****

BERZIET UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

**Review Spam Detection Based on Hybrid Features**

Prepared by

Muhannad Shamasneh

Mohammad Hussein

Mohammad Abu Mahmoud

Supervised by

Hanna Bullata, Aziz Qarush.

A graduation Project submitted to the Department of Electrical

and Computer Engineering in partial fulfillment of the requirements

for the degree of B.Sc. in Electrical Engineering

BIRZEIT

April – 2018

الإهداء

إلى من علمونا معنى الحب والتضحية فلم يبخلوا يوماَ علينا بشيء، إلى من شدو على ايدينا لنصبح يوماً ما نطمح اليه، إلى والدينا الأعزاء ..

إلى من يكللون لنا دربنا بالحرية، وإلى كل اسرة فلسطينية ضحت وناضلت وقدمت الشهداء، الى شهدائنا الابرار وعائلاتهم ..

إلى من انار لنا طريقنا الى الصواب، الى من انتفعنا بخبرته فأظهر لنا بسماحته تواضع العلماء، الى استاذنا الكريم عزيز قرعوش..

إلى من عشنا معهم اجمل الذكريات وشاركنا المشاعر نفسها بحلوها ومرها، الى من دعمونا وساندونا وتركوا فينا بصمة لا تنسى، الى اصدقائنا الاعزاء..المستخلص

**ازدادت بشكل ملحوظ مؤخراً المواقع التجارية التي تسمح للمستخدمين شراء السلع او الخدمات عن طريق الانترنت. هذه المواقع تحتوي على معلومات ذات أهمية كبيرة للمشترين و أصحاب السلع أو الأعمال، ألا وهي الآراء التي يكتبها المستخدمون على هذه السلع أو الخدمات بحيث تعكس تجربتهم مع هذه السلعة ورأيهم فيها من حيث ايجابياتها أو سلبياتها. هذه الآراء تعد مهمة جداً للمستخدمين وأصحاب الأعمال، فمن ناحية المستخدم فانه قبل شراء اي منتج فإنه غالبا ما يلجأ الى قراءة آراء الناس حول هذا المنتج وان هذه الأراء تؤثر على قرار الشراء عنده. أما بالنسبة لأصحاب الأعمال فان حجم مبيعاتهم يتأثر زيادةً او نقصاناً من خلال هذه الآراء ان كانت ايجابية أو سلبية. ولذلك فقد لجأ بعض المستخدمين أو أصحاب الأعمال الى كتابة أراء مزيفة لترويج لمنتجاتهم أو لنقد منتجات المنافسين وذلك للتأثير على القرار الشرائي للمستخدم.**

**في هذا التقرير فاننا نعرض لكم طريقة كشف الآراء المزيفة والمستخدمين الذين يكتبون آراء مزيفة باستخدام هجين من المميزات التي تميز هذه الآراء الزائفة عن غيرها من الآراء، وايضاً كشف الذين يقومون بكتابة آراء مزيفة بحيث اننا قمنا بتجهيز البيانات المتوفرة لدينا ومن ثم استخلصنا منها عدة ميزات تميز النص المزيف والأشخاص المزيفين. ثم استخدمنا ثلاثة طرق لتصنيف البيانات: الأولى ، وهي الطريقة التقليدية لتصنيف البيانات باستخدام التعليم الآلي. الطريقة الثانية تعتمد على نموذج احتمالي الذي يعتمد على خليط بين مميزات الأشخاص المزيفين والنص المزيف، والطريقة الثالثة تعتمد على نظام النتيجة بحيث يتم اعطاء نتيجة للنص أو الشخص بناءً على الميزات له.**

## Abstract

Commercial websites which allow people to purchase products and services online - through public domain - are rapidly grown recently. These websites contain a lot of user reviews regarding the products and services they provide. These reviews contain valuable information about the products themselves and customer experience. Therefore, this source of information has become very critical for customers and businesses. For a customer, his purchase decision is highly affected by other customers’ reviews as he tends to read them and compare them with other products’ reviews before making any purchase decision. For businesses, their sales can be significantly affected by these reviews. Positive reviews attract more customers while negative reviews affect the number of customers who intend to buy the product. Due to that huge impact; some users and businesses illegally tend to write fake reviews to promote their products or demote competitors’ products; the thing that directed researches to take this topic into consideration nowadays.

In this report, we will introduce a methodology based on hybrid features for detecting spam reviews and spammers. It focuses on extracting a set of informative features of both review text and reviewer account after preprocessing the reviews. Feature selection methods will be then used to select the most appropriate feature from the set of all extracted features. Later on, we will define the models for detecting spam reviews or spammers including: classical classification using supervised machine learning, probabilistic model optimized by genetics and score-based model optimized by genetics.

Table of Contents

[Chapter 1 1](#_Toc514670011)

[1.1. Introduction 1](#_Toc514670012)

[1.1.1. Introduction and motivation 1](#_Toc514670013)

[1.1.2. Problem Statement 3](#_Toc514670014)

[1.1.3. Methodology 3](#_Toc514670015)

[1.1.4. Report Outline 4](#_Toc514670016)

[Chapter 2 5](#_Toc514670017)

[2.1. Related Work 5](#_Toc514670018)

[2.1.1. Spam Review Detection 5](#_Toc514670019)

[2.1.2. Spammers Detection 10](#_Toc514670020)

[Chapter 3 15](#_Toc514670021)

[3.1. Proposed Work 15](#_Toc514670022)

[3.1.1. Notation 15](#_Toc514670023)

[3.1.2. Proposed Work 17](#_Toc514670024)

[3.1.3. Features Extraction 17](#_Toc514670025)

[3.1.4. Review Classification 25](#_Toc514670026)

[Chapter 4 28](#_Toc514670027)

[4.1. Datasets, Tools and Evaluation Measures. 28](#_Toc514670028)

[4.1.1. Dataset 28](#_Toc514670029)

[4.1.2. Tools 30](#_Toc514670030)

[4.1.3. Evaluation Framework 31](#_Toc514670031)

[Chapter 5 35](#_Toc514670032)

[5.1. Conclusion 35](#_Toc514670033)

[Chapter 6 36](#_Toc514670034)

[6.1. References 36](#_Toc514670035)

## List of Figures

[Figure ‎1.1: Type of Spams. 1](#_Toc514669747)

[Figure ‎1.2: Example of Fake Review Type1. 2](#_Toc514669748)

[Figure ‎2.1: Spam detection techniques based on used features. 5](#_Toc514669749)

[Figure ‎3.1: Flow-Chart of our proposed work. 17](#_Toc514669750)

[Figure ‎3.2: Review preprocessing stages. 19](#_Toc514669751)

[Figure ‎3.3: Semantic feature extraction stages. 20](#_Toc514669752)

[Figure ‎3.4: Genetics algorithms basic steps. 27](#_Toc514669753)

[Figure ‎3.5: Chromosome of two genes representation. 27](#_Toc514669754)

[Figure ‎4.1: Confusion Matrix. 32](#_Toc514669755)

[Figure ‎4.2: ROC curve evaluation measure. 34](#_Toc514669756)

# 

## Introduction

### Introduction and motivation

The Internet is now considered to be the most demanded service all around the world. People now can communicate and share their knowledge very easily through different ways such as writing comments in the social sites or review the products they have used describing their experience using that product. As a result, the reviews became very important for the customer and the business --as many online shopping paradigms have spread out which help users to review products and share their reviews on social media. Such reviews impose a huge impact on product sales since people tend to check the reviews before deciding to purchase products. Since reviews can be provided by anyone on the public domain, users misuse the reviewing services by providing fake reviews to promote or demote particular products. Later in this research, fake reviews will be referred as ‘spam’ and individuals who provide spam reviews will be referred as spammers.

Figure ‎1.1: Type of Spams.

Spams are mainly classified as web spam, Email spam and spam reviews as shown in the Figure ‎1.1. Web spam is defined as any web page content that used to improve the ranking of this page more than the others. Web spam is also classified as content spam which means adding some phrases to the web page document to increase its rank and link spam by posting the link of site that the spammer wants to promote or demote. While the E-mail spam is defined as unwanted or unrelated which may be used for advertisement or for different purposes. However, the E-mail spam can be detected easily because it has a special component so it has less detriment than web spam. Finally, the spam reviews (i.e. fake reviews provided by spammers to promote or demote particular product) is the type to be covered in this research. Spam reviews are classified by Jindal et al [1] into mainly three types:

1. **Type 1 Untruthful reviews:** Providing fake opinions about products to promote or demote products by undeserving positive and negative opinion respectively which is written with hidden motives. This type of reviews is also called fake reviews. Figure ‎1.2 shows an example on type 1.

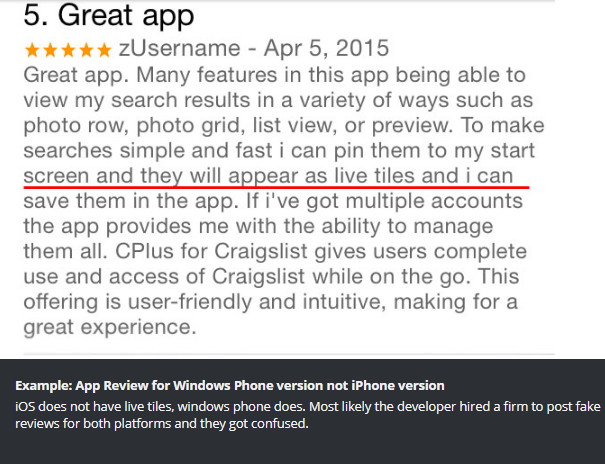


Figure ‎1.2: Example of Fake Review Type1.

1. **Type 2 Review on brands:** this type of spam not sufficient to concentrate on the products or services, since it can be on the cost of product, the sellers, the manufacturers of the products and other aspects. Although it supposed to be on true, but they are considered as spam as they cannot provide any helpful about the concerned product. (e.g. “I hate Dell product and I’ll never buy their products”).
2. **Type 3 Non-Reviews:** it also called Non-related reviews which are considered as spam since it can be an advertisement or unhelpful opinions such as answers, question and any random text. (e.g. “<3 <3 <3 <3 <3” and “how to make a deliver request?”).

Due to the big challenge in the business today some companies illegally tend to hires employees to write fake reviews to promote their products or demote other competitor product. So, there is a big concern from the customer and the business to have real reviews, for these reasons many sites tend to filter spams and display only the real reviews in order to maintain credibility. Handling the problem of review spam can be done in two ways manually filtering which is time consuming and will not be effective enough because the spammers now use new methods to write spam reviews that’s hard to detect by human since the ordinary people cannot measure the semantic similarity, also reliance of multiple features of identifying spams is limited in case of manually filtering, the second way is automated filtering based on a set of features extracted from review text and reviewer account to identify the spam reviews which is fast and more efficient than manually filtering. One of the examples from websites that uses automated filtering is YELP[[1]](#footnote-1) since they use filters to detect the spam reviews made on the website and just showing to the user’s honest reviews.

### Problem Statement

Spam reviews problem is a challenging problem; as most people nowadays use online shopping to make purchase decisions based on product reviews, especially those reviews that come first in a website. Therefore, it is crucial for customers to find honest reviews in order to make the right decision.

Our goal is to develop an automated solution that filters the reviews out based on extracted features from the review content, reviewer behaviors and information. This application will help the customers to trust the websites that use spam filters to display the honest reviews only.

### Methodology

There are four major steps to create our spam detection model which is based on a set of features extracted from review text and reviewer account.

1. Reviews gathering and annotation.
2. Preprocessing the reviews text.
3. Extracting hybrid features from both review text and reviewer account and representing them into a features vector.
4. Classifying reviews into spam and non-spam using three approaches.

a. Classical supervised machine learning techniques.

b. Probabilistic model optimized by genetics.

c. Score-based model optimized by genetics.

### Report Outline

***Chapter 2:*** Introduces the related work of other people in this field and the different approaches they used to solve the problem.

***Chapter 3:*** Discusses our proposed work, features extraction from reviews information, and reviews classification models.

***Chapter 4:*** Compares between different datasets and tools that will be used to evaluate and validate the results.

Lastly, ***chapter 5***: Concludes all the overall results compared with related work, introduces the future work, with possible improvements and enhancements.

# 

## Related Work

In this chapter we introduce the work of other people in this field. We classify this chapter into two main sub sections: First section is about spam review detection; which identifies spam reviews based on the review data and metadata only. Second section is about identifying spammers (i.e. reviewers who involved in spamming activities) as in the Figure ‎2.1 below.

Figure ‎2.1: Spam detection techniques based on used features.

### Spam Review Detection

These studies are about detecting spam reviews using the features extracted from the review text and its metadata. We divide the work here into 3 subsections: Identifying fake reviews (Type 1), review on brands (Type 2), non-related reviews (Type 3) and hybrid from them.

#### Fake Reviews

**Jindal et al.** (1)focused on identifying the unexpected patterns of reviewers which in turn used to detect the spam reviews, so this paper divides the dataset into two classes spam or not spam by states if a reviewer wrote only a highly negative reviews on a specific product while many others wrote in general some positive reviews so this review is highly expected to be spam. Hence, this paper is to find an application that is based on a heuristic program to find the unexpected behavior, so this approach firstly defining several types of expectation based on the data itself by giving joint probabilities to determine the unexpectedness (deviation from expectations), then to give the unexpectedness a measure to rank the review. It is then based on different measurements to identify if this review is spam or not.

**Ott et al.** (2)adopt lexical model and used Bag of words which are the most common features used in this model. This consists of group of words that is extracted from the reviews text from which n-gram features can be extracted, the researchers developed a model relying on the part of speech, Bigram and Linguistic Inquiry and Word Count features, they also used the frequency of n-gram in their model to detect the un-truthful reviews (Type 1). They used a golden standard dataset from amazon Mechanical Turk about hotels, and also a data of honest reviews from Tip Adviser, then they used unsupervised classifier using SVM classifier to divide all the opinions (deceptive and honest reviews) into positive and negative. This step is used to separate the negative and positive data to achieve 86% accuracy and when they remove the separation they achieved 84% accuracy. This implies that separation of data is improve the accuracy from 84% to 86%.

**Qian & Lie et al.** (3)Used conceptual level Similarity Measure-based Review Spam Detection in this approach of spam detection, the customer reviews is collected then it passed through a preprocessing stage then to a stop word removal stage and it is then try to find the duplicate of near duplicate of reviews and this is done by finding a matched features between the reviews and give a high score as the review is more matching other reviews. Then according to this score, we determine if the review is spam or not since of this score of matched features is less than a specific threshold its classified as not spam. There is a second case of not spam if the review is unique. and these two types of not spam classification is done by three steps: the first step is to extract features from reviews and storing them in features database, then the matched features that extracted previously used to construct a feature matrix, then to matching the feature calculation between reviews and that is done by calculating a score for different review pairs, and then to categorized as spam or non-spam based on threshold value.

**Ahmed** (4)proposed an approach to detect opinion spam based on text analysis using semantic similarity metrics and n-gram features since it effective at detecting content similarity, then to validate the effectiveness of the n-gram features at detecting spams. They used the unique reviews that written by different users (singular reviews), they preprocessing the data, remove the stop words and transform the tokens into standard form(stemming). Four types of n-gram features was extracted (bigrams, trigram, and four-gram). Then these features were used in machine learning classification by applying many classification methods namely, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Linear Support Vector Machine (LSVM), Decision tree (DT) and Stochastic Gradient Descent (SGD). they use a semantic similarity metrics between the text of reviews to identify the near-duplicate or duplicate reviews. They also combined the n-gram features with the benefit of keystroke dynamic features to identify the fake reviews. Such as the writing speed per character of number based on pauses between strokes. Since purses in typing and time while the reviewers editing their reviews patterns can help to identify the truthful and deceptive writings.

**Peng & Zhong et al** (5) proposed a sentiment analysis technique of spam detection believed from the authors that rating will not necessarily represent the sentiment accurately. Reviews dataset was extracted from Resellerrating.com, nearly 2500 reviews was used. Information gain theory was chosen for feature selection. The information gain of the sentiment words is calculated after preprocessing. 220 positive words and 150 negative words were chosen. These lexicons have been built by SentiWordNet plus. Then to determine the sentiment strength of each review by extract features of each product. (i.e. Sentiment Score (SS) which means the sentiment polarity of review, Sentiment Ratio, SR which is a ratio of sentiment sentence to all sentences and Difference of Sentiment Polarity and DSP which represent if there are a conflict between rating score and sentiment score. Then find unexpected patterns in product review by set of discriminative rules, besides establishing a time series to detect spams which depends on rating score as indicator. Finally, they combine the time series method with discriminative rules to get more accurate results.

#### Non-Related Reviews

**Mevada et al.** (6)developed system to detect the three types of review spam using supervised machine learning techniques. The SVM classifier builds models for the training data to predict any future input. They divided the system into components and the first component is data processor which is responsible for collecting and cleaning the data before using the system. Then the data proceed to the next stage of text mining by remove, sentence splitter, stop words, tokenization, and POS tagging. The last component is the classifier which is trained by the data collected and tested by test data. The input for the system is review text, after text mining component the SVM classifier calculate to any model this review is closer then give the result as spam or not.

**Long et al.** (7)proposed a method to find non-reviews based on four features. Firstly, they tend to find unusual patterns. Unusual patterns defined as links to other sites, phones, advertisements, email address and prices. Since as unusual patterns appears more in the review, the review much like to be a spam of non-related type. Secondly, used the opinion word ratio since true spam review needs to show a useful opinion of product, so each review should have the minimum number of key words that related to the product itself, so non-reviews contains little number of related keywords. Thirdly, they based on ontology word ratio which is a base knowledge about the product since the review that cannot achieve this knowledge maybe identified as non-review. Finally, they claimed that non-review spam can be identified based on sentence ratio since there is number of non-reviews can be written with standard words without a special characters and unexpected patterns mentioned previously, then to make a combination of these words that is usually meaningless. Then they used a grammar parser to know if the test of review is review or not based on the sentences list that generated.

#### Review on Brand

**Long et al.** (7)this model used the ontology that mentioned in section 2.1.1.2, each product branch consist of origin classes describes its product information (product name, place of manufacturing, name of distributor and other information), i.e. the class origin of phone product branch maybe the different type of the phones that the entity support (Samsung, Apple and other types). This module is mainly used to count the words in the text of reviews that related to the branch entity to calculate a percentage result.

**Sushant & Bharat** (8)Proposed model of detection the first and the second type of spam reviews, the system talked about in this paper uses the n-gram model to detect the fake reviews and the feature selection for the detection of the brand spam reviews, this system also calculates a spam degree for each reviewer depending on the metadata provided in the dataset. according to the authors the system lets the user search for the product they want and APIs fetch the reviews of that product from that website (amazon.com), the system then clusters the reviews in groups, after that a generated ARFF file that contains features of an original review goes to the J48 classifier, which is used for training and testing in that system. this paper actually finds the first type of review spam and by using a classifier and applies the output of the classifier after doing the preprocessing (getting rid of the stop words and stemming) to a method which checks the **word support count**and compare it with predefined threshold, deciding that the review with support count greater than the threshold is a brand spam.

**Rashmi & Vivekanand et al.** (9) designed a model of spam detection reviews on an amazon dataset about a movies, the proposed technique used in this paper is similar to the technique used in (8). After getting the review, clustering is used into groups generating an ARFF file based on a set of features, which in turn will enter the classifiers. This paper divided the work into two classification tasks, the sentiment classification and the spam classification, the sentiment classifier depends on semantic orientation calculation algorithm using an NLP parser, the spam classifier depends not only on the text of the review but also on the rating of that review. After finishing classification, the classified fake reviews enter more calculations using the association rule to calculate a threshold for the support count if the review gets a value higher than that threshold it will be considered as a brand spam review.

#### Hybrid Method Of Spam Reviews Detection

**Algur et al.** (10)followed Language Modeling Approach for Consumer Review Spam Detection to detect fake reviews (type1) and non-related reviews (type3). They show an approach of spam detecting the spam reviews by measuring its trustworthiness. The researchers used in their work the pre-processing techniques like POS, stop word removal and stemming the data, they also used their own POS tagger based on the word-net lexicon and the availability of Word-net API. Then they used the unsupervised probabilistic language model to detect the un-truth review spam (type 1) and a supervised classification to detect the non-review spam (type 3). For the un-truth reviews they design a computational model using KL (Kullback-Leibler) divergence. For (type 3) they identify features which were used in detecting the web spam such as Lexical, Stylistic and Syntactical features. Then they used the logistic regression and SVM (Support Vector Machine).

### Spammers Detection

#### Based On Review Detection

**Shojaee et al** (11)developed a stylometric-based model, this study focused on the writing style of the reviewers to detect spam reviews. Stylometric features can be divided into lexical and syntactic features. Total length or average length of reviews, ratio of the characters in word and ratio of a non-character or number in reviews is an example of lexical features. While the frequency of function words and punctuation is an example of syntactic features. They used the same gold standard dataset that created and used by **Ott et al.** (2). Then they extracted 234 of stylometric-features (lexical and syntactic), and they used WEKA machine learning tool for conducting their experiments. Different algorithm was adopted, such as SVM and Naive Bayes models, they separated the data based on features so they put the lexical and syntactic features and calculate the F-measure for this model, and also calculate the F-measure for the combined dataset to find the was 84% which is higher than using separated dataset.

**Li et al.** (12) focused on heuristic features. They used data from Epinions.com and labeled them manually. They considered that reviews with high rank and helpful comments is more to be trustworthy, while the others review with low helpful are suspected to be spam. They ignored reviews from anonymous users and removed the duplicate reviews. Then they used supervised learning such as naive Bayes, SVM and logistic regression to detect the spams. Two groups of features were extracted from the data, the first group was content similarity features, sentiment features and content features from text (unigrams & bigrams), while the second group is more reviewers related features (reviewer behavior & profile features). The result of the model shows that the naive Bayes model achieved much better accuracy using only behavior features. It scored 0.583 F-measure using all features which is not good. But they conclude from the model that the behavioral features are very important since the accuracy is dropped significantly excluding them.

**Mukherjee & Kumar et al.** (13) This paper avails other’s work, criticize other’s work, and improve the techniques used by them, this paper depends on a unsupervised Bayesian machine learning technique, that clusters the reviewers into two classes spammer and non-spammer based on a set of features that can expose spammer through studying the behavior of reviewers, and another set talking about detecting the fake reviews taking into consideration the author of this review and his behavior and history that was explored in the previous set of features. Almost all the features in this paper depends on pre-defined thresholds, these thresholds is found using a small labeled dataset (spammer/ non spammer) and this represent some super visioning in the model, but as they consider themselves working on an unsupervised technique, they exclude this dataset from the final evaluation. The novel ASM (Author spamicity model) which gives the authors a degree in range in [0, 1] of being spamming, the authors who are taking into consideration are only the reviewers who have 3 or more reviews, the results evaluation is done using a supervised classification depending on the results of the unsupervised spam opinion model without the need to manually labeling the data. The main dataset used in the ASM model is a real-life data on a manufactured product on amazon, this large dataset consists of about (50K reviewers with 1 million reviews), with all the needed metadata used in the features proposed in this paper.

#### Based On Reviewer Account

**Fei et al.** (14) focused on behavior of reviewer for detecting the spam reviews in short time burst. So that burst can be due to the sudden popularity of the product or it is a spam. They built a network of reviewers which appears in many different bursts, then to represent this network into a graph named as (Markov Random Field). Then they used a statistical model that helps the classification of reviewers as spammers or not. As we mentioned that the researchers depend on the behavioral features of the reviewers such as the rating deviation and bursts of reviews, these features are normalized as a range [0, 1]. As the previous model the authors ignore single reviews. Since we need more than one review for one reviewer to build statistically reasonable reviewer behavior. This model achieved 77.6% accuracy, however this model is used to detect the fake reviews in short bursts.

**Lim et al** (15)focused on the behavior of reviewers to detect the fake reviews. This model assumed that the spammers target certain products or brand to make the most impact. They also depend in their model on the rating behavior since the spammer will give a product rating different from the average rating of the all reviews. The dataset on which this model used is collected from Amazon that is contains the review text with rating in rage [0, 1]. Then the reviews were filtered by keep the reviews of users who are active or at least wrote three reviews. Then different models were designed, i.e., targeting product (TP), targeting group (TG), general rating deviation (GD) and early rating deviation(ED). All these models based on many spammer behaviors since each model producing a separate score, then to combined these scores to calculate the final spam score. However, they focused in their model on few products believed from them that the spammers will monitor only few products. And then they will write a spam review when the time is suitable to change the rating. So, they rely on targeting products in short period since the rating will be completely different from the others, and they will try to affect the other reviewer’s opinions so early.

**Liy et al** (16) working with Dianping, first large-scale analysis of restaurant reviews filtered by Dianping fake review filtering system in this model. Since Dianping is a company that developed an efficient fake review filtering system. This study works on large scale real life restaurant reviews. Since it is differing than the previous works by the data volume since it is shared by Dianping with user’s ID which is around 6 million reviews, also it differs by the data richness compared to other datasets since it contains of users IP address, users profile which are allows authors to create more features used in machine learning models and also with feature novelty since this is the first to give ideas of temporal and spatial features, such as reviews, user profiles and IPs of users. For this proposed method the authors implement some novel temporal and spatial features for supervised spam reviews detection. Temporal Patterns show some longitudinal studies along the time dimension. For example, spammers are active during specific days and less active in weekends and this is what the studies of temporal patterns shows. However, spatial patterns used to study the arrangement of spammers on earth and include the space between them, so these authors rely on these features on their model of spammer detection.

**Xie et al.** (17) focused in the singleton reviews in constraint of other researchers who works in multiple reviews per reviewer, since more than 90 percent of reviewers write only one review. Also, the authors noticed that the size of singleton review is big compared to non-singleton reviews. So, these reviews have more impacts at the rating of product. Then they used to extract a relationship between volume of singleton class and rating of store claimed that if there is an intense increase in the volume of singleton and sharp either high or low of the store rating, we conclude that the spammers attempt to control the store rating early to get more impact. For each window time the authors calculate three metrics that is the ratio of singleton reviews with average rating and average number of reviews. Next to find a multidimensional time series using the previous metrics by fitting the time curve on each dimension. Then they applied the LCS (longest common substring) algorithm, then they measured the spam burst on each point of the three dimensions, then they design a multi-scale anomaly detection algorithm. Finally, they used human judges to define data as spam or not and compared the labeled dataset by human by the output of this algorithm, to find that the accuracy was 75.86%.

#### Based On Graph Features

**Wang et al.** (18)proposed one of the first studies that used graph representation in fake review and reviewer detection in their methodology. This approach is different than the other approaches, since they didn't used the text of review, also it is able to detect more ambiguous cases of spam reviews activities. The authors study training to detect kinds of spammers who can manipulate their behavior to act just like genuine reviewers and think that they cannot be detected by other techniques. Author proposes a novel based method of heterogeneous review graph to detect the relationships between the reviewers, store and reviews. They represent each element previews as graph, since relationships between these nodes can be extracted by explore the interactions between these nodes. Then to identify the suspected reviewer. They also developed a mathematical method to identify the trustiness of reviewers, how much the reviews are honest and the reliability of stores.

**Mukherjee et al.** (19)designed a multi-edge graph to detect spam reviewers, which is based on extract the relationship between reviews, reviewers and store. The considered that is very normal for reviewers to have multiple reviewers due to the different purchases, hence it is expected for stores to have duplicate reviews since different stores may provide same product. The heterogeneous (multi-edge) graph contains mainly three nodes: reviewer, review and store, since each node is linked to another node. Since the reviewer’s node is linked to the reviews which them wrote, and each of these reviews is linked to a store linked. Moreover, each node is attached to set of features such as each reviewer has number of reviews, average rating and rating for each review. They believed that if there similar rating supports the same idea so there is no conflict of opinions, and there is a conflict if the rating is different. Three metrics are introduced in these processes which are trustiness of reviewers, reliability of the store and honesty of review, this three metrics affected by each other such as disease in honesty of review will affect the reviewer’s trustworthiness, while the trustworthy reviewers will have amount of honest reviews. They calculate these three metrics measurements and used them to rank reviewers, reviews and stores, and according to the evaluation results, the reviews/reviewers with highest ranking are most likely to be spam/spammer.

**Akoglu et al.** (20)proposed a fast and efficient method Fraud Eagle to detect a fraudsters and fake reviews. This model has several advantages since it used the network effect between products and reviewers and don't base on the text of reviews. The dataset mainly consists of set of reviewers, products and reviews, since each review is written by a giving reviewer about a particular product with a star rating as integer in range of [1-5]. Whilst the nodes (reviews, reviewers and products) are connected to each other as network graph and each link between two nodes associated with a rating review. The proposed model consists of mainly two steps, firstly to give a score to the reviews and reviewers for spam detection purposes, then to grouping these scores to be more representative and helpful. This method adopted unsupervised learning techniques that used unlabeled datasets. Also, it is a scalable for larger datasets with linearly grows time with network size. The authors hired propagation-based algorithm that in rule employed the advantage of the network effect to detect the relations between nodes. Then to infer the classification for the reviewers as spammer or honest, reviews as fake or true and products as good or bad by the giving score to each node and updating the score iteratively, then to propagate it to the neighbor’s nodes via the connectors of the graph.

# 

## Proposed Work

In the recent years, online reviews played an important role in the purchasing decisions, so many of researchers as discussed in chapter 2 proposed many methods to identify these fake reviews. Many of them depended in their approach in the opinion mining and emotion using text analysis which includes extracting lexical features (n-gram, POS, content similarity and semantic similarity) from the text itself to determine if the review is spam or not. While others researchers used to identify fake reviews by extracting spammer behavioral features. However, a few of them used to combine these two types of features together in their model.

The first approach of detecting fake reviews is restricted by the content of the review only. Moreover, it is possible for a company to employ different people to write fake opinions that cannot be detected by text analysis since each spammer had a specific style in writing. It should also be taken into consideration that spammers can reuse their reviews, and they can alter their reviews entirely so that not all of these fake reviews will be detected.

The second approach is utilizing the spammer behavioral features by analyzing the attitude of reviewers to detect fake reviews. This is done by extracting behavioral features that are related to the spammer himself, such as: IP address, rating of the product, window time, and number of posted reviews. A behavioral model is used to detect spammer or a group of spammers who work together in the same area using same IP, same account or same product. These methods are good indicators to determine if a review is a spam or not but that is not necessary true all the time. However, these researchers ignored that reviewers not necessarily have duplicate or semi-duplicate reviews since they can change their attitude and style of writing in order to increase the spam effectiveness and deceive the operating filtering algorithms. Also, spammers not necessarily post their reviews in different time bursts with different machines so can avoid the proposed techniques that depend on time features and IP address.

### Notation

Let Rh= {r1, r2, …} be a set of reviews for a particular product or place H, where r representing the review object which consist of 2 sets r• =<Reviews Attributes,User Attribute>**.**

**Review Attributes.** Each review consist of a set of 3 attributes, Review Attributes={Text ,Rating ,Date}, where text  represent the review text which consist of set of words and denoted as ‘Rtext’, Rating represent the star rating of the review which values are in the range from zero to five, high rating give the reader an idea of how the reviews are helpful, and denoted as Rrate’, and the Date represents the date and time when the review was posted as form of year-month-dayand denoted as ‘Rdate.

**User Attributes.** Each user attributes consist of a set of 3 attributes, User Attributes= {ID, Name, AvgRating, Count}, Where ID represent the unique user id that maps to the user since it is helpful in identifying the duplicate submissions from same id since it is forbidden to have a conflict opinion in the same product donated as RId. Name representing the name of reviewer as a group of characters donated as ‘RName’. AvgRating represent the average overall rating of a particular user in all brands that give a general idea about the rating behavior of a reviewer donated as ‘Rrating’. And Count which represent the all numbers of reviews that related to a particular user donated as ‘Rcount’.

### Proposed Work

The main goal in our proposed is to incorporate the review text and reviewer account information to identify spam reviews. The first step of our work is to collect a dataset of real reviews from Amazon/Yelp for learning and evaluating purposes, despite the absence of ground-truth annotated data which is a major challenge in our proposed work. The second step is to extract the hybrid features from the reviews which depends on reviews text and reviewer account. Since we will rely on text features, we will need to preprocess the reviews texts in order to make the model is more efficient and accurate. The third step is to build the features vector of each review to be used by the proposed spam detection algorithm, where we will propose different algorithms to identify the spam reviews. The Figure ‎3.1 shown below summarize these steps.

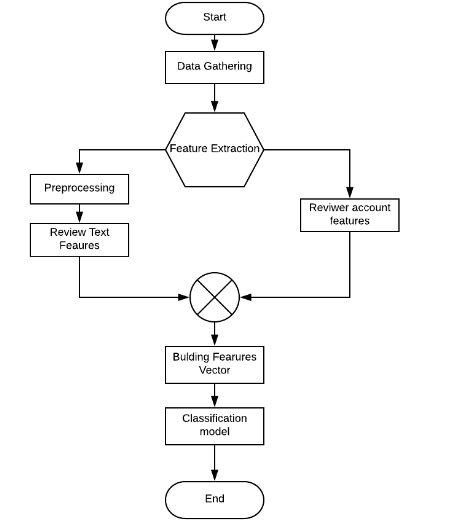


Figure ‎3.1: Flow-Chart of our proposed work.

### Features Extraction

This section explains the features that will be used in this project, and mention at least one method to observe the related feature with the mathematical model to find it, all depending on the notations section.

#### Text Features

Text features are the features that can be extracted from the text of reviews. Such as duplicate and sentiment similarity and other Stylistic features which includes a lexical attributes (e.g. word length distribution and special characters frequencies) and structural attributes (e.g. vocabulary richness measures, digit n-grams and function words).

**Duplicate**

As many spammers tend to write many reviews on the same product, and this process is time consuming and the imagination of spammers is limited, so they tend to copy the reviews or part of it, due to this content similarity is one of the main features in exposing spamming activities since this feature can measure content similarity of the reviews which is used to detect the duplicate reviews. If there is a review is similar to another review in content, we can consider these two reviews as considered to be duplicate of each other’s and thus can be considered as spam. However, this feature will not be able to detect unique spam reviews since it is designed to detect the duplicate reviews.

As we want to compute similarity of two reviews so we need to identify the similarities between their words. So, the first step is to represent the review as bag of word (vector space model). Given two review texts and, each review can be modeled as a n-dimensional vector where each dimension is representing one word using TF/IDF to get and as two vectors which is one of the ways to uniform the term frequencies is to weight a term by the inverse of document frequency. TF/IDF is given by the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑1 |

Where

Cosine similarity is measure that calculate the cosine of the angle of two non-zero vectors, where the outcome is neatly bounded in [0, 1]. Their cosine similarity that gives can be formulated as

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑2 |

As this feature based on the text of reviews so the text need to be cleaned. Preprocessing is the process of cleaning the data from unwanted elements and also to normalize the reviews since the same words sometimes written in different form to perform accurately and get more efficiency in the proposed model. Preprocessing techniques is summarized in the Figure ‎3.1 below.

Figure ‎3.2: Review preprocessing stages.

Cosine similarity gives a high advantage of being simple to compute since there is many tools used to compute the cosine similarity between two documents, also many researchers as Salton and Leask (21) who used the lexical matching using cosine similarity get good results in the smart IR system. However, cosine similarity was able to identify the few duplicates of exact matches (e.g. ‘NIKON d7100 is a great camera’ and ‘nikon-d7100 is a great camera’) but couldn’t able to find the similarity between a little different duplicate since it failed to identify the synonyms of the words (e.g. ‘hotel was awesome’ and ‘hotel was great’). Thus, we will use a semantic feature to be able to identify more difficult duplicates.

One of the cons of content similarity in finding similar sentences that have the same meaning (contains synonym), this problem can be solved using a semantic feature that depends on the WordNet semantic dictionary, this feature is considered as a natural language processing feature (NLP), that measures the sentence similarity (i.e. meaning similarity) of two reviews.

“” WordNet is a lexical database for English Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations””

The lexical similarity approach will be used in to implement this feature, this process has 6 processing steps are shown in the Figure ‎3.3.

Figure ‎3.3: Semantic feature extraction stages.

We will calculate each of semantic and duplicate features between two reviews separately, and then combine them into one feature according to the following equation.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑3 |

Where r1, r2 are the reviews and is a constant that will be set experimentally.

**Length of the review**

Sometimes non-spammers prefer to write short review texts as people tend to skip long reviews. According to the previous work Akoglu et al (20), they argued that the average length of non-spammers should fall in a specific range. This feature can be measured by the following equation.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑4 |

If then it may be an indicator of a review spam.

**Advertisement**

Spammers sometimes use to embed a URL-link within the review that not related to the product itself maybe to promote the other site, as the URL-link consist of two components which is base and path. For example for the next URL-link “[https://www.amazon.com/  
ExerciseThought-Extra-Graphic-T-Shirt/dp/B0174B8HNO/](https://www.amazon.com/ExerciseThought-Extra-Graphic-T-Shirt/dp/B0174B8HNO/)”, the base URL is [www.amazon.com](https://www.amazon.com/Exercise-Thought-Extra-Graphic-T-Shirt/dp/B0174B8HNO/) while the path of the site is “[Exercise-Thought-Extra-Graphic-T-Shirt/dp/B0174B8HNO/](https://www.amazon.com/Exercise-Thought-Extra-Graphic-T-Shirt/dp/B0174B8HNO/)”. We consider that if the base URL that mentioned in the review it is not matched with the site base URL then it should be a good indicator that the review is spam. (e.g. if the URL-link of the site is “<https://www.amazon.com/Exercise-Thought-Extra-Graphic-T-Shirt/dp/B0174B8HNO/>” while the URL-link mentioned is “[https://www.aliexpress.com/item/detail/32818868551.html?spm=  
2114.10010108.1000023.10.41f9505dtYRfaZ](https://www.aliexpress.com/item/detail/32818868551.html?spm=2114.10010108.1000023.10.41f9505dtYRfaZ)” then the noticed is these two base URL’s are not matched thus this type of review maybe a spam review, since we want to use Amazon dataset so any review contains a URL with base address that doesn’t matched “<https://www.amazon.com/>” may be a spam review. This type of spams is an indicator that the review is spam of non-related type since these types of reviews does not give helpfulness to the peoples.

**Brand or Price Mention**

Spam review can be specified as review on brand only, although it can be a true review, but it doesn't helpful for the others who see the review since it is not related to the product but talks about the brand of the reviewed product or the price of the product or maybe about the seller of the product. This feature can give the review a score (BScore; which is brand score) by each times the brand name, price or to check if there are key words such as manufacturing, seller, delivery, etc. are mentioned in the review text. Bscore can be calculated by the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑5 |

Where Wordi is the special word which donates that the review on brand or not, n is the number of special words, then to check If Bscore>Bthreshold where Bthreshold represent the review on brand threshold, then it is good indicator that the review is spam review of review on brand type.

**Average Word Length**

Most of the web pages have an average word length between 4 and 6 and that what the studies show as in the Figure ‎3.4 since if the average length is out of this range then it is an indicator of spam activities.

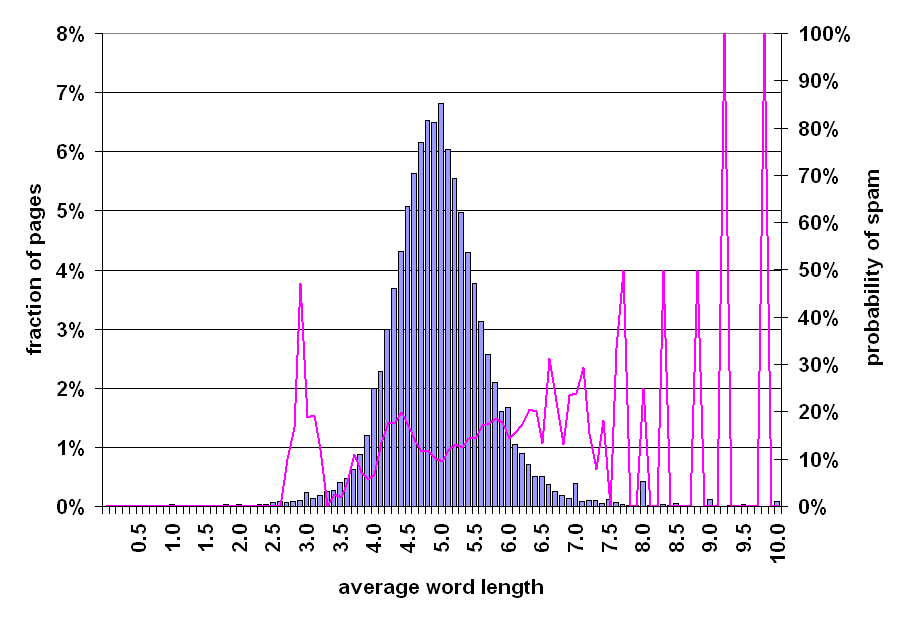


Figure .: Average length of reviews[[2]](#footnote-2).

Average Word Length (AWL) of reviews can be calculated by the length summation of all words divided by the number of words in that review according to the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‑ |

Where n is the number of words in review.

**Percentage of Numerals**

**--**

**POS percentages**

**--**

**N-gram features**

#### Spammers Features

**Normalized Maximum Number of Reviews**

Analyzing the behavior for the spammers they note that posting many reviews from the same user within time specified period is good indication for spamming activity because spammers usually tend to write more than one fake review on the same product or different products so the number of reviews for him became larger. Since we have the posting time for the review we can extract Maximum Number of Reviews (MNR) as follow.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑6 |

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑7 |

So, when the result is greater than or equal predefined threshold then the reviewer may be considered as spammer.

**Extreme Rating**

Spammers usually tend to give extremely high or low rating to the products in order to promote. This feature alone is not sufficient to decide that the reviewer is spammer so this feature combined with other features to help indicating spammers. This feature is extracted from the rating on the product as follow.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑8 |

Since if the output of the above function is 1 so the reviewer may be a spammer else no indicator that the reviewer is spammer.

**Helpfulness Rating for the Review**

The websites provide the opportunity for users to vote for other reviews from other people by choose if it is helpful or not then automatically calculate the helpfulness rating for the review and this is good indicator that the reviewer is spammer or not. Helpfulness Rating (HR) can be calculated by taking the reviews and calculates the average for helpful reviews that he has and compare it with predefined threshold as the equation shows.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑9 |

If the result is greater than the threshold it is more likely to be spammer but if it less than the threshold leads to less likely be a spammer.

**Account Freshness**

One of the properties that can expose spammers that they don’t use their accounts for a long time. They write reviews in a short period of time. In contrast the genuine users write reviews from time to time. Depending on this, the freshness of the activities of the reviewers’ accounts can be taken into consideration when exposing the spammer reviewers.

The method used for observing such features is the review burstiness, which is the date difference between the first and the last review issued by the same user Account.

For example, if an account issued all its reviews in a period of 3 or 4 months, that is an ordinary activity for a reasonable number of reviews, but let’s say an account takes 15 days only to write all its reviews, this is a footprint of spamming activity, we can define the degree of **review burstiness** (RB) as the date difference between the first and the last review issued by the same user. This method needs a threshold to done the calculations, which is can be set experimentally.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑10 |

Where last () is the last date the user u issues a review and first () is the first date the user u issues a review. This feature gives an output in the range of [0,1] which represents the degree of spamicity.

However, there are users who use their account so frequently, hence this feature will give them a high degree of spam and this can affect the final results, but with the other features we provided in our methodology this problem can be overwhelmed.

**Reviewing Early**

Early reviews are the most important reviews on any product and can be an influential element on people decisions to try the product or not, so the initial reviews affect the sales of the product significantly. So, spammers always try to be among the 0first reviewers for the products.

This feature can be observed by a method called the **Ratio of First Reviews** **(ROFR)** by finding the ratio of the first reviews to the total reviews for the same reviewer and compute it for every user to track the spammer reviewers. ROFR is given by the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑11 |

Where SFR is the size of the first reviews set that the user u posted while the is the total size of reviews of that user. However, if the dataset provides more than one product, this feature can be helpful.

**Rating Deviation Score**

Spammer maybe gives different reviews in different brands; hence spammer will have a rating on the review itself and the average of other overall reviews in different brands. This feature gives the variance between the review rating and average rating of the reviewer, which given by the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑12 |

Where is the brand, is the probability of the reviews on brand divided by the total reviews, is the average rating for the brand, is the average overall rating. This feature is helpful in identifying spammers since having a high deviation on a product with ratio of overall rating reviews is a good indicator for the reviewer to be a spammer.

### Review Classification

#### Classical Approach

The first approach of spam detection is used the classical supervised machine learning techniques which requires a labeled training dataset classified as spam and not spam to build the model that will be used later to predict the unlabeled data as testing dataset. Then this labeled dataset used to extract a combination of text features and the reviewer features searching for good result which then represented as feature vectors as a set of 9 features F= {f1, f2, …, f9}. Depending on these features using the classifiers (e.g. SVM, Naive Bayes, KNN, DCT, etc.) we can build and training the model. Hence, applying testing (unlabeled) dataset and comparing the classes given from each classifier with the true classes to calculate the confusion matrix and all accuracy measures.

#### Probabilistic Modeling

Generally, talk, the probabilistic model (Bayesian modeling) is a way to describe the data that can be observed from a system using the mathematics of probability theory (Bayesian rules), to express all forms of uncertainty of that model. Since our system depend on a set of features this set consists of two heterogeneous subsets, a subset of review (text) features f ϵ {DUP, ADV, RLEN} = Ft, and the other subset of the reviewer features f ϵ {MAXR, ER, HLP, AF, RERLY, RDS} = Fr. our system will classify the reviews into one of two classes spam or not spam depending the on the mentioned features set according to the following equation.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑13 |

Where wt, wr is the weight of the text features and the reviewer features respectively, their sum equal to 1.

As shown in the previous statistical equation, classification of a review will not depend on the two subsets of features in an equally manner, instead it will give each subset a weight that will be calculated using the genetics algorithm.

#### Score Based Model

Score based ranking will give every feature a score or weight according to its importance and effect on classification the reviews. In decision theory the weighted sum model will give every feature in our features set a weight, the following equation describes the weighted sum model criterion.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎3‑14 |

Where fi ϵ all features mentioned, wi is the weight of the feature i in the decision of classifying the review as spam or not where where n is the number of features. Again, these weights are calculated and optimized using the genetics algorithm.

**Genetics Algorithm Optimization**

A genetic algorithm is a simulation of the genetic state of a population of individuals, their combined chromosomes represents their genetic state. This algorithm also simulates the operations happen in the genes such as mutation and crossover. Here we will apply the genetic algorithm to optimize the feature weighting in both the probabilistic and the score-based model. Since in these two approaches we will assign real-valued weights to each feature in order to reduce the impact of week features and increase the impact of strong features to maximize the classification accuracy.

First a set of initial weights is given based on our expectations and knowledge of each feature, then the algorithm processes to find better solutions through an iterative manner. The Figure ‎3.4 below shows basic stages of genetics algorithm.

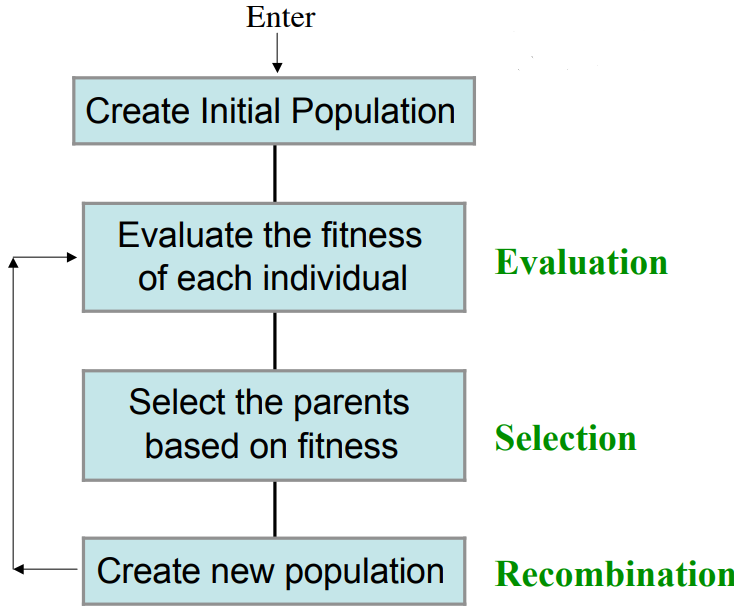


Figure ‎3.5: Genetics algorithms basic steps.

As the optimization of weights is a genetics problem, we need to formalize the problem; chromosomes are the basis of creation the population. Since each chromosome represented by a vector of positive real-values between 0 and 1 inclusive by the number of features such that the sum of all weighs is up to 1. Each weight represent a gene (e.g. if we need to optimize weights of 2 features we need to represent each chromosome as in the Figure ‎3.5).

|  |  |
| --- | --- |
| Gen1 | Gen2 |
| Weight1 | Weight2 |

Figure ‎3.6: Chromosome of two genes representation.

Algorithm will start with a set of chromosome as initial population taking care that the sum of all weights should be in range of 0 to 1 inclusive. Then this initial population should be passed to fitness function to evaluate the chromosomes.

Crossover operation which called Continuous Uniform Crossover (CUCO) that proposed by (22) is used which assure that the sum of all weights after offspring is equal 1. Given two vectors of features weights and the offspring of those chromosomes are defined as and . Where

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‑ |
|  |  | Equation ‑ |

Where s is called , which is a constant through a single crossover operation and should be . Since if then , proved by the following expression.

## Datasets, Tools and Evaluation Measures.

### Dataset

One of the most difficulties in machine learning problems is to collect and choose the dataset that will be used for training and testing the proposed model, hence we are looking for informative data in our work to extract the desired features to be dependable in our approach of detecting the spams reviews. However, there is another difficulty which is how to collect dataset with the true labels for the model training and testing. There are four sources to get datasets such as Yelp, Amazon, Google My Business API and datasets from previous related work in this field, since most of researches uses these sources to get the data. However, referring to the related works it may easy to have negative reviews (non-fake reviews) since there is sites like Yelp and Dianping that filters the reviews and display just honest reviews and used different ways for data annotation.

Most of the work was trying to create a manually labeled dataset and that what happens with **Li et al.** (12) since they a principled way to have an annotated dataset. They found more others related work about the spam reviews. They then employed 10 college students to annotate 60,000 reviews manually after reading the collected related work to know what the spam and the way to label spam reviews. Each student should independently label the data by determining its spam or not, each review should be labeled by different two students and if there a conflict in labeling then it will be resolved by a third student. **Sun et al. [25]** proposed a synthesis process to generate fake reviews from a pool of truthful reviews, then to measure the performance according to the existing detection algorithm and human readers; they suggested that the fake reviews can be automatically generated by replacing the sentences of a truthful review with other reviews sentences.

**Shojaee et al.** (11)proposed a novel technique to annotate the dataset that collected from amazon.com in this research by providing an information and metadata about reviews and reviewers. The authors developed an online annotation system to improve the accuracy of annotation process. It’s necessary to have a labeling dataset for spam detection approaches to evaluate the proposed method. However, it is difficult to manually annotate the dataset since it is confusing for humans and consuming a lot of time. The authors hire an existing works and proposed some rules to increase the accuracy of labeling dataset. They proposed a method to manually labeling be simultaneously considering clues by design an algorithm of asking questions, 11 questions to detect spam reviews and 5 for spammer detection, since all of these questions are extracted from previous existing work which is selected based of features selection. The algorithm calculates the two numbers (k, k’). Since if there a clue of spammer detection the algorithm increment k by one. And for every review it calculates a number of spam detection clue and then increment k’ by one for each found clue. Then in case of k’ is greater than a determined threshold (k’) the review will be set to 1 or spam since the algorithm detect the reviewer as spammer. Otherwise to check the if k is greater than a threshold (k). And set the review to spam too. And if the two numbers are less than the related threshold the review will be set to 0.

We have four available choices for the dataset. The first one is Google Api[[3]](#footnote-3), this dataset is provided by google for business so you must have business account and pay money for google to get the dataset so, we discarded it after trying to get these reviews we found that they provide just the top five reviews about products or places. The second choice is the public datasets from other related work in this field, we contact some persons via email to get the dataset but they didn't respond to us so we can’t depend on the datasets they have. The third choice is Yelp dataset, this dataset provided by Yelp[[4]](#footnote-4) website and it is large enough about (174000) review. This dataset contains the desired attributes we looking for to extract our features from it such as rating, helpful, user information, timestamp and plain text review. But the main disadvantage that it is not labeled as spam or non-spam which will make an extra step for us to label these data manually or by supervised learning. The last choice is Amazon data provided by Amazon[[5]](#footnote-5) website, it has the required metadata we are looking for and scalable but it is not labeled as spam or not. In general, it was difficult to get labeled dataset so we have to choose among these available choices; as a result we determine to work on Amazon dataset.

For our dataset we will follow a procedure to annotate the data which is first giving every review a score indicates if it is spam or not then determine a threshold for classifying the data in two parts, the first part will be the reviews higher than the threshold which represents non-spam reviews so, it labeled as non-spam reviews. The second part will be the reviews with score less than the threshold for this part of reviews will classify them manually as spam or non-spam.

There is another challenge among with getting labeled data and desired metadata that is the difficulty to have [[6]](#footnote-6)fake reviews since it is hard to know the spammer writing behavior of writing fake reviews. since there is sites like Yelp Dianping that filters the reviews so getting fake became harder.

As initial stage we will use an available and annotated dataset described in (2), this dataset consist of:

1. 400 truthful positive reviews from TripAdvisor.
2. 400 fake positive reviews from Mechanical Truk.
3. 400 truthful negative reviews from Yelp, TripAdvisor,Expedia and Hotels.com
4. 400 fake negative reviews from Mechanical Truck

This described dataset is consistent and contains a good testing dataset, however this dataset is helpful in case features extracted from the reviews text since it contains text only.

### Tools

Detecting spam reviews is considered a text classification problem and people used different tools for this problem such as Weka, Knime, Orange.

The tool we will used is Weka since it has a big collection of machine learning techniques such as naive Bayes, Decision tree, K-means, KNN, neural networks, etc. Also, this tool has a lot of preprocessing mechanisms for the text such as transformation, loading, and extraction. Moreover, it takes the sentence and represent it in featured vector using TF/IDF. it’s an open source and implemented in java platform.

The other tool we will use is Stanford NLP which provides a lot of processing techniques for the text including part of speech tagging(POS), removing stop words, removing punctuation marks, and named entity recognizer(NER). This tool is open source and supports many languages including the English (in which we work). Also, its written in java. So, we will use this tool for preprocessing for the data before sending it to Weka.

The input for the Weka is a set of labeled dataset. Firstly, we will do preprocessing for the data such as removing stop words in the Stanford NLP tool. Then, we represent each review by a word vector using TF/IDF in Weka. Here, the data is ready to be classified so, we will use four different algorithms to classify these data which are Naive Bayes, Neural Network, KNN, and Decision Tree. we will see which classifier gives the best result so we can depend on it for the classification work in this project.

### Evaluation Framework

As we develop a solution to solve problem using machine learning then we need to choose metrics to evaluate our machine learning model, since these metrics influence how the performance and accuracy of the implemented machine learning algorithms. However, these metrics is standard in the machine learning domain which is precision, recall and f-measure.

#### Confusion Matrix

Is an easy way to represent the resulted values from the algorithm and it contains four values as shown in the Figure ‎4.1 below

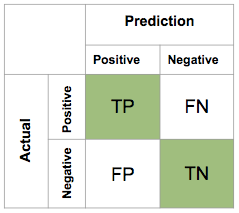


Figure ‎4.1: Confusion Matrix.

1. **True Positives (TP):** means how the actual class of the data point is classified true as predicted.
2. **True Negatives (TN):** means how the actual class of the data point is classified false as predicted.
3. **False Positives (FP):** means how the actual class of the data point is classified true as predicted while the predicted class is negative.
4. **False Negatives (FN):** means how the actual class of the data point is classified negative as predicted while the predicted class is true.

#### Accuracy

This type of accuracy measurements used to show how many correct classification points related to the all prediction operation. This type gives an overall idea of how the model is accurate. It represented by the following relationship.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎4‑1 |

#### Precision

A measure that shows how many TP operation related to the number of all points that classified as positive (TP and FP). This type of measures is useful in critical cases. Precision is represented by the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎4‑2 |

#### Recall

A measure that tell us how many TP classes which TRUE is and classified as TRUE related to the total number of actually TRUE. Recall is represented by the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎4‑3 |

#### F-measure

A measure which is the harmonic mean of precision and recall which is in the best case 1 and 0 at worst case. The f-measure is given by the following formula.

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎4‑4 |

#### ROC curve

the receiver operating characteristic curve is a plot of the true positive rate(sensitivity) as a function of the false positive rate(specificity) for different cut-off points, every point on the curve represents a (sensitivity, specificity) pair corresponding to specific decision threshold [1]

|  |  |  |
| --- | --- | --- |
|  |  | Equation ‎4‑5 |
|  |  | Equation ‎4‑6 |

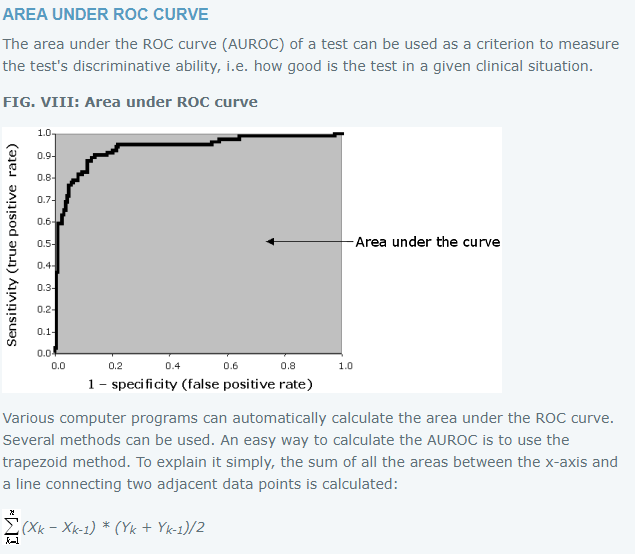
**

Figure ‎4.2: ROC curve evaluation measure.

# 

## Conclusion

Due to the increasing of consumer dependence on online reviews, spammers tend to write manually or by proposed review system fake reviews on a product or organization. Since customers’ decisions are affected by the posted review, fake reviews become a dangerous prospect for online reviews. However, customers find that is difficult to identify bad from good products. According to this the posted reviews should be filtered as spam or not spam to be more confident.

In this project, we have focused on identifying spam reviews problem using many features that extracted from the reviews after specify them as text which includes duplicate and semantic similarity and spammers behavior features which includes number of rating, account freshness, reviewing early and rating deviation and other features. As the major challenge in identifying fake reviews is come from how to extract best set of features that will be used in the model. However, these features are very limited. Many researchers who suffers from lack of accuracy depends on their work on textual features and the others depends on the behavioral features while a few of them relied on their work on a combination of these features. We will use a hybrid features to boost the performance and accuracy of our model.

We will depend on three approaches of our spam detection model. Firstly, we will classify reviews as spam depends on the combination of extracted features using classical supervised machine learning techniques (SVM, KNN, Naive Bayes, DCT, etc.). Secondly, using semi-supervised learning depends on probabilistic based on genetics. Finally, using unsupervised learning that depends on score-based optimized by genetics.

## References

1. *Finding Unusual Review Patterns Using Unexpected Rules.* **Jindal, Nitin , Liu , Bing and Lim, Ee-Peng.** s.l. : CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 1549-1552.

2. *Finding Deceptive Opinion Spam by Any Stretch of the Imagination.* **Ott, Myle, et al., et al.** s.l. : HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, Vol. 1, pp. 309-319 .

3. *Identifying Multiple User Ids of the Same Authors.* **Qian, Tieyun and Liu, Bing .** s.l. : Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1124-1135.

4. *Detecting Opinion Spam and Fake News Using N-gram.* **Hadeer, Ahmad .** 2012.

5. *Detecting Spam Review through Sentiment Analysis.* **Peng, Qingxi and Zhong, Ming .** s.l. : JSW, 2014, pp. 2065-2072.

6. *An Opinion Spam Analyzer for Product Reviews Using Supervised Machine Learning Method.* **Mevada, Daya and Daxini, Viraj.** 2015.

7. *Opinion Spam Recognition Method for Online Reviews Using Ontological Features.* **Long, Nguyen Hoang , Nghia , Pham Hoang Trong and Vuong, Ngo Minh .** s.l. : Opinion spam recognition method for online reviews using ontological features, 2014, p. 44.

8. *Fake Review and Brand Spam Detection using J48 Classifier.* **Kokate, Sushant and Tidke, Bharat .** 2015.

9. *Detection of Fake Review and Brand Spam Using Data Mining Technique.* **Adike, Rashmi Gomatesh and Reddy, Vivekanand.** s.l. : Int J Recent Trends Eng Res, 2016, pp. 251-256.

10. *Conceptual Level Similarity Measure Based on Review Spam Detection.* **Algur, Siddu P, et al., et al.** s.l. : Signal and Image Processing (ICSIP), 2010 International Conference on, 2010, pp. 416-423.

11. *Detecting Deceptive Reviews Using Lexical and Syntactic Features.* **Shojaee, Somayeh, Azrifah , Masrah and Azreen, Azmi Murad.** s.l. : Intelligent Systems Design and Applications (ISDA), 2013, pp. 53-58.

12. *Learning to Identify Review Spam.* **Li, Fangtao, et al., et al.** 2011.

13. *Spotting Opinion Spammers using Behavioral Footprints.* **Mukherjee, Arjun, et al., et al.** s.l. : Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 632-640.

14. *Exploiting Burstiness in Reviews for Review Spammer Detection.* **Fei, Geli, et al., et al.** s.l. : Icwsm, 2013, pp. 175-184.

15. *Detecting Product Review Spammers Using Rating Behaviors.* **Lim, Ee-Peng , et al., et al.** s.l. : Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 939-948.

16. *Analyzing and Detecting Opinion Spam on a Large-Scale Dataset via Temporal and Spatial Patterns.* **Liy, Huayi, et al., et al.** s.l. : CWSM, 2015, pp. 634-637.

17. *Review Spam Detection via Time Series Pattern Discovery.* **Xie, Sihong, et al., et al.** s.l. : Proceedings of the 21st International Conference on World Wide Web, 2012, pp. 635-636.

18. *Review Graph Based Online Store Review Spammer Detection.* **Wang, Guan, et al., et al.** s.l. : Data mining (icdm), 2011, pp. 1242-1247.

19. *Spotting Fake Reviewer Groups in Consumer Reviews.* **Mukherjee, Arjun , Liu, Bing and Glance, Natalie .** s.l. : Proceedings of the 21st international conference on World Wide Web, 2012, pp. 191-200.

20. *Opinion Fraud Detection in Online Reviews by Network Effects.* **Akoglu, Leman, McGlohon, Mary and Faloutsos, Christos.** s.l. : ICWSM, 2013, pp. 2-11.

21. *Computer Evaluation of Indexing and Text Processing.* **Salton, Gerard and Lesk, Michael .** s.l. : Journal of the ACM (JACM), 1971, pp. 8-36.

1. Yelp: <https://www.yelp.com/dataset> [↑](#footnote-ref-1)
2. <http://www2006.org/programme/files/xhtml/3052/xhtml/www2006.html> [↑](#footnote-ref-2)
3. Google API: <https://developers.google.com/my-business/content/review-data> [↑](#footnote-ref-3)
4. Yelp: <https://www.yelp.com/dataset> [↑](#footnote-ref-4)
5. Amazon: <http://jmcauley.ucsd.edu/data/amazon> [↑](#footnote-ref-5)
6. [↑](#footnote-ref-6)