Web Browser Plug-in to Detect Fake Reviews in E-commerce Sites

A dissertation submitted for the Degree of Master of Computer Science

D.P.Jayathunga
University of Colombo School of Computing
2019

S	
E1	
E2	
For Office Use Only	

Masters Project Final Report
(MCS)
2019

Project Title	Web Browser Plug-in to Detect Fake Reviews in E-commerce Sites
Student Name	D.P.Jayathunga
Registration No. **& Index No.**	2016/MCS/046 16440467
Supervisor's Name	Dr. H E M H B Ekanayake

For Office Use ONLY

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: D.P. Jayathunga

Registration Number: 2016/MCS/046

Index Number: 16440467

_____

Signature: Date:

This is to certify that this thesis is based on the work of

Mr./Ms.

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name:

_____

Signature: Date:

Abstract

With the recent proliferation of online shopping customer usually publish reviews to the store and product after online shopping. At the same time, the online reviews become an important approach for the potential customer to know about the store and product. They usually check the online reviews to make decision whether to buy the product or not. Meanwhile, sellers and manufacturers are carrying out investigation of online reviews for decision making. Positive opinions can result in significant financial gains and/or fames for organizations and individuals. But to affect customers' buying decisions, fake opinions are generated for purpose to promote special targets and/or denounce their competitors. Filtering out of untruthful information becomes an important issue in current situation. Throughout the period of project, I have analysed this issue and was able to introduced an enhanced method to identify the untrusted reviews as well as untrusted sellers and customers. I have introduced a web browser plug-in with an API which is capable of extracting the data (customer reviews, seller details, customer details) from e-commerce website and analysed the data using semi-supervised learning. After evaluating the results of the extracted comparing the already evaluated data set from the amazon website if shows the 80% of accuracy.

Acknowledgement

I would like to express my heartfelt gratitude and appreciation to my project supervisor Dr. H. E.M. H. B. Ekanayake who has been tremendously helpful throughout the project and also for his constant help and expert supervision over the project.

I cannot express enough thanks to Mr. H.M.S.N Ariyadasa, Lecturer, Uva Wellassa University for the tremendous support given from the initial stage of this work.

My completion of this project could not have been successful without the support of my dearest family. I want to dedicate the success of this project to my father, mother, and sister who were very influential at every bit of time.

Table of Contents

Table of Figures

Chapter 01- Introduction

1.1 Background

The development of Information and Communication Technology (ICT) has a severe impact in day to day life of the human. People tend to use many ICT tools and techniques in order to do their activities efficiently, effectively and accurately. Thus, ICT has become a major requirement for many fields such as Business, Governance, Healthcare, Engineering, Agriculture and Education [1].

The invention of the Internet and the World Wide Web (WWW), has become the turning point of ICT technologies since it facilitates to connect people each other from anywhere of the world in a very short period of time. Among all, this has a great influence on business which striate away effect to the growth of the economy of a country. The internet is very popular among almost every individual in the modern world due to the many advanced features of it such as web of services, web of documents and webinar services etc.

Among all these advanced features of the Internet, e-commerce became extremely popular and has a great impact on economic growth and international marketing. This platform facilitates people to sell and purchase products or services over the Internet. Thus, developing of e commerce websites became a high demand trend in IT industry. One intention of developing an e-commerce website is to gather a huge community and be more productive in selling products to the consumers. Thus, many e-commerce websites have standardized practices in order to accomplish their goals. This has lead them to identify and concern more on the key quality factors such as usability, conceptual and representative reliability in order to make their e commerce website a success.

One standard practice in developing e-commerce websites is, facilitating consumers to leave a review and read others' reviews on many products with the aim of providing the genuine and

important source of information to the consumer as well as to determine the success of the business [2].

In online shopping, many consumers prefer to read the reviews in order to make the best decisions on the product before purchasing it. Products which have the positive reviews give a good impression to the consumer regarding the product and also many negative reviews have high potential to take up consumers away. Therefore, many online merchants tend to manage a good online reputation while enhancing the popularity of the product [3].

1.2 Motivation

Since consumers highly consider fellow consumers' review and trust them as personal recommendations when purchasing a product, many fake reviews are made by an organization or by an individual to mislead the consumer due to many reasons [5]. Even merchant himself may hire internet ghostwriters to post positive reviews in order to promote their reputation. In the other hand, competitors may give or hire internet ghostwriters to post negative reviews in order to damage the reputation which is unethical. Due to these matters, consumers cannot take correct judgement based on the reviews which emphasis false assessment regarding the product.

Since consumers have to spend a considerable amount of time to get the correct decision based on the reviews, the need of an assistance has arisen. Thus, many institutes, association and scholars have introduced many algorithms, models and tools to detect the fake reviews [4,6] such as Review skeptic (www.reviewskeptic.com). But still a significant amount of fake reviews is present today which emphasize the important of assisting the consumers to help detecting fake reviews and also self-protecting [7].

Also, consumer cannot only rely on those [8], since many of them not clearly producing the truthfulness of each review separately and also only capable of functioning on some selected websites and producing the ratings [5,8]. Similarly, no projects were carry out to evaluate the customers who has added reviews and rating the seller along with the fake review detection.

Therefore, this research project is mainly focusing on developing a browser plugin which is capable of detecting the each given review individually either a fake or genuine also it will rate the seller and the customer who has given the review as well.

1.3 Goal and Objectives

The main aim of the project is to develop a web browser plug-in in order to detect and make visible the trustworthiness of each review and the customer of online shopping websites as well as providing overall rate for the product and the seller to protect the genuine interest of the consumer.

In the purpose of detecting fake reviews, an enhanced model/algorithm is developed and applied. The ultimate goal in this study is not only to identify the fake reviews but also to make visible the trustworthiness of each review and the customer who has added the review to the consumer that help every consumer to categorize which review to consider and which review not to consider. Also, it provides an overall rate for the entire reviews of the product with the actual seller rating on the e-commerce website itself.

Mainly the project can be divided into four parts based on the functionalities of the projects:

1. Extraction of the review content
2. Fake review detection
3. Rating the seller and the product
4. Identify the truthfulness of the customer who has added the reviews

This lead to define the objectives of the project as mentioned below:

1. To extract the review content from the plug-in
2. To provide a model to detect fake and genuine reviews using NLP
3. To provide the overall rate of the reviews and sellers
4. To display the result to the consumer from the same e-commerce website
5. To display the truthfulness of the customer who has added reviews to the website.
6. To display the trustworthiness of the seller

The limitation of the project is that this plug-in has been developed only for the google chrome browser due to time limitation and thus this has only been tested with google chrome.

1.5 Scope of the Project

The research project focuses on developing a software component specific only to one browser due to the limitation of time which can be used to detect the fake and genuine reviews on e-commerce websites. Thus, the consumer can avoid the fake reviews and the customers who has given the fake reviews so that they can make the correct decision by only rely on genuine reviews and the overall rate of the reviews and the seller. Therefore, a web browser plug-in is proposed to develop which is specific to one web browser.

Since consumers have to spend a considerable amount of time to get the correct decision based on the reviews, the need of a plug-in has arisen. This plug-in will help the consumer to make their decision on the particular product in a short period of time since it will add the result to the same e-commerce web page which contains the identified fake/genuine reviews as well as overall rate of the trustworthiness reviews.

When considering the existing works there are several models available to detect the fake and genuine reviews [6]. Considering the limitation and drawbacks of the existing models/algorithms an enhanced model/algorithm will be suggested for the purpose of detecting the fake and genuine reviews and applied to the plug-in.

1.6 Structure of the dissertation

Following is the structure of the dissertation with various chapter including a summary.

Chapter 1: Introduction

Within this chapter, I have explored the background of the study, motivation, goal and objectives including the scope of the project as well as the achievement of the study on brief. Furthermore, this chapter include the structure of the dissertation.

Chapter 2: Literature Review

In this chapter, I have discussed various works and research efforts related to this research project done by other researches. It includes the technologies they have adapted, accuracy and the limitation of their work. Ultimately, I also have explained the significance of the study with respect to the existing related studies.

Chapter 3: Methodology

Under this chapter I have explained the adapted technologies and approach of the study.

Chapter 3: Design and implementation

In here, I have explained overall design overview and its functional view which include a high-level architecture, Flow chart and Data Flow Diagram. Also, I have mentioned all the assumptions and limitations which I have considered throughout the study. Also, this chapter include the implementation of the project with the details of adapted technologies.

Chapter 4: Evaluation

In this chapter, I have discussed about the results of the project, evaluation plan, perform testing, user evaluation and performance evaluation.

Chapter 5: Conclusion and future work

In here, I have explained the results of the study and achievements in detail. Further, it contains the conclusion of the project after thoroughly analyzing the results from the evaluation which has describe in the chapter 4. I addition to that, I suggested some works which can be done in the future.

List of Reference

Appendix: It is contained study resources.

Chapter 02- Literature Review

2.1 Overview

With rapid increase of online shopping, e-commerce websites have been popularized among people. Thus, customer usually tend to share their experience on online shopping via reviews to the store and product after their purchases is done. At the same time, the online reviews become an important approach for the potential customer to know about the seller as well as the product. Customers usually tend to check the online reviews when they want to make a decision whether to buy the product or not. Meanwhile, sellers and manufacturers are carrying out investigation of online reviews for decision making in order to increase their business with a great profit because positive opinions regarding them and their products can result in significant financial gains and/or fames for organizations and individuals. Similarly, negative opinions result the opposite.

Since reviews have a great impact on the product and the seller, some people make many fake reviews which is an illegal activity to mislead the customers for purpose to promote special targets and/or denounce their competitors. Thus, avoiding these fake reviews and the people who add the fake reviews became a significant issue in e-commerce websites. In this chapter, I discuss many works related to fake review identification which address this issue in many ways in e-commerce websites and compare them with the study I have carried out.

Fake Review has many names such as Spam Review, Fake Review, Bogus Review, Deceptive review Opinion Spammer, Review Spammer, Fake Reviewer, Shill (Stooge or Plant) [reference]. There are many related research works, basically done in order to identify these spam reviews. Analysis of online opinions became a popular research topic recently. In this chapter I have discuss similar studies done by other researches. This literature review was done throughout the project in order to identify the problem defined in the previous chapter and to gain update knowledge regarding the study.

2.2 Related works

Many related research works were carry out under two classifications based on several identified characteristics such as word use, word duplication and the number of sentences and different reviewer and behaviour related characteristics; content-based spam detection and Behavior-based spam detection [9].

Ott Myle and his group were able to develop an algorithm, a model and a tool called Review Skeptic (www.reviewskeptic.com) based on the research they conducted on content-based spam detection using machine learning [10]. Users have to copy and paste the relevant review of the hotel on their website to scan whether the given review is fake or not. The Accuracy of the tool is nearly 90%.

Archana Battarai Vasile Rus and Dipankar Dasgupta investigated the characteristic of comment spam in blogs. They were able to introduce a framework under content-based spam detection to extract the features of the blog spam and classify the comments using semi-supervised classifiers and supervise classifiers. Though the framework is more realistic and more flexible their framework did not evaluate on large-scale dataset [11].

Ee-Peng Lim and his colleagues have proposed a behaviour based detection mechanism to detect review spammer who was trying to control the rating of the products. They have derived an aggregated behaviour scoring method to rank only the fake reviewer. Further, they have applied regression model from user-labeled ground truth spammers to score only the spammers but not the review [12].

Fakespot is another freely available web tool which is used to analyze the reviews of certain product or service which is only provided by Amazon, Yelp, TripAdvisor, and Apple. It will analyze and present the result after copying and pasting the relevant URL of the product page or business page. It provides an overall rating from A to F, an overview of the analysis, a summary of reviews, and a list of unreliable reviewers [5,7].

Moreover, many other web tools are available online such as ReviewMeta (www.reviewmeta.com) and Review Reveal [8] to identify fake reviews on certain e-commerce web site. But the accuracy of the results generated from those tools is not yet proof by any source and sticks on evaluating either one or few e-commerce websites reviews though those tools are success into some extent.

Current studies are mainly focused on mining opinions in reviews and/or classify reviews as positive or negative based on the sentiments of the sentences. Perhaps, the most extensively studied topic on spam is Web spam. The objective of Web spam is to make search engines to rank the target pages high in order to attract people to visit these pages. Review spam is quite different. Spammers write undeserving positive reviews to promote their target objects and/or malicious negative reviews to damage the reputation of some other target objects. These faked opinion information is called opinion spam [13]. For academic researchers, they have conducted various research studies on sentiment analysis tasks. If their acquired opinion resources contain many opinion spams, it is meaningless to provide any sentiment analysis results.

Therefore, detecting such spurious reviews and spammers have become a pressing issue. The problem of spam review detection is presented for the first time [14]. In their research, duplicate and near duplicate reviews are assumed to be fake reviews. Supervised learning has been employed to detect spam review. In another research [15], the impact of single reviewer to the online store, and anomaly pattern of rating are analyzed to detect the spam reviews. In another study [12], several characteristic behaviors of review spammers are identified, and these behaviors are modelled to detect the spammers. In one study, the unusual review patterns which can represent suspicious behaviors are identified, and unexpected rules are formulated [16]. A novel concept of review graph is proposed by Wang and others in their research [15], which capture the relationships among all reviewers, reviews and stores that the reviewers have reviewed as a heterogeneous graph.

Then an iterative computation model is proposed to identify suspicious reviewers. In one research, three approaches to detect deceptive opinion spam by integrating work from psychology and computational linguistics [10]. The inconsistency problem between the

evaluation score and review content is studied, and credibility of customers is detected in another study [17]. In a study [18], a frequent item set mining method is proposed to find a set of candidate spam groups. Several behavioral models derived from the collusion phenomenon among fake reviewers and relation models are been used to detect the spam groups. The previous studies mentioned above generally detect the spam review by means of rating score.

2.3 Summary

As a conclusion though several models and tools have been introduced, still considerable amount of fake reviews can be discovered due to the several weaknesses of those. Thus the importance of assisting the consumer to distinguish between fake and genuine reviews has arisen. No any previous works done to identify the real customers or the trustworthiness of the customer who added the reviews in the website along with identifying the fake reviews. Similarly, though many works carried out to calculate the overall rating excluding fake reviews, there were no any study found to identify the sellers' truthfulness. Thus, the problem of this study was defined where the primary aim of the project is to introduce a mechanism to detect fake reviews, to identify the truthfulness of the customers and to calculate genuine sellers' rating. Therefore, from this project I have introduce a web tool with enhancing accuracy and suitable model of detecting fake and genuine reviews, the identifying the truthful customers and identifying the trustful sellers.

Chapter 03- Methodology

3.1 Overview

Under this chapter I have explained the approach of the study in detail which was used in developing the entire system.

Similarly, this chapter contain the technology adapted in the development the software component: plug-in and the model which is capable of identifying fake reviews and based on the results identifying the trusted customers who has given the reviews and also the truthfulness of the seller.

3.2 Approach

The main task of the project is divided in to three subtasks as mentioned bellow which were much convenient steps in developing the entire system smoothly.

1. Plugin development with Data extraction
2. Detecting whether the review is fake or not
3. Identifying the truthfulness of Customer and Seller

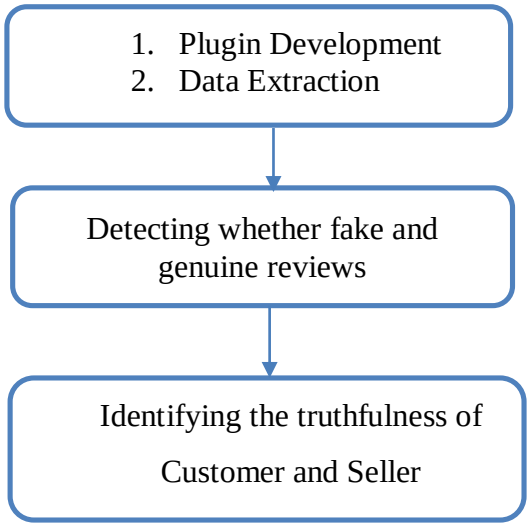

Figure 3.2.1. subtask of the project

3.2.1 Data extraction

The first subtask was tricky since it should not be focusing on single source. It should work with any type of e-commerce web site where the reviews can be founded. While thoroughly analysing e-commerce websites I was able to identify that there are some websites which has the feature of providing the needed component (in my study needed data was reviews and customer ID) through an API (such as Amazon and eBay). But there was a problem in extracting data from other e-commerce websites in which this feature is not available. Therefore, to overcome this obstacle I had to come up with a solution which is cable of extracting data from any type of e-commerce websites which have an API and the websites which doesn't have. Therefore, as a solution I have used web scraping for retrieving data (reviews alone with the customer ID/name and the seller ID/name) to identify the customer from the website.

3.2.2. Detecting fake and genuine reviews

The second subtask is to classify the reviews in to two categories which are fake or not and based on that results, identifying the truthfulness of Customer and Seller as the third subtask. In order implement the subtask two I have built a classifiers using semi supervised learning methods. Semi-supervised learning is a class of machine learning task and technique which is capable of making use of unlabelled data for training. If there are sufficient labelled data to construct classifiers, supervised learning methods will be effective. Labelled instances are often too difficult, too expensive, or time consuming to get, due to they require research. When it comes to e-commerce website product reviews, I have a large set of unlabelled data. Most of the times semi supervised learning gives a better accuracy than supervised learning which is only trained on the labelled data. The data which has been trained using semi-supervised learning is sent to a centralized database with customer details and the seller details in order to use in further implementation of the system.

3.2.3. Identifying the truthfulness of Customer and Seller

In order to identify the truthfulness of the customer and the trustworthiness of the seller I have used the results of the trained data set from the centralized database.

Identifying the truthfulness of the customer was done simply analysing the identified fake reviews along with customer ID/name of a particular e-commerce website. The condition for that was if the customer has continuously added a significant amount of reviews which were identified as fake reviews through the model compare to the all reviews done by him/her for a particular product or for a set of products or for particular seller then that customer can be a suspicious person. Thus, the particular suspicious customer was added to a separate table in the database. As the final step of identifying truthfulness of the customer the plug-in will continuously monitor his/her next two reviews. If those reviews are also detected as fake, then the customer will be identified as an untrusted customer and display customer ID/name in red.

For identifying the seller trustworthiness, the fake reviews are analysed alone with the seller ID/name and the product. Here, the condition is that if the 30% from reviews are identified as positive fake reviews relevant to the seller ID/name. By considering the above condition and the overall rate of the seller, the seller can be identified either a suspicious seller or not.

3.3 Adapted Technology

In order to implement the above three subtasks, I have used following technologies.

3.3.1. Google chrome plug-in development

I have developed and tested the plug-in in google chrome due to time limitation and used following technologies to build the Google chrome plug-in

1. HTML
2. JQuery
3. JavaScript
4. Bootstrap CSS

3.3.2. Data Extraction

Method used for data extraction from website is web scraping which is a simple and effective method for extracting data from a given URL. For implementing web scraping technology I have used following tools

1. Node JS
2. Cheerio JS
3. Javascript
4. Jquery

3.3.3. Data processing

I have used Python scripting language to implement following in data processing.

1. Pre Processing of data
2. Calculate behavioral dimension
3. Sampling
4. To implement Machine Learning algorithm
5. To implement semi supervised learning

Chapter 04 – Design and Implementation

4.1. Overview

As the initial step in designing of the project I have drafted the high-level architecture. Then based on the high-level architecture I have designed the flow chart and ER diagram which are included in this chapter under the section of Design. Also, in this chapter I have mentioned the assumptions and limitations which were considered throughout the project. Similarly, in this chapter I have discussed implementation of the project in detail with the adapted technologies under the section of implementation.

4. 2 Design

4.2.1 High-level architecture

Figure 4.2.1 shows the high-level architecture of the project having three levels which include three sub-components namely Browser Plug-in, API and the centralized server. Arrows has been used to show direction of the communication among the three levels.

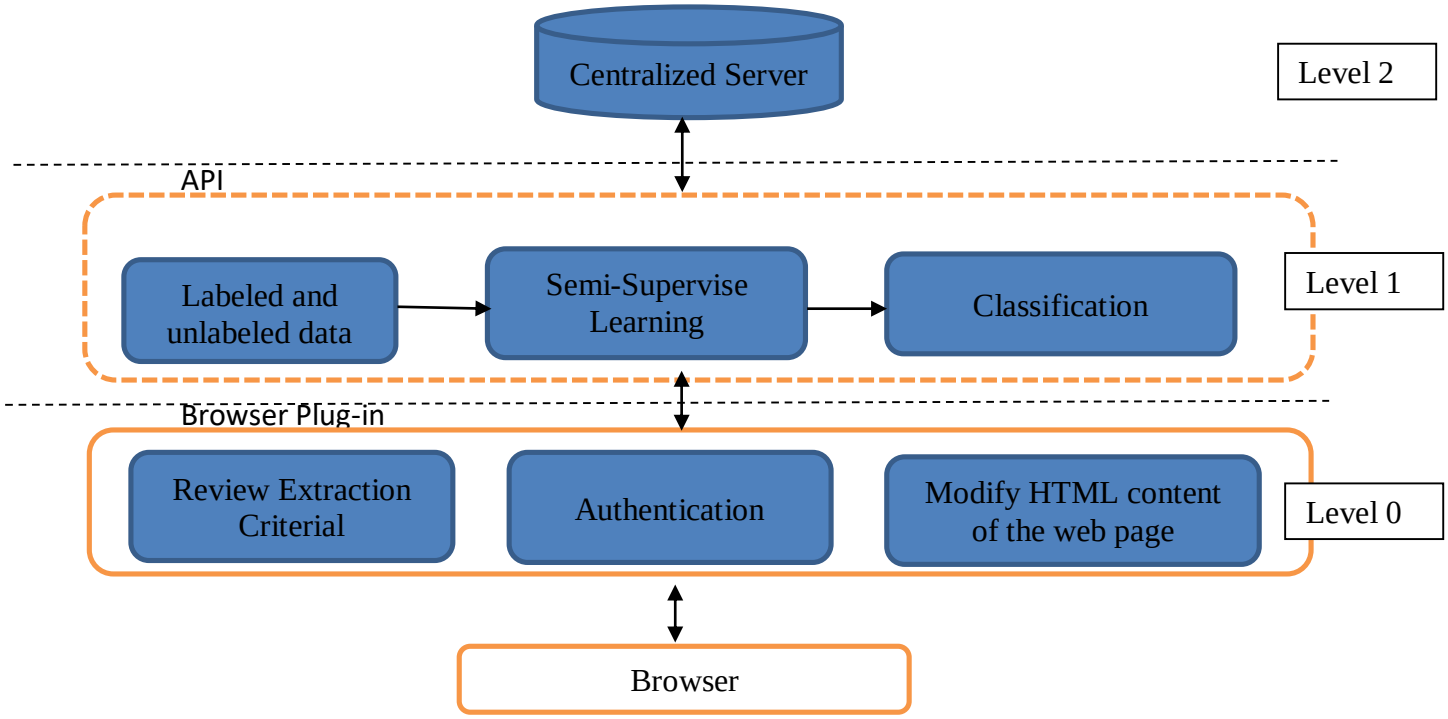

Figure 4.2.1 High-Level Architecture

4.2.2 Flow Chart

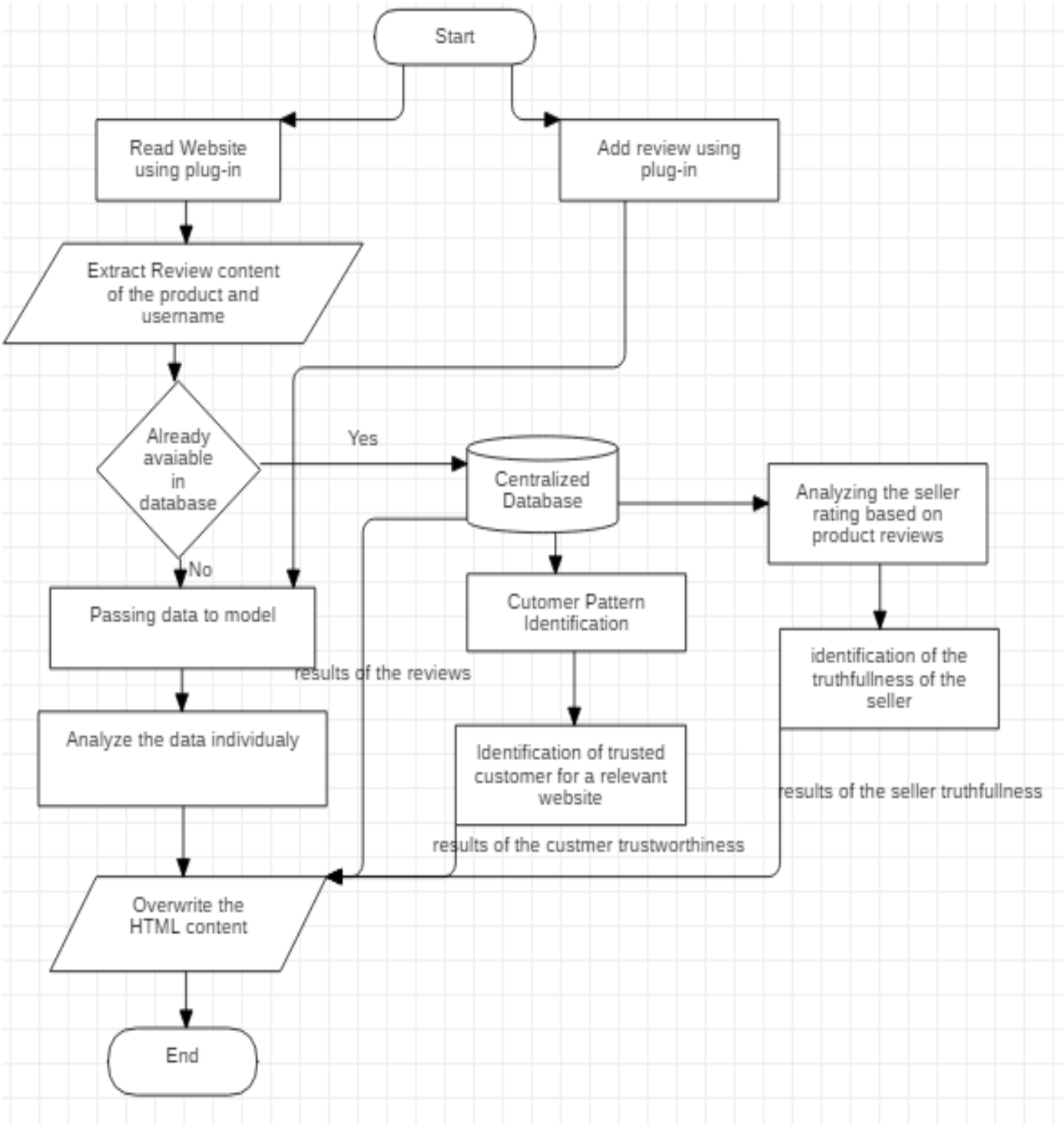

Figure 4.2.2: Flow Chart of the fake review detection project

Figure 4.2.2 show the flow chart of the project. As the initial step plug-in will be able to identify the content of the reviews in an e-commerce web page. After identifying the content of the reviews it will extract the content of each review one by one.

Once it is extracted it will check whether the data is already available in the centralized database or not. If it is available in the database straightway plug-in will display the results. If it is not available in the database, the data will be sent for analysing the review using machine learning algorithm. For the purpose of analysing semi- supervised learning were adapted. After analysing every review of the web page overall rate also will be calculated based on the number of fake reviews identified and sent back to the database which will be needed for further calculations.

The results of the trained data set from the centralized database will be used to identify the truthfulness of the customer and the trustworthiness of the seller. Four category of results will be display using the plug-in.

There are:

1. Results of the review identified as either fake or genuine
2. Result of the truthfulness of the customer who has added reviews.
3. Results of the trustworthiness of the seller.
4. Overall rating for the product.

Also, the plug-in will take the reviews which were entered by a person after installing it in the browser and analyse.

4.3 Assumptions and Limitations

Assumption

1. Reviews are posted in the mode of texture.
2. Large-scale review data sets which contain fake reviews can be used to test the system.

Limitation

1. Plug-in will be developed specific to one browser.

4.3 Implementation

Initially I have implemented the Google chrome browser plugin which can be used to extract data and pass data to the data processing model and finally display the result of the processed data. To implement this, I have used HTML, Bootstrap.CSS, Cheerio JS, NodeJs, Jquery.

I have used the initial template of the Google chrome browser plugin and extended the plugin with the required modification. For the basic plugin I have used HTML, and bootstrap CSS. Then for the web scraping data from web pages I have used NodeJS and Cheerio Js.

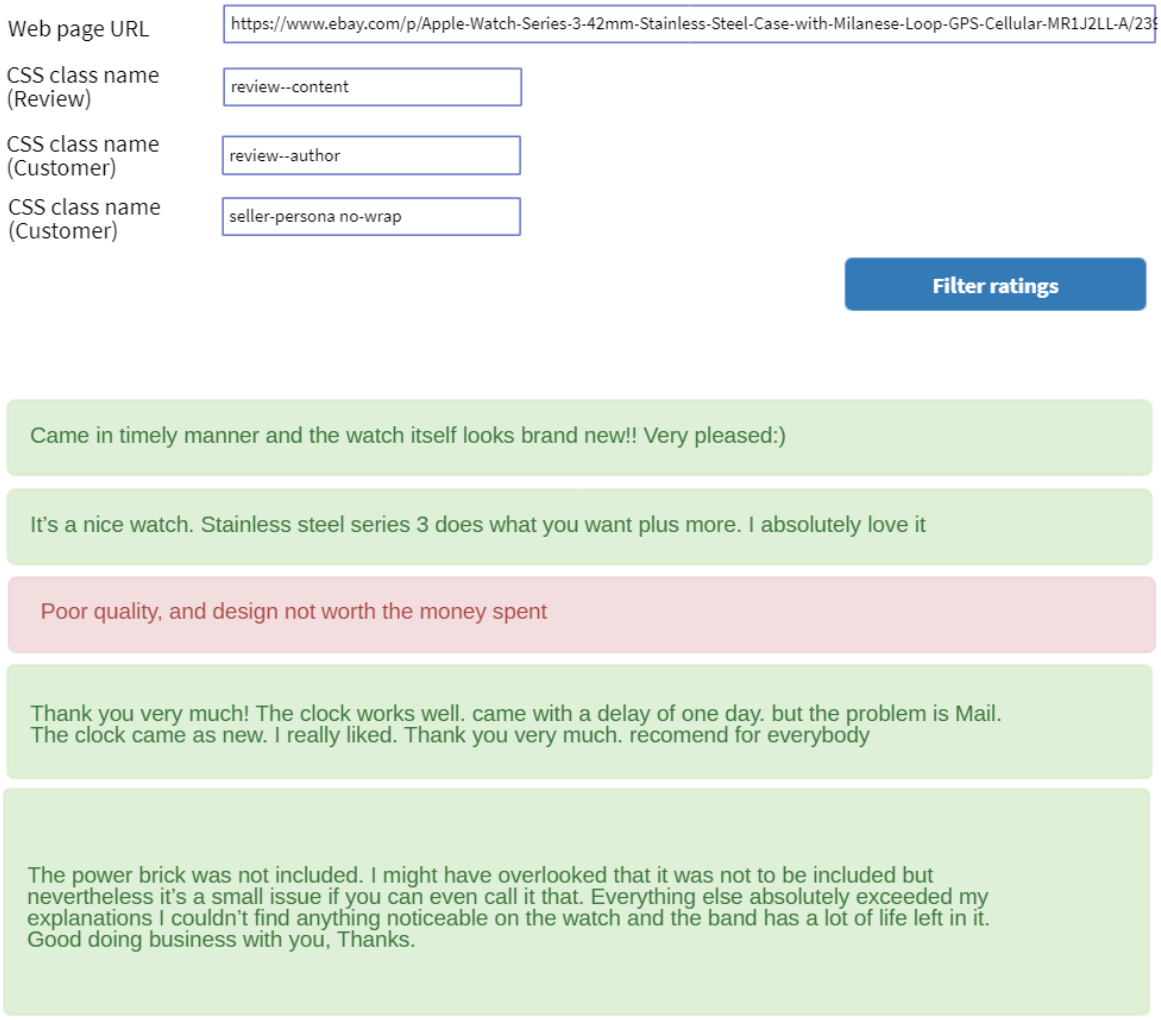

Figure 4.3.1.: Google chrome plugin

Next phase of the implementation is to development of model for data processing to identifying fake reviews. In order to identify fake reviews, I have used Semi-supervised learning method and implemented it using Python language.

Collection of Data

I have used data set from amazon for processing, and I have assumed the collection of data is valid.

Preprocessing of Data

- **Cleaning of Data**

The information that I gathered had loads of copy records and the initial step was to expel these. Following this, I adjusted the date field of the considerable number of records to guarantee that the designing was steady.

- **Handling of Text reviews**

The initial step here was to expel all the Stop Words. Stop Words will be words which don't contain significant essentialness to be utilized in inquiry questions. These words are sifted through in light of the fact that they return tremendous measure of superfluous data. At that point I changed over the content to lower case and evacuated accentuations, unique characters, void areas, numbers and regular word endings. At last, I made the Term Document Matrix to discover comparability between the content audits. Following is the word haze of the content audits:

- **Calculating Behavioral Dimensions**

Utilizing the properties that I separated, I distinguished the accompanying four social highlights that could be utilized to fabricate the classifier

Maximum Number of Reviews: This element processes the most extreme number of surveys in multi day for a creator and standardizes it by the greatest incentive.

Review Length: This feature is basically the number of words in each preprocessed text review

Rating Deviation: This component finds the deviation of analyst's evaluating for a specific eatery from the normal rating for that café (barring the commentator's appraising) and normalizing it by the most extreme conceivable deviation, 4 on a 5-star scale.

Maximum Content Similarity: For figuring this element, I previously processed the cosine likenesses for each conceivable pair of surveys that are given by a specific analyst. At that point, I pick the limit of these cosine likenesses to speak to this component

Sampling

Utilizing arbitrary inspecting, I split the informational collection into preparing and testing sets in the proportion of 70:30 individually. At that point I separated the preparation set to such an extent that around 60 % of the records were unlabeled and the remaining were named. Following this, I utilized subsets of expanding sizes from the marked information to prepare the base student (Naïve Bayes). To produce the subsets of named information, I utilized both straightforward irregular examining and stratified inspecting approaches. The consequences of these methodologies are talked about in the Experiment and Results' area.

Machine Learning Algorithm

In this project I have used semi-supervised learning with self-training. Also the use of Self-learning enabled to use many classifiers such as Naïve Bayes, Decision Trees, Logistic Regression and compare the performance.

Semi-supervise setting

To start with, I use Naïve Bayes as a base learning to prepare few labeled information. The classifier is then used to anticipate for unlabeled data dependent on the arrangement certainty. At that point, subset of the unlabeled data was taken, together with their expectation marks and train another classifier. The subset for the most part comprises of unlabeled models with high-certainty expectations over an explicit edge esteem.

In addition to use of Naïve Bayes, Decision Trees and Logistic Regression also were used. The performance of every model of the semi-supervised learning then be compared and contrasted among each other.

Chapter 05 – Evaluation

During the all the evaluation process, I have kept up a similar proportion of data sets.

In order do the experiment beforehand I have initiate a null hypothesis.

Naïve Bayes experiment

Stratified sampling for both semi-supervised model and the supervised model gave similar results. Naïve Bayes with simple random sampling for less number of data performed better with the semi-supervised model than the supervised model. For better outcomes I have expanded the quantity of labeled data, yet shockingly semi-supervised approach was not give dependably a better outcome as supervised model. This is a deviation from initial prediction.

Decision Tree experiment

With the increment of labeled data set, accuracy of both the models set to a stable value (Approximately 80%). Decision Tree works better with the semi-supervised model contrasted with the supervised model. This coordinated with the hypothesis I have made.

Logistic Regression experiment

With the increment of labeled data set, accuracy of both the models set to a stable value (Approximately 80%). But semi-supervised learning model was not recommending preferred outcome over supervised model. This is a deviation from the initial hypothesis.

Browser plugin development

To read review content, customer name, seller nameI have used the CSS class property assigned for the element. Which makes easier to read all the related information at a glance. Since various websites maintain different CSS classes I ad to come up with a general solution

which can be applied to all the sites, which is manually checking the CSS class using inspect element. Most of the cases this works fine but in some cases due to the Javascript errors in the website or may be not having uniformity within the site make difficult to get the 100% success rate.

Chapter 06 –Conclusion and future work

5.1 Conclusion

This study primarily was designed to detect fake review of e-commerce website and identify the real customers and sellers. Therefore, this study mainly divided into three part: one is to develop a web browser plug-in, second one is to detect fake reviews and third part is to identify the trustworthiness of the customers and sellers. For extraction of the reviews I have used web scraping. Once each review is sent for analyzing it was done by the use of semi-supervised learning in machine learning.

In order to identify the truthfulness of the customer and the trustworthiness of the seller I have used the results of the trained data set from the database in web server.

Identifying the truthfulness of the customer was done by analysing the identified fake reviews along with customer details of a particular e-commerce website. After analysing, if the particular customer is suspicious his/her details was added to a separate table in the database. As the final step of identifying truthfulness of the customer the plug-in will continuously monitor his/her next two reviews. If those reviews are also detected as fake, then the customer will be identified as an untrusted customer and display the customer in red.

For identifying the seller trustworthiness, the fake reviews are analysed alone with the seller details and the product. Here, the condition is that if the 30% from reviews are identified as positive fake reviews relevant to the seller details. By considering the above condition and the overall rate of the seller, the seller can be identified either a suspicious seller or not.

As a summary the rating which was calculated during the study by analyzing the user reviews and avoiding the fake reviews has a significant difference from the existed overall star rating.

Therefore, we can reach to conclusion that the star ratings are not much reliable, and it would not be the most suitable factor to get to know the user experience and satisfaction of a product.

5.2 Future Work

Today, many people are using social media marketing and selling to increase their selling rate of the products. Since there is no any comment or feedback found in some products or some brands, combining the API extraction of social media data such as Facebook and youtube review comments would be much beneficial to enhance the accuracy of the plug-in. Thus, as the future work I would suggest to use social media data in order to have more accurate overall rating for the product. This will useful for business intelligent to decide market penetration of the product throughout the world.

I would suggest to that the technology of the fake detection model can be changed into deep learning model, which can learn on its own. Also, for the extraction of the reviews from the e-commerce web page I would suggest to use selenium automation for web scraping.

Reference

[1] M. M. Yunus and A. Suliman, “Information & Communication Technology (ICT) tools in teaching and learning literature component in Malaysian secondary schools,” Asian Soc. Sci., vol. 10, no. 7, pp. 136–152, 2014

[2] Hu, N., Liu, L., and Zhang, J.J., "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects", Information Technology and management, 9(3), 2008, pp. 201-214.

[3] Chen, R.Y., Guo, J.Y., Deng, X.L.: Detecting fake reviews of hype about restaurants by sentiment analysis. In: Chen, Y., Balke, W.-T., Xu, J., Xu, W., Jin, P., Lin, X., Tang, T., Hwang, E. (eds.) WAIM 2014. LNCS, vol. 8597, pp. 22–30. Springer, Heidelberg (2014)

[4] Malbon, J. (2013). Taking fake online consumer reviews seriously. Journal of Consumer Policy, 36, 139-157. doi:10.1007/s10603-012-9216-7

[5] Fakespot | About Fakespot
https://www.fakespot.com/about

[6] Munzel, A. 2016. “Assisting Consumers in Detecting Fake Reviews: The Role of Identity Information Disclosure and Consensus.” Journal of Retailing and Consumer Services 32:96–108.

[7] Fakespot: uncover fake reviews on Amazon, Yelp and Tripadvisor - gHacks Tech News
https://www.ghacks.net/2018/01/03/fakespot-uncover-fake-reviews-on-amazon-yelp-and-tripadvisor/

[8] Review Reveal - Identify Fake Amazon Reviews
https://chrome.google.com/webstore/detail/review-reveal-identify-fa
/iildffljigfpnmfkhkfmglknmihbihlg?hl=en

[9] Long NH, Nghia PHT, Vuong NM (2014) Opinion spam recognition method for online reviews using ontological features. Tap chi KHOA HoC DHSP TPHCM (61) 44

[10] Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock (2011 (Published). "Finding Deceptive Opinion Spam by Any Stretch of the Imagination,") Conference Finding Deceptive Opinion Spam by Any Stretch of the Imagination. 309–319, http://reviewskeptic.com/.

[11] Bhattarai, A., V. Rus, D. Dasgupta (2009 Published). "Characterizing Comment Spam in the Blogosphere through Content Analysis," Conference Characterizing Comment Spam in the Blogosphere through Content Analysis.

[12] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, Hady Wirawan Lauw, Detecting product review spammers using rating behaviours, Proceedings of the 19th ACM international conference on Information and knowledge management, October 26-30, 2010, Toronto, ON, Canada

[13] Jindal Nitin, Liu Bing. Opinion spam and analysis[C]//Proceedings of the 2008 International Conference on WebSearch and Data Mining. New York: ACM Press,2008:219-230.

[14] Jindal Nitin, LIU Bing. Review spam detection[C] //Proceedings of the 16th international conference on WorldWide Web. Canada. New York: ACM Press ,2007:1189-1190.

[15] LIN Shuyang , WANG Guan ,XieSihong et al. Reviewspam detection via temporal pattern discovery [C] //Proceedings of the 18th ACM SIGKDD InternationalConference on Knowledge Discovery And Data Mining.New York:ACM Press , 2012:823-831.

[16] Jindal Nitin, Lim Ee-Peng, Nguyen Viet-An, et al.Detecting product review spammers using rating behaviors[C]// Proceedings of the 19th ACM international conferenceon Information and knowledge management. New York:ACM Press , 2010:939-948.

[17] Rong, Zhang et al. Exploiting shopping and reviewingbehavior to re-score online evaluations [C]//Proceedings ofthe 21st international conference companion on World Wide Web. ACM, 2012.

[18] Arjun, Bing Liu Mukherjee, and Natalie Glance. Spottingfake reviewer groups in consumer reviews.[C]//Proceedings of the 21st international conference on WorldWide Web. ACM, 2012.

Appendices

Following code explains the use of Cheerio JS and Jquery in web scrapping

```
'use strict';
const cheerio = require('cheerio');
const request = require('request');
var util = require('util')
var Reviews = require('../model/appModel.js');

exports.list_all_reviews = function(req, res) {
   Reviews.getAllReviews(function(err, task) {
   if (err)
     res.send(err);
   console.log('res', task);
   res.send(task);
 });
};

exports.getJSON = function (req, res, callback){
 // Set the headers
 var headers = {
 'User-Agent':       'Super Agent/0.0.1',
 'Content-Type':     'application/x-www-form-urlencoded'
 }

 var options = {
   url: req.param('url'),
   method: 'GET',
   headers: headers,
   }
 console.log("hhh : "+req.param('url'));
 request(options, function (error, response, body) {
 if (!error && response.statusCode == 200) {
     let $ = cheerio.load(body);
     var reviews = [];
     let title = $(req.param('class'), body);
```

```
        title.each(function(i, elm) {
            reviews.push({"review": $(this).text()});
        });
        callback(res.send(JSON.stringify(reviews)));

    }
    else
        console.log(error);
    })
}
```

Node Js for create web server

```
const express = require('express'),
 app = express(),
 bodyParser = require('body-parser');
 port = process.env.PORT || 3000;


const mysql = require('mysql');
// connection configurations
const mc = mysql.createConnection({
   host: 'localhost',
   user: 'root',
   password: '',
   database: 'fake_review_detection'
});
// connect to database
mc.connect();

app.listen(port);

console.log('API server started on: ' + port);

app.use(bodyParser.urlencoded({ extended: true }));
app.use(bodyParser.json());

var routes = require('./app/routes/approutes'); //importing route
routes(app); //register the route
```