Analysing and evaluating customer feedback on products on the Shopee e-commerce platform

Khoa Le-Anh Hoang, Ky Vinh Le, Khanh Nhat Ngo
Faculty of Information Science and Engineering, University of Information Technology
Vietnam National University Ho Chi Minh City
Ho Chi Minh City, Vietnam
{22520667, 22520730, 22520640}@gm.uit.edu.vn

Abstract

The development of e-commerce platforms is very fast, especially in the new digital era. For this reason, online shopping has continued to be popular. Of all the platforms, Shopee is the most popular online shopping platform in Southeast Asia, offering its users a wide range of different products and services. For shops and businesses, understanding customer feedback has become increasingly important to evaluate product quality and foster customer loyalty when they buy their products. In this context, we focus on analyzing and evaluating customer feedback on products on the Shopee e-commerce platform. Our data had been collected and synthesized based on comments and rating stars in customers' reviews of some products using natural language processing (NLP) techniques and sentiment analysis. Then we developed a machine learning model that categorizes customer emotional levels based on their comments. Furthermore, we have created a website to show our analysis results for a product, which helps customers know the rating of positive and negative comments and the best and worst responses to the product

1 Introduction

In the rapidly developing process of e-business and m-business, the assessment of the customer feedback has turned out to be one of the vital factors for the business and shops to survive in the online market. Consumers are provided with a variety of goods and customer services offered within an e-commerce site, so they can obtain a particular kind of good. Furthermore, from a buyer's perspective, they often rely on previous feedback from prior customers purchasing a product to decide whether to buy that product or not, but the problem is that there are lots of comments; some comments don't express the emotions of customers about the product. In this context, to clarify these issues from a buyer's view, we analyzed and evaluated buyer feedback on products listed on Shopee to help them make a good decision about buying products and also help businesses and shops gain deeper insights into product quality, customer sentiments, preferences, and satisfaction levels.

Firstly, we collected data on the comments and rating stars of some products in Shopee e-marketplace. After annotating according to the guidelines, we annotated the comments into positive and negative. Subsequently, we conducted a experiment to pre-process the comments using NLP techniques, namely text normalization and word tokenizing. After that, we performed word weighting method by employing TF-IDF vectorizer method, Next, the positive and negative feedbacks after being pre-processed were further classified by using classification machine learning models, namely Logistic Regression, SVM, Naive Bayes, and Random Forest. Of all the selected models, SVM gave the best results of F1-score and accuracy score. Last of all, we chose this model to predict positive and negative feedbacks of a specific product, and then we posted our results on our website, including the ratio of positive, negative and spam comments, as well as some of the best and worst comments. Moreover, users could find comments that related to the word given in the search box. Figure 1 describes our framework of this project.

For us to undertake this study, Section 2 comprises the task definition and the previous studies concerning this topic. Next, Section 3 presents the process of building our dataset, including three steps: data collection, data annotation, and data evaluation. Thus, Section 4 holds the experiment conducted on the

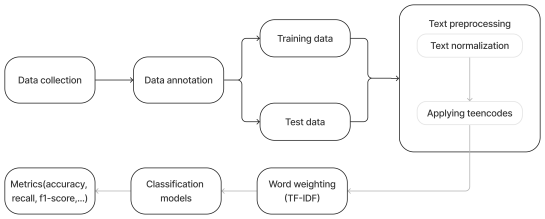

Figure 1: Research framework.

dataset, include text pre-processing along with the results of some machine learning models using given data to find the best model has the highest accuracy and f1-score. which assist customers in their decisions whether to purchase the product or not. Last but not the least, in Section 5, we bring out the final conclusions and pointers to future research.

2 Fundamental of Analysing and Evaluating feedback

2.1 Task definition

To analyse the customer sentiments through their comments, the goal of this task is to classify the comments into 2 main labels, positive and negative

- Input: Given data of customer comments about product.
- Output: One of the two labels below:
    1. **Positive**: The categories of words used in the review relate to how much the customer loves, likes or rates the product. This may mean that it has a good quality and may have features that can provide value that may meet or even exceed the expectation of customers.
    2. **Negative**: This shows that the complaint presents an unfavorable attitude towards the product or the client has some issue with the product. This is because the product that is produced may be of wrong quality or standards as expected by the customer.

2.2 Existing method for Analysing and Evaluating feedback

Wulandari et al.(Wulandari et al., 2022) investigated the impact of social media marketing on the purchasing behaviour of using the products at the Shopee marketplace for the Shopee users at Muhammadiyah Tangerang University, Indonesia with the help of product reviews. Yin et al. (Yin et al., 2022) conducted a research which helped them identify the attitude of the users in the Twitter regarding two e-commerce firms which are Lazada and Shopee. This enabled Shopee and Lazada in recognizing what was being said about them and their strategies, their advantages and disadvantage so that they could enhance their marketing activities. Saputri and Februariyanti (Saputri and Februariyanti, 2022) intended to conduct a study on the sentiment analysis of both the positive and the negative sentiment regarding the Shopee e-commerce app through the application of the RStudio and the Naive Bayes Classifier Algorithm. It makes use of Information Retrieval techniques that include text mining and text processing like stemming, tokenizing, case folding, normalization and selective filtering. Using the same databases, Hantoro et al. (Hantoro et al., 2022) also used SVM, and a sentiment analysis for understanding customer opinion and categorizing the positive and negative feedbacks. The paper of Sentiment analysis of e-commerce site: A survey of current techniques and possible future research by Huang et al. (Huang et al., 2023) reviews current sentiment analysis techniques on the e-commerce site and also the study suggests the future direction of research based on the analysis of 271 papers and selected only 54 experimental papers. Sagarino et al. (Sagarino et al., 2022), the researcher selected these reviews with an aim of identifying customer recommendations on Shopee Philippines as influenced by sentiment analysis of the reviews. The following research steps were carried out for the analysis of the reviews; Scrapping of reviews, Preprocessing of the scraping data and Sentiment analysis of the reviews Using VADER. The prevalence of sentiments was determined through MNB and SVM analysis.

3 Dataset creation

3.1 Data collection

Obtaining customer feedback data is the most important part for analyzing their sentiments, also their perspectives on products quality

on Shopee. In this study, we collected a dataset which contains customer reviews and corresponding rating stars from the Shopee e-commerce platform.

Here's a breakdown of the data collection process:

- Data Source: The data was directly crawled from Shopee using Shopee's API. It's important to ensure compliance with Shopee's terms of service and ethical data collection practices during the scraping process.
- Data Format: The collected data consists of two key components:
  - Customer Reviews: This refers to the textual comments left by customers about the products they purchased.
  - Rating Stars: This represents the numerical star rating (typically ranging from 1 to 5) assigned by customers to the products.

3.2 Data Annotation Process

Metric For Inter-Annotator Agreement In this context, we used Cohen's Kappa to compute inter-annotator agreement. Cohens Kappa (Cohen, 1960) was first introduced as a measure of agreement between observers of psychological behavior.The original intent of Cohen's Kappa was to measure the degree of agreement or disagreement of two or more people observing the same phenomenon. Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories. (Vieira et al., 2010). The equation for k is:

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{1}$$

where k represents an inter-annotator agreement, $Pr(a)$ is the relative observed agreement among raters or the total agreement probability, and $Pr(e)$) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If $k = 1$, it means that all the raters are in complete agreement. If $k \leq 0$, there is no agreement among raters(other than what would be expected by chance).

All the members of our group have participated in data annotation. First, we annotated 200 samples based on our annotation guidelines, then we read carefully and followed the given guidelines. After that, we annotated 1000 random samples in collected data. We repeated the steps 5 times to computed IAA using Cohen's Kappa, which mesured the agreements between annotators classifying labels. The result is that IAA of the annotation was 0.9, nearly perfect agreement.

Annotation Guidelines: a detailed instruction for annotators to ensure that labels could be annotated perfectly and efficiently. The guideline includes some of the following:

1. Possible comments suggest that the product is usable and may be fulfilling the basic needs of a customer in the sense of the usable product as described, good finish or that it may be influenced by other aspects such as slow delivery speeds or poor shop attitude but the overall view is good for the product. This is the information that we assume that the quality of the product and service is good. It will be labeled as **Positive**.

2. Comments that can contribute to the assertion that the product is not bound to sufficiently address the need of the customer, the shop has an indication of fraud, negative reviews, or simply pointing to factors surrounding it without paying reference to the said product. This implies that the product and service offered are not of good quality, enough or standard. These will be labeled as **Negative**.

3. Comments that do not refer to product or service quality, mostly spam, will be eliminated from the dataset.

After annotating 1000 samples based on annotation guidelines and the result of the inter-annotator agreement, we continued to apply these guidelines to annotate the remaining data.

3.3 Dataset evaluation

Our dataset includes 19,579 customers' comments, of which 12,222 comments are positive feedbacks and the rest is the negatives. Figure

2 show the number of positive and negative feedbacks in our dataset.

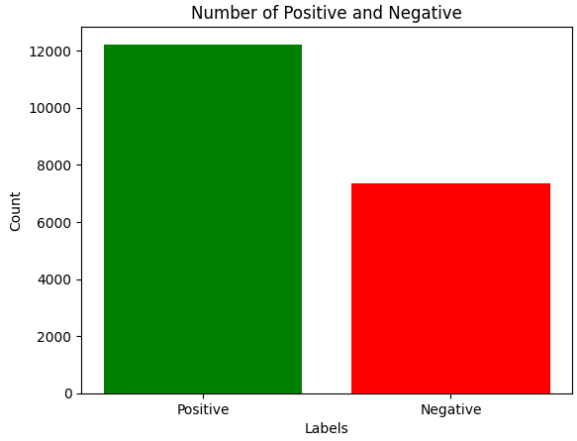

Figure 2: Number of Positive and Negative comments

Then, we divided the whole dataset into two subsets: the training set and the test set, with a ratio of 8:2. Figure 3 shows the number of positive and negative comments in the two given subsets.

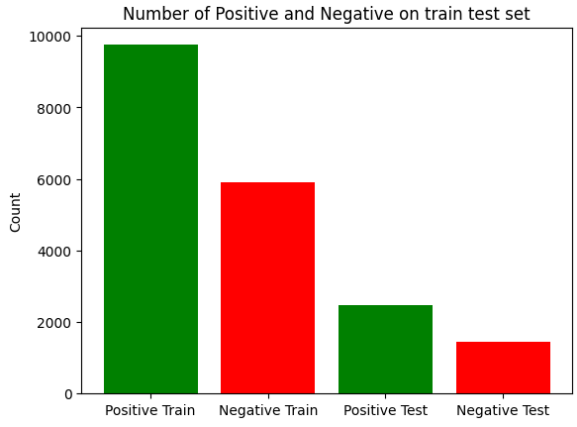

Figure 3: Number of positive and negative comments in training set and test set

These responses mostly talk about good or bad product quality, some of which is about the service and the attitude of the shop on customers. This suggests that many shops focus on product quality, while buyers usually focus on both product and service quality. This is a problem that makes many shops underestimated because the quality of service does not satisfy customers.

4 Experiment and Results

4.1 Data Pre-processing

In this context, we applied text processing to clean the data from noise, including changing the text to lower case, removing punctuation marks and extra space characters, shortening the words at the end of the text, and translating the Vietnamese's teencode bases on our dictionary. Table 1 show the data before and after text preprocessing.

Table 1: Data before and after pre-processing

Raw data	**Pre-processed data**
Shop đóng gói cẩn thận , khi sạc thì quạt ko có mạnh mà nó nhẹ nhẹ , sản phẩm cầm tay , nhẹ giá cả z thì chắc cg ổn áp	shop đóng gói cẩn thận khi sạc thì quạt không có mạnh mà nó nhẹ nhẹ sản phẩm cầm tay nhẹ giá cả vậy thì chắc cũng ổn áp
Chất lượng sản phẩm : quạt tốt Tính năng nổi bật : quạt mát Shop phục vụ tốt lắm ạ , hàng quốc tế mà giao hàng nhanh quá , Trời bị bị hủy nhầm , hên là vẫn chưa hoàn về lại Cảm ơn shop nhé	chất lượng sản phẩm quạt tốt tính năng nổi bật quạt mát shop phục vụ tốt lắm ạ hàng quốc tế mà giao hàng nhanh quá trời bị bị hủy nhầm hên là vẫn chưa hoàn về lại cảm ơn shop nhé

4.2 Word Weighting

The pre-processed data had been implementing Term Frequency - Inverse Document Frequency (TF-IDF) method to assess the relevance of a word in a comment. This basic concept can be conceived as that in the classification when a particular word or a phrase which has frequently been used in an article and it is not commonly or frequently used in other articles, then the word or the phrase is said to have a good class distinction capability and thus useful for classification(Liu et al., 2018). It is the most commonly used feature word weight calculation function in the current vector space model. It is mainly composed of two parts, namely Term Frequency (TF) measures the number of times each word appeared in each document, and Inverse Document Frequency (IDF), on the other hand, is used to identify whether a word or phrase is frequent or uncommon throughout a dataset. Compared to words that are used seldom, common words are less valuable (Kabra and Nagar, 2023). The

formulars of TF and IDF are as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{i=1} n_{k,j}} \quad (2)$$

where n_{i_j} is the number of occurrances of the word t_i in feedback d_j, $\sum_{i=1} n_{k,j}$ is the sum of the all of them in the feedback d_j

$$idf_i = \frac{|D|}{|\{j : t_j \in d_j\}|} \quad (3)$$

where $|D|$ is the total number of feedback in the dataset $|\{j : t_j \in d_j\}|$ is the number of documents containing the word t_i. If the word is not in the dataset, it will cause the dividend to be zero.(Liu et al., 2018). So in general use $1 + |\{j : t_j \in d_j\}|$:

$$tfidf_{i,j} = \frac{tf_{i,j} \times idf_i}{\sqrt{\sum_{j:t_j \in d_j} [tf_{i,j} \times idf_i]^2}} \quad (4)$$

where $tfidf_{i,j}$ is the weight of word t_i in the whole corpus of feedback.

4.3 Machine learning classification model for sentiment analysis

In this context, we applied some machine learning models for sentiment analysis based on customers' feedback, including **Logistic Regression, Naïve Bayes, Support Vector Machine(SVM)** and **Random Forest**. After training these models with the training data and predicting the test data, we could find the best model to predict customers' sentiments.

- **Logistic Regression** is a supervised machine learning algorithm commonly used for binary classification. It employs a logistic function to model the probability of a data point belonging to a specific class. This algorithm is a good choice for simple classification tasks with linear data.
- **Naïve Bayes** is a supervised machine learning algorithm based on Bayes' theorem and the assumption that features of a data point are independent. It's primarily used for classification tasks. Naive Bayes is a good choice for classification problems with sparse data and a reasonable assumption of feature independence.
- **Support Vector Machine (SVM)** is a supervised machine learning algorithm used for classification and regression. It finds a hyperplane in the data space that optimally separates data into different classes. SVM is a good choice for complex classification problems with nonlinear data.
- **Random Forest** is a supervised machine learning algorithm that utilizes an ensemble of decision trees to perform classification or regression. It's built by creating multiple random decision trees from a randomly chosen subset of the data. This tree-based algorithm is a good choice for complex classification and regression tasks with high accuracy and interpretability demands.

4.4 Evaluation metrics

In this study, we used common metrics to evaluate the performance of the models using the data created in Section 3, including Accuracy, Precision, Recall and F1-score. We chose macro F1-score is our optimal metrics, which is the harmonic mean of Precision and Recall,taking both false positives and false negatives into account.

4.5 Experimental results

Table 2 shows the result of these machine learning models performances using the data created by steps in Section 3.

Table 2: Analysing sentiments based on feedbacks results

Model	Accuracy	F1	Recall	Precision
Logistic Regression	0.97	0.976	0.978	0.975
Naïve Bayes	0.926	0.943	0.971	0.918
SVM	0.974	0.98	0.983	0.977
Random Forest	0.953	0.963	0.973	0.956

Next, we use cross-validation method scoring by F1-score to find the average results for each metrics. The results are in Table 3.

Table 3: Analysing sentiments based on feedbacks results after cross-validation with F1-score

Model	F1-score
Logistic Regression	0.958
Naïve Bayes	0.907
SVM	0.961
Random Forest	0.933

According to Table 2, SVM model has the performance outperforms the others on all metrics. Also in Table 3, this model still has the highest F1-score of all the given models for predicting customer sentiment, at 0.961. For these reasons, we decided to choose SVM as our analysing model for our study.

4.6 Error analysis and discussion

Figure 4 shows the confusion matrix of SVM. In general, this model could classify the positive and negative feedback effectively.

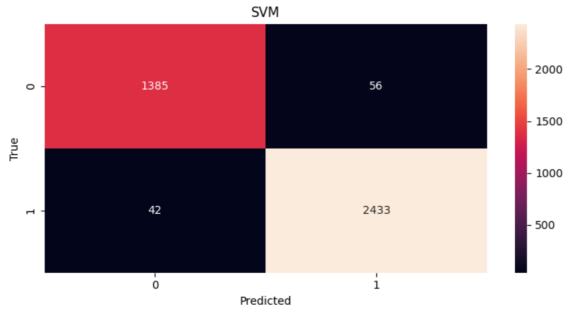

Figure 4: Confusion matrix of the SVM model.

After carrying out the sentiment analysis based on customers' feedback in our dataset, we realized that there are some feedbacks that are in **False Positive** and **False Negative** categories in the confusion matrix. They are typically comments with spelling mistakes, some of them contain teencodes that the pre-processing step could not normalize; or feedback could have the user's positive and negative feelings, etc. These comments could not or could hardly be predicted by the given models, especially SVM.

5 Conclusion and Future works

This study shows how we implement machine learning and text processing in sentiment analysis. After experimentation, we have developed our web application, which helps users or consumers summarize a large amount of feedback using machine learning for sentiment analysis and also helps them understand more detail about a specific product by knowing how many comments have positive and negative feelings. Regarding the model, it effectively performs its basic function of summarizing a large amount of customer feedback based on their sentiments. These models also have high performance on the test set, especially SVM. However, TF-IDF vectorization has many limitations when implemented in practice; the selected models are incapable of contextual consideration and do not really understand the meaning of words or the linking between words like modern NLP models or transformers. In the near future, we will use these NLP models and apply deep learning in this context.

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Kusdarnowo Hantoro, Dwipa Handayani, and Siti Setiawati. 2022. A implementation of text mining in sentiment analysis of shopee indonesia using svm. *Bulletin of Information Technology (BIT)*, 3(2):115–120.

Huang Huang, Adeleh Asemi, and Mumtaz Begum Mustafa. 2023. Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. *IEEE Access*.

Bhavna Kabra and Chetan Nagar. 2023. Convolutional neural network based sentiment analysis with tf-idf based vectorization. *Journal of Integrated Science and Technology*, 11(3):503–503.

Rifki Kosasih and Anggi Alberto. 2021. Sentiment analysis of game product on shopee using the tf-idf method and naive bayes classifier. *LKOM Jurnal Ilmiah*, 13(2):101–109.

Cai-zhi Liu, Yan-xiu Sheng, Zhi-qiang Wei, and Yong-Quan Yang. 2018. Research of text classification based on improved tf-idf algorithm. In *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, pages 218–222. IEEE.

Muhammad Eka Purbaya, Diovianto Putra Rakhmadani, Maliana Puspa Arum, Luthfi Zian Nasifah, et al. 2023. Implementation of n-gram methodology to analyze sentiment reviews for indonesian chips purchases in shopee e-marketplace. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(3):609–617.

Vin Myca C Sagarino, Jennen Isabelle M Montejo, and Angie M Ceniza-Canillo. 2022. Sentiment analysis of product reviews as customer recommendations in shopee philippines using hybrid approach. In *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, pages 1–6. IEEE.

Yulinar Rizkyani Saputri and Herny Februariyanti. 2022. Sentiment analysis on shopee e-commerce

using the naïve bayes classifier algorithm. *Jurnal Mantik*, 6(2):1349–1357.

Shuyan Sun. 2011. Meta-analysis of cohen's kappa. *Health Services and Outcomes Research Methodology*, 11:145–163.

Susana M Vieira, Uzay Kaymak, and João MC Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. In *International conference on fuzzy systems*, pages 1–8. IEEE.

Iwuk Wulandari, Abdul Rauf, et al. 2022. Analysis of social media marketing and product review on the marketplace shopee on purchase decisions. *Review of Integrative Business and Economics Research*, 11(1):274.

Jenny Yow Bee Yin, Nor Hasliza Md Saad, and Zulnaidi Yaacob. 2022. Exploring sentiment analysis on e-commerce business: Lazada and shopee. *Tem journal*, 11(4):1508–1519.