Jurnal Infotel
Vol. 16, No. 1, February 2024, pp. 1–16.
doi:10.20895/infotel.v16i1.1081

Research Article

Topic Sentiment using Logistic Regression and Latent Dirichlet Allocation as a Customer Satisfaction Analysis Model

Puji Winar Cahyo[1,*], Ulfi Saidata Aesyi[2], and Bagas Dwi Santosa[3]

[1]Department of Informatics, Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta 55293, Indonesia
[2,3]Department of Information System, Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta 55293, Indonesia

*Corresponding email: pwcahyo@gmail.com

Received: November 11, 2023; Revised: November 24, 2023; Accepted: January 22, 2024.

Abstract: Buying and selling goods now is more interesting through e-commerce or marketplaces because of the ease of carrying out online transactions. Each transaction usually generates a response from the customer. The transaction response on the Shopee platform is still in paragraph form and needs to be more specific. Therefore, this research aims to build a model analysis of customer satisfaction using the best algorithm between support vector machine (SVM), random forest, and logistic regression. This research method uses sentiment classification with logistic regression because the logistic regression algorithm has the best accuracy, with an accuracy of 90.5. Meanwhile, the SVM algorithm achieved an accuracy of 90.4, and random forest reached 90.2. The three algorithms were tested three times, splitting data train:test at 80:20, 70:30, and 60:40. The best results were obtained by splitting data at 60:40. The best model is used to predict data without labels. The prediction produces 12,844 positive sentiment comment data, 112 negative sentiment comment data, and 70 neutral sentiment comment data. The results of this research continued to topic modeling using latent dirichlet allocation (LDA) to generate a trending topic of customer satisfaction on sales products. Implications of discussing each trend topic can be used as a reference for improving products and services, especially in communicating with customers.

Keywords: customer satisfaction, logistic regression, sentiment analysis, Shopee, topic modeling

ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Rapidly evolving technology has quite a big influence on people's daily lives, one of which is the influence felt in the economic sector. The influence that is visible today is seen in the process of buying and selling goods or services. Buying and selling goods or services are identically carried out through e-commerce or marketplaces. This commercial transaction process utilizes information technology and can be carried out by organizations or individuals [1]. In current developments, there are many Shopee platform users in Indonesia, and the Shopee platform has become one of the most well-known e-commerce sites because of the ease of carrying out online transactions. Online transactions have become quite attractive due to the convenience and shortening of time in the transaction process, so people do not need to come to the shop in person [2]. This phenomenon has triggered a significant increase in online transactions, causing online stores to experience high competition due to the increasing number of online stores [3].

Selling products from the marketplace has increased quite rapidly [4]. On the Shopee platform, the development of several shops can be analyzed based on the satisfaction response of their respective users. The analysis can be adjusted by identifying satisfaction, which will be measured through customer satisfaction comments that have been given. Customer satisfaction comments on the Shopee platform are still in paragraph form and do not specifically indicate comments on services, goods, or anything else [5]. Through customer satisfaction comments that still need to be specific, it is necessary to analyze text for customer satisfaction based on frequently discussed topics. This method of analyzing customer satisfaction topics based on text has often been done. However, to implement this analysis, it is necessary to study text-mining algorithms designed to provide appropriate results on topic trends [6]. Topic analysis can be developed in more depth by looking for topics that negatively and positively influence services or products sold in an online shop. Meanwhile, sentiment analysis can be used as a classification method to analyze negative and positive influences on text data [7]. Classification methods of sentiment analysis and topic analysis are combined, and we will get topics with a positive or negative trend regarding the comments given [8] by online shop customers.

Sentiment classification using support vector machines (SVM) is analyzed based on the level of algorithm performance using varying kernel parameters, using a comparison of three SVM parameters, including radial kernel, linear grid, and radial grid. This comparison found that SVM with a radial grid kernel achieved better performance than the radial kernel and linear grid [9]. Meanwhile, a different test using the random forest algorithm for classifying Indonesian sentiment with weighting using bag of word (BOW) and term frequency-inverse document frequency (TF-IDF) resulted in an out of bag score reaching 0.829. From the testing process, it was stated that the use of weighting variations resulted in an increase in value that was not too significant, so it can be concluded that the use of weighting variations in sentiment classification using random forest has not had a sufficiently positive impact [10].

Apart from that, a comparison of classification algorithms between SVM and logistic regression was carried out to analyze sentiment among the population in India during COVID-19 by taking data from March to June. The results of these two algorithms showed that the SVM algorithm achieved an accuracy of 91.50 %, and logistic regression achieved a test accuracy of 87.75 % [11]. Another study related to sentiment analysis using SVM and latent dirichlet allocation (LDA) on Kitabisa application reviews via the Google Play

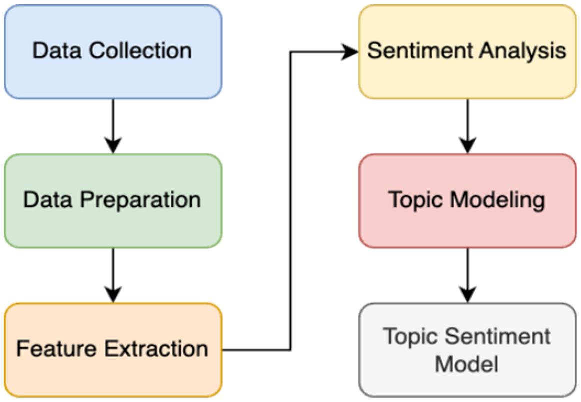

Figure 1: Research steps.

Store found that the sentiment classification model achieved 72 % accuracy, 76 % precision, and 72 % recall due to the imbalanced training data set. Therefore, SMOTE-Tomek links was implemented to handle unbalanced data, increasing accuracy to 98 %, precision to 98 %, and recall to 98 %. The previous research has differences in the method used to build the model, especially in the sentiment classification algorithms and feature extraction methods. The SVM algorithm shows a relatively high level of accuracy compared to other algorithms. Combining BOW and TF-IDF for feature extraction does not provide significant classification results. Meanwhile, the SMOTE-Tomek links method is used for unbalanced data. Therefore, based on the results of several tests from previous researchers, the best algorithm can be obtained by comparing the accuracy of each test result.

This research aims to build a model from sentiment analysis by testing some algorithms to obtain the best algorithm. The best algorithm will be applied to customer satisfaction comments at stores. Meanwhile, topic modeling focuses on the optimal number of topics used through testing using the coherence score [12].

The result contributes to forming a customer satisfaction analysis model for the Shopee platform through topic trends, which can be used to inform shop owners' decisions about offering services or selling products. This research uses several stages, including data collecting, data preparation, feature extraction, sentiment analysis, topic modeling, and topic sentiment model. The research stages used can be seen in Figure 1.

1.1 Data Collection

The first stage in this research is data collection using the comment and rating data extraction method originating from the Shopee marketplace website [13]. Website data is searched and analyzed to identify each component and specific information is extracted or retrieved to find the desired or sought data [14]. The data collection process was carried out using web scraping using the Scrapy and Selenium libraries in Python programming.

This research used three popular data seller accounts on the Shopee marketplace with the shop names Dyr_id, Dinalulhijab, and NajwaFashionMuslim for customer comment data. The number of comment data that was successfully retrieved was Dyr_id with 1,725 comment data, Dinnalulhijab with 3,840 comments, and NajwaFashionMuslim with 74,413

Table 1: Example of comments

Rating	Comment
5	*bahannya halus, sesuai ekspektasi sih, hehehe puas dengan harga murah*
3	*Pengiriman dpt tapi respon nya agak kurang memuaskan Neng oke sip puas*
1	*Bergo Mariam Instan, malah Dikasih Pasminaa Jelek Gini, Bahan Jilbadnya Jelek2 Tipis, Panjang Pasmina Nya Jg Gak Sesuai Penjangnya*

Table 2: Data preparation steps

Preparation	Results
Original Data	*Bahannya halus, sesuai ekspektasi sih hehehe puas dengan harga murah*
	Pengiriman dpt tapi respon nya agak kurang memuaskan oke sip puas
	Bergo Mariam Instan malah Dikasih Pasminaa Jelek Gini, Bahan Jilbadnya Jelek Tipis Panjang Pasmina Nya Jg Gak Sesuai Panjangnya
Case Folding	*bahannya halus, sesuai ekspektasi sih, hehehe puas dengan harga murah*
	pengiriman dpt tapi respon nya agak kurang memuaskan oke sip puas
	bergo mariam instan, malah dikasih pasminaa jelek gini, bahan jilbadnya jelek2 tipis, panjang pasmina nya jg gak sesuai panjangnya
Tokenizing	*[“bahannya", “halus", “sesuai", "ekspektasi", “sih", “hehehe", “puas", “dengan", “harga", “murah"]*
	[“pengiriman", “dpt", “tapi", “respon", “nya", “agak", “kurang", “memuaskan", “oke", “sip", “puas"]
	[“bergo", “mariam", “instan", “malah" “dikasih" “pasminaa", “jelek", “gini", “bahan", “jilbadnya", “jelek", “tipis", “panjang", “pasmina", “nya", “jg", “gak", “sesuai", “panjangnya"]
Stopword Removal	*[“bahan", “halus", “sesuai", "ekspektasi", “puas", “harga", “murah"]*
	[“kirim", “dapat", “tapi", “respon", “agak", “kurang", “puas", “oke", “sip", “puas"]
	[“bergo", “mariam", “instan", “malah" “kasih" “pasmina", “jelek", “gini", “bahan", “jilbab", “jelek", “tipis", “panjang", “pasmina", “tidak", “sesuai", “panjang"]

comments. Meanwhile, these comments were taken from three Shopee seller accounts. Examples of comment data can be seen in Table 1.

1.2 Data Preparation

Data preparation stage is used to clean and prepare the data needed to be processed at the next stage. The pre-processing carried out is technically divided into four steps. All steps in this pre-processing stage are carried out using the Python programming language with the NLTK and Sastrawi libraries. These stages include the following [15]:

a. Remove duplicate comments;
b. Case folding, namely the process of changing capital letter characters to lowercase;
c. Tokenizing, namely the process of breaking a sentence into a group of words;
d. Stopword removal, namely the process of obtaining important words from the Tokenizing results by eliminating words that have no meaning in the text;
e. Stemming, namely searching for basic words using the Sastrawi library;
f. Data filtering, namely by removing duplicate comments that have similarities after processes a-e at the pre-processing stage are carried out.

Table 3: Sentiment labeling

Rating	Comment	Label
5	*bahan halus sesuai ekspektasi puas harga murah*	Positive
3	*kirim dapat tapi respon agak kurang puas neng oke sip puas*	Neutral
1	*bergo mariam instan malah kasih pasmina jelek gini bahan jilbab jelek tipis panjang pasmina tidak sesuai panjang*	Negative

The pre-processing results produced clean data for each shop, including Dyr_id with 1,119 comments, Dinnalulhijab with 2,079 rows, and NajwaFashionMuslim with 22,848 comments. An example of the pre-processing stages carried out in stages 2-5 can be seen in Table 2.

After carrying out the pre-processing stage, the comment data from each shop is given positive, neutral, and negative sentiment labels according to predictions of human ability to determine the rating given by customers. It is given a positive label if the ratings given are 4 and 5, a neutral label if the ratings given are 3, and a negative label if the ratings given are 1 and 2, as seen in Table 3.

1.3 Feature Extraction (TF-IDF)

The data resulting from data preparation is continued towards feature extraction using term frequency-inverse document frequency (TF-IDF). TF is the frequency of the number of occurrences of a word in a document, which can be seen in (1). Meanwhile, IDF is the inverse value of records containing these words, as seen in (2). TF and IDF will be multiplied for each word to produce a weight value. The formula for calculating TF-IDF is shown in (3) [16].

$$tf(w,d) = \log(1 + fw, d) \tag{1}$$

$$idf(w,D) = \log\left(\frac{N}{f(w,D)}\right) \tag{2}$$

$$tfidf(w,d,D) = tf(w,d) \cdot idf(w,D) \tag{3}$$

Description of each equation notation: $tf(w,d)$ is the number of occurrences of the word w in document d. Subsequently, N is the total number of all documents, while d is the document whose tf will be searched for. Finally, $idf(w,D)$ is the log calculation of the total of all N documents divided by the results of $tf(w,d)$ for each d document to produce the TF-IDF value.

1.4 Sentiment Analysis

Sentiment analysis searches for sentiment using a classification method [17]. The sentiment search process will run when feature extraction on the training and test data has been completed. Meanwhile, the sentiment analysis used to form a customer satisfaction model uses three experimental algorithms, including SVM, random forest, and logistic regression, which are technically all calculated using the Scikit-Learn library.

From these three experiments, an appropriate algorithm will be selected to predict customer satisfaction through the comments given. As is known, these three algorithms are suitable for solving problems in the form of high-dimensional data [18].

1.5 Topic Modeling

Topic Modeling is the stage where the topic of discussion will be formed from data resulting from customer satisfaction sentiment. Formation of conversation topics uses the LDA algorithm, which is a method in unsupervised learning and a generative probability model that describes documents as random combinations of latent topics where specific topics are characterized by word distribution [19]. Meanwhile, to determine the optimal number of topics, use the coherence score value from the resulting topic coherence [20]. Topic Coherence is a set of words considered to have the same suitability in human interpretation. Meanwhile, the coherence score is the semantic similarity value of each word that forms many different topics, so to calculate the coherence score, it can be seen in (4) [21]. Each resulting topic can be used to improve store products or services.

$$score(vi, vj, \in) = \log \frac{D(vi, vj) + \in}{D(vj)} \tag{4}$$

Description of each notation in (4): $score(vi, vj, \in)$ is the coherence score for the word vi and vj, $D(vi, vj)$ is the number of documents containing the words vi and vj, $D(vj)$ is the number of documents containing the word vj, vi is the frequency of the word vi, vj is the frequency of the word vj, while $\in$ is a variable that guarantees the result to be positive.

1.6 Topic Sentiment Model

The sentiment classification process produces three groups of prediction data: positive, negative, and neutral. The result of sentiment classification can be used to produce trend topics that are often discussed. Positive sentiment is used as a reference for customer data stating satisfaction. Negative sentiment is used as a reference for customer data expressing dissatisfaction. Meanwhile, neutral sentiment topics are used to find neutral customer comments. The sentiment results for each topic are used to evaluate efforts to increase sales in online stores.

2 Result

The comment data, initially 79,978, was normalized, and a total of 26,046 comment data was obtained. The data that has been cleaned is 50 % labeled with 11,756 positive sentiments, 536 negative sentiments, and 730 neutral sentiments, which are then used to form a sentiment model by testing three models. All models were tested to find the best model to apply to sentiment analysis of customer comments.

The model tested uses several algorithms, including SVM, random forest, and logistic regression. The evaluation process is carried out by training the model according to the data distribution, starting from the training data distribution, with testing data including 80:20, 70:30, and 60:40 [22]. The test results of the three algorithms produced varying levels of precision, recall, and accuracy, as seen in Table 4.

Table 4: Evaluation result

Algorithm	Evaluation	80:20	70:30	60:40
SVM	Precision	82.9	85.7	86.0
	Recall	90.1	90.4	90.4
	Accuracy	90.1	90.4	90.4
Random Forest	Precision	83.6	83.3	85.0
	Recall	89.8	89.7	90.1
	Accuracy	89.8	89.7	90.2
Logistic Regression	Precision	86.5	86.4	87.2
	Recall	90.4	90.3	90.5
	Accuracy	90.4	90.3	90.5

The testing results, which were carried out three times by splitting the data, produced the highest accuracy value achieved by logistic regression, delivering a precision value of 87.2, recall of 90.5, and accuracy of 90.5. Meanwhile, the SVM algorithm is in second place with a precision value of 86.0, recall of 90.4, and accuracy of 90.4. The random forest algorithm is in third place with a precision value of 85.0, recall of 90.1, and accuracy of 90.2. Through the results of the tests that have been carried out, the logistic regression algorithm has been determined as an algorithm that forms a sentiment model for predictions on 26,046 unlabeled data. The predictions produced 12844 comment data labeled positive, 112 comment data labeled negative, and 70 comment data labeled neutral. Each group of positive and negative comments was then subjected to topic modeling using LDA by determining the optimal number of topics that would be applied to each group.

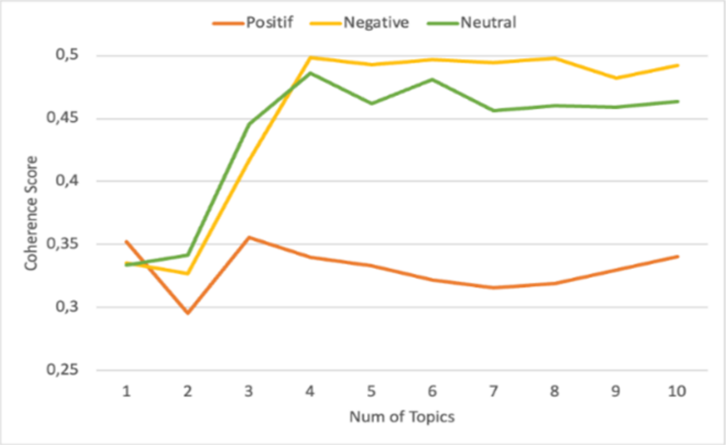

Figure 2: Coherence score.

It can be seen in Figure 2 that the optimal topic value for comments with the positive sentiment group has a total of 3 optimal topics with a coherence score of 0.36, comments with a negative and neutral sentiment group have an optimal number of 4 topics, coherence score of negative sentiment is 0.50 and neutral 0.49. The optimal number of topics for each sentiment has been obtained. Then, the topic formation has been proceeded to discuss the

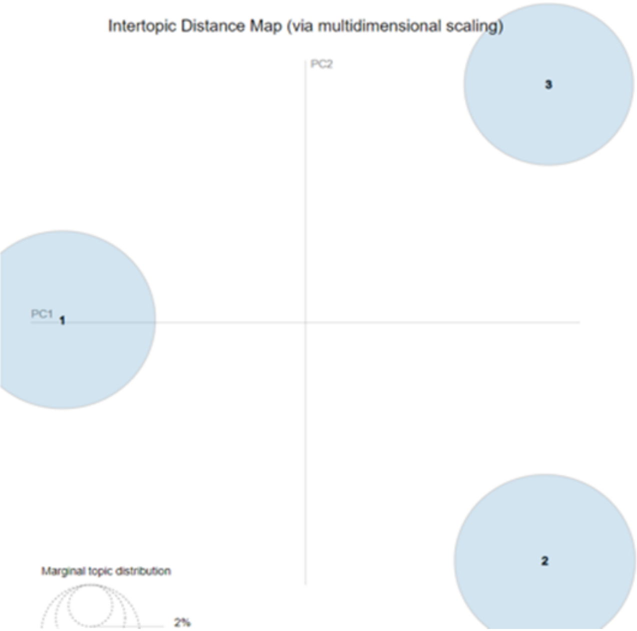

Figure 3: Positive distance.

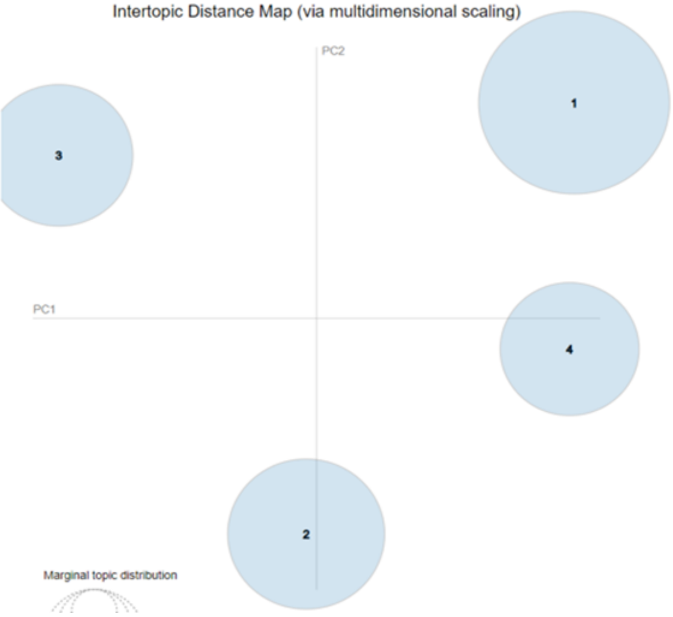

Figure 4: Negative distance.

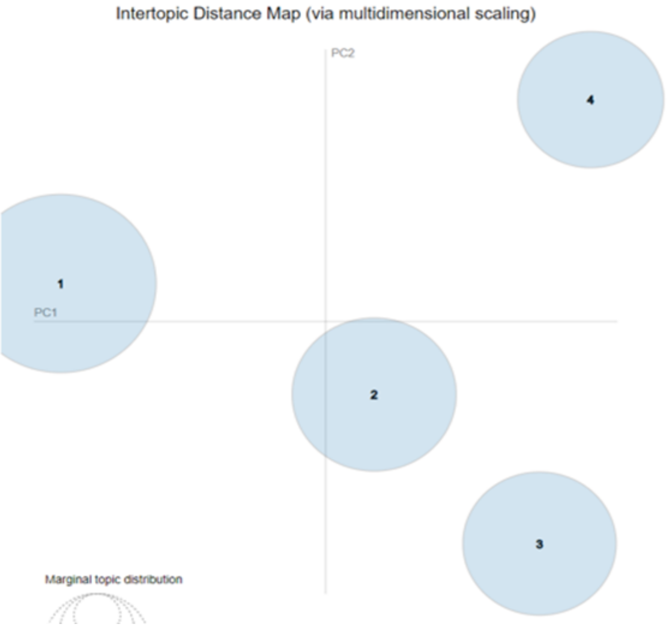

Figure 5: Neutral distance.

contents of each customer satisfaction sentiment. The results of topic formation resulting in positive sentiment with inter-distance can be seen in Figure 3, negative sentiment in Figure 4, and neutral sentiment in Figure 5.

The interdistance results for each topic have a reasonably far distance to discuss the topic, so it is possible that the discussion of customer satisfaction through the comments provided needs a closer intersection. This relatively large inter-distance occurs in all comment sentiments. Discussion of the topic on positive sentiment can be seen from the words that make up it, which can be seen in Figure 6 to Figure 8.

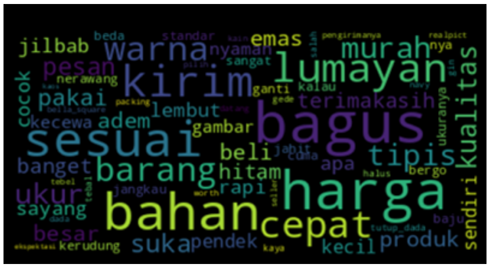

Figure 6: Positive topic #1.

The topic on positive sentiment discusses comments that show positive customer satisfaction through analysis of words that correspond to each topic, including the first topic, which discusses positive responses to the products being sold; this is proven by the words in it (*barang, bahan, hitam, lembut, adem, pakai, sesuai*) can be seen in Figure 6; while the

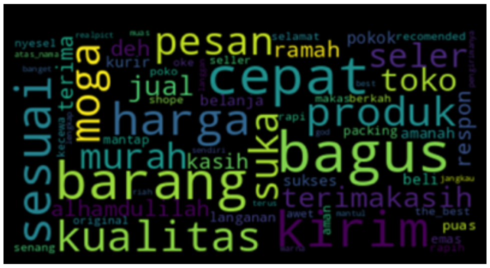

Figure 7: Positive topic #2.

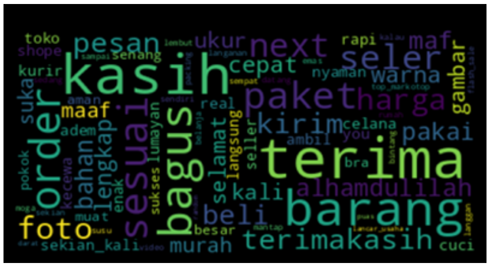

Figure 8: Positive topic #3.

second topic discusses the price of goods with the words in it (*harga, jual, murah, kualitas*) which can be seen in Figure 7; Meanwhile, the third topic discusses product delivery (*ekspedisi paket, kirim, cepat*) which can be seen in Figure 8. Meanwhile, the topic of negative sentiment shows comments of disappointment, which can be seen in Figure 9 to Figure 12.

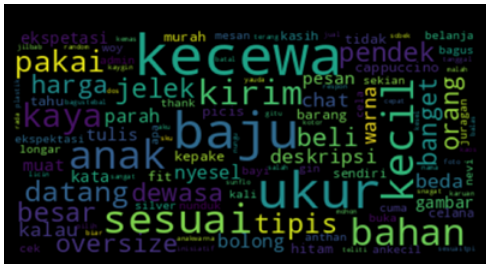

Figure 9: Negative topic #1.

Topics on negative sentiment tend to focus more on discussing materials, sizes, communication methods, and delivery. The following is the content of the discussion of each topic: the first topic, the second topic, and the fourth topic discuss materials and sizes with the most dialogue on the words (*ukur, bahan, tipis, warna oversize*) which can be seen in Figure 9, Figure 10 and Figure 12; Meanwhile, the third topic discusses more communication and delivery methods with the most discussion being on the words (*kirim, tunggu, konfirmasi, balas, itikad, pesan layanan*) which can be seen in Figure 11. Meanwhile, the topic of neutral sentiment can be seen in Figure 13 to Figure 16.

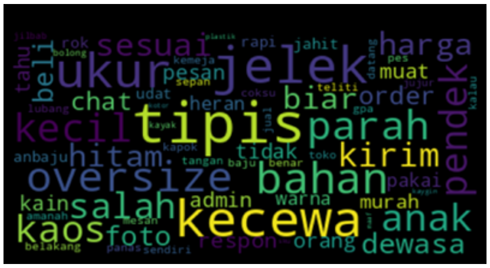

Figure 10: Negative topic #2.

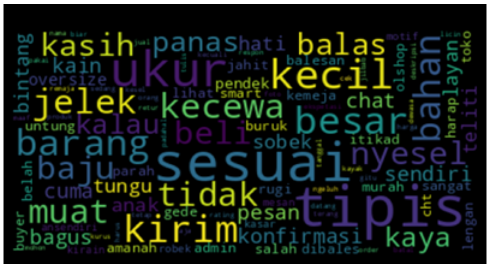

Figure 11: Negative topic #3.

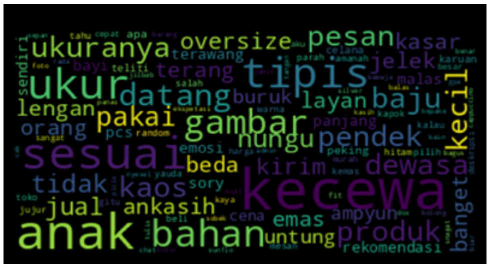

Figure 12: Negative topic #4.

The topic of discussion on neutral sentiment tends to be more about shipping, color, product quality, price, and size. As can be shown in Figure 13, these are the words that form the first topic with a lot of discussion related to color with the words in it (*kirim, warna, hitam, biru*), while Figure 14 is the second topic, and Figure 16 is the 4th topic with more discussion of material quality with words in it (*bahan, panas, tipis, bolong*). Meanwhile, the third topic in Figure 15 discusses price and size with the words in it (*harga, ukur, kecil, oversize, melar*).

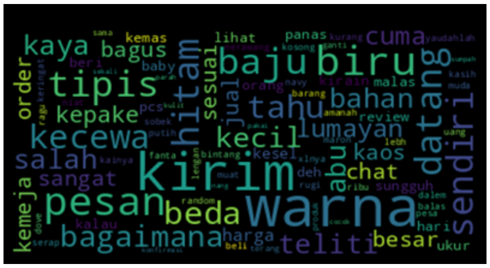

Figure 13: Neutral topic #1.

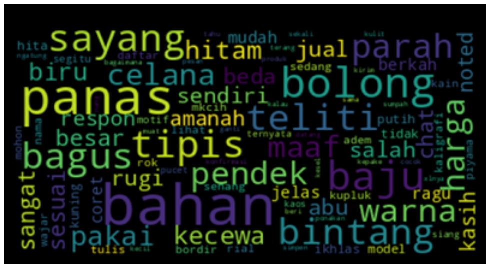

Figure 14: Neutral topic #2.

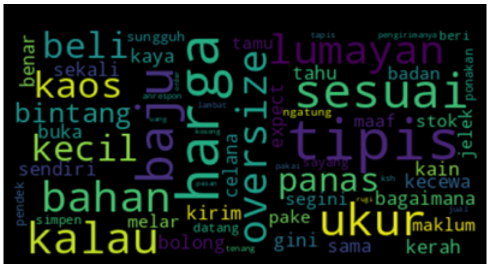

Figure 15: Neutral topic #3.

The results of the topic discussion for each sentiment can be analyzed in more depth. Positive sentiment concerns customer satisfaction, which often responds to product quality, price, and delivery. Meanwhile, the discussion on the topic of negative sentiment, which is customer dissatisfaction, mainly discusses the quality of materials, sizes, and methods of

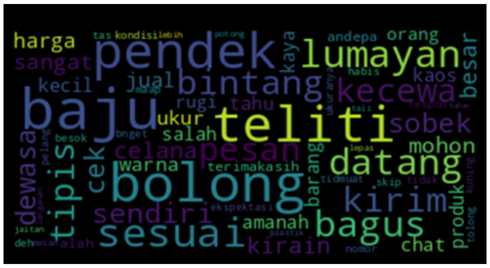

Figure 16: Neutral topic #4.

communication. This research generates the same discussion regarding customer satisfaction with the delivery process and customer dissatisfaction with product quality based on the materials used [23]. Meanwhile, neutral responses were seen when talking about the topics of delivery, product quality, price, and size.

3 Discussion

This research was successfully carried out by analyzing customer comment data regarding the sale of Muslim clothing on the Shopee marketplace. These results are shown in creating the best sentiment analysis model using the logistic regression algorithm. The best accuracy was obtained through the testing process of three algorithms, namely logistic regression with an accuracy value of 90.5, SVM achieved an accuracy of 90.4, and the random forest algorithm reached 90.2. The three algorithms achieved the highest accuracy when splitting 60 data for training and 40 for testing. Meanwhile, the prediction results produced more positive sentiment, with 12,844 comment data labeled positive, 112 comment data labeled negative, and 70 comment data labeled neutral. Positive sentiment discusses customer satisfaction, mostly product quality, price, and delivery. Negative sentiment discusses customer dissatisfaction, mostly product quality, size, and communication methods. Meanwhile, neutral sentiment discusses ordinary things in product quality, price, and size.

4 Conclusion

The satisfaction analysis model can be formed using the method topic sentiment models for customer satisfaction, which can be built using the logistic regression algorithm and Latent Dirichlet Allocation topic modeling. The logistic regression algorithm obtained an accuracy of 90.5. Meanwhile, for topic modeling, it produced 3-4 optimal topics for each classification result. Discussing each topic can be used as a reference for improving products and services, especially in communicating with customers. This research can be developed to analyze buyer characteristics based on comments given after purchasing a product.

Acknowledgments

We want to express our sincere thanks to Universitas Jenderal Achmad Yani Yogyakarta for the internal funding scheme research support, which allowed this research to be completed on time.

References

[1] J. H. Norian, A. M. Jama, M. H. Eltaieb, and A. A. Adam, "Data-driven e-commerce techniques and challenges in the era of the fourth industrial revolution," *Sci. J. Informatics*, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/sji

[2] K. Aswini, "Advantages and challenges of e-commerce customers and businesses: In Indian perspective," *SHANLAX Int. J. Manag.*, vol. 6, no. 1, pp. 173–176, 2018, [Online]. Available: http://www.shanlaxjournals.in

[3] T. J. Gerpott and J. Berends, "Competitive pricing on online markets: a literature review," *J. Revenue Pricing Manag.*, vol. 21, no. 6, pp. 596–622, 2022, doi: 10.1057/s41272-022-00390-x.

[4] S. Suparman, "Tokopedia's Muslim fash forward 2022 to boost sales of Muslim fashion items - Companies - The Jakarta Post," TheJakartPost. [Online]. Available: https://www.thejakartapost.com/business/2022/09/08/tokopedias-muslim-fash-forward-2022-to-boost-sales-of-muslim-fashion-items.html

[5] U. S. Aesyi and P. W. Cahyo, "Cuscoma: Platform peningkatan penjualan produk berdasarkan analisis komentar pelanggan di marketplace Shopee menggunakan metode metode rule-based," *J. Sains dan Inform.*, vol. 9, no. November 2022, pp. 1–8, 2023, doi: 10.34128/jsi.v9i1.539.

[6] P. W. Cahyo and M. Habibi, "Entity profiling to identify actor involvement in topics of social media content," *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 14, no. 4, p. 417, 2020, doi: 10.22146/ijccs.59869.

[7] E. R. Kaburuan, Y. S. Sari, and I. Agustina, "Sentiment analysis on product reviews from Shopee marketplace using the naïve Bayes classifier," *Lontar Komput. J. Ilm. Teknol. Inf.*, vol. 13, no. 3, p. 150, 2022, doi: 10.24843/lkjiti.2022.v13.i03.p02.

[8] Praveen, SV, J. M. Lorenz, R. Ittamalla, K. Dhama, C. Chakraborty, D. V. S. Kumar, and T. Mohan, "Twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people's perspectives regarding COVID-19 booster vaccine shots in India: Crucial to expanding vaccination coverage," *Vaccines*, vol. 10, no. 11, 2022, doi: 10.3390/vaccines10111929.

[9] L. K. Ramasamy, S. Kadry, Y. Nam, and M. N. Meqdad, "Performance analysis of sentiments in Twitter dataset using SVM models," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, pp. 2275–2284, 2021, doi: 10.11591/ijece.v11i3.pp2275-2284.

[10] M. A. Fauzi, "Random forest approach fo sentiment analysis in Indonesian language," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 46–50, 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[11] S. Majumder, A. Aich, and S. Das, "Sentiment Analysis of People During Lockdown Period of COVID-19 Using SVM and Logistic Regression Analysis," SSRN Electron. J., pp. 1–10, 2021, doi: 10.2139/ssrn.3801039.

[12] K. H. Musliadi, H. Zainuddin, and Y. Wabula, "Twitter social media conversion topic trending analysis using latent dirichlet allocation algorithm," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 390–399, 2022, doi: 10.37385/jaets.v4i1.1143.

[13] P. W. Cahyo and L. Sudarmana, "A comparison of k-means and agglomerative clustering for users segmentation based on question answerer reputation in brainly platform," *Elinvo (Electronics, Informatics, Vocat. Educ.)*, vol. 6, no. 2, pp. 166–173, 2021, [Online]. Available: https://journal.uny.ac.id/index.php/elinvo/article/view/44486

[14] R. R. Fayzrakhmanov, E. Sallinger, B. Spencer, T. Furche, and G. Gottlob, "Browserless web data extraction: Challenges and opportunities," in *2018 IW3C2 (International World Wide Web Conference Committee)*, 2018, pp. 1095–1104. doi: http://doi.org/10.1145/3178876.3186008.

[15] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *J. Big Data*, vol. 8, no. 1, pp. 1–16, 2021, doi: 10.1186/s40537-021-00413-1.

[16] A. Addiga and S. Bagui, "Sentiment analysis on twitter data using term frequency-inverse document frequency," *J. Comput. Commun.*, vol. 10, no. 08, pp. 117–128, 2022, doi: 10.4236/jcc.2022.108008.

[17] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: Approaches, datasets, and future research," *Appl. Sci.*, vol. 13, no. 7, 2023, doi: 10.3390/app13074550.

[18] A. Novianto and M. D. Anasanti, "Autism spectrum disorder (ASD) identification using feature-based machine learning classification model," *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 17, no. 3, p. 259, 2023, doi: 10.22146/ijccs.83585.

[19] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne, "Effective comparison of LDA with LSA for topic modelling," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 13-15 May 2020, Madurai, India. doi: 10.1109/ICICCS48265.2020.9120888.

[20] P. W. Cahyo, M. Habibi, A. Priadana, and A. B. Saputra, "Analysis of popular hashtags on instagram account the ministry of health," in *Proceedings of the International Conference on Health and Medical Sciences (AHMS 2020)*, vol. 34, no. Ahms 2020, pp. 270–273, 2021, doi: 10.2991/ahsr.k.210127.062.

[21] D. L. C. Pardede and M. A. I. Waskita, "Analisis pemodelan topik untuk ulasan tentang peduli lindungi," *J. Ilm. Inform. Komput.*, vol. 28, no. 1, pp. 17–26, 2023, doi: 10.35760/ik.2023.v28i1.7925.

[22] R. P. Pratama and A. Tjahyanto, “The influence of fake accounts on sentiment analysis related to COVID-19 in Indonesia,” *Procedia Comput. Sci.*, vol. 197, no. 2021, pp. 143–150, 2021, doi: 10.1016/j.procs.2021.12.128.

[23] X. Liu and Z. Kao, “Research on influencing factors of customer satisfaction of e-commerce of characteristic agricultural products,” in *Procedia Computer Science, Elsevier B.V.*, 2021, pp. 1505–1512. doi: 10.1016/j.procs.2022.01.192.