Review Spam Detection Using Machine Learning Techniques

Rohit Narayan

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.
May 2016

Review Spam Detection Using Machine Learning Techniques

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Technology

in

Computer Science and Engineering

(Specialization: Information Security)

by

Rohit Narayan

(Roll: 214CS2155)

under the guidance of

Prof. Sanjay Kumar Jena

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.
May 2016

Declaration by the Student

I certify that:

- The work enclosed in this thesis has been done by me under the supervision of my project guide.
- The work has not been submitted to any other Institute for any degree or diploma.
- I have confirmed to the norms and guidelines given in the Ethical Code of Conduct of National Institute of Technology, Rourkela.
- Whenever I have adopted materials (data, theoretical analysis, figures or text) from other authors, I have given them due credit through citation and by giving their details in the references.

Name: Rohit Narayan
Date:
Signature:

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India.

May 16, 2016

Certificate

This is to certify that the work in the thesis entitled ***Review Spam Detection Using Machine Learning Techniques*** by ***Rohit Narayan*** is a record of an original research work carried out under my supervision and guidance in partial fulfilment of the requirements for the award of the degree of **Master of Technology** in Computer Science and Engineering with specialization Information Security. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Sanjay Kumar Jena
Professor
Department of CSE, NIT Rourkela

Acknowledgment

I owe deep gratitude to the ones who have contributed greatly in completion of this thesis.

Foremost, I would also like to express my gratitude towards my project advisor, Prof. Sanjay Kumar Jena, whose mentor-ship has been paramount, not only in carrying out the research for this thesis, but also in developing long-term goals for my career. His guidance has been unique and delightful. I would also like to thank my mentor, Jitendra Rout Sir, who provided his able guidance whenever I needed it. He inspired me to be an independent thinker, and to choose and work with independence.

I would also like to extend special thanks to my project review panel for their time and attention to detail. The constructive feedback received has been keenly instrumental in improvising my work further.

I would like to specially thank my friend Sagarika Behura for his profound insight and for guiding me to improve the final product, as well as my other friends for their support and encouragement.

My parents receive my deepest love for being the strength in me.

Rohit Narayan

Abstract

Nowadays with the increasing popularity of internet, online marketing is going to become more and more popular. This is because, a lot of products and services are easily available online. Hence, reviews about these all products and services are very important for customers as well as organizations. Unfortunately, driven by the will for profit or promotion, fraudsters used to produce fake reviews. These fake reviews written by fraudsters prevent customers and organizations reaching actual conclusions about the products. Hence, fake reviews or review spam must be detected and eliminated so as to prevent deceptive potential customers. In our work, supervised and semi-supervised learning technique have been applied to detect review spam. The most apt data sets in the research area of review spam detection has been used in proposed work. For supervised learning, we try to obtain some feature sets from different automated approaches such as LIWC, POS Tagging, N-gram etc., that can best distinguish the spam and non-spam reviews. Along with these features sentiment analysis, data mining and opinion mining technique have also been applied. For semi-supervised learning, PU-learning algorithm is being used along with six different classifiers (Decision Tree, Naive Bayes, Support Vector Machine, k-Nearest Neighbor, Random Forest, Logistic Regression) to detect review spam from the available data set. Finally, a comparison of proposed technique with some existing review spam detection techniques has been done.

Keywords: Review Spam; Opinion Mining; Sentiment Analysis; Machine Learning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Nowadays e-commerce sites have become very popular because a lot of products and services and their reviews are easily available online. Online reviews have become a good way for users for their decision making while making any purchase from these sites. Today because of the popularity of e-commerce sites, spammers have made their target to these sites for review spam apart from other spam like email spam or web spam. Review spam means basically fake review that is written by fraudsters. Mostly e-commerce sites give section for review in order that users can write their opinion about products. There are also many review sites available like *TripAdvisor.com* which allows customer to write review for different hotels, *Zomato.com* which allows to write review about different restaurant, *Amazon.com* which allow users to write their opinion about their products and services, *Flipkart.com*, *Yelp.com* etc. Such type of content provided by web is named as user-generated content. User-generated content contains a lot of valuable and important information about the products and services. Since there is no control on the quality of this content on the web and hence, these promote fraudsters to write fake and wrong information about the products. These fake and wrong information written by fraudsters is called as review spam. Fake reviews prevent customers and organizations reaching actual conclusions about the products. Hence, it highly affects the e-commerce business. Hence, over the last few years, these review sites have been removing fake reviews about from their websites using their own spam detection technique.

Machine learning techniques have been more popular for spam detection. They uses supervised (required all data set labelled), semi-supervised (require very few data set labelled) and unsupervised (works for unlabelled data set) learning technique.

Generally, fake reviews are written for two purposes one for promoting some target objects (positive fake review or positive spam) and another for damage the reputation of other targets (negative fake review or negative spam).

Review 1: *We stay at Hilton for 4 nights last march. It was a pleasant stay. We got a large room with 2 double beds and 2 bathrooms, The TV was Ok, a 27' CRT Flat Screen. The coincierge was very friendly when we need. The room was very cleaned when we arrived, we ordered some pizzas from room service and the pizza was Ok also.The main Hall is beautiful. The breakfast is charged, 20 dollars, kinda expensive. The internet access (WiFi) is charged, 13 dollars/day. Pros: Low rate price, huge rooms, close to attractions at Loop, close to metro station. Cons: Expensive breakfast, Internet access charged. Tip: When leaving the building, always use the Michigan Av exit. Its a great view.* [1]

Review 2: *My husband and I satayed for two nights at the Hilton Chicago,and enjoyed every minute of it! The bedrooms are immaculate,and the linnens are very soft. We also appreciated the free wifi,as we could stay in touch with friends while staying in Chicago. The bathroom was quite spacious,and I loved the smell of the shampoo they provided-not like most hotel shampoos. Their service was amazing,and we absolutely loved the beautiful indoor pool. I would recommend staying here to anyone.* [2]

There are no clear indication from above two reviews that which review is fake and which are actual. But Review 1 is actual however Review 2 is fake. This can be only identified by data mining and machine learning technique.

[1] http://myleott.com/

[2] http://myleott.com/

1.2 Challenges in Review Spam Detection

- The fake reviews look like genuine reviews with a lot of similar keywords.
- Reviews are very subjective in nature and therefore can vary from a small paragraph to a long description.
- There are a number of review sites are available which provide space for writing reviews to reviewers, so it is very difficult to find out that reviewer has actual used the product and wrote the actual review or fake review.
- Both witty and sarcasm reviews present on a common place and hence, it is a very tough task to analyze such reviews.
- There is no labelled data set available online to train spam model. Even when people were asked to label reviews as spam, the concurrence rate was around 60% [9].

1.3 Problem Statement

Our main aim is to develop a model to detect review spams from review websites using review text. We have used the most apt data sets in the area of review spam detection research work. Both supervised and semi-supervised learning technique have been applied to obtain spam (reviw) from the data set. For supervised learning, we try to obtain some feature sets from different automated approaches that can best distinguish the spam and non-spam reviews. Along with these features, sentiment analysis and data mining technique have also been used. For semi-supervised learning, *PU-learning* algorithm along with different classifier are used to detect review spam from the data set. Finally, a comparison of proposed technique with some existing review spam detection techniques has been done.

1.4 Motivation and Objective

1.4.1 Motivation

From the last few years, e-commerce sites have become very popular because a lot of products and services and their reviews are easily available online. Online reviews have become a good way for users for their decision making while making any purchase from these sites. Today because of the popularity of e-commerce sites, spammers have made their target to these sites for review spam. Mostly e-commerce sites give section for review in order that users can write their opinion about products. There are also many review sites available like *TripAdvisor.com, Zomato.com, Amazon.com, Yelp.com* which allow users to write their opinion about their products and services. Such type of content provided by web is named as user-generated content. User-generated content contains a lot of valuable and important information about the products and services. Since there is no control on the quality of this content on the web and hence, these promote fraudsters to write fake and wrong information about the products. Fake reviews prevent customers and organizations reaching actual conclusions about the products. Hence, it highly affects the e-commerce business. Hence, over the last few years, these review sites have been removing fake reviews from their websites using their own spam detection technique. Machine learning techniques have been more popular for spam detection and hence, maintenance team of these websites use supervised (required all data set labeled), semi-supervised (require very few data set labeled) and unsupervised (works for unlabeled data set) learning technique.

1.4.2 Objective

Our main objectives are following:

- To develop a model to detect review spams from review websites using review text.
- To obtain some feature sets from different automated approaches that can best

distinguish the spam and non-spam reviews.

- To detect spam (review) from both labeled and partially labeled data set.
- Apply the concept of machine learning (supervised and semi-supervised learning), opinion mining, data mining and sentiment analysis.

1.5 Thesis Organisation

The thesis is organised into seven chapters. ***Chapter 1*** contains overview of review spam and the various challenges that occurs during review spam detection. ***Chapter 2*** highlights a literature review on review spam and its detection. It explains various types of reviews, spam, spammers and spam detection techniques. It includes some related work that has been done in the area of review spam detection. ***Chapter 3*** presents a brief description about supervised learning technique for review spam detection. It includes various feature sets and classifiers used in supervised spam detection. ***Chapter 4*** highlights semi-supervised learning techniques in review spam detection. ***Chapter 5*** presents proposed work for both supervised and semi-supervised learning technique. ***Chapter 6*** displays the results obtained using both supervised and semi-supervised learning technique. It also includes comparison of proposed model with some existing model of review spam detection in term of accuracy. ***Chapter 7*** presents conclusion of both the techniques and and their possible future directions.

Chapter 2

Literature Survey

2.1 What is Reviews?

A review is a feedback or evaluation of a service, a company or a product such as a movie (a movie review), a book (a book review), a mobile phone (a mobile phone review), a hotel (a hotel review), a restaurant (a restaurant review) etc. There are many review sites available (like TripAdvisor, Zomato, Yelp etc.) which allow users to write their opinion about the products and services. Anyone who writes review is called as reviewer.

2.2 Types of Reviews

2.2.1 Positive Reviews

If reviewers write positive things about the product or services, such review is called as Positive Reviews.
e.g. *The hotel is very nice. Room and services are too good. That is the awesome place to stay whole day and night. Rent is also affordable.*

2.2.2 Negative Reviews

If reviewers write negative things about the product or services, such review is called as Positive Reviews.
e.g. *Do not buy Samsung Galaxy S6. It is the worst mobile among all that i have used. No battery backup. Very bad camera quality. Touch pad is very hard.*

2.3 Types of Spams

2.3.1 Email Spam

If the sender sends unwanted and unsolicited email either directly or indirectly to user and there is no relationship of this email to the user is called as email spam. It is also called as junk email or unsolicited email. Email spam comes under the category of electronic spam. Example of such type of spam is phishing email [4] [5].

2.3.2 Web Spam

Web spam (also called as search spam) refers to the action of the deceptive search engine so that the rank of a specific website becomes more than it deserves [6].

2.3.3 SMS Spam

If someone transmits unwanted and unsolicited messages over communication media (i.e. cell phone) is called as SMS spam [7]. It comes under the category of electronic spam.

2.3.4 Comment Spam

Comment spams are generally written by spammers by posting their fake comments about the products and services.

2.4 What is Review Spam?

Today because of the popularity of e-commerce sites, spammers have made their target to these sites for review spam. Mostly e-commerce sites give section for review in order that users can write their opinion about products. There are also many review sites available which allow users to write their opinion about the products and services. Such type of content provided by web is named as user-generated content. User-generated content contains a lot of valuable and important information about the products and services. Since there is no control on the quality of this content on the web and hence, these promote fraudsters to write fake reviews.

2.5 Types of Review Spam

Review spams are generally categorized in three categories [1]:

Type 1 (Untruthful opinions): It is also divided into two sub-categories:

i. *Hyper spam:* Fraudsters write positive fake opinions to promote some targets.

ii. *Defaming spam:* Fraudsters write negative fake opinions to damage the reputation of some targets..

Type 2 (Reviews on brand only): Such type of review only focuses on brand name. Fraudsters write only about the brand, i.e. the manufacturers of the products rather than the products.

Type 3 (Non-reviews): Fraudsters write something that is totally unrelated to the products i.e. junk, such type of review spam comes under non-reviews. They have two forms:

i. advertisements, and

ii. irrelevant opinion.

Table 2.1 shows a basic idea about the nature of review spam and the quality of product. Hence from Table 2.1, we conclude that cell 1 and 4 promoting the target product. Cell 5 and 6 show neither promoting nor damaging the reputation of product, however cell 2 and 3 totally damaging the reputation of product.

Table 2.1: Nature of Review Spam with Respect to Quality of Product

Products	**Review Spam(Positive)**	**Review Spam(Negative)**
Good Quality	1	2
Bad Quality	3	4
Average Quality	5	6

2.6 Types of Spammers

A spammer is a person or a machine who writes spam (spam may be either email spam, web spam, review spam etc.). While finding fake review (spam) we can find two types of spammers. These are:

Individual Spammer:

- A single reviewer who uses different user-ids to register several times at a site for writing fake review.
- They write either only positive reviews about a product for promotion or only negative reviews for damage the reputation of competitors product.
- They give too high rating for the products.

A group of spammers:

- A group of reviewers who divide group in sub-group and each of these sub divisions work on different sites for writing fake reviews.
- Every spam member give lower rating to the product.
- The spammers write spam during launch time so that they can take the control over the sale of the product.

2.7 Spam Detection Techniques

Basically three machine learning techniques are used to detect spam. These are:

Supervised Learning Technique: In supervised learning technique, we need labelled reviews or data set. We extracted a set of features from these data set. These features are generally LIWC, POS tagging, N-gram and sentiment score. After these steps differnt classifiers like SVM, decision tree, logistic regression, Naive Bayes etc., are trained and accuracy is calculated. This is very simple form among all spam detection techniques.

Semi-supervised Learning Technique: Semi-Supervised learning technique is same as supervised learning technique with slightly differnce is that we do not need to label all the data set. If we have a very few labelled data set, then we can use such learning technique. Very few works have been done in this area.

Unsupervised Learning Technique: If we have unlabelled data set then we go for unsupervised learning technique where we find some hidden pattern. It includes k-mean clustering and mixture models etc.

2.8 Related Work

In the past, a lot of work has been done in the area of spam detection (email spam, web spam, SMS spam). If the sender sends unwanted and unsolicited email either directly or indirectly to user and there is no relationship of this email to the user is called as email spam. A very common attack in the area of email spam is Phishing attack. Phishing means attempts to steal personal information such as login id, password, credit card details etc. for malicious uses. Fette *et al.* [4]) have shown in their work that phishing attacks can be easily detected with high accuracy. They applied the concept of machine learning on some user generated feature sets such as IP based URLs, age of linked to domain names, non matching URLs, links to non-modal domain, HTML emails, number of links, number of domains, number of dots, contains javascript, spam-filter output. Proposed method was able to detect phishing websites or the emails those are used to direct victims to those websites. Authors evaluated their method on set of 860 phishing emails and, 6950 non-phishing emails, and achieved the accuracy over 96%.

The proposed work by Li *et al.* [5] is also based on email spam in which they investigated how to mix multiple email filters supported multivariate analysis so that they can provide a barrier to spam. Authors have shown that multiple emails filter for providing a barrier is more powerful than a single filter barrier alone. They have introduced an algorithm named W-Voting for calculating the accuracy. The algorithm consists of mainly two phases: training and filtering. The training phase is used to filter all multiple filters and the fitering phase is used for classification which classify new emails. The experiment was performed on two dataset *PU1* and *Ling Spam Corpus*. Author concluded that PU1 Corpus contained 43.77% spam however Ling Spam contained 16.63% spam.

Another type of spam that we studied is web spam. Web spam refers to the action of the deceptive search engine so that the rank of a specific website becomes more than it deserves [6]. Abernethy *et al.* [17] provided a graph based approach for web spam detection. They presented *WITCH* algorithm to detect web spam and also compared this algorithm to many existing algorithms and found that it is better than all those proposed techniques. Witch algorithm detects spam hosts or such pages on the Web. The datasets that have been used are collected from *WEBSPAM-UK2006*. These datasets contained 11,402 hosts out of which 7,473 were labeled and all are in *.uk* domain. Author have used 236 features in their proposed work such as: average length of word, total number of words in the title, PageRank, total number of neighbors and others proposed in [18] and [19]. Maximum accuracy achieved by this method is 95.3% using SVM classifier.

If someone transmits unwanted and unsolicited messages over communication media (i.e. cell phone) is called as SMS spam. Karami *et al.* [7] have used various content based features and LIWC features in their work to detect SMS spam. Content based features include capital words, spam words, SMS segments, unique words, URL, SMS frequency, using word ”call”, the rate of URL to SMS segments, the rate of spam words to unique words, the rate of spam words to SMS segments, the rate of capital words to unique words, the rate of capital words to SMS segments, the rate of spam words to SMS segments, the rate of unique words to SMS segments, the rate of URL to unique words. The Linguistic Inquiry and Word Count (LIWC) is a text analyzing tool which analyzes 80 different types of features like texts functional aspects, psychological concerns like emotion, perception and personal concerns like money, religion etc. [16]. They have taken 20 LIWC features include the rate of score of verbs to the score of all words, the difference between the scores of "Money" and the score of "Death", number of punctuations, number of pronouns, number of exclamation marks etc. Author collected datasets from Grumbletext website. The dataset consists 5,574 labeled short messages out of which 747 are SMS spam and 4,827 are non-spam. They applied different classifiers and concluded that accuracy varies from 92% to 98%.

Detection of opinion spam was first introduced by Jindal & Liu [1] in 2008. They categorized the review spam into 3 categories: *Untruthful opinions* (if fraudsters write positive fake opinions to promote some targets is called as hyper spam and if fraudsters write negative fake opinions to damage the reputation of some targets is called as defaming spam), *reviews on brands only* (fraudsters write only about the brand, i.e. the manufacturers of the products rather than the products) and *non-reviews* (fraudsters write something that is totally unrelated to the products, this may be either advertisements or irrelevant opinion). Authors introduced three types of feature in their proposed work i.e., review centric features, reviewer centric features and product centric features. On the basis of these features they built different models for detecting different types of review spam using different supervised learning techniques.

A behavioral approach was proposed by Lim *et al.* [8] to detect review spammers. They tried to find out some behaviors of spammers like they target products and try to maximize their impact. Proposed method are based on: single product having multiple reviews behavior, single product group having multiple reviews behavior, general deviation behavior and early deviation behavior. On the basis of these behaviors of the spammers they proposed a model to detect review spammers.

The first gold standard data set for study of review spam was created by Ott *et al.* [9] [10]. Ott *et al.* [9] in 2011 created data set containing 800 positive reviews out of which 400 are truthful and 400 are deceptive reviews. These reviews are taken from *tripadvisor.com*, *yelp.com* and Amazons popular Mechanical Turk crowdsourcing service. Also, Ott *et al.* [10] in 2013 generated data set containing 800 negative reviews out of which 400 are truthful and 400 are deceptive reviews through Amazons popular Mechanical Turk crowdsourcing service (mturk.com). Authors assigned three human judges and two meta judges for detection of spam on dataset described in [9] and got maximum accuracy 61.9% with f-score 69.7 for truthful and 48.7 in deceptive review. However same process when repeated on data set described in [10] then they achieved maximum accuracy 69.4% with f-score 68.8 for truthful and 69.9 in deceptive review. Authors have also applied some standard features like n-gram and linguistic features on same data set using supervised learning techniques to detect fake reviews.

Algur *et al.* [13] in their proposed work used the concept of similarity measure based on conceptual level and features of product that have been written by reviewers to detect a given review is spam or non-spam. According to them there are mainly three types of format uses to write review. These are: pros and cons in which reviewers only write pros and cons about the product, pros, cons and description in which reviewers write all the details about the pros and cons of the product, and free format in which reviewers can write anything (such type of format is used by Amazon.com). Authors first extracted features from the reviews and then created confusion matrix. After this, they calculated similarity measure on the basis of features and calculated accuracy. Proposed techniques give the accuracy of 57.29% for pros reviews and 30.00% for cons reviews or average 43.64% accuracy.

Feng *et al.* [14] showed in their work that product reviews contain a natural distribution of opinions and on the basis of this, they built a model to detect review spam. They collected data set from *Amazon.com* containing 400 reviews and took 80% of the data for training set and 20% of the data for test set and natural distribution of opinions was taken as features, achieved the maximum accuracy of 72.5% using SVM classifier.

Liu *et al.* [11] and Mukherjee *et al.* [12] have used the concept of frequent pattern mining in their work to detect reviewers group. Liu *et al.* [11] in their proposed work, they used basically three steps: (1) frequent pattern mining to find candidate groups in which they extracted review data and generated a set of transaction. Each transaction is treated a unique product containing reviewer id of belonging reviewer. After that they applied frequent pattern mining and output was group of candidate spammer. (2) Second steps containing calculation of spam indicator values using time window, group deviation, group content similarity, member content similarity early time frame, ratio of group size, group size and support count. (3) Last step was calculating rank using SVM Rank [20]. Following these three steps they detected group of spammer who work together to write spam. Mukherjee *et al.* [12] also followed the same steps but instead of SVM Rank, they used GSRank to detect a group of spammers.

Lim *et al.* [8] proposed a model that is based on behavior of spammers. They used to assign a rank to spammer on the basis of behavior scoring method and they detect spammers according to that rank. Authors collected data set from amazon.com and applied the concept of both behavior scoring method and supervised learning technique to detect review spammers.

A lot of work has been done in supervised learning technique. But the drawback is we need to label all the data set. To overcome such problem Fusilier *et al.* [21] applied the concept of semi-supervised learning technique to detect review spam detection. Authors used the data set created by ott *et al.* [9] containing 800 positive reviews out of which 400 are deceptive and 400 are truthful. They took 160 data set as a test set which is labelled and and for training took 520 unlabelled data set and combination of 20, 40, 60, 80, 100, 120 as a positive instances. After that they applied *PU-learning* algorithm on these positive and unlabelled instances to calculate accuracy. They used one-class, naive bayes and SVM classifier in their work.

Liu *et al.* [22] in their proposed work also used the concept of semi-supervised learning technique to detect spam. They divided data set into two set of classes. A prticular data set comes into a class named P, a large number of data set come into an another class called M. Such technique is called as *partially supervised classification.* They used *Expectation-Maximization* or *EM* algorithm to identify class P from class M. *EM* algorithm generates a sequence of solutions. For each solution they used naive bayes classifier to calculate accuracy.

Karimpour *et al.* [32] used both *PU-learning* and *EM* algorithm along with semi-supervised technique but to detect web spam. They used *WEBSPAM-UK2007* data set which is publically available. It is based on *.uk* domain which is done in May 2007. It consists 105 millions pages and 3 billions link in 114,529 hosts. Their training set consists 3,848 hosts. Authors applied both algorithm on these data set and achieved F-score 0.86 and also compared their result with other exixting techniques like Naive Bayesian, Bayesian Network etc.

Chapter 3

Supervised Learning Technique

3.1 Feature Sets from Different Automated Approaches

3.1.1 Linguistic Inquiry Word Count

The Linguistic Inquiry and Word Count (LIWC) is a text analyzing tool which analyzes 80 different types of features like linguistic dimension (i.e. words count, words per sentence etc.), psychological processes (i.e. positive emotion, negative emotion, perceptual processes, biological processes etc.), personal concerns (i.e. home, money, religion, death etc.) and spoken categories (i.e. assent, nonfluencies, fillers etc.) [33].

3.1.2 POS Tags

Work in linguistics has already proved that the distribution of frequency of parts of speech (POS) tagging of any text is directly dependent on the genre of that text [Biber et al., 1999; Rayson et al., 2001]. Hence, according to this approach, feature made for every review is primarily based on the frequency of every POS tag for testing relationship this feature and actual and fake reviews.

3.1.3 N-gram Feature

In n-gram feature, we select n contiguous words from a text as a feature. If one word at a time is being considered as a feature then, it is called as unigram; if two contiguous words at a time is being selected then, it is bigram and similarly if we select three contiguous words at a time as a features then, it is called as trigram. These features help us to model all the content and its context. In this work, only unigram as a feature has been used [9].

3.1.4 Sentiment Score

The negative spammers generally used to write more negative words in their review like horrible, disappointed etc. and hence, show more negative sentiment than a truthful negative review. Similarly, positive spammers used to write more positive words like beautiful, great etc. and show more positive sentiment than an actual positive review.

3.2 Classification Techniques

Features from above approaches are used to train 6 classifiers i.e. Decision Tree, Naive Bayes, Support Vector Machine (SVM), k-NN, Random Forest and Logistic Regression.

3.2.1 Decision Tree

Decision tree is one of the simplest classification algorithm used in machine learning technique. It is based on tree structure in which internal nodes represent test sets and leaves represent class label (decision that is taken after calculating all attributes). Each branch represents output of test. A decision tree contains three types of node i.e. root, branch and leaf node [26]. These are basic steps of decision tree algorithm:

Steps:

1. Construct the tree in top-down divide and conquer recursive manner.

2. Initially, put all training set at root node.

3. Partition the input data recursively based on selected attributes.

4. Select test set at each node based on statical measure i.e. information gain.

5. These are terminating conditions:
   - All input are member of same class.
   - There are no input for partitioning.
   - No sample is left.

3.2.2 Naive Bayes

It is a probabilistic classifier based on Bayes theorem with strong assumption that all the features are not dependent on each other. Such assumption is known as class conditional independence. An important advantage of Naive Bayes is that it requires a very less amount of training data set for classification. It is one of the fast classifier since it works in a single scan [23].

Bayes theorem give a way of finding posterior probability $P(c \mid x)$ from $P(x \mid c)$, P(c) and P(x). Naive bayes classifier consider that effect of a predictor x (only value of x) on a give class c is not dependent on other predictors. Following is the formula for calculating posterior probability:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

where:
$P(c \mid x)$: posterior probability of target class on given attribute.
$P(x \mid c)$:probability of predictor on given class (likehood).
P(c): prior probability of the class.
P(x): prior probability of the predictor.

3.2.3 Support Vector Machine

Support Vector Machine (SVM) also known as Support Vector Network in machine learning is a supervised learning technique used for classification and regression. In simple, given a training examples set, each of them marked belonging to one of two categories. SVM training algorithm constructs a model that decides and assigns a new example falls into one category or the other. Hence SVM classifier is represented by a separating hyperplane. This hyperplane generated from training set then classifies data from test set [24] [25].

Suppose we have two classes shown in Figure 3.1, denoted by square and circle and two axises x and y denoting features. SVM finds a hyperplane that classify all the training set into two classes.

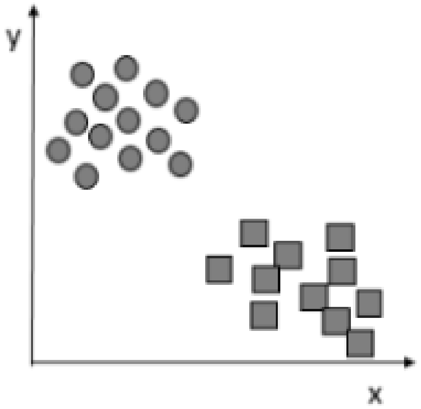

Figure 3.1: SVM Classifier with Two Classes

Figure 3.2 denotes some separable hyperplane according to SVM classifier. Among all hyperplanes, the best choice will be the hyperplane that leaves maximum margin from both the classes.

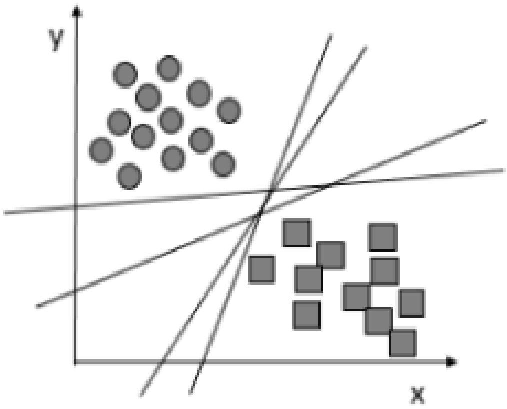

Figure 3.2: SVM Classifier with Hyperplane

3.2.4 k-Nearest Neighbor

k-Nearest Neighbor (k-NN) classifier is the simplest among all the classifiers and is used for both classification and regression. In this, input consists of k closest training sets in the feature space. Its output is class membership. If k=1, then object is directly assigned to single nearest neighbor class else object is assigned to that class in which object is most common in its k nearest neighbor.

3.2.5 Random Forest

Random Forest classifier works where Decision Tree fails. In other word, if trees are grown very deep or taken irregular shape i.e. overfit training set then for averaging multiple deep decision tree, random forests work on different part of same training set by generating multitude of decision trees during training time. The major belief with random forest method is that most of the tree can provide correct prediction of class for most of the data. Figure 3.3 shows that three having node Y provide correct prediction because of their majority and tree having node N provide wrong prediction [23].

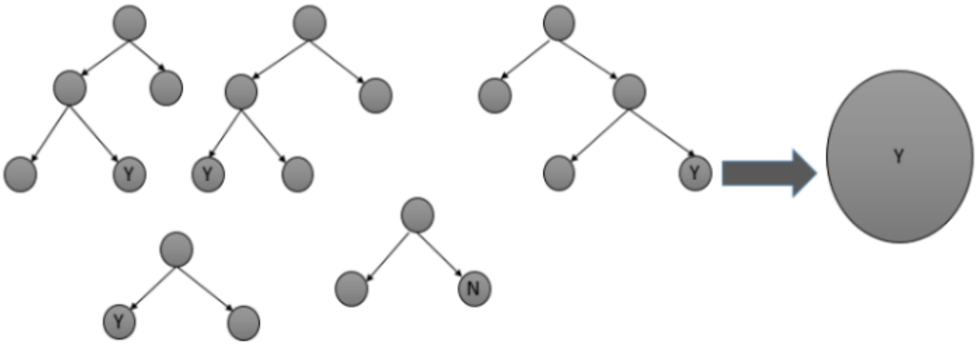

Figure 3.3: Random Forest Classifier

3.2.6 Logistic Regression

Logistic Regression, also known as logit regression is very popular technique used for classification and regression. This is simple and provides good performance. It is a discriminative probabilistic model that operates over vector inputs which are real valued and predicts the probability of an outcome that can have only two values (i.e. a dichotomy). The dimension of input vectors are features having no restriction against them being correlated.

Logistic Regression produces a logistic curve, which values lies between 0 and 1 as shown in Figure 3.4. Logistic regression is similar to linear regression, but the curve is constructed using natural logarithm rather than probability. The predictors do not have to be normally distributed or equal varience in each group.

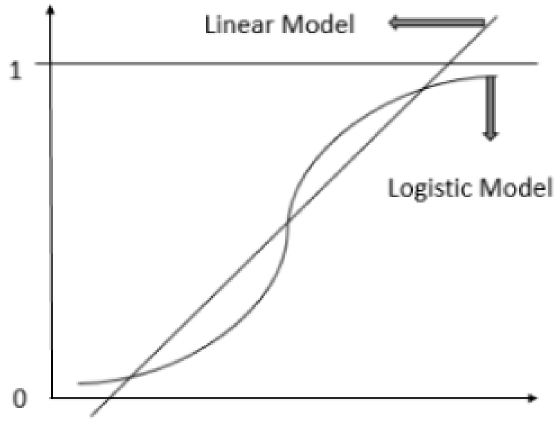

Figure 3.4: Logistic Model

Chapter 4

Semi-supervised Learning Technique

Semi-supervised learning technique is a machine learning technique that uses a large amount of unlabeled data and a very few labeled data set for training. Semi-supervised learning lies between supervised learning (completely labeled data) and unsupervised learning (completely unlabeled data). Many researchers found that if a large amount of unlabeled data, when used with a few labeled data set, can produce good accuracy in term of learning problem.

4.1 Assumptions in Semi-supervised Technique

There are three main assumptions in semi-supervised learning technique which make it simpler and easier. These are:

4.1.1 Smoothness Assumption

In the case of supervised learning, output varies smoothly with the distance on the basis of prior belief. In case of semi-supervised learning, density of input is also taken into account. Hence, we can say that if two points x and y are in a high density region are considered to be close rather than x is in high density region and y is in low density region or vice-versa. Figure 4.1 shows that x and y are close since they in high density region and x and z or y and z are not close since one is in high density region, others in low density.

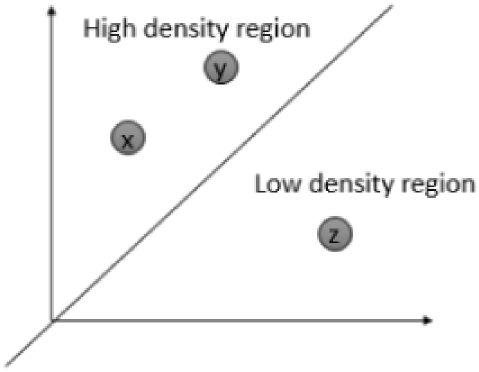

Figure 4.1: Smooth Assumption

4.1.2 Cluster Assumption

Since points of each class form a cluster. Now assumption is that if two points x and y are in same cluster are considered to be in the same class, however two points in different cluster are not considered in the same class. Figure 4.2 shows x and y are member of same class however x and z are not.

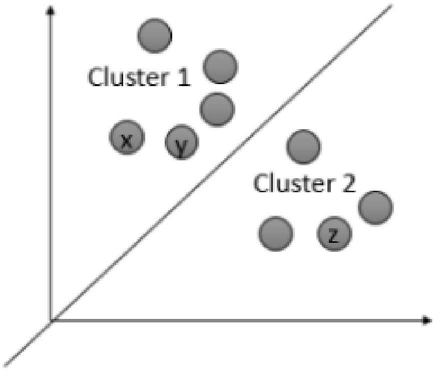

Figure 4.2: Cluster Assumption

4.1.3 Manifold Assumption

Manifold assumption is different from above two assumptions. The assumption is that, a high dimensional data lies in approximately low dimension manifold. Such assumption is useful when we have a high dimensional data and it is hard to model.

4.2 Semi-supervised Learning Methods

There are many semi-supervised learning methods are used in the area of machine learning. Some of them are generative method, self training method, co-training method, graph based method etc [27].

4.2.1 Generative Method

Generative methods are one of the oldest semi-supervused learning method. This method is based on p(x,y) = $p(x \mid y)$p(y), where $p(x \mid y)$ is a recognizable distribution. In this, first mixture component of large volume of unlabeled data is recognized then perform labeling. It is an inductive mixture with very less parameter.

4.2.2 Self-Training

Self-Training is very common method used in semi-supervised learning method. In this, first classifier is trained with few selected labeled data set and then classifier classifies unlabeled data sets. Now predicted data sets are append with selected label data and then classifier is retrained with this data set and the process is repeated. The process of retraining the data again and again is called as bootstrapping or self-teaching. These are the basic steps for Self-Training method [29].

Steps:

1. The classifier is trained with few labeled data (completely positive and completely negative).
2. The classifier is run with that data set which is weak label on the basis of maximum likelihood ratio.
3. Unlabel data set is label with the output of detector.
4. A subset is selected from these labeled data set using some features metric.
5. Process is repeated until all data set to be trained.

4.2.3 Co-Training

Co-Training method is based on different features containing by data. It is assumed that each sample consists two different feature sets that give different information about the instances. These two views should be conditional independent. From each view, class of instances are predicted accurately. Co-training begins with learning an individual classifier for each view. With the help of these classifiers, we label unlabeled data set [30].

4.2.4 Multiview Learning

It is extended version of Co-Training method in which we use multiple views rather than two views. Rest steps are same as Co-Training method.

4.2.5 Graph Based Method

Graph based method is totally based on graph where each node represents data set (labeled and unlabeled) both and edges represent similarity between data. This method follows smoothness assumption. One important advantage with this method is that it does not require any parameter. This method is transductive and discriminative in nature.

Chapter 5

Proposed Work

5.1 Supervised Learning Technique

5.1.1 Dataset Description

The contents are under publication.

5.1.2 Features Used

The contents are under publication.

5.1.3 Proposed Model

The contents are under publication.

5.2 Semi-supervised Learning Technique

5.2.1 Dataset Description

The contents are under publication.

5.2.2 Proposed Model

The contents are under publication.

Chapter 6

Results and Discussions

6.1 Supervised Learning Technique

The contents are under publication.

6.1.1 Performance Analysis

The contents are under publication.

6.2 Semi-supervised Learning Technique

The contents are under publication.

Chapter 7

Conclusion and Future Work

7.1 Supervised Learning Technique

In this work, three sets of features i.e. LIWC, POS Tag and N-gram from different automated approaches along with the sentiment score have been used. These feature sets have been applied individually as well as in some combinations to train different classifiers. Six classification algorithm were employed such as: Decision Tree, Naive Bayes, SVM, k-NN, Random Forest and Logistic Regression. Our experimental results reveals that Logistic Regression outperforms other classifiers. In the case of individual feature set, unigram gives maximum accuracy of 75.62% with F-score 76.07. However, for combinations unigram and LIWC along with sentiment score gives accuracy of 86.25% with F-score 86.72 and that is maximum. At last, we have compared our proposed technique with some existing review spam detection techniques on the basis of their accuracy which shows our technique gives better result than others.

7.2 Semi-supervised Learning Technique

For semi-supervised learning, we applied *PU-learning* algorithm along with six different classifiers (Decision Tree, Naive bayes, SVM, k-NN, Random Forest, Logistic Regression) to detect review spam from the data set. Different sub-corpa from data sets have been taken. For building test set, first we randomly selected 160 opinions, out of which 80 are deceptive and 80 are truthful. The rest 640 opinions have been used for 3 different size of training sets. They consist 40, 80 and 120 positive instances (deceptive opinion) respectively. In all the cases, 520 unlabeled instances are fixed.

Now, PU-learning algorithm has been used for review spam detection. Maximum accuracy we have achieved is of 78.12% with F-score 76.67 when used 80 examples of deceptive opinions from datasets as training set with 520 unlabeled dataset using k-NN classifier.

7.3 Future Work

In this work, we have used supervised learning technique where required all the data set to be label and semi-supervised learning technique where few data set are supposed to be labeled . But in the research area of review spam detection a very few label data set are available and hence, in future the same work can be extended for unsupervised learning technique to overcome the unavailability of labeled data sets.

Bibliography

[1] Jindal, Nitin, and Bing Liu. "Opinion spam and analysis." *Proceedings of the 2008 International Conference on Web Search and Data Mining.* ACM, 2008.

[2] Crawford, Michael, et al. "Survey of review spam detection using machine learning techniques." *Journal Of Big Data* 2.1 (2015): 1-24.

[3] Heydari, Atefeh, et al. "Detection of review spam: A survey." *Expert Systems with Applications* 42.7 (2015): 3634-3642.

[4] Fette, Ian, Norman Sadeh, and Anthony Tomasic. "Learning to detect phishing emails." *Proceedings of the 16th international conference on World Wide Web.* ACM, 2007.

[5] Li, Wenbin, Ning Zhong, and Chunnian Liu. "Combining multiple email filters based on multivariate statistical analysis." *Foundations of Intelligent Systems.* Springer Berlin Heidelberg, 2006. 729-738.

[6] Spirin, Nikita, and Jiawei Han. "Survey on web spam detection: principles and algorithms." *ACM SIGKDD Explorations Newsletter* 13.2 (2012): 50-64.

[7] Karami, Amir, and Lina Zhou. "Improving static SMS spam detection by using new content-based features." (2014).

[8] Lim, Ee-Peng, et al. "Detecting product review spammers using rating behaviors." *Proceedings of the 19th ACM international conference on Information and knowledge management.* ACM, 2010.

[9] Ott, Myle, et al. "Finding deceptive opinion spam by any stretch of the imagination." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.* Association for Computational Linguistics, 2011.

[10] Ott, Myle, Claire Cardie, and Jeffrey T. Hancock. "Negative Deceptive Opinion Spam." *HLT-NAACL.* 2013.

[11] Mukherjee, Arjun, et al. "Detecting group review spam." *Proceedings of the 20th international conference companion on World wide web.* ACM, 2011.

[12] Mukherjee, Arjun, Bing Liu, and Natalie Glance. "Spotting fake reviewer groups in consumer reviews." *Proceedings of the 21st international conference on World Wide Web.* ACM, 2012.

[13] Algur, Siddu P., et al. "Conceptual level similarity measure based review spam detection." *Signal and Image Processing (ICSIP), 2010 International Conference on.* IEEE, 2010.

[14] Feng, Song, et al. "Distributional Footprints of Deceptive Product Reviews." *ICWSM* 12 (2012): 98-105.

[15] Peng, Qingxi, and Ming Zhong. "Detecting spam review through sentiment analysis." *Journal of Software* 9.8 (2014): 2065-2072.

[16] Harris, C. "Detecting deceptive opinion spam using human computation." *Workshops at AAAI on Artificial Intelligence.* 2012.

[17] Abernethy, Jacob, Olivier Chapelle, and Carlos Castillo. "Graph regularization methods for web spam detection." *Machine Learning* 81.2 (2010): 207-225.

[18] Castillo, Carlos, et al. "A reference collection for web spam." *ACM Sigir Forum.* Vol. 40. No. 2. ACM, 2006.

[19] Ntoulas, Alexandros, et al. "Detecting spam web pages through content analysis." *Proceedings of the 15th international conference on World Wide Web.* ACM, 2006.

[20] Joachims, Thorsten. "Optimizing search engines using clickthrough data." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2002.

[21] Hernndez, D., et al. "Using PU-learning to detect deceptive opinion spam." *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.* 2013.

[22] Liu, Bing, et al. "Partially supervised classification of text documents." *ICML.* Vol. 2. 2002.

[23] Han, J. and Kamber, M. "Data Mining: Concepts and Techniques." *2nd Edition. Morgan Kaufmann Publishers,* San Francisco, USA. (ISBN-55860-901-6), 2006.

[24] R. Duda, P. Hart, and D. stork, "Pattern Classification." *2nd Edition, Wiley Interscience,* 2001.

[25] E. Frank, and I. Witten, "Data Mining: Practical Machine Learning Tools and Techniques." *2nd Edition, Morgan Kaufmann, San Francisco,* 2005.

[26] Galathiya, A. S., A. P. Ganatra, and C. K. Bhensdadia. "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning." *International Journal of Computer Science and Information Technologies* 3.2 (2012): 3427-3431.

[27] Prakash, V. Jothi, and Dr LM Nithya. "A Survey On Semi-Supervised Learning Techniques." *arXiv preprint arXiv:1402.4645*, 2014.

[28] Zhu, Xiaojin. "Semi-supervised learning literature survey." 2005.

[29] C. Rosenberg, M. Hebert, and H. Schneiderman, Semi-Supervised Self-Training of Object Detection Models, *Proc. Seventh Workshop Applications of Computer Vision,* vol. 1, pp. 29-36, Jan.2005.

[30] Blum, Avrim, and Tom Mitchell. ”Combining labeled and unlabeled data with co-training.” *Proceedings of the eleventh annual conference on Computational learning theory.* ACM, 1998.

[31] Zhou, Dengyong, Jiayuan Huang, and Bernhard Schlkopf. ”Learning from labeled and unlabeled data on a directed graph.” *Proceedings of the 22nd international conference on Machine learning.* ACM, 2005.

[32] Karimpour, Jaber, Ali A. Noroozi, and Somayeh Alizadeh. ”Web Spam Detection by Learning from Small Labeled Samples.” *International Journal of Computer Applications 50.21,* 2012.

[33] James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. ”The Development and Psychometric Properties of LIWC2007.” *LIWC.net, Austin, Texas 78703 USA in conjunctin with the LIWC2007 software program.*, 2007.

Dissemination of Work

Accepted

1. Rohit Narayan, Jitendra Kumar Rout and Sanjay Kumar Jena, ”Review Spam Detection using Opinion Mining”, *4th International Conference on Advanced Computing, Networking, and Informatics (ICACNI)* (SPRINGER), 2016.

2. Rohit Narayan, Jitendra Kumar Rout and Sanjay Kumar Jena, ”Review Spam Detection using Semi-supervised Technique”, *4th International Conference on Advanced Computing, Networking, and Informatics (ICACNI)* (SPRINGER), 2016.