Supervised/ Support Vector Machine(SVM)	Review: POS, LIWC, Unigrams, Bigrams, Trigrams	Text
Li J, 2014	Borrowed from Ott 2011(hotel), 720 deceptive reviews by AMT, customer, expert(restaurant) 432 deceptive reviews by AMT, customer, expert(doctor) matching a set of truthful reviews (TripAdvisor .com)	Hotels, Restaurants, Doctors	NA/Sparse Additive Generative Model(SAGM)	Review: LIWC, POS, Unigram	Text
Rayana 2015	Yelp Chi, YelpNYC, YelpZip (Yelp.com)	Hotels, Restaurants	Unsupervised/ FraudEagle	Reviews: behavior，text Reviewers & Product: behavior，text	Network
Li H 2014	9,765 reviews, 9,067 reviewers, 5,535 IP (Dianping.cn)	Restaurants	Semi-supervised/ Collective Positive and Unlabeled learning (CPU)	Reviews, Reviewers, IP	Network
Hammad	2,848 reviews	Hotels,	Supervised/	Reviews:	Text

2015	(TripAdvisor.com. eg, booking.com, agoda.ae)	Books	/NB, SVM, ID3, K-NN with K=3	rate,date, isHelpful, etc. Reviewers name, age, location,etc.	
Wang 2011	408,470 reviews, 343,603 reviewers, 14561 stores (Resellerratings .com)	Snapshots	NA/ Define reviewer's trustiness, a store's reliability, and a review's honesty	Review Reviewers Products	Network
Shojaee 2013	Borrowed from Ott 2011	Hotels		Review: Stylometric	Text
Lau 2011	2,318,989 reviews (Amazon.com)	Automotive, Beauty, Grocery, Cameras, Computers, Books, DVDs, Music, Software	NA/SLM	Review: Syntactical, lexical, and stylistic	Text
Fei 2013	Borrowed from Jindal 2008	Books, Music, DVDs, Manufactured products	Supervised/K-NN, Markov Random Field (MRF), Loopy Belief Propagation (LBP)	Reviews: content similarity, meta data features, etc. Reviewers: behavioral features, etc.	Burst
Akoglu 2013	1, 132, 373 reviews 966, 842 reviewers, 15, 094 apps (SWM dataset)	Games, Movies, News, Sports	unsupervised/ Loopy Belief Propagation (LBP)	Reviews，Reviewers，Products	Network
Günnemann 2014a	400k reviews, (Aamazon.com) 230k reviews (Yelp.com) 250k reviews (TripAdvisor. com)	Food, Restaurants, Hotels	NA/Robust Latent Autor-egression (RLA)	Reviews, Reviewers	Burst

2.6 Summary

Chapter 2 presented an overview of approaches that have been proposed in the

review spam domain. Textural features (e.g., LIWS, POS tags) are often extracted to be used for training detection classifiers. In order to develop robust classifiers, researchers tried to extract other features related to the metadata of reviews or the behavior of users as building models. Experiments have shown that combining multiple features yields a high performance compared to using single feature (Mukherjee 2013, Shojaee 2013). Additionally, multiple features give more directions to detect fake reviews. However, different type of features was selected and used in all the current research but few studied what type of features can achieve a better performance and how many features should be selected. In future work, we need to study features selection that can be decided to give a better performance for online review spam detection.

Based on our study in current researches, for the detection from language, most of them used supervised learning techniques like SVM or logistic regression. For the spam detection problem, supervised learning techniques are used to separate reviews as truthful or fake. All data are required to be labeled. As we discussed in the chapter 1, most of the labeled datasets used as a training input are either labeled fake reviews manually or hired users to write fake ones. However, it is a problem to build classifiers based on those synthetic datasets since the datasets may not represent the real-world review spam. (Ott 2011, Ott 2013), using the same methods on AMT dataset and Yelp's filtered reviews dataset but different features，have achieved different results, which implies that artificial fake reviews and real world fake reviews have different distinguishing features. Mukherjee (2013a) have shown that synthetic datasets give a poor indication of performance.

It is difficult to label real world dataset accurately, researchers realized that experiments should work on the real dataset. Therefore, semi-supervised and unsupervised methods attract more interest due to their advantages, requiring less labor to label fake reviews and reducing the noise data due to mislabeling by human judgement. For the detection from network， the unlabeled data or unlabeled data with a small amount of labeled data are used in many models. However, the comparison work between unsupervised and semi-supervised with supervised learning methods has not been done, which makes it hard to give a conclusion and limits the research.

Based on these findings from the literature, we explore sentiment features which ascertain the attitudes and opinions expressed in the review texts and provide a new direction to detect fake reviews. In most review datasets, fake reviews are obtained by using a method of manually labelling reviews. Yelp provides labeled review based on its automatic review recommendation algorithm. Therefore, this thesis adopts the YelpZip[6] dataset for developing and testing algorithm.

[6] http://odds.cs.stonybrook.edu/yelpzip-dataset/

CHAPTER 3

YELP DATASET

Begun over a decade ago, Yelp has been growing to be a website and mobile app to connect people with lots of local businesses. Yelp has rooted in multiple countries across the globe, making it the leading local guide for real word-of-mouth on everything from boutiques and mechanics to restaurants and dentists. Currently, Yelp is the home of more than 148 million reviews and receiving more than 170 million unique visits monthly from its mobile app, mobile website and desktop (Yelp Metrics, 2017).

3.1 Introduction to Yelp and its Users' Reviews

On Yelp, Yelpers (users of Yelp) can search for local businesses, e.g., nearby restaurants, and read their reviews. Besides, Yelpers can also share their opinions by leaving reviews for certain businesses. In order to post a review, a user must open a free account with Yelp, which requires the user to register a valid email address and some additional information, such as local address, gender, age and photo. Since inception, Yelp reviews have been growing rapidly, especially within past several years. Yelp reviews cover a variety of business categories, such as shopping stores, restaurants, hotel and local services, beauty and fitness. In Yelp, high-score reviews, such 5-star and 4-star reviews, take nearly 70% of total reviews, which indicates Yelp is a high-score review system. Different from other online review websites, Yelp runs

an automated software to recommend reviews for readers, and more than 20% of reviews are marked as "Not Recommended". A brief summary of Yelp reviews is given in Fig. 1.

When a user looks for a local business, e.g., the Mews Tavern restaurant in Fig. (a), on Yelp, it will show how many reviews have been posted for this business along with other this restaurant's information. For instance, Mews Tavern restaurant received 318 reviews before Feb. 18, 2018. These reviews will be publically available and free to any Yelp reader with or without an account. These reviews include diverse messages for readers to learn about the quality and service of this local business.

Generally, a review on Yelp is the personal experience or comment that a customer receives directly from the business and wants to share within Yelp community. Yelp requires users to post their personal, accurate, and timely reviews. Normally, a Yelp review is mainly composed of following features, 1) the *User* who posts the review and her/his profile, 2) the *Date* of receiving or updating the review, 3) the numerical *Rating* ranging from 1 star to 5 stars, 4) the *Text* that user describes the experience literally, and 5) the *Vote* from other users for the review, e.g, Useful, Funny and Cool. An example of Yelp review can be found in Fig. 2 (b).

Among all features, the text feature is of critical importance to a review. The text often offers a rich narrative and a wealth of detail about the review. For instance, when a reader reads a review that has rating score as 5-star, he/she might ask "why does this user give this high-score review?". It is normal for the reader to look for more supporting clues by reading the describing text. If the text contains lots of details, e.g.,

why and how the experience/service was, to support the 5-star posting, readers will think this review is more useful. Otherwise, they may be suspicious of this review.

3.2 Recommended Reviews vs Not-recommended Reviews

Yelp users may choose to share their experiences/reviews for different reasons. For instance, a local business owner could invite their customers to post a positive Yelp review to boost their business. Different social incentives have been adopted in different reviewing systems, including Yelp, to encourage people to submit a review (Wang, 2010).

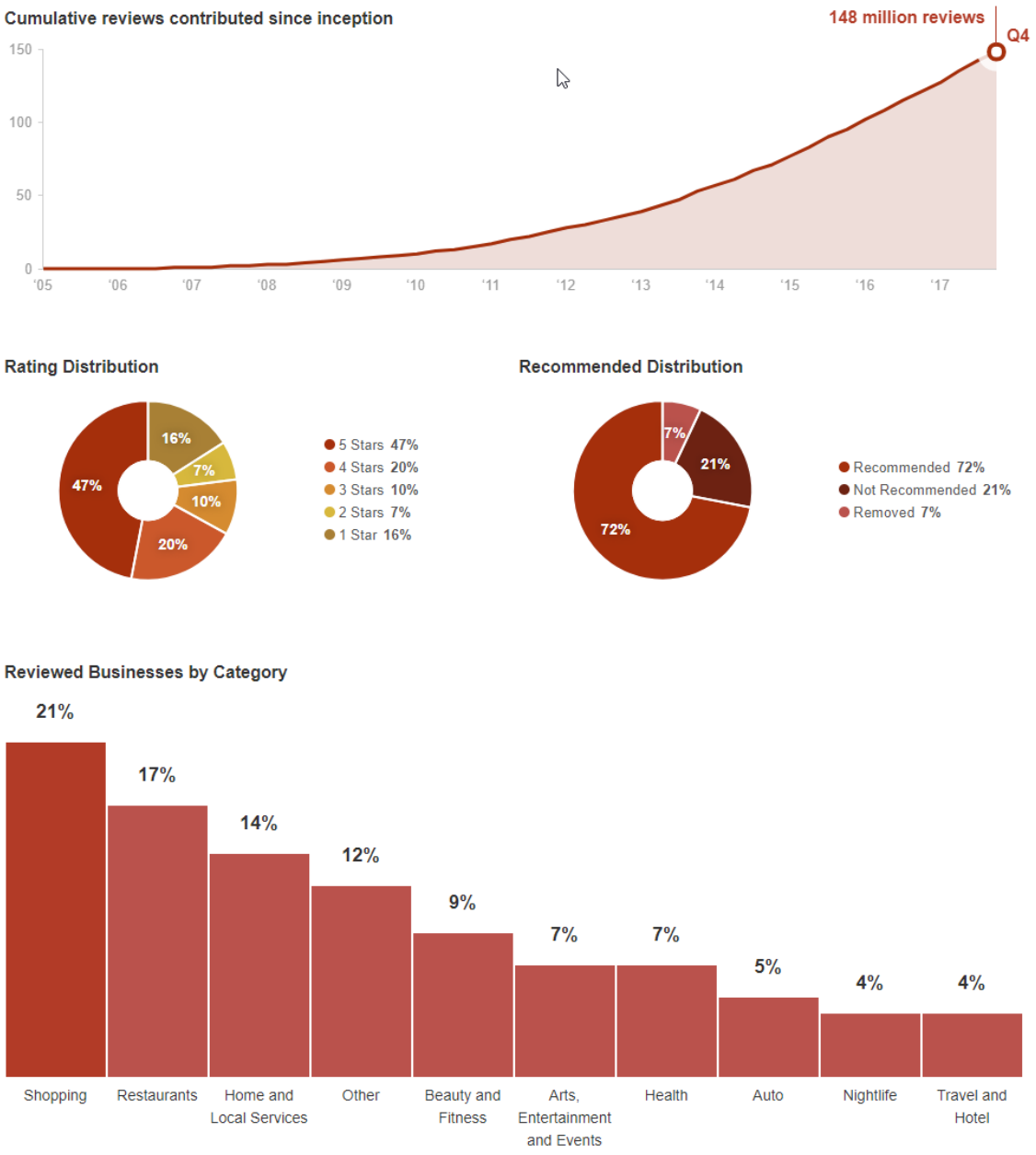

Fig. 1 Yelp Review Statistics as of December 31, 2017

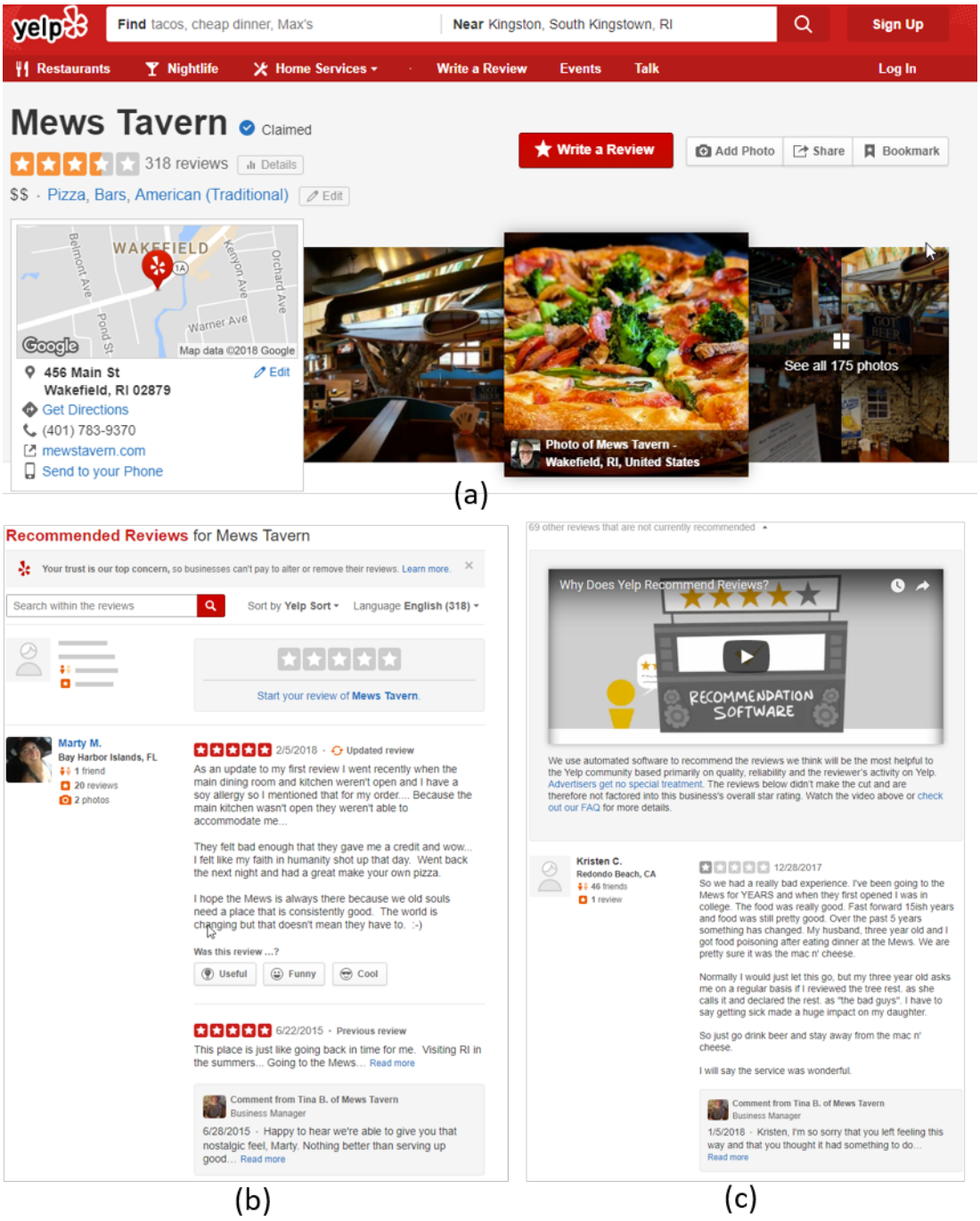

Fig. 2 A local business example on Yelp: (a) Mews Tavern restaurant, a local restaurant in Kingston RI, and received 318 reviews in total, 69 of which are not recommended for users by Yelp, before Feb. 18, 2018; (b) an example of the recommended reviews; (3) an example of the not-recommended reviews;

Reviews are the critical components of Yelp. People usually refer to the Yelp reviews to judge the quality of the local businesses, which drives Yelp to carefully filter reviews for readers. In Fig. 1, it clearly shows that there are 21% of total reviews that are filtered out and put into the section of "Not Recommended". Those reviews are not actually removed from Yelp, but are put into a light gray area at bottom of review page and are hard for Yelp visitors to find and read, since it's believed that those flagged reviews are fake, unhelpful, or biased (Nesler 2017). An example of not-recommended review is showing in Fig. 2 (c).

What actually determines whether a review is recommended by Yelp's automated recommendation algorithm? Yelp does not officially release how the recommendation algorithm works due to the worries of penetrating the machinations of the algorithm. As a platform, however, Yelp knows much more information about the reviewers and reviews. There are several basic principles that Yelp follows to NOT recommend a variety of reviews, 1) reviews from users who are not active, 2) reviews that may be biased because of being solicited from family, friends, or favored customers, 3) reviews that may have been written or purchased by business owners to help themselves or hurt a competitor, and 4) reviews that are unhelpful rants and raves.

In reality, it is really hard for individual readers to judge whether a review is good or not due to following reasons. First, online review systems, including Yelp, have no ground truth. No one can really tell whether a review is true or fake. Second, it is extremely difficult for a reader to look through all reviews of a business, since the number of reviews might be hundreds of, even thousands of. Finally, an individual reader might not have expertise to judge the quality of reviews. However, Yelp system can be in a good place to recommend reviews, as the system itself knows much more about users' profiles and

activities, and can hire experts in both linguistics and engineering to develop algorithms that can be run automatically nearly 150 million reviews and filter out suspicious ones. Yelp's automatic review recommendation algorithm is more robust and scalable than normal readers.

The "Recommended" feature is the unique feature of a review on Yelp. It has not been seen that other review systems other than Yelp provide such functionality. Yet, this "Recommended" feature has been widely adopted in the current literature that takes Yelp review dataset and label reviews. Normally, a review is labeled as 1, if the review is recommended; and it is labeled as -1, if the review is not recommended. In this thesis, it takes the same method to label Yelp review data.

3.3 Yelp Review Open Datasets

About five years ago, Yelp announced its Dataset Challenge program, which encourages students and researchers from around the globe to conduct research and analysis using Yelp data set and to discover insights hidden in the data set. Over the years, there has been incredible interest and usage of the dataset for educational purposes. For instance, teachers use it to teach their classes about databases, engineers use it learn graph databases, students use it to understand machine learning, and researchers use it to conduct natural language processing & sentiment analysis (Yelp Dataset Challenge 2017). Hundreds of academic papers have been written using Yelp's dataset. Yelp open dataset is publically available on Yelp (Yelp Open Dataset 2017).

Beside data sets published by Yelp, lots of other searchers have collected Yelp reviews and also make them publically available (Mukherjee 2013, Rayana 2015). Those data sets have fewer reviews than those published by Yelp, and normally

require less computation and are very useful for verifying algorithms. This thesis mainly adopts small-scale Yelp review data sets, and focuses on studying the relationship between the sentiment features of a review's text and the "Not-recommended" feature of the review.

CHAPTER 4

SENTIMENT FEATURES OF REVIEW TEXTS

The text feature is of critical importance to reviews. Normally, texts include rich sentiment elements that literally explain the reasons of giving such a review. It may also provide the clues that certain reviews are not recommended by Yelp automatic recommendation algorithm.

4.1 Modeling of Yelp Reviews

As is discussed in Section 3, a Yelp review (R) is composed of different attributes. This thesis, it adopts the following attributes:

- User ID or uID: Yelp system assigns each registered user a unique ID for tracking the activities, e.g., submitting a review for a business, in the system. User IDs are unique strings on Yelp.
- Product ID or pID: Yelp system assigns each registered product (hotels or restaurants) a unique ID for tracking the information left by users. e.g., information about the hotel, rating posted by users. Product IDs are unique strings on Yelp.
- Rating or r: it is the numeric score for a review, and is normally seen as stars. It includes 1 star, 2 stars, 3 stars, 4 stars and 5 stars. In short, the scores are used from 1 to 5. The higher the score is, the more favorite the user means to like the business.
- Date or d: the day that Yelp receives the review.
- Text or t: the literal content that the user submitted to explain the review in details.

- Label or l: Whether Yelp automated recommendation algorithm recommends the review. It has two label values, literally as *Recommended* or *Not Recommended*, and numerically as 1 or -1. This thesis uses the numeric values for l.

Briefly, a review is written as $R(uID, pID, r, d, t, l)$. If a business receives m reviews, those reviews are written as $(R_1, R_2, \dots, R_m)$.

4.2 Sentiment Features of Review Texts

In the current literature, Sentiment Analysis (SA) refers to adopting Natural Language Processing (NLP) and Machine Learning to identify and extract subjective information in a piece of writing context. The technique of SA is extremely helpful, because SA explores the general opinions or attitudes towards certain topics, products or services. Specifically, on Yelp reviews, SA demonstrates the opinions that users posted on certain local businesses.

Typically for sentiment analysis of online reviews, first a set of seed words is adopted to determine whether a piece of text contains positive or negative sentiments. Then, the positive or negative direction (positive, negative or neutral) of an opinion is determined based on the words that were present in the review text. Finally, the semantic classification algorithm (Dave 2003) or machine learning approach (Turney 2002) could be taken to mine the sentiment opinion from all reviews to classify the products/services as recommended or not recommended.

Different from traditional approaches, this thesis adopts another text mining method similar to the method used for sentiment analysis of online reviews in (Hu, 2012). The adopted text mining method is an efficient and standard term frequency

measure, which has been widely adopted in the Information Retrieval community (Salton, 1983). Briefly, for any given review R, it develops six numerical sentiment features from its text t, which are named as *Strong Positive* (SP) or R^{SP}, *Strong Negative* (SN) or R^{SN}, *Ordinary Positive* (OP) or R^{OP}, *Ordinary Negative* (ON) or R^{ON}, *Ordinary* (O) or R^{O}, and *Strong* (S) or R^{S}. The examples found by each category are shown in Table 2. This thesis adopts four widely-used word dictionaries to evaluate those six sentiment feature values. The strong positive word dictionary, represented as $Dict^{SP}$, includes 44 strong positive words (Archak, 2007); the strong negative word dictionary, represented as $Dict^{SN}$, includes 30 strong negative words (Archak, 2007); the ordinary positive and negative word dictionaries, represented as $Dict^{OP}$ and $Dict^{ON}$, include 2,006 positive words and 4,783 negative words, respectively (Hu, 2004; Liu, 2005).

Particularly, for any review's text, those six sentiment features are calculated as follows.

- Step 1: Tokenize the review text t into words.
- Step 2: Compare the tokenized words with $Dict^{SP}$, and calculate the number of occurrences of strong positive words in the review text, which is represented as N^{SP}. Conduct similar calculations against $Dict^{SN}$, $Dict^{OP}$ and $Dict^{ON}$, and get the number of occurrences of strong negative words, represented as N^{SN}, ordinary positive words, represented as N^{OP}, and ordinary negative words, represented as N^{ON}, respectively.
- Step 3: Calculate the total number of occurrences of the sentiment words in the review text, represented as N^{Total}.

$$N^{Total} = N^{SP} + N^{SN} + N^{OP} + N^{ON} \quad (1)$$

- Step 4: Calculate the scores of these six sentiment features.

$$\begin{cases} R^{SP} = N^{SP}/N^{Total} \\ R^{SP} = N^{SN}/N^{Total} \\ R^{OP} = N^{OP}/N^{Total} \\ R^{ON} = N^{ON}/N^{Total} \\ R^{S} = (N^{SP} + N^{SN})/N^{Total} \\ R^{O} = (N^{OP} + N^{ON})/N^{Total} \end{cases} \quad (2)$$

These sentiment scores are used to represent a review text for to study the differences in terms of sentiment between not-recommended reviews and recommended reviews on Yelp in this thesis.

Table 2 Examples of lexicon for each category

Strong Positive (SP)	*Strong Negative* (SN)	*Ordinary Positive* (OP)	*Ordinary Negative* (ON)
awesome	awful	abound	abnormal
best	bad	accessible	abolish
easy	cancelled	acclaim	babble
excellent	disappointed	acclamation	backaches
favorite	forever	backbone	backbite
great	horrible	bargain	backward
outstanding	misleading	calmness	cackle
professional	never	capable	calamitous
…	…	decisive	calamitously
		easy	calamity
		fairness	damage
		good	emaciated
		honoring	fat
		ideal	glum
		…	…

CHAPTER 5

REVIEW BURSTS ON YELP

Review bursts are abnormal behaviors, but they really exist, in online review systems. Study of review bursts is of critical importance to understand the review recommendation mechanism on Yelp.

5.1 Introduction to Review Bursts

Generally speaking, the reviews that are posted about online products and services should arrive in the system randomly, which means the arrivals of reviews should not have obvious correlations among each other. However, it has been reported that there are review busty behaviors in existing online review systems, e.g., Amazon.com (Xie 2012; Fei 2013). The *review bursts* mean that there are certain time periods, when there are sudden concentrations of reviews, meaning more reviews are posted in these periods than other normal periods. In reality, there are different reasons that cause review bursts, e.g., a sudden increase of popularity caused by successful commercial Ads, and spam attacks, e.g., injecting fake high-score reviews to boost a product.

Being the online platform to collect and share nearly 150 million of reviews for hundreds of thousands of local businesses, it is natural to study review bursts on Yelp. It is a common idea for local business owners to post more reviews on Yelp to show the popularity of their businesses.

For instance, a restaurant owner may encourage its customers to post more reviews by giving certain amount of discounts after the restaurant is open. A dentist can leave the customers a reminding card about posting more reviews on different platforms, e.g., Google and Yelp, after every half-year visit. Review bursts root

naturally on Yelp as well. In this thesis, it is another focus to study review bursts on Yelp.

A kernel density method was proposed to detect the bursts in a product's reviews (Xie 2012); another general method of counting the review number was adopted to report bursts (Fei 2013). This thesis adopts similar methods to calculate the burst periods in Yelp reviews.

5.2 The Density of Burstiness

Suppose a local business on Yelp has received a set of m reviews $\{R_1, R_2, \dots, R_m\}$, which are sorted on basis of the arrival dates $\{d_1, d_2, \dots, d_m\}$ ($d_i \le d_j,\ 1 \le i < j \le m$). And, the sliding widow has the size of W days, e.g., 30 days. The burstiness density value for review R_i, represented as $f_B(d_i)$, is calculated as follows.

- Obtain a subset of review dates restricted to $\left\{d_j, \left|d_j - d_i\right| \le \frac{W}{2}, 1 \le j \le m\right\}$, or Set_{d_i}. In other words, Set_{d_i} is composed of all review dates that are half of window size either ahead or behind of day d_i.
- Calculate the summation of mutual closeness between dates in Set_{d_i} as $f_B(d_i)$.

$$\begin{cases} f_B(d_i) = \sum_{d_j, d_k \in Set_{d_i};\ j<k} Dist(d_j, d_k) \\ Dist(d_j, d_k) = \frac{1}{|d_k - d_j| + 1} \end{cases} \tag{3}$$

where $|\cdot|$ is absolute value operation. In equation (3), the distance function, $Dist(d_j, d_k)$, aims to show the closeness of two dates, i.e., d_j and d_k. When $d_j = d_k$, which means two reviews are posted on the same day, the distance between this pair is the largest value, i.e., 1; when $\left|d_j - d_k\right| = W$, which means that one review is posted at the beginning of the sliding window, and the other is posted at the end of the sliding

window, the distance between this pair is the smallest value, i.e., $\frac{1}{W+1}$. In addition, both the number of reviews and the difference between each pair of review dates have been considered to calculate the final $f_B(d_i)$ value in equation (3). Generally, the more dates Set_{d_i} has and the closer each pair of dates are, the larger the cumulative $f_B(d_i)$ value is.

Eventually, every review is associated with a density value $f_B(d_i)$, $d_i \in \{d_1, d_2, \dots, d_m\}$, and the burstiness density curve looks the one in Fig. 3(b), from which it is clearly seen that the curve has obvious peaks and valleys. Peaks show there are more reviews arriving in corresponding time periods. As a result, Yelp also has the behaviors of review bursts.

5.3 Selection of Density Periods

Normally, review bursts cause that reviews are not evenly distributed on the life span. In some periods, there are more coming reviews than rest of other periods. Besides, detecting such bursts (Fei 2013), it is also of importance to how the dense reviews are different from other reviews. This thesis defines *Density Period* to study represent the period that has dense reviews.

From Fig. 3(b), it is clearly seen that the burstiness density values changes dramatically. The maximum value could be more than 90 and the minimum value could be smaller than 10. And, the burstiness density curves are dramatically different for different local businesses' reviews. This thesis defines a density period as follows.

- Calculate the *normalized burstiness density*, represented as $f_B{}'(d_i)$.

$$f_{min}=\min_{1\le i\le m} f_B(d_i)$$

$$f_{max}=\max_{1\le i\le m} f_B(d_i)$$

$$f_B{}'(d_i) = \frac{f_B(d_i) - f_{min}}{f_{max} - f_{min}} \tag{4}$$

where f_{min} and f_{max} represent the minimum and maximum burstiness.

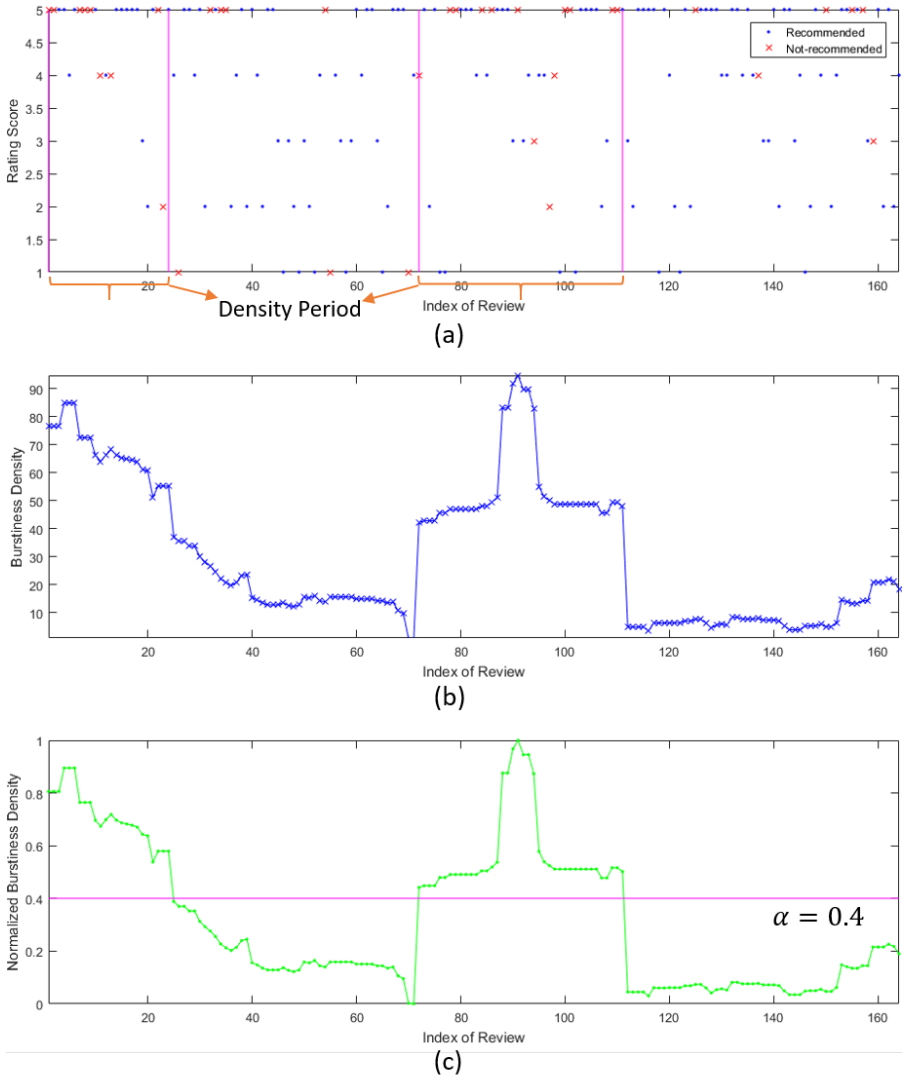

Fig. 3 A demo of review bursts and calculation of density periods. The raw review data is from a restaurant in YelpZip Dataset, and the restaurant id is 4414 in the dataset. (a) Rating scores distribution of the restaurant, and two density periods with are identified as $\alpha = 0.4$. (b) Burstiness density values for reviews, with the maximum value as 94.66. (c) Normalized burstiness density values. When setting α as 0.4, there are two density periods with index ranges as [1, 24] and [72, 111].

density values of the reviews for a local business. After normalization, $f_B{}'(d_i)$ ranges in $[0, 1]$. An example of the normalized burstiness density can be found in Fig. 3(c).

- Search for a consecutive period, where $f_B{}'(d_i) \geq \alpha$ and α is the threshold, as a density period. For instance, α is set to be 0.4, there are two density periods identified from Fig. 3(c), and are shown in Fig. 3(a) correspondingly.

The selection criteria are slightly different from detecting of review bursts (Fei 2013), where detection criteria focused on discovering the occurrences of bursts and detecting the abnormal behaviors. In this thesis, however, it focuses on discovering the periods of occurrences of dense reviews. A period has the start time, when reviews start to burst, as well as the end time, when the burst ends. It also needs a flexible threshold to obtain different levels of density periods. If α is set to a larger value, e.g., close to 1, a found density period should have reviews that are much closer to each other in terms of arrival dates.

CHAPTER 6

EXPERIMENTS AND DISCUSSIONS

In this section, the review dataset used in this thesis is first introduced, followed by giving the experiments to apply sentiment features and density periods to study reviews on Yelp.

6.1 Review Dataset Description

In Section 3.3, several publically available Yelp review data sets were briefly introduced. In this thesis, the Yelp review dataset used for experiments is the restaurant reviews that were first used by Rayana and Akoglu (Rayana 2015). This dataset includes 608,598 Yelp reviews for restaurants in the states of New York, New Jersey, Vermont, Connecticut and Pennsylvania in USA. Reviews include local restaurants, user information, timestamp, ratings, and a plaintext review. Briefly, the dataset includes online reviews from 5,044 restaurants posted by 260,277 reviewers. Rayana and Akoglu took Yelp's automatic recommendation algorithm and label collected reviews. If the algorithm identifies a review as 'fake or suspicious', the review was labeled as *Not-Recommended* or -1; otherwise, the review was labeled as *Recommended* or 1.

This thesis aims to apply the sentiment features to study *Not-recommended* reviews as well as study the density periods. Additional criteria have been used to select restaurants, which are 1) the number of reviews of a restaurant is no less than 50 and no more than 2,000; 2) among all review, there are at least 10% of reviews as not-recommend reviews. Those two criteria help to reduce biases in selecting subset for

experiments. Finally, the subset selected for experiments includes 1,387 different restaurants.

In total, there were 278,138 reviews collected from 1,387 restaurants, among which 44,894 reviews were filtered out by Yelp as Not-recommended reviews. Table 3 shows the percentages that every rating score takes in All Reviews category (including recommended and not-recommended reviews) and Not-recommended Reviews category. For instance, 5-star reviews take 38.78% of all reviews and 44.98% of not-recommended reviews, respectively. In Table 3, it can been clearly seen that the not-recommended reviews have more sentiment biases. In not-recommended category, the polarized reviews, e.g., 5-star and 1-star reviews, take apparently larger percentages, and the neutral reviews, e.g., 3-star reviews, takes much less percentages.

Table 3. Percentage comparison of review rating scores between All Reviews category and Not-recommended Reviews category

	5-Star	**4-Star**	**3-Star**	**2-Star**	**1-Star**
All Reviews	38.78%	33.88%	13.17%	7.69%	6.48%
Not-recommended Reviews	44.98%	24.21%	7.45%	8.36%	15%

6.2 Sentiment Analysis: Recommended Reviews vs Not-recommended Reviews

The rating value is the numerical feature of online reviews, which is the representation of a review's sentiment. Normally, a 5-star review left for a restaurant on Yelp means the reviewer very much likes this restaurant and would recommend this restaurant to readers. Although, the rating feature is widely adopted to study the

online reviews, it still has limitations. For instance, if there are two 5-star reviews, and one is labeled as Recommended and another is labeled as Not-Recommended. It would be very difficult for readers to differentiate them only based on values. Comparing with rating feature, the text feature of a review contains rich sentiment information to support the reviewer's opinion. It is of importance to conduct sentiment analysis against recommended reviews and not- recommended reviews on Yelp.

In Section 4.2, six sentiment features were introduced, i.e., *Strong Positive* (SP) or R^{SP}, *Strong Negative* (SN) or R^{SN}, *Ordinary Positive* (OP) or R^{OP}, *Ordinary Negative* (ON) or R^{ON}, *Ordinary* (O) or R^{O}, and *Strong* (S) or R^{S}. In this subsection, these six features are adopted to investigate the differences in terms of sentiment between recommended reviews and not-recommended reviews as follows.

- Divide the reviews of any restaurant selected for experiments into three categories. *Positive Category* includes all 4-star and 5-star reviews, which represents positive attitude; *Neutral Category* includes all 3-star reviews, which represents neutral attitude; *Negative Category* includes 1-star and 2-star reviews, which represents negative attitude.
- Divide the reviews in each category into two groups. *Recommended group* contains the recommended reviews in a group, and *not-recommended group* contains not-recommended reviews in a group.
- Apply a sentiment feature, e.g., R^{SP}, to study the reviews in two groups for certain category, e.g., Positive Category. Specially, calculate the average sentiment feature values of the reviews in the recommended group, represented as Ave_R and in not-recommended group, represented as Ave_{NR} , respectively.

where

$$Ave_R = \frac{\sum_1^n R^{SP}}{n}$$

$$Ave_{NR} = \frac{\sum_1^m R^{SP}}{m} \quad (5)$$

- Compare Ave_{NR} with Ave_R. If $Ave_{NR} > Ave_R$, it means that not-recommended reviews have stronger sentiment polarity than that of recommended reviews in terms of the given sentiment feature, e.g., R^{SP}. If $Ave_{NR} < Ave_R$, recommended reviews have stronger sentiment polarity. Otherwise, both subgroups have similar sentiment polarity.

In summary, the reviews for a restaurant are divided into three categories, i.e., Positive Category, Neutral Category, and Negative Category, and each category includes two groups, i.e., recommended group and not-recommended group. For two groups in each category, instead of applying rating values, it adopts six sentiment features to conduct the comparison.

The selected 1,387 selected restaurants are adopted to conduct experiments. The reviews for each restaurant are divided into three categories. Take Positive Category as an example. One sentiment feature, e.g., R^{SP}, is used to calculate the average sentiment feature values for not-recommended group, i.e., Ave_{NR}, and the average sentiment feature values for recommended group, i.e., Ave_R. It is found that there are 80.75% of 1,387 restaurants, whose Ave_{NR} is larger than Ave_R regarding to Positive Category. The comparison results using R^{SP} are given in Table 4, which also includes the comparisons for Neutral Category and Negative Category.

The studies using other five sentiment features, i.e., R^{OP}, R^{SN}, R^{ON}, R^S and R^o are conducted similarly, and the comparison results are listed in Tables 5, 6, 7, 8 and 9,

respectively. The following conclusions can be drawn based on these experiment results.

First, for reviews in positive category, not-recommended ones usually have stronger positive sentiment polarity than recommended ones. In Table 4, it shows that there are 80.75% of 1,387 restaurants, whose Ave_{NR} is larger than Ave_R in terms of using R^{SP} for analysis. Such percentage value is 61.72% in Table 5, where R^{OP} is adopted as analysis. These high percentages imply that there are more positive words, both strong positive words and ordinary positive words, which are given in the texts of not-recommended reviews.

Second, for reviews in negative category, not-recommended ones usually have stronger negative sentiment polarity than recommended ones. It is found in Table 6 that there are 59.96% of 1,387 restaurants, whose Ave_{NR} is larger than Ave_R in terms of using R^{SN} for analysis. Such percentage value is 56.31% in Table 7, where R^{ON} is adopted as analysis. Similarly, these high percentages imply that there are more negative words, both strong negative words and ordinary negative words, which are left in the texts of not-recommended reviews.

Finally, Table 8 shows the analysis results using the strong sentiment feature, i.e., R^S, which indicates that not-recommended reviews usually contain more polarized (positive or negative) words. In contrast, Table 9 indicates that the recommended reviews normally contain more ordinary words than not-recommended reviews.

Table 4. Comparison among three categories adopting *Strong Positive* feature: R^{SP}.

Category	$Ave_{NR} > Ave_R$	$Ave_{NR} = Ave_R$	$Ave_{NR} < Ave_R$
Positive Category	**80.75%**	0	19.25%
Neutral Category	39.44%	0.72%	59.84%
Negative Category	43.55%	1.44%	55.01%

Table 5. Comparison among three categories adopting *Ordinary Positive* feature: R^{OP}

Category	$Ave_{NR} > Ave_R$	$Ave_{NR} = Ave_R$	$Ave_{NR} < Ave_R$
Positive Category	**61.72%**	0	38.28%
Neutral Category	35.4%	0.29%	64.31%
Negative Category	32.08%	0.58%	67.34%

Table 6. Comparison among three categories adopting *Strong Negative* feature: R^{SN}.

Category	$Ave_{NR} > Ave_R$	$Ave_{NR} = Ave_R$	$Ave_{NR} < Ave_R$
Positive Category	48.95%	0	51.05%
Neutral Category	35.4%	0.65%	63.95%
Negative Category	**59.96%**	0.65%	42.39%

Table 7. Comparison among three categories adopting *Ordinary Negative* feature: R^{ON}

Category	$Ave_{NR} > Ave_R$	$Ave_{NR} = Ave_R$	$Ave_{NR} < Ave_R$
Positive Category	44.63%	0	55.37%
Neutral Category	36.84%	0.72%	62.44%
Negative Category	**56.31%**	0.79%	42.9%

Table 8. Comparison among three categories adopting *Strong* feature: R^{S}.

Category	$Ave_{NR} > Ave_R$	$Ave_{NR} = Ave_R$	$Ave_{NR} < Ave_R$
Positive Category	**69%**	0	31%
Neutral Category	40.37%	0.36%	59.26%
Negative Category	**56.38%**	0.58%	43.09%

Table 9. Comparison among three categories adopting *Ordinary* feature: R^O.

Category	$Ave_{NR} > Ave_R$	$Ave_{NR} = Ave_R$	$Ave_{NR} < Ave_R$
Positive Category	38.72%	0	61.28%
Neutral Category	35.9%	0.29%	63.81%
Negative Category	39.73%	0.5%	59.77%

6.3 Study of Densities Periods

It is common sense that business owners, especially small business owners, would like to increase the number of reviews to their businesses on public platforms for various purposes, e.g., increases of public popularity or promotion of the businesses. Those deliberate behaviors normally result in review bursts and occurrence of density periods in review's lifespan, which has been discussed in Section 5. For experiments, the density periods are further chosen as follows.

- For any restaurant, calculate its burstiness density function (i.e., f_B) and normalized burstiness density function (i.e., $f_B{}'$).
- Select the density periods on base of the criteria $f_B{}' \geq \alpha$ and α is the threshold for all restaurants (e.g., $\alpha = 0.4$). At least one density period can be found, as $f_B{}'$ always ranges from 0 to 1. Each found density period has starting index[7]

[7] It refers to the first day of sliding window.

and end index[8]. (In experiments, the reviews for a restaurant are arranged by their arrival dates ascendingly.)

- For every found density period, obtain the maximum burstiness density value in the corresponding f_B.

Table 10 shows a list of the selected density periods with top 20 largest burstiness density values on the 1,387 test restaurant set. Following observations can be obtained.

First, the maximum burstiness density values for density periods vary significantly. In Table 10, the largest burstiness density value of selected density periods could be 3,923.34, while the smallest one is also 157.94. This means there do exist periods in the Yelp review system where lots of reviews were posted for some local businesses within a short period of time, e.g., 60 days. (The sliding window size W is set to be 30 days in this thesis.)

Second, the review bursts often occur at the beginning days of the reviews' lifespan. For instance, 9 of 20 selected density periods have the start index as 1, which means those periods happen at the very beginning of the reviews' lifespan[9]. This observation indicates that review bursts are suspicious as deliberate behavior. It is understandable that reviews can make a business by boosting a new restaurant's rating to popularize it or can break a business by being defamed by its competition when it first opens.

Finally, some restaurants have multiple density periods. For instance, the restaurant with ID as 151 in Table 10 has three density periods. One period is at the

[8] It refers to the last day of sliding window.
[9] It refers to the time period of our data collection, during which reviews were posted for a particular restaurant. Indeed, the restaurant may have other reviews, but these are not in our dataset.

beginning and the other two are in the middle of lifespan. This observation can also be seen in Fig. 3. Multiple density periods found for one restaurant indicate reviews bursts could happen multiple times for local businesses on Yelp.

Table 10. 20 selected density periods with the top largest burstiness density values on test set.

Restaurant ID	**Start and end indexes of the density period**	**Maximum burstiness density value of the density period**
2879	[789, 941]	3,923.34
1597	[1857, 1911]	647.55
3332	[1, 56]	627.47
1100	[1, 59]	600.24
3962	[140, 199]	589.22
247	[1, 57]	578.58
151	[852, 1262]	432.37
4183	[570, 616]	418.77
3882	[1, 35]	296.55
151	[675, 764]	251.42
1344	[1, 32]	239.24
1005	[1, 48]	233.48
3778	[1, 35]	223.81
2397	[70, 151]	216.2
2386	[1, 31]	206.72
4558	[21, 49]	189.86
1401	[935, 1074]	183.53
899	[655, 803]	179.86
2174	[1, 32]	178.56
151	[656, 670]	157.94

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

There is a growing trend that people rely on the reviews on Yelp to learn the qualities of the services from local businesses. Normally, local businesses that have a large percentage of positive reviews tend to attract more customers than those having lots of negative reviews. In order to improve the popularity for the businesses, it is highly possible that some business owners try to cheat Yelp review system by leaving fake or deceptive reviews to deliberately mislead potential customers. Yelp has developed a review recommendation algorithm to automatically categorize reviews into recommended group and not-recommended group. In this thesis, there are six sentiment features, i.e., Strong Positive, Strong Negative, Ordinary Positive, Ordinary Negative, Ordinary, and Strong, which are developed from the texts of reviews. And those sentiment features have been used to investigate the differences in terms of sentiment between recommended reviews and not-recommended reviews. It has been found that not-recommended reviews normally contain more polarized (positive or negative) words than recommended reviews on Yelp.

Furthermore, it has been shown that reviews bursts also occur on Yelp review system. In this thesis, it explores the *Date* component of a review to develop the Density of Burstiness for the reviews of a business and then develop *Density Periods* to study Yelp's review bursts. In summary, Yelp reviews have following concentration observations.

- The maximum burstiness density values for density periods vary significantly. And, the largest burstiness density value of detected density periods is larger than 3,923 with giving sliding window size as 30 days.

- The review bursts often occur at the beginning days of the reviews' lifespan. 9 of 20 selected density periods happen at the very beginning of the reviews' lifespan.
- Some restaurants have multiple density periods. One restaurant with ID as 151 has been detected with three density periods.

7.2 Future Work

In the future, there are several possible directions along this topic. First, it is of importance to adopt the larger and latest Yelp review dataset to conduct studies. Yelp launched its dataset challenge program (Yelp Dataset Challenge 2017). The dataset published in every round is a very large dataset and contains rich review information. It will be promising to adopt this dataset to study the topics addressed in this thesis. Second, another direction to study is which sentiments can help to further detect the review manipulation behaviors on Yelp. Sentiment implies the reasons of giving such a review and should correlate to the rating given. For example, analyzing the 5-star rating may be hard to tell if it is genuine or fake. A true 5-star review should contain some details as to what makes that product worth buying or summarize the content of features of the product. However, a fake one may merely describe the product as wonderful, great or amazing, etc. Based on experiments, it has been found that not-recommended reviews normally contain more polarized (positive or negative) words than recommended reviews. It is possible to conduct a statistical study of polarized word distributions between recommended reviews and not-recommended reviews. This study may indicate hints to differentiate fake reviews. Third, it can incorporate sentiment features to study Density Periods. For instance, it is of interest to investigate

whether concentrated reviews have stronger polarity tendency. Finally, many other methods, e.g., machine learning algorithms, neuronal network methods and regression models, could be incorporated with the sentiment features and density periods that are developed in this thesis to further investigate the Yelp reviews.

BIBLIOGRAPHY

Akoglu L, Chandy R, and Faloutsos C, (2013). "Opinion fraud detection in online reviews by network effects." *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM* 2013. 2-11.

Archak, N.; Ghose, A.; and Ipeirotis, P.G., "Show me the money! deriving the pricing power of product features by mining consumer reviews", *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 56–65.

Banerjee S and Chua A. Y, "Authentic versus fictitious online reviews: A textual analysis across luxury, budget, and mid-range hotels," *Journal of Information Science*, vol. 43, no. 1, pp. 122–134, 2017.

Bryan Hooi, Neil Shah, Alex Beutel, Stephan Gunneman, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. 2015. *Birdnest: Bayesian inference for ratings-fraud detection,* arXiv preprint arXiv:1511.06030.

Dave, K.; Lawrence, S. and Pennock, D.M., "Mining the peanut gallery: opinion extraction and semantic classification of product reviews", *Proceedings of the 13th International World Wide Web Conference*, 2003, pp. 519–528.

D'Onfro Jillian "A whopping 20% of yelp reviews are fake," 2013. [Online]. Available: http://read.bi/1M03jxl

Feng S, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection," in

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 171–175. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390665. 2390708.

Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R (2013) Exploiting Burstiness in reviews for review spammer detection. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 13:175–184.

Gu¨nnemann N, Gu¨nnemann S, and Faloutsos C, "Robust multivariate autoregression for anomaly detection in dynamic product ratings," in *Proceedings of the 23rd*

International Conference on World Wide Web, ser. WWW '14. New York, NY, USA: ACM, 2014 a, pp. 361–372. [Online]. Available: http://doi.acm.org/10.1145/2566486.2568008

Günnemann S, Günnemann N, and Christos Faloutsos. 2014b. Detecting anomalies in dynamic rating data: a robust probabilistic model for rating evolution. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14).* ACM, New York, NY, USA, 841-850. DOI: https://doi.org/10.1145/2623330.2623721

Hammad A.A and A. El-Halees, An approach for detecting spam in arabic opinion Reviews. *The International Arab Journal of Information Technology*, Vol. 12, pp. 9–16, 2015.

Hernández D, Guzmán R, Móntes y Gomez M, Rosso P (2013) Using PU-learning to detect deceptive opinion spam. In: *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.*, pp 38–45.

Hooi B, Shah N, Beutel A, Gu¨nnemann S, Akoglu L, Kumar M, Makhija D, and C. Faloutsos, Birdnest: Bayesian inference for ratings-fraud detection, in *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016.

Hu N, Bose I, Koh N. S, and Liu L, "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," *Decision Support Systems*, vol. 52, no. 3, pp. 674 – 684, 2012.

Hu, M. and Liu B., "Mining and Summarizing Customer Reviews." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, Aug 22-25, 2004, Seattle, Washington, USA,

Jiang M, Cui P, Beutel A, Faloutsos C, and Yang S. Catchsync: catching synchronized behavior in large directed graphs. In KDD, pages 941–950, 2014.

Jindal N and Liu B, "Review spam detection," in *Proceedings of the 16th International Conference on World Wide Web, ser. WWW '07. New York, NY, USA:ACM, 2007*, pp. 1189–1190. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242759

Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230. ACM, Stanford, CA.

Jindal N, Liu B, Lim EP (2010) Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM international conference on Information and knowledge management,* pp. 1549–1552. ACM, Toronto, ON, Canada.

Lim E.-P., Nguyen V.-A, Jindal N, Liu B, and Lauw H. W, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ser. CIKM '10. New York, NY, USA: ACM, 2010*, pp. 939–948. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871557

Li F, Huang M, Yang Y, and Zhu X, "Learning to identify review spam," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three, ser. IJCAI'11*. AAAI Press, 2011, pp. 2488–2493. [Online]. Available: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-414.

Li H, Chen Z, Liu B, Wei X, and Shao J, "Spotting fake reviews via collective positive-unlabeled learning," in *Proceedings of the 2014 IEEE International Conference on Data Mining, ser*. ICDM '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 899–904. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2014.47

Li, H., Chen, Z., Mukherjee, A., Liu, B., & Shao, J. (2015). Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015* (pp. 634-637). AAAI Press.

Li J, Ott M, Cardie C, Hovy E (2014) Towards a general rule for identifying deceptive opinion spam. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1566–1576, Baltimore, Maryland, USA, June 23-25 2014.

Liu, B.; Hu, M.; and Cheng, J., "Opinion Observer: Analyzing and Comparing Opinions on the Web." *Proceedings of the 14th International World Wide Web conference* (WWW-2005), May 10-14, 2005, Chiba, Japan.

Liu, Y., Sun, Y., and Yu, T., "Defending multiple-user-multiple-target attacks in online reputation systems," in *Privacy, Security, Risk and Trust (PASSAT) and Social Com, 2011 IEEE Third International Conference on*, Oct 2011, pp. 425–434.

"Local consumer review survey 2014," 2018. [Online]. Available: https: //www.brightlocal.com/learn/local-consumer-review-survey-2014/

Luca, M. and Zervas, G., "Fake it till you make it: Reputation, competition, and yelp review fraud," *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

Malbon, J., "Taking fake online consumer reviews seriously," *Journal of Consumer Policy*, vol. 36, no. 2, pp. 139–157, Jun 2013.

Mayzlin, D., Dover, Y., and Chevalier, J., "Promotional reviews: An empirical investigation of online review manipulation," *American Economic Review*, vol. 104, no. 8, pp. 2421–55, 2014.

Mukherjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews. In: *Proceedings of the 21st international conference on World Wide Web*. (pp. 191–200). ACM, Lyon, France.

Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N., "What Yelp Fake Review Filter Might Be Doing", *Proceedings of The International AAAI Conference on Weblogs and Social Media* (ICWSM-2013), July 8-10, 2013a, Boston, USA.

Mukherjee A, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)*, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, and Jingrui He (Eds.). ACM, New York, NY, USA, 632-640. DOI: https://doi.org/10.1145/2487575.2487580, 2013b.

Nesler, John, "Why Are Some Yelp Reviews Not Recommended?", https://www.postmm.com/yelp-reviews-not-recommended-data-analysis/, Jul. 1, 2017.

Nikou Günnemann, Stephan Günnemann, and Christos Faloutsos. 2014a. Robust multivariate autoregression for anomaly detection in dynamic product ratings. In *Proceedings of the 23rd international conference on World wide web (WWW '14)*. ACM, New York, NY, USA, 361-372. DOI: http://dx.doi.org/10.1145/2566486.2568008

Ott M, Choi Y, Cardie C, Hancock JT (2011) 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, pp, 309–319. Association for Computational Linguistics.

Ott M, C. Cardie, and J. T. Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 201-210. DOI=http://dx.doi.org/10.1145/2187836.2187864.

Ott M, Cardie C, Hancock JT (2013) Negative Deceptive Opinion Spam. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*., pp 497–501.

Rayana S and Akoglu L. 2015. "Collective opinion spam detection: Bridging review networks and metadata," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '15*. New York, NY, USA: ACM, 2015, pp. 985–994. [Online]. Available: http://doi.acm.org/10.1145/2783258.2783370

Wang G, Xie S, Liu B, Yu PS (2011) Review graph based online store review spammer detection. In: *Data mining (icdm), 2011 ieee 11th international conference* on (pp. 1242–1247). IEEE, Vancouver, Canada.

Wang G, Xie S, Liu B, Yu PS (2012) Identify online store review spammers via social

review graph. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(4):61

Wang, Z., Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews," *The B.E. Journal of Economic Analysis and Policy*, Vol. 10, No. 1, 2010.

Salton, G. and McGill, M.J., "Introduction to Modern Information Retrieval", *MCGRAW HILL COMPUTER SCIENCE SERIES*, 1983.

Schwarz Yoav "Review sites like yelp can now legally manipulate reviews for cash. but should they?" 2014. [Online]. Available: https://venturebeat.com/2014/10/24/review-sites-like-yelp-can-now-legally-manipulate-reviews-for-cash-but-should-they

Sen P, Namata G, Bilgic M, Getoor L, Gallagher B, and Eliassi-Rad T, Collective classification in network data, *AI Magazine*, vol. 29, pp. 93–106, 2008.

Shojaee S, Murad MAA, Bin Azman A, Sharef NM, Nadali S (2013) Detecting deceptive reviews using lexical and syntactic features. In: *Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on (pp. 53–58)*. IEEE, Serdang, Malaysia.

Turney, P.D., "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews", *Proceedings of the 40th Annual Meeting on Association for Computational Lingistic*, 2002, pp. 417–424.

Xie S, Guan Wang, Shuyang Lin, and Philip S. Yu. 2012. "Review spam detection via temporal pattern discovery," in *Proceedings of the 18th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 823–831. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339662

Xu C, Zhang J, Chang K, and Long C, "Uncovering collusive spammers in Chinese review websites," in Proceedings of *the 22nd ACM International Conference on Information & Knowledge Management, ser*. CIKM '13. New York, NY, USA: ACM, 2013, pp. 979–988. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505700

Ye J and Akoglu L, "Discovering opinion spammer groups by network footprints," in *Proceedings of the 2015 ACM on Conference on Online Social Networks, ser. COSN '15*. New York, NY, USA: ACM, 2015, pp. 97–97. [Online]. Available: http://doi.acm.org/10.1145/2817946.2820606

Yelp Metrics, "An Introduction to Yelp Metrics as of December 31, 2017", [Online]. Available: https://www.yelp.com/factsheet.

Yelp Open Dataset and Dataset Challenge, https://engineeringblog.yelp.com/2017/08/yelp-open-dataset-and-dataset-schallenge-round-10.html, 2017.