Product Quality Classification on Shopee through User Review Sentiments

Edgar Aaron A. Go and Rodolfo C. Camaclang III

***Abstract*— This study explores the integration of machine learning into e-commerce platforms for automated product legitimacy detection, focusing on Shopee's Philippine site. By developing a Chrome browser extension that applies TF-IDF for text vectorization and logistic regression for classification, the system analyzes user reviews to determine the legitimacy of products. The study aims to protect consumers from counterfeit products while fostering an authentic marketplace for legitimate sellers. Data was collected from Shopee reviews, which were used to train and evaluate the model. The model achieved an accuracy of 84.5%, along with high precision, recall, and F1-scores. In terms of usability, a System Usability Scale (SUS) score of 76.72 confirmed the extension as user-friendly and effective, providing a reliable solution to improve trust and transparency in e-commerce.**

***Index Terms*— machine learning, sentiment analysis, product classification, e-commerce**

I. Introduction

A. Background of the Study

In recent years, online marketplace shopping or e-commerce is undergoing a high rate of development. It has revolutionized the way consumers access products globally, offering convenience and diversity of results. Since the ease of access of the platforms is recognized by most users, it has brought forth several challenges with it.

Counterfeit items, deceptive product descriptions, and fraudulent practices are a common encounter in these sites. These encounters have been classified by the Internet Crime Complain Center (IC3) as online fraud and can be non-delivery of items, item description, representation mismatch, and so on [1]. Since it has become prevalent, it has prompted the exploration of innovative solutions in the field of machine learning to protect consumers engaging in online transactions.

Reviews are an important factor in determining product legitimacy. It may be accompanied by a star rating and their feedback regarding the product, commenting on quality, accuracy to the description, timeliness of delivery, customer services, and others. These can be gathered to be used for machine training and analysis to create a generalization whether the item listed is counterfeit or not.

There have been several approaches developed and used in past studies to analyze product reviews. A survey of the performance of the following algorithms—Naïve Bayes (NB), Multinomial NB, Bernoulli NB, Logistic Regression, Stochastic Descent learning, Support Vector Machine Classifier (SVC), Linear SVC, and Nu SVC—was considered by [2] to classify reviews, with Logistic Regression Classifier receiving the highest mean accuracy at 65.61%. Ensemble learning method or Random Forest was tested against Decision Trees by [3], finding that decision trees have better performance than the Random Forest method. In the study of [4], on large datasets, the support vector machine classifier provided the best classifying accuracy.

However, the previously mentioned methods and approaches only analyzed the reviews of a product, providing a Boolean output of positive or negative review or classifying it in a scale between 1 and 5. These findings are unable to provide an accurate decision on whether the product that the reviews are taken from is legitimate or not.

This study seeks to design and implement a browser extension that can automatically classify a product that the user is searching for. Machine learning technologies will be used to analyze the review section of the product, creating an accurate assumption whether the product is a legitimate item through review sentiments.

The paper is structured as follows. Following the background of the problem is the section for the objectives, significance, scope and delimitations, and definition of terminologies. This is followed by the conceptual literature and review of related literature and is concluded by the methodology section.

B. Significance of the Study

The integration of machine learning into e-commerce platforms for automated product legitimacy detection will allow for restoring and keeping trust within online marketplaces. By deploying a precise model that could accurately classify listed products, this study will be able to provide a proactive defense against counterfeit products and false advertisements. Beyond protecting consumers, the study recognizes the importance of developing an authentic environment that benefits legitimate sellers, contributing to the overall health and sustainability of the e-commerce ecosystem.

C. Objectives of the Study

The general objective of the study is to develop a Chrome browser extension that uses TF-IDF to create vectors to be used to train a logistic regression model. This is used to classify a product accurately from an ecommerce platform in its legitimacy through sentiment analysis from user reviews. The specific objectives of the study are the following:

1) Annotate the data manually given a set of criteria

Presented to the Faculty of the Institute of Computer Science, University of the Philippines Los Baños in partial fulfillment of the requirements for the Degree of Bachelor of Science in Computer Science

2) Preprocess and create vectors using TF-IDF from the annotated data to train a logistic regression model
3) Develop an application to perform API requests to retrieve user reviews with text from a Shopee product page
4) Develop a Chrome extension to automate product quality classification based on the results of the model; and
5) Evaluate the performance of the model given the evaluation metrics.

D. Scope and Limitations

This study will only focus on the multinational e-commerce platform Shopee, mainly on the Philippine version, as the website to conduct the data gathering and experiments required. The reviews will be retrieved from Shopee products to be used for training and sentiment analysis. The use of other e-commerce websites will not be considered in this study. The predictions and results of the model will be heavily reliant on the data set it will be trained on, which may affect the overall accuracy of the findings.

E. Definition of Terms

1) Machine learning – a branch of artificial intelligence (AI) that allows for computers to learn and make predictions and decisions based on the data it has processed.
2) Sentiment analysis – a technique that involves determining information based on text data to classify it whether it is positive, negative, or neutral.
3) NLP – natural language processing. A branch of AI that allows processing and understanding of text such as user reviews.
4) TF-IDF - term frequency–inverse document frequency; creates vectors and measures its importance in a document relative to a corpus
5) Logistic Regression Model - supervised machine learning algorithm used to perform binary classifications once vectors are fitted in the model

II. Review of Related Literature

A. E-commerce

[5] defines e-commerce as the conduct of sale, purchase, transfer, or exchange of products, services, and information through electronic means and technologies, which can be within business, business-to-business, and business-to-consumer interactions.

Companies perform their businesses online in four main avenues: direct marketing, selling, and services; online banking and billing; secure distribution of information; and value-chain trading and corporate purchasing [6].

B. Online Fraud

An online fraud can be any invitation, request, notification, or offer, that is spread across the Internet to obtain a victim's information or money, or otherwise obtain a financial benefit through deceptive means [7]. Swindling is a term used in e-commerce to describe a deliberate action to obtain money from customers by offering false services and resources [8].

Online fraud has been classified by the Internet Crime Complaint Center (IC3) into the following six categories: non-delivery of goods; misrepresentation of the items; triangulation; fee staking; selling of black-market goods; and multiple bidding and shill bidding [1].

C. Machine Learning

Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms [9]. It is also a dynamic field of computational algorithms created to mimic human intelligence through the process of learning from the surrounding environment [10]. Every interaction and action executed by the machine becomes valuable information for the system, enhancing its capabilities for future use [11]. It is a fast-growing area in today's technology that combines computer science and statistics. It plays a central role in both artificial intelligence and data science [12].

A machine learning algorithm typically follows three steps, starting with a decision process, an error function, then the model optimization process (IBM, n.d.).

D. Related Studies

E-commerce websites are being taken advantage of by listing fraudulent products and preying on misinformed consumers. In recent years, the growth of e-commerce has transformed the way people shop, offering high convenience and accessibility. While this has undoubtedly brought numerous benefits, it has also given rise to a concerning trend – the proliferation of fraudulent products in online platforms. There are sellers that take advantage of the vastness of the internet to peddle substandard or non-existent items, leaving unsuspecting consumers at a significant risk of financial loss. The surge in e-commerce fraud has become a pressing issue that demands attention and effective solutions. Machine learning methods and sentiment analysis are solutions that can be used to identify fraudulent products, safeguarding consumers, and maintaining the integrity of online marketplaces.

Machine learning methods involving natural language processing and ensemble learning have been utilized for the detection of fraudulent products. [2] employed various techniques, such as Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Stochastic Descent Learning (SGD) Classifier, Support Vector Machine Classifier (SVC), Linear SVC, and Nu SVC, to create a review-based rating prediction system. Through testing on five Yelp datasets, the logistic regression classifier exhibited the highest mean accuracy at 65.61%, with a peak accuracy of 65.95%.

A study by [3], explored the effectiveness of machine learning algorithms in detecting fake goods, focusing on decision trees and random forest. The ensemble learning algorithm Decision Tree outperformed the Random Forest algorithm, showcasing its superior performance in identifying counterfeit

goods. The evaluation involved a separate validation dataset, reinforcing the algorithms' effectiveness beyond the training data. Notably, the decision tree algorithm exhibited significantly higher accuracy than the random forest algorithm, correctly classifying 2000 more products, as indicated by a confusion matrix [3]. [13] used review polarity to assess product legitimacy, exemplified by the M3 Smart Health Watch. Review polarity was determined by aggregating scores from customer ratings, comments, and media content. The sentiment analysis indicated that the M3 Smart Health Watch, with predominantly positive reviews, may not require fraud detection analysis. This approach provides potential buyers with confidence in purchasing the watch without concerns [13].

Sentiment analysis is a branch of machine learning that can use user reviews to understand their sentiments towards a product which can help with determining product legitimacy. The Naïve Bayesian, Support Vector Machine Classifier (SVC), Stochastic Gradient Descent (SGD), Linear Regression (LR), Random Forest, and Decision Tree are the algorithms used by [4] in their experiment. To determine the best accuracy among the algorithms, cross-validation methods and 10-fold accuracy have been conducted. There are three categories of product reviews where each algorithm will classify and provide results based on evaluation metrics. TF-IDF and bag-of-words feature selection processes provided the best results for all the datasets. The support vector machine provided the greatest accuracy in each data set as SVMs perform better in working with large scale data sets without the risk of overfitting. Several experiments are conducted to determine the performance of two machine learning models. To measure performance, evaluating metrics will be used, and measuring accuracy is the most suitable for this purpose. The precision of a classifier on a specific test dataset is determined by the percentage of instances correctly categorized by the classifier. The system's performance is assessed using three widely used statistical metrics: recall, precision, and F-measure, which are derived from a confusion matrix. The confusion matrix categorizes data into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP signifies accurate predictions of the positive class, while FP indicates incorrect positive predictions. TN represents accurate predictions of the negative class, and FN denotes incorrect negative predictions by the system. The SVM classifier shows a precision of 82.853%, a recall of 82.884%, and an F1 score of 82.662%. In contrast, the Naive Bayes classifier achieves a precision of 83.990%, a recall of 83.997%, and an F1 score of 83.993%. Consequently, the SVM and Naive Bayes models exhibit accuracies of 84% and 82.875%, respectively, outperforming traditional techniques [14].

According to [15], the combination of word vector, word frequency, and FCNN methods exhibits superior performance across various metrics, boasting a precision of 0.97, recall of 0.81, F1 score of 0.92, and a training time of 10.82 seconds. In their study, a performance comparison with other algorithms further supports the efficacy of this combined approach, revealing a recall of 0.81, precision of 0.97, and an F1 score of 0.92 [15].

Moving to the experiment conducted by [16], which focused on sentiment analysis reviews using the Naive Bayes Algorithm within Shopee, the researchers utilized Knime to measure experimental accuracy. The dataset comprised 200 reviews, half positive and half negative, aiming to assess the algorithm's effectiveness in accurately classifying sentiments within Shopee reviews. Out of the 60 testing data in the second partition related to product reviews, 28 positive reviews were accurately predicted, with 2 negative reviews incorrectly included in positive predictions. Notably, no positive reviews were misclassified as negative. Additionally, 30 negative reviews were accurately predicted, resulting in an accuracy value of 96.667%, precision of 100%, and recall of 93.3%. These metrics underscore the analysis's effectiveness in correctly identifying and categorizing sentiments within Shopee reviews [16]. The accuracy testing results for the Naive Bayes algorithm, utilizing partitioning techniques, yielded an AUC value of 1.00 for the positive class, indicating excellent classification performance and emphasizing the algorithm's effectiveness in accurately identifying and categorizing the positive class within the dataset [16].

Other than evaluating product ratings and reviews, [17] propose a method for detecting deceptive reviews that influence consumer buying patterns, considering various attributes of a reviewer to create a unique profile for each customer. The method demonstrates an effective capability to identify spam activities under specific assumptions, with preliminary experiments showing promising results, highlighting the potential effectiveness of the proposed strategy in addressing the impact of fake reviews [17]. While existing literature primarily focused on addressing spam words through various techniques, deficiencies persisted, particularly concerning elements such as IP addresses, MAC addresses, and email accounts, with limited application of machine learning approaches. In response, [17] proposed approach covers these aspects, incorporating IP addresses, MAC addresses, and email accounts and leveraging machine learning techniques for enhanced efficiency. This comprehensive approach aims to improve detection accuracy, especially in cases of less common fraud scenarios. Testing the method on a set of 90 reviews demonstrated an accuracy rate of up to 75%, showcasing its potential to enhance fraud detection compared to prior techniques [17].

III. Materials and Methods

A. Research Design

This study is a quantitative research as it follows a descriptive evaluative design. It is descriptive because the model that is developed is evaluated and trained through text descriptions from product reviews.

B. Devices and Technologies Used

This study was conducted on an ASUS ROG Strix G531GT laptop with a processor of Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz clock rate and 8.0 GB of RAM.

The programming language that was used for developing the model is Python as there are several readily available libraries that assists on the study, such as scikit-learn for the model, and

Flask to create a local server for the extension to communicate with to the model. Here is a list of all standard library imports and external packages used in the classifier along with what they were used for:

1) re: Provides support for regular expressions. Used to check if the provided link is valid.
2) csv: Reading and writing of CSV (Comma Separated Values) files. All data that the model uses for training and classifying are in CSV format.
3) datetime: Used to retrieve date and time for the file name of the downloadable CSV file.
4) warnings: Allows control over warning messages in the code.
5) flask: A lightweight web application framework for Python. Hosts the local server that the extension communicates with.
6) flask-cors: Provides Cross-Origin Resource Sharing (CORS) support to Flask for sending requests.
7) sklearn
    a) sklearn.feature_extraction.text.TfidfVectorizer: Converts a collection of raw documents to a matrix of TF-IDF features.
    b) sklearn.linear_model.LogisticRegression: Implements logistic regression classifier.
    c) sklearn.metrics.accuracy_score: Measures the accuracy of a classification model.
    d) sklearn.metrics.classification_report: Generates a text report showing main classification metrics.
    e) sklearn.metrics.confusion_matrix: Computes the confusion matrix to evaluate the accuracy of a classification.
8) joblib: Allows saving and loading of the trained model.

TF-IDF was used to measure the importance of a word in a document relative to a corpus, highlighting terms that are significant in a specific document but not in the entire corpus. It is commonly used in information retrieval to rank the relevance of documents to a given query.

The model used by the study is the logistic regression model as it is used for binary classification. This model learns to classify the reviews into positive or negative sentiment based on the feature representations derived from the vectors generated by TF-IDF. Then, the model is able to generalize patterns in the data and make predictions on new reviews.

C. Model Training Process

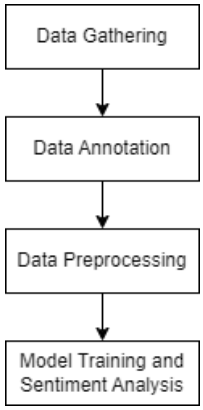


Fig. 1: Steps in the Model Training Process

1) Data Gathering Procedure: A published data set was used for training the model and was retrieved from Hugging Face and [18]. The data set contains 13,000 Shopee reviews for training and 2,250 for testing. This data contains the review itself and the rating the user have provided, and the language of the review is in Tagalog but contains a blend of English words.

A Python program was created to perform requests to Shopee API in order to retrieve the reviews of the product that the user wishes to classify. These reviews were then processed through the trained model.

2) Data Annotation: In the process of annotating the training dataset, two respondents, which are both university students who have a proficiency in the English and Tagalog language, along with the researcher independently provided annotations for each data point as inaccuracies may exist from the ratings. A rubric is created below by the researcher to be used as a guideline on how to annotate the reviews. The reviews were annotated as either negative (0) or positive (1).

- Negative (0): Users that show dissatisfaction with the product's performance, features, quality, price, or seller service. Negative experiences such as difficulties in using the product. Criticisms and recommendations to not purchase the product.
- Positive (1): Users that show satisfaction with the product's performance, features, quality, price, or seller service. Positive experiences such as product reliability and ease of use. Recommendations to other users and likelihood to purchase from seller again.
- Neutral: These are removed from the data set. Provides mixed feedback and shows ambivalence about the received product.

The annotations from the respondents and the researcher were compared, and the majority decision was chosen as the final annotation for each data point. A total of 5,329 positive reviews and 5,171 negative reviews were annotated in the final data set.

3) Data Preprocessing:

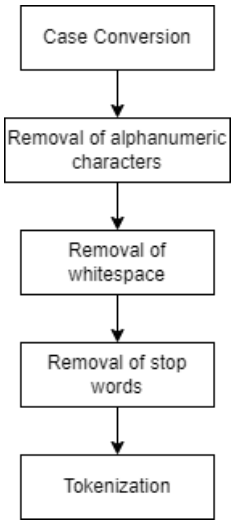


Fig. 2: Steps in the Data Preprocessing

The reviews in the data set underwent multiple data preprocessing steps to improve performance, reduce the training time of the model, and to remove unnecessary information. The preprocessing steps that will be used are the following:

- Case conversion: The reviews are transformed into lowercase.
- Removal of non-alphanumeric characters: Alphanumeric characters are retained. The removed characters are then substituted with a single space.
- Removal of whitespace: Tabs, newlines, and multiple spaces are replaced into a single space.
- Removal of stop words: Words such as ”the”, ”are”, and ”is” are removed from the list of tokens. Tagalog stop words were also removed from the list of tokens [19].
- Tokenization: TF-IDF was used to convert the reviews into tokens with vector representation.

4) Model Training and Sentiment Analysis: The model was trained through a logistic regression model classifier by sklearn given the processed reviews. A testing data set was also preprocessed to be used to determine the accuracy of the model's classifications. Once trained with good accuracy, reviews will then be retrieved from a product page to be processed through the model. A support vector machine (SVM) and Logistic Regression model in combination with bag-of-words and Word2Vec for creating vectors was also used to compare performances.

D. Statistical Analysis

In this study, several evaluation metrics was used to determine the overall performance of the machine learning model. These are as follows:

- Accuracy: The ratio of correctly predicted instances to the total instances. May or may not be suitable depending on the balancing of the data set.
- Precision: The ratio of true positive predictions to the total predicted positives. It is the accuracy of positive predictions.
- Recall: The ratio of true positive predictions to the total actual positives. This metric shows if the model can capture all positive instances.
- F1 Score: The mean of precision and recall. It provides a balance between precision and recall.
- Confusion matrix: A table that helps visualize and summarize the model's performance, displaying true positives, true negatives, false positives, and false negatives.

In order to classify the product based on the reviews, the percentage of positive reviews was taken from the output. Classification of products are based on the following criteria following the study of [20]:

- 0-25% Positive Reviews: Poor Quality and Possibly Fraudulent
- 26-40% Positive Reviews: Below Average Quality
- 41-60% Positive Reviews: Average Quality
- 61-80% Positive Reviews: Above Average Quality
- 81-95% Positive Reviews: Excellent Quality and Legitimate Product
- 96-100% Positive Reviews: Excellent Quality but Unsure on Legitimacy on Product

E. Back-end

A local Flask server was used to facilitate communication between the Chrome extension and the classification model, enabling the extension to send data for classification and receive results in real-time. The server contains four primary routes that can be utilized by the Chrome extension to perform various functions:

1) /total (POST): This route accepts a URL in the request body, extracts the shop ID and item ID from the URL, fetches the item's total ratings from Shopee's API, and returns a summary of the total ratings and the product name in JSON format. This simply shows the user the number of reviews to be retrieved.
2) /retrieve (POST): This route accepts a URL in the request body, extracts the shop ID and item ID from the URL, and sends requests to Shopee API to retrieve the product reviews, saving the data to a CSV file. Requests are sent to get 20 reviews at a time to reduce the number of requests and a header specifying that the request is from a browser is required to the desired response.
3) /classify (GET): This route initiates the classification process of product reviews, reading the results from the CSV file saved from /retrieve and passing it through the model to calculate the positive and negative review counts and percentages. It also categorizes the quality of the product based on the percentage of positive reviews and returns this information in JSON format.
4) /download (GET): This route allows the user to download the classified review data as a CSV file, including a timestamp in the filename for easy reference.

Hosting the Python server online was considered instead of running a local server. However, free hosting services block API requests to websites that are not on their whitelist. The domain for shopee.ph and their API was not listed and paid subscriptions are needed to be allowed to send requests.

Due to the server being run locally, there are more information displayed on the server terminal that the users can utilize. The terminal shows the shop id and product id of the URL sent, the individual requests to the API, and the step it is currently in while classifying.

F. Front-end

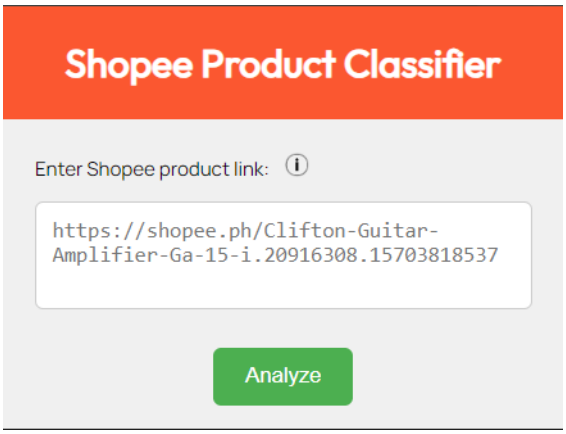


Fig. 3: Chrome extension main screen

A Chrome extension was developed using HTML and CSS for the user interface (UI) design and layout, and JavaScript for the functional aspects of the extension which is mainly performing API calls to the server. The main screen contains a text box wherein the user can enter a Shopee product link. A valid product link must be provided before it can be analyzed. A valid link consists of three main sections: the domain name, shop id, and item id.

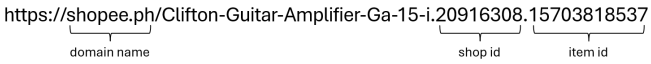


Fig. 4: An example of a valid link

After the button has been pressed and the Flask server is running, it sends the POST request to the server to classify the reviews in the product page. During this time, the server returns the number of total reviews of the product while it attempts to retrieve them, then finally returns the product name, classification, and number of positive and negative reviews that the product has. There is also a button to download a CSV file that contains the product reviews with its corresponding classification if the user wishes to check the output.

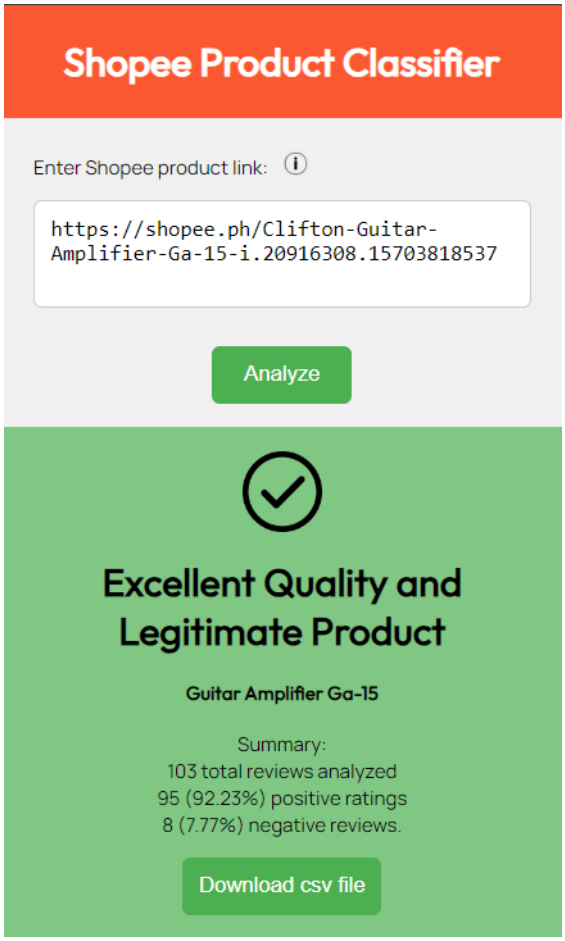
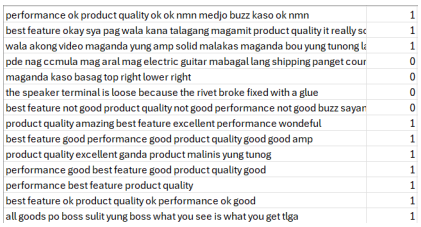


Fig. 5: A product with excellent quality



performance ok product quality ok ok nmn medjo buzz kaso ok nmn	1
best feature okay sya pag wala kana talagang magamit product quality it really sc	1
wala akong video maganda yung amp solid malakas maganda bou yung tunong l	1
pde nag ccmula mag aral mag electric guitar mabagal lang shipping panget cour	0
maganda kaso basag top right lower right	0
the speaker terminal is loose because the rivet broke fixed with a glue	0
best feature not good product quality not good performance not good buzz sayar	0
product quality amazing best feature excellent performance wondeful	1
best feature good performance good product quality good good amp	1
product quality excellent ganda product malinis yung tunog	1
performance good best feature good product quality good	1
performance best feature product quality	1
best feature ok product quality ok performance ok good	1
all goods po boss sulit yung boss what you see is what you get tlga	1

Fig. 6: Downloaded CSV file from the result

G. Feedback Collection

A system usability scale (SUS) survey was conducted to evaluate whether the Chrome extension developed for classifying products in Shopee met user expections and was user-friendly. The system usability scale serves as a versatile tool for assessing the overall usability of a product or experience. The scale was tailored to include 10 questions with five response options, ranging from strongly disagree to strongly agree. Each question alternates between positive and negative statements about the usability of the system being evaluated. The following are the questions within a SUS survey:

1) I think that I would like to use this extension frequently.
2) I found the extension unnecessarily complex.
3) I thought the extension was easy to use.
4) I think that I would need the support of a technical person to be able to use this extension.
5) I found the various functions in this extension were well integrated.

6) I thought there was too much inconsistency in this extension.
7) I would imagine that most people would learn to use this extension very quickly.
8) I found the extension very cumbersome to use.
9) I felt very confident using the extension.
10) I needed to learn a lot of things before I could get going with this extension.

The sampling technique that was used is purposive sampling. This allowed the researcher to find the most suitable respondents for the survey to assess the usability of the extension. The respondents chosen are university students between the ages of 20 and 24 and have prior experience in ordering and interacting with online marketplaces and are also frequent users of the platforms. A total of sixteen respondents was selected for the survey.

The survey was conducted through Google Forms for easier data gathering and visualization. A Zoom meeting was also scheduled with each respondent to guide them with installation of necessary technologies and loading of the extension. The respondents are able to give feedback as they use the extension and answered the survey during the meeting.

IV. Results and Discussion

A. Performance Evaluation of Sentiment Analysis Models

TABLE I: Performance Metrics for Different Methods

Method	Avg Precision	Avg Recall	Avg F1-score	Accuracy
W2V + SVM	0.80701	0.80667	0.80661	0.80667
W2V + LR	0.80470	0.80444	0.80440	0.80444
TFIDF + SVM	0.83293	0.83278	0.83276	0.83278
TFIDF + LR	**0.84523**	**0.84500**	**0.84497**	**0.84500**
BoW + SVM	0.79502	0.79500	0.79500	0.79500
BoW + LR	0.82400	0.82389	0.82387	0.82389

For the method Word2Vec and SVM, it achieves a precision of approximately 80.7%, indicating that about 80.7% of the instances predicted as positive are actually positive. The recall is also around 80.7%, meaning that 80.7% of the actual positive instances are correctly identified. The second method performs similarly to Word2Vec + SVM, with slightly lower precision, recall, F1-score, and accuracy, all around 80.4%. TF-IDF and SVM outperforms the Word2Vec-based methods, with a precision, recall, and F1-score of around 83.3%. The accuracy is also approximately 83.3%. TF-IDF and logistic regression shows the best performance among all methods, with precision, recall, and F1-score of around 84.5%. The accuracy is also approximately 84.5%. Bag-of-words and SVM performs slightly worse than the TF-IDF-based methods, with precision, recall, F1-score, and accuracy around 79.5%, and Bag-of-words with logistic regression performs better than BagOfWords + SVM and similar to Word2Vec + SVM, with precision, recall, F1-score, and accuracy around 82.4

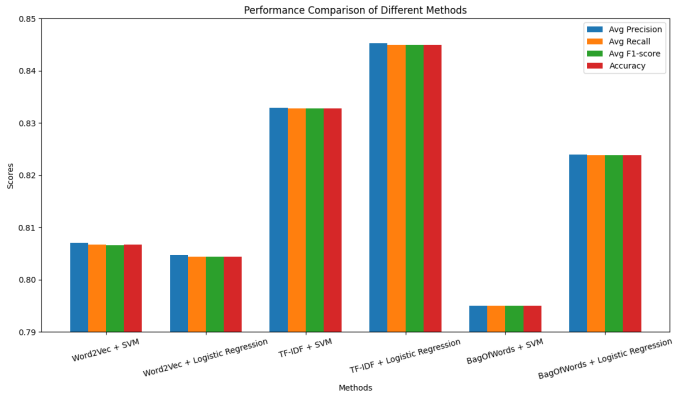


Fig. 7: Bar graph interpretation of Table 1

As shown in the Table 1 and Fig. 7, TF-IDF-based methods, especially when combined with Logistic Regression, show the highest performance across all metrics. Word2Vec-based methods and BagOfWords-based methods perform slightly lower but still achieve reasonable results. Logistic Regression generally outperforms SVM in this task, regardless of the feature extraction method used.

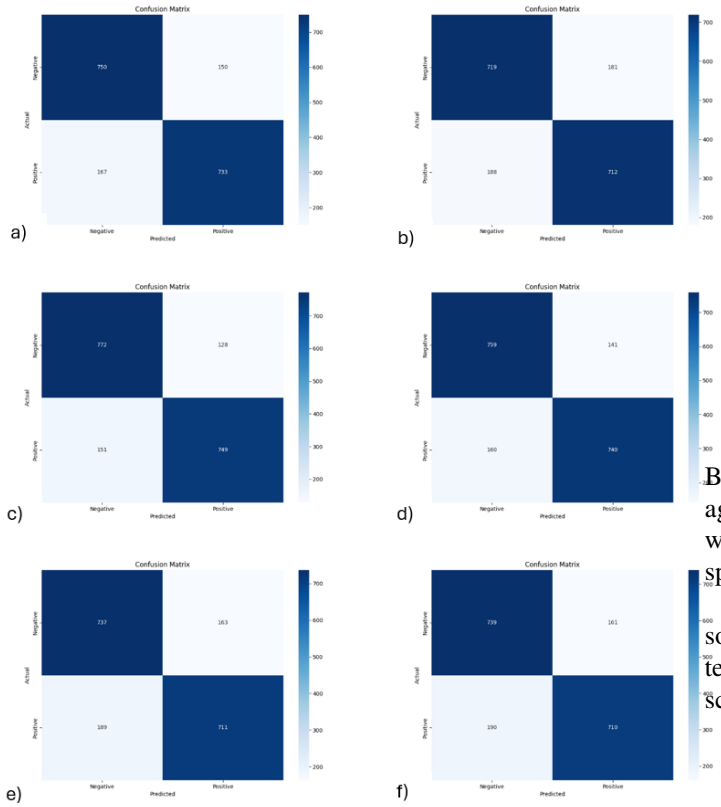


Fig. 8: Confusion matrices for a) BoW + LG, b) BoW + SVM, c) TFIDF + LG, d) TFIDF + SVM, e) W2V + LG, f) W2V + SVM

B. Usability Assessment Results

The following are the results of sixteen system usability scale surveys. Each question are to be answered in a range of 1 to 5, with 1 being strongly disagree and 5 being strongly agree.

Q1: I think that I would like to use this extension frequently:

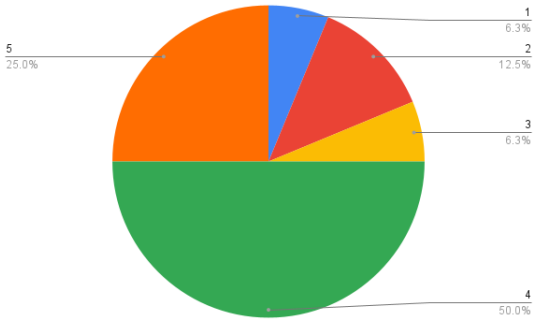


Fig. 9: Survey responses for question 1

Based on the results, the majority of the responses are 'agree'. 50% of the respondents agree that they will be using the extension frequently while 25% of the respondents strongly agree.

This shows that the extension can prove helpful in aiding the user in selecting the products that are legitimate based on the classifications. There are respondents that are neutral or disagrees with the statement. This may be about having to launch the local server every time one wishes to use the extension.

Q2: I found the extension unnecessarily complex:

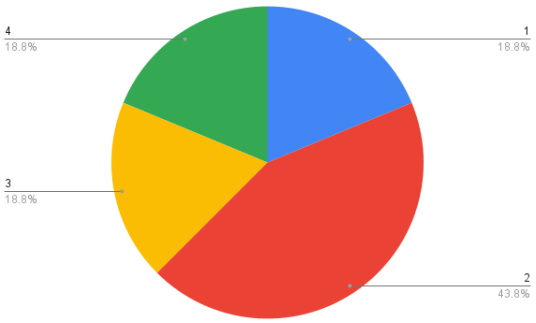


Fig. 10: Survey responses for question 2

Based on the results, the majority of the responses are 'disagree'. 43.8% of the respondents disagree that the extension was unnecessarily complex. There are equal percentages of responses (18.8%) for 'neutral', 'strongly disagree' and 'agree'.

The discrepancy on the 18.8% responses may show that some among the respondents are not well experienced in technical processes since they had to install everything from scratch. There are no respondents that chose 'strongly agree'.

Q3: I thought the extension was easy to use:

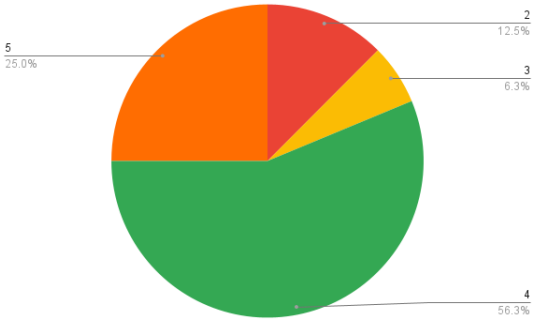


Fig. 11: Survey responses for question 3

Based on the results, the majority of the responses are 'agree'. 56.3% of the respondents agree and 25% of the respondents strongly agree that the extension was easy to use.

The extension does not require much effort to traverse, where at most requiring two clicks to classify and download. This explains the high number of respondents selecting 'agree' or 'strongly agree'. There are no respondents that chose 'strongly disagree'.

Q4: I think that I would need the support of a technical person to be able to use this extension:

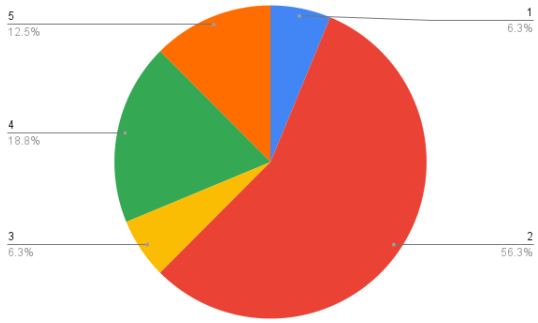


Fig. 12: Survey responses for question 4

Based on the results, the majority of the responses are 'disagree'. 56.3% of the respondents disagree that they would need support of a technical person to use the extension.

It can be understood as using the extension does not require any support for another person as it is identified to be easy to use, which is different to the difficulty encountered in installation in which the 31.3% that voted for 'agree' may have thought of.

Q5: I found the various functions in this extension were well integrated:

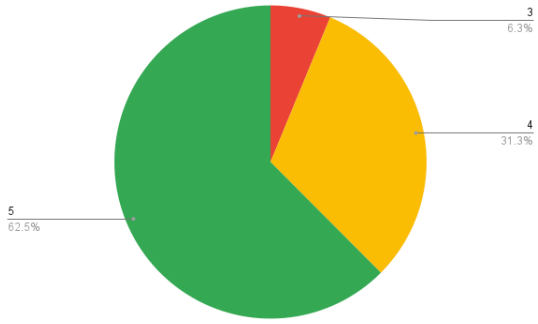


Fig. 13: Survey responses for question 5

Based on the results, the majority of the responses are 'strongly agree'. 62.5% of the respondents strongly agree that the functions of the extension are well integrated. 31.3% agree to this statement and 6.3% of the respondents are neutral. There are no respondents that chose either of the 'disagree' options.

This could be tied into the ease and simplicity of the extension wherein there is only four major functions that it performs, which is to retrieve URL, retrieve reviews, classify the product, and download functionality.

Q6: I thought there was too much inconsistency in this extension:

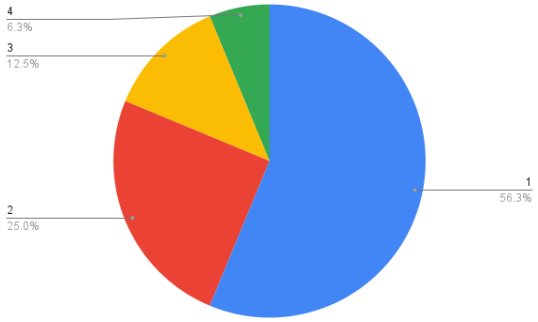


Fig. 14: Survey responses for question 6

Based on the results, the majority of the responses are 'strongly disagree'. 56.3% of the respondents strongly disagree that there was too much inconsistency in the extension. 25% also disagree about this statement. There are no respondents that chose 'strongly agree'. The same interpretation in Question 5 can be observed here.

Q7: I would imagine that most people would learn to use this extension very quickly:

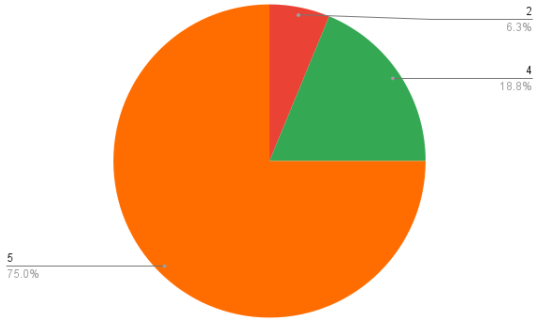


Fig. 15: Survey responses for question 7

Based on the results, the majority of the responses are 'strongly agree'. 75% of the respondents strongly agree that most people would be able to learn to use the extension very quickly.

More that a fourth of the respondents chose either the 'agree' options. This can be related to the ease of usage while the 18.8% of respondents who chose 'disagree' may have accounted the installation portion of the extension again. There are no respondents that chose 'neutral' and 'strongly disagree'.

Q8: I found the extension very cumbersome to use:

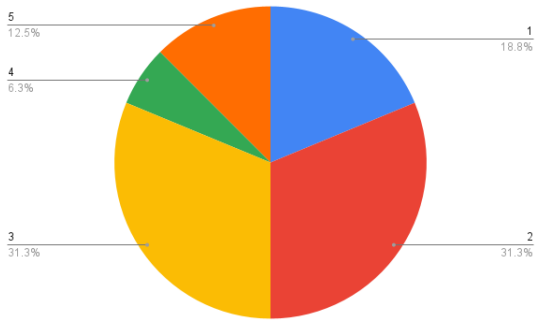


Fig. 16: Survey responses for question 8

Based on the results, the highest number of responses are shared between 'disagree' and 'neutral'. 31.3% of the respondents are neutral and 31.3% of the respondents disagree that the extension is very cumbersome to use.

This is another observation regarding the difficulty of installation and the process of having to start up the local server to use the extension, which resulted to a high percentage of 'neutral' responses. Disregarding these responses, 50% of the respondents chose either of the 'disagree' options and 18.9% of the respondents chose either of the 'agree' options.

Q9: I felt very confident using the extension:

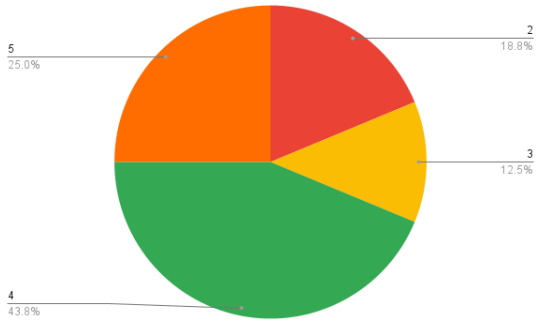


Fig. 17: Survey responses for question 9

Based on the results, the majority of the responses are 'agree'. 43.8% of the respondents agree that they felt confident in using the extensions. 25% of the respondents strongly agree about this statement.

After being instructed on how to use the extension, majority of the respondents are able to use the extension without support or guidance which lead to the 68.8% of responses being either of the 'agree' options. There are no respondent that chose 'strongly disagree'.

Q10: I needed to learn a lot of things before I could get going with this extension:

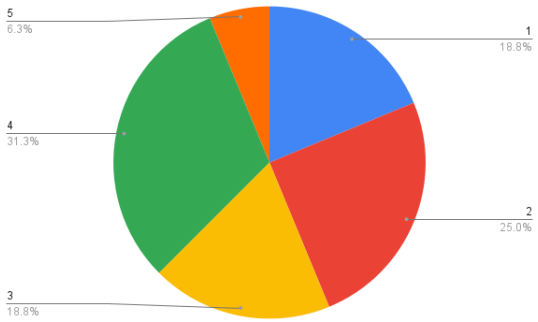


Fig. 18: Survey responses for question 10

Based on the results, the majority of the responses are 'agree'. 31.3% of the respondents agree that they needed to learn a lot of things before being able to use the extension. 6.3% of the respondents strongly agree about this statement. This totals to 37.6% of respondents choosing an 'agree' option and a total of 43.8% of the respondents chose a 'disagree' option.

C. System Usability Scale Scores

A SUS score is a metric that can be used in measuring the overall usability of applications or products. There are four steps in calculating the score:

1) Convert the responses of each respondent from each question into points
   - For odd numbered questions: User response subtracted by 1
   - For even numbered questions: 5 subtracted by user response
2) Add up all the points from each question
3) Multiply total points by 2.5 to get their user score
4) Repeat steps 1-3 until all user scores are computed, then get their average for the SUS score.

Respondent	SUS Score
1	70
2	90
3	60
4	80
5	90
6	70
7	85
8	70
9	72.5
10	60
11	75
12	75
13	85
14	85
15	80
16	80
Avg SUS Score	**76.71875**

TABLE II: User Scores and SUS Score

Table II shows the computed SUS score of each respondent of the survey along with the average SUS score. According to [21], an average of 68.2 mean score is expected for a

web interface type to indicate a good level of usability. [21] has created a scale in order to interpret a SUS score into an adjective rating, acceptability scores and school grading scales.

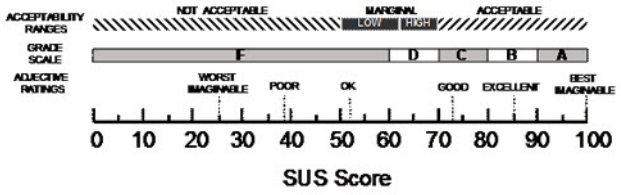


Fig. 19: SUS Scale

Using the scale shown on Fig. 19, by averaging with a 76.72 SUS Score, the Chrome extension lands in between the good and excellent adjective ratings, receiving a C grade scale, and is within acceptable range.

V. Conclusion

The main objective of this study is to develop a Chrome browser extension that uses TF-IDF to create vectors to be used to train a logistic regression model in order to classify a product accurately from Shopee in its legitimacy through sentiment analysis from user reviews. Based on the findings, the model is able to outperform different combinations of text vectorization methods and classifiers and it is able to achieve an accuracy of 84.5%, average precision of 84.52%, average recall of 84.5% and an average F1-score of 84.49%.

In terms of usability, the system usability scale (SUS) surveys revealed that users found the extension generally easy to use, well-integrated, and not overly complex. The SUS scores averaged to 76.72, which falls between the ”good” and ”excellent” categories, indicating that the extension is well-received by users and meets usability standards.

References

[1] A. Abdallah, M. Maarof, and A. Zainal, “Fraud detection system: A survey,” *Journal of Network and Computer Applications*, vol. 68, pp. 90–133, 2016.

[2] M. Shahid, A. Chaudhary, and . K. A. Gupta D., “Review based rating prediction using machine learning techniques,” *2022 11th International Conference on System Modeling Advancement in Research Trends (SMART)*, pp. 118–122, 2022.

[3] H. Gunawardhana, B. Kumara, R. K., and P. Jayaweera, “Effectiveness of machine learning algorithms on battling counterfeit items in e-commerce marketplaces,” *2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 1–7, 2023.

[4] T. Haque, S. N., and F. Shah, “Sentiment analysis on large scale amazon product reviews,” *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, pp. 1–6, 2018.

[5] A. Whinston, S. Choi, and D. Stahl, “The economics of electronic commerce,” *ResearchGate*, 1997.

[6] R. Goel, *E-Commerce*. New Age International, 2007.

[7] C. Cross, R. Smith, and K. Richards, “Challenges of responding to online fraud victimization in australia,” *Trends and Issues in Crime and Criminal Justice*, vol. 474, 2014.

[8] B. Thomas, J. Clergue, A. Schaad, and M. Dacier, “A comparison of conventional and online fraud,” *2nd International Conference on Critical Infrastructures*, 2004.

[9] (n.d.) What is machine learning? [Online]. Available: https://www.ibm.com/topics/machine-learning

[10] I. Naqa and M. Murphy, “What is machine learning?” *Machine Learning in Radiation Oncology*, pp. 3–11, 2015.

[11] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: An overview,” *Journal of Physics: Conference Series*, vol. 1142, 2018.

[12] M. Jordan and T. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, 2015.

[13] M. Misiran, S. Tan, P. Saw, N. Subri, N. Darus, Z. Yusof, and N. Ahmad, “Early detection method for money fraudulent activities on e-commerce platform via sentiment analysis,” *Journal of Entrepreneurship and Business*, vol. 9, pp. 121–142, 2021.

[14] S. Dey, S. Wasif, D. Tonmoy, S. Sultana, S. J., and M. Dey, “A comparative study of support vector machine and naive bayes classifier for sentiment analysis on amazon product reviews,” *2020 International Conference on Contemporary Computing and Applications (IC3A)*, pp. 217–220, 2020.

[15] L. Rong, W. W., and H. Debo, “Sentiment analysis of ecommerce product review data based on deep learning,” *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp. 65–68, 2021.

[16] D. Pratmanto, R. Rousyati, F. Wati, A. Widodo, S. Suleman, and R. Wijianto, “App review sentiment analysis shopee application in google play store using naïve bayes algorithm,” *Journal of Physics: Conference Series*, vol. 1641, 2020.

[17] M. Saeed, F. Yousaf, O. Khalid, M. Gilani, Q. Nawaz, and I. Hamid, “Fraud detection in e-commerce using machine learning,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, 2021.

[18] N. Riego, D. Villarba, A. Sison, F. Pineda, and H. Lagunzad, “Enhancement to low-resource text classification via sequential transfer learning,” *United International Journal for Research Technology (UIJRT)*, 2023.

[19] “spacy: Industrial-strength nlp,” 2019. [Online]. Available: https://github.com/explosion/spaCy/blob/master/spacy/lang/tl/stop_words.py

[20] T. Collinger, “From reviews to revenue,” *Northwestern University*, 2016.

[21] A. Bangor, P. Kortum, and J. Miller, “Determining what individual sus scores mean: Adding an adjective rating scale,” *Journal of User Experience*, 2009.