International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

Predicting Customer Satisfaction through Sentiment Analysis on Online Review

Anas Zakaria[1], Manahan Siallagan[2]
[1,2] School of Business and Management, Institut Teknologi Bandung, Indonesia

ABSTRACT: User-generated content, such as user reviews, posts, tags, ratings, and opinions on the internet, can be used as a business indicator if collected and appropriately analyzed. One of the examples is predicting customer satisfaction through implementing big data analytics on online reviews. In analyzing the user-generated content to predict customer satisfaction, the author implements machine learning approach using the Sentiment Analysis method. Five-fold cross-validation was performed to train the classification model. The training was performed with a combination of tokenization methods: term frequency-inverse document frequency (tf-idf) and bag-of-words; n-gram types: unigram, bigram, trigram, and combination of unigram, bigram, and trigram; and machine learning algorithms: linear support vector classification (LinearSVC) and multinomial naïve bayes (MultinomialNB). The result was then evaluated using classification performance metrics such as precision, recall, F1 measure, and AUC score.

The result shows that the tf-idf vectorizer performs similarly to the bag-of-words method. A similar result was also observed for machine learning algorithm selection. Both MultinomialNB and LinearSVC produce the same performance. Low-level n-grams (such as unigrams and bigrams) tended to have higher precision, recall, F1 measure, and AUC score than high-order n-grams (such as trigrams). The best results were achieved by combining unigrams, bigrams, and trigrams, resulting in an average performance score of 0.94 for all measurements. From the result and analysis, the author finds that predicting customer satisfaction using text and sentiment analysis methods on user-generated content is possible. The model's performance in this experiment is decent, with high precision, recall, F1, and AUC score.

KEYWORDS: Customer Satisfaction, Sentiment analysis, Classification, Machine learning, User-generated content.

INTRODUCTION

The recent digital environment development drives the generation of a vast amount of user-generated content. User-generated content, such as user posts, user reviews, tags, ratings, and opinions on the internet, can be used as a business indicator if collected and appropriately analyzed. Converting user-generated content into information can also provide organizations with detailed and credible information about their customers' opinions and perceptions of their services [1]. Hence, a firm or manager's ability to convert insight from user-generated data into valuable information could drive business success [2].

One example of user-generated content is a customer review of a service or a product. In a customer review, customers able to quantitatively evaluate the products or services they have purchased and illustrate the reason in writing [3]. The current trend demonstrates that before making a purchase of goods or services, consumers look for pertinent information to lessen their uncertainty of choice. Because in the information search process, ratings and reviews are the most trusted data sources of consumers [3], people tend to rely on this information before making a purchase decision. Thus, firms may benefit from mining and analyzing user-generated content data such as comments and sentiments [2].

Analyzing user-generated content to drive business decisions could be seen as an implementation of big data analytics for business. It is critical in big data environments to process and act quickly on available data. Although mining data from big data is a challenging task, big data has the potential to revolutionize all areas of science [4]. The implementation of big data analytics could be implemented to understand customer needs better. One company's customer relationship management performance could be improved by better understanding customer needs [5].

An example of the application of big data analytics in business is the measurement of a customer's perception or experience with a product or service using unstructured data such as user reviews or social media posts [6]. The need to monitor customer experience arises as a result of customers interacting with businesses through multiple touch points across a variety of channels and media,

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

deriving in more complex customer journeys [7]. Due to the immense diversity and size of social media data, it is difficult for humans or businesses to gather the most recent trends and summarize the situation as it stands about products; this necessitates the need for automated opinion mining [8]. Sentiment Analysis can tackle this challenge as it can extract opinions from enormous datasets promptly.

Measuring customer satisfaction is critical for a company because it is strongly linked to financial performance [9]. As the customer journey has become more complex, measuring customer satisfaction through big data analytics, particularly natural language processing, is essential [10].

The usage of user-generated content as a source for determining customer sentiment had performed by [11]. Using Twitter data, [11] performed sentiment analysis to gather insight from public opinion by classifying the tweets based on their positive or negative sentiment value. Sentiment analysis or opinion mining is the study of opinions, attitudes, and emotions toward an entity [12]. The entity can be topics, individuals, or events. In the scientific community, the two terms are interchangeable, refering to the same thing [4].

Another work by [13] developed a framework for measuring costumers satisfaction towards mobile application products by analyzing online review data using sentiment analysis combined with VIKOR method (*ViseKriterijumsa Optimizacija I Kompromisno Resenje*). The utilization of opinion mining and sentiment analysis for analyzing online reviews written by customers can also be implemented in determining key attributes affecting customer satisfaction within hospitality service [14].

METHODS

A. Data Collection and Pre-processing

The general overview of the design methodology can be seen in Fig.1. The data gathered are users' reviews about a point-of-sales company application from the Google Play Store platform. The data were then pre-processed to remove unnecessary characters that could poorly affect the analysis result, such as punctuation, emoticon, numeric value, and white spaces. Another pre-processing that is performed on the data is the stemming process. Stemming is transforming a word into its stem (initial) form. The stemming process in this research was performed by utilizing Sastrawi, a python library that specialized in stemming Indonesian words.

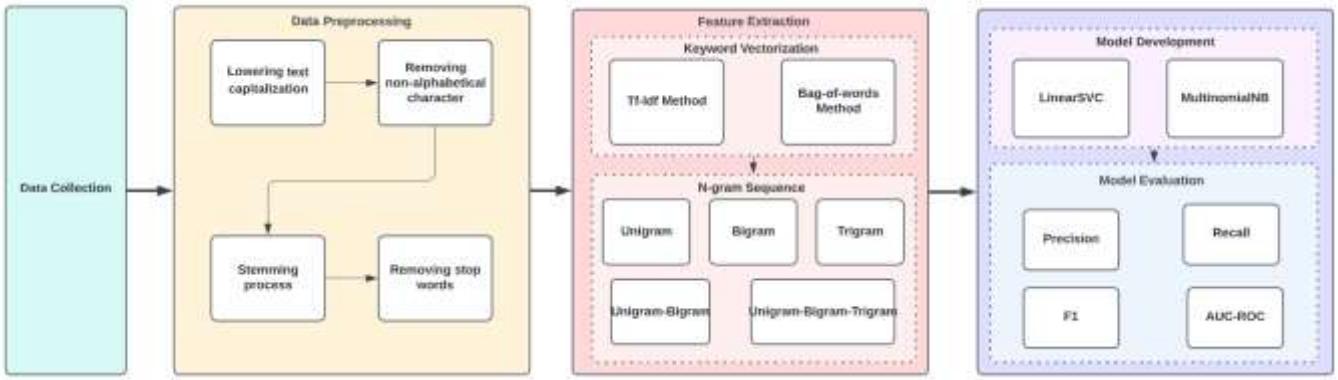

Fig.1 Research design implemented in this paper.

After the stemming process, the author continues the data pre-processing by removing stop words from the document. Stop words are a list of words commonly used as pronunciation or particle. These words frequently appear in a document but do not have sentimental value. Removing the stop words is necessary since it could alter the analysis result due to their high number of appearances on the documents. Text data transformation during data pre-processing is illustrated in Table 1.

B. Feature Extraction

When building a classification model, selecting the feature vector is critical task to do. Selecting a suitable feature vector could hugely impact in the success level of our classifier. The feature vector is used to construct a model from which the classifier learns and can classify previously unseen data [11]. In this research, tokenization is performed to the review data to convert the text data

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

into a vector. The tokenization or vectorization builds a bag of words based on their frequency in the document. The author implements two vectorization methods: bag-of-words and term frequency–inverse document frequency (tf-idf) vectorization. Five n-grams variations are used during vectorization process. The five n-grams variations are: unigram, bigram, trigram, unigram-bigram (mixed of unigram and bigram), and unigram-bigram-trigram (mixed of unigram, bigram, and trigram).

Table 1. Text data transformation process.

Pre-processing Step	Sample
Initial text data	*Aplikasinya sangat membantu usaha klinik saya..*
Lowering capitalization	*aplikasinya sangat membantu usaha klinik saya..*
Removing non-alphabetical characters	*aplikasinya sangat membantu usaha klinik saya*
Stemming	*aplikasi sangat bantu usaha klinik saya*
Removing stop words	*aplikasi sangat bantu usaha klinik*

Table 2. Tokenization based on n-gram types.

N-gram	Sample Token
Unigram	*'aplikasi', 'bantu', 'usaha'*
Bigram	*'aplikasi bantu', 'usaha klinik'*
Trigram	*'aplikasi bantu usaha', 'tidak bisa masuk'*
Unigram-Bigram	*'aplikasi', 'bantu usaha'*
Unigram-Bigram-Trigram	*'aplikasi', 'bantu usaha', 'tidak bisa masuk'*

C. Model Development and Evaluation

The top 100 tokens from the vectorizations are then used as a feature in building the machine learning model using the Linear Support Vector Classification (LinearSVC) and Multinomial Naïve Bayes (MultinomialNB) algorithm. The labeling process of each review as positive or negative was done by using the rating score. The author grouped the reviews with one and two-star ratings as negative (dissatisfied) and four and five-star reviews as positive (satisfied). The training and assessment of classifier performance are performed by implementing five-fold cross-validation. The model's performance was evaluated using classification performance metrics such as precision (1), recall (2), F1 measure (3), and area under the receiver operating curve (AUC).

$$precision = \frac{tp}{tp+fp} \quad (1)$$

$$recall = \frac{tp}{tp+fn} \quad (2)$$

$$F1 = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (3)$$

With tp is the true positive, fp is false positive, and fn as false negative.

D. Opinion Mining on Satisfied or Dissatisfied Review

The author uses bag-of-words vectorization combined with trigram tokens to assess the opinion about the services. The bag-of-words vectors are chosen because it is easier to interpret since it is more intuitive for humans. Trigram was chosen following the general structure of an Indonesian sentence that commonly consists of three attributes: subject, verb, and object or adjective. The opinion mining process was then performed for positive (satisfied) and negative (dissatisfied) reviews to determine the main concern for each review category. The top ten of the most frequent tokens are then analyzed.

RESULTS

The result is from 6,994 reviews written in Indonesian received by the company during 2019 and 2022. The reviews sentiment's distribution is pictured in Fig 2.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

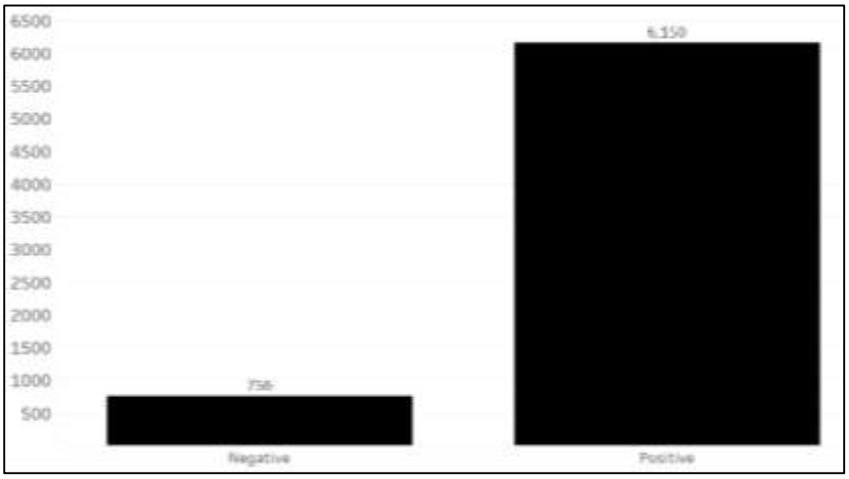

Fig.2 Distribution of reviews' sentiment.

It can be seen that the data distribution is imbalanced between positive and negative classes. Reviews with positive sentiment dominate the distribution with 6,150 data points or approximately 87% of the total reviews. The rest, or 13%, is the reviews with negative sentiment.

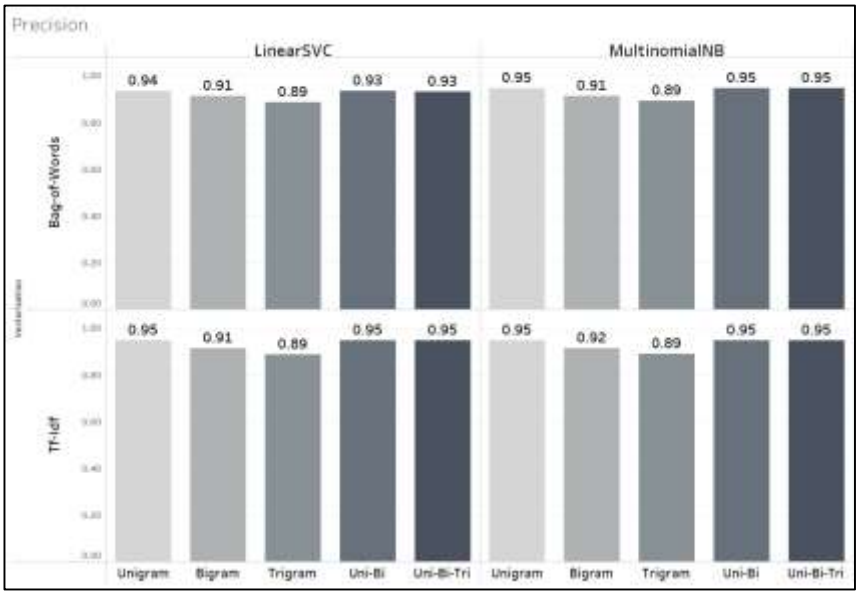

Fig.3 Precision performance of each combination.

The precision score for this experiment ranges from 0.89 to 0.95 (Fig.3). The lowest happens in all trigram token combinations with a precision value of 0.89. While the highest precision score, 0.95, is achieved by unigram, mixed of unigram-bigram (uni-bi), and mixed of unigram-bigram-trigram (uni-bi-tri) tokens in combination with tf-idf vectorizer, both for LinearSVC and MultinomialNB algorithm. The highest score was also achieved by unigram, unigram-bigram, and unigram-bigram-trigram tokens in combination with the bag-of-words vectorizer and MultinomialNB algorithm.

A similar result pattern was also observed for the recall score, where trigram combination along all vectorizers and model algorithms gave the lowest performance score. Generally, the recall score for each combination is high, with a minimum score of 0.90 and the highest score of 0.95 (Fig.4). The highest score is achieved by combining unigram, unigram-bigram, and unigram-bigram-trigram with tf-idf vectorizer for both LinearSVC and MultinomialNB algorithms. A combination of unigram along with bag-of-words vectorization and MultinomialNB algorithm also produces similar results.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

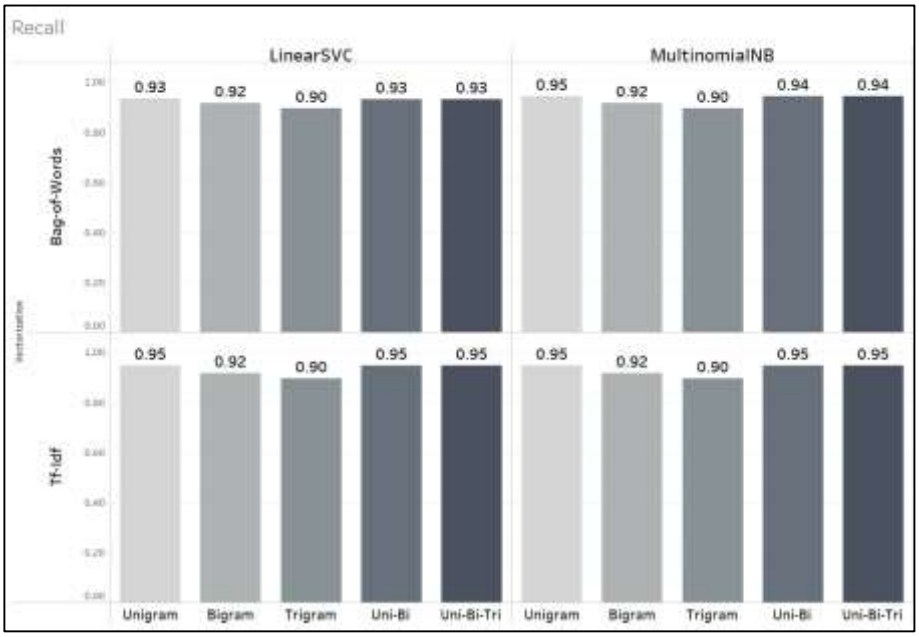

Fig.4 Recall performance of each combination.

F1 measure as the corresponding mean between precision and recall also shows the same pattern as the two previous measurements. The F1 scores in this experiment range from 0.86 to 0.95 (Fig.5). Trigram tokens have the lowest performance across all vectorizers and model algorithms combinations with F1 scores of 0.86. A combination of unigram, unigram-bigram, and unigram-bigram-trigram with tf-idf vectorizer for both LinearSVC and MultinomialNB algorithms resulted in the highest result with 0.95 F1 scores.

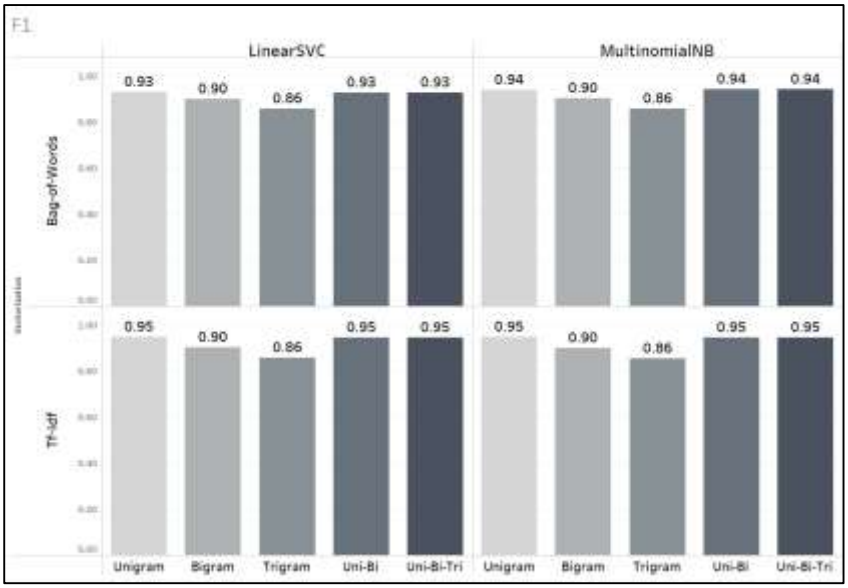

Fig.5 F1 measure of each combination.

The last performance evaluation is the area under the receiver operating curve (AUC). Fig.6 shows that bigram and trigram tokens perform poorly compared to unigram and mixed of the three tokens type. The trigram tokens, the worst-performing tokens, received a 0.58-0.60 AUC-ROC score. The best performing combination with AUC scores of 0.94 is achieved by mixed of unigram-bigram or unigram-bigram-trigram token, tf-idf vectorizer, and MultinomialNB algorithm.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

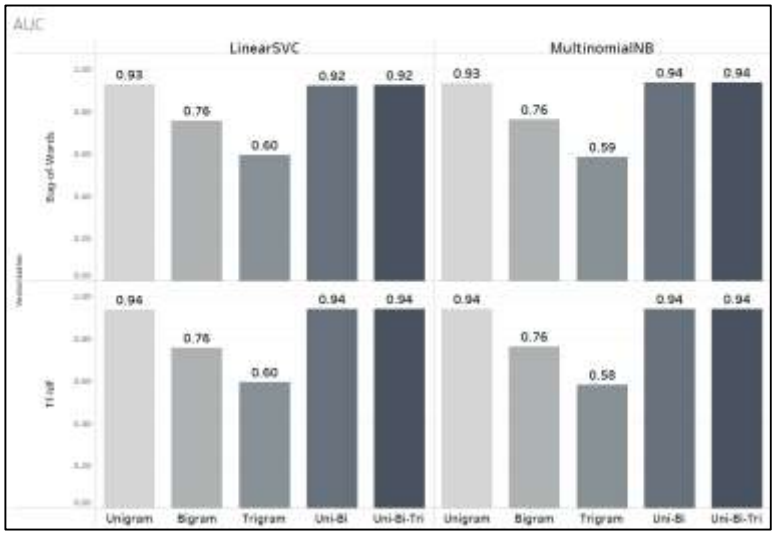

Fig.6 AUC scores of each combination.

The result of opinion mining from user reviews is shown in Table 3. The main concern for the negative (dissatisfied) review is the application that often gets an error *(aplikasi sering error*) along with the inability to log in to their account (*tidak bisa login, ga bisa buka, gak bisa masuk*). Meanwhile, for the positive review, they express their opinion about how well the application helps the user's business (*aplikasi sangat membantu, sangat membantu usaha*) and how easy to operate the application (easy to use, *bagus mudah gunakan*).

Table 3. Top ten most frequent terms for positive and negative reviews

Negative		**Positive**	
Term	**Frequency**	**Term**	**Frequency**
aplikasi sering error	10	*aplikasi sangat membantu*	74
tidak sesuai dengan	10	*sangat membantu usaha*	72
ketika tutup kasir	8	easy to use	58
laporan laba rugi	8	*terima kasih* ***	38
pakai aplikasi ini	8	*sangat membantu dalam*	36
tidak bisa login	8	*aplikasinya sangat membantu*	34
aplikasi kasir ini	6	*nice pos kasir*	28
ga bisa buka	6	*bagus mudah gunakan*	24
ga bisa dibuka	6	*bagus sangat membantu*	22
gak bisa masuk	6	*aplikasi sangat bagus*	20

***company name

DISCUSSIONS

In measuring the model performance, weighted average scoring was applied for precision, recall, and f1 metrics. The weighted average is performed in order to take into account class imbalance in the dataset. Without considering the class imbalance, the performance might be biased since it could perform excellently in one type of class but not perform well in the other class.

The poor performance of high-order n-grams (bigram and trigram) in several measurements is aligned with previous research by [15]. According to their findings, sentiment analysis of documents at the sentence level using unigrams outperforms higher-order n-grams [15]. This phenomenon could be explained by the fact that the frequency of bi and trigrams per sentence is even lower than that of unigrams at the sentence level [16].

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

Previous research found that machine learning models performed significantly better when tested using the tf-idf approach than bag-of-words [17]. Even though, for this experiment, the difference is insignificant. The similar result between bag-of-words and tf-idf performance is likely due to the almost similar result of tokens extracted by bag-of-words and tf-idf vectorizer.
The same thing was also observed for the relative performance of machine learning. Both MultinomialNB and LinearSVC have similar performance in general. The result shows that Naïve Bayes and Support Vector Machine are decent sentiment analysis classifiers. This result aligns with experiments from [18], [19], [20], which show that both Naïve Bayes and Support Vector Machines have decent performance as a classifier in text-based data.

CONCLUSIONS AND RECOMMENDATIONS

The author finds that predicting customer satisfaction using text and sentiment analysis methods on user-generated content is possible. The model's performance in this experiment is decent, with high precision, recall, F1, and AUC score.
The critical thing to consider is the n-grams token used to create the feature. High-order n-grams tend to perform worse compared to low-level n-grams. Trigrams tokens have the lowest performance in precision, recall, F1 measure, and AUC score, with the average scores as follows: 0.89, 0.90, 0.86, and 0.59. The best performing n-grams were achieved by combining unigram, bigram, and trigram with an average performance score of 0.94 for all measurements. Hence author suggests the model built using a combination of unigram, bigram, and trigram tokens as feature vectors.

The best combination of n-grams, vectorization methods, and machine learning algorithms to achieve the highest performance score is a unigram-bigram-trigram token, tf-idf vectorizer, and the MultinomialNB algorithm. This combination will result in 0.95 precision, 0.95 recall, 0.95 F1 measure, and 0.94 AUC score.
The recommendation from the author is to expand further the research on the different source types of user-generated content. This is important since the interaction channel between the company and its customer currently varies. The author also encourages technical improvement, especially in pre-processing Indonesian language data. The transformation of slank words or abbreviations could help improve the result of future research.

REFERENCES

1. J. Kim and C. Lim, "Advanced Engineering Informatics Customer complaints monitoring with customer review data analytics : An integrated method of sentiment and statistical process control analyses," Adv. Eng. Informatics, vol. 49, no. April, p. 101304, 2021, doi: 10.1016/j.aei.2021.101304.
2. E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, "A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making," Ind. Mark. Manag., vol. 90, no. November 2018, pp. 523–537, 2019, doi: 10.1016/j.indmarman.2019.08.003.
3. S. W. Lee, G. Jiang, H. Y. Kong, and C. Liu, "A difference of multimedia consumer's rating and review through sentiment analysis," Multimed. Tools Appl., vol. 80, no. 26–27, pp. 34625–34642, 2021, doi: 10.1007/s11042-020-08820-x.
4. R. Jena, "An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach," Ind. Mark. Manag., vol. 90, no. January, pp. 605–616, 2020, doi: 10.1016/j.indmarman.2019.12.012.
5. S. F. Wamba, A. Gunasekaran, S. Akter, S. J. fan Ren, R. Dubey, and S. J. Childe, "Big data analytics and firm performance: Effects of dynamic capabilities," J. Bus. Res., vol. 70, pp. 356–365, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.009.
6. H. Yakubu and C. K. Kwong, "Forecasting the importance of product attributes using online customer reviews and Google Trends," Technol. Forecast. Soc. Change, vol. 171, no. June, p. 120983, 2021, doi: 10.1016/j.techfore.2021.120983.
7. K. N. Lemon and P. C. Verhoef, "Understanding customer experience throughout the customer journey," J. Mark., vol. 80, no. 6, pp. 69–96, 2016, doi: 10.1509/jm.15.0420.
8. E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," in Procedia Computer Science, 2013, vol. 17, pp. 26–32. doi: 10.1016/j.procs.2013.05.005.
9. P. Williams and E. Naumann, "Customer satisfaction and business performance : a firm-level analysis," vol. 1, no. May 2009, pp. 20–32, 2011, doi: 10.1108/08876041111107032.
10. Y. Piris and A. C. Gay, "Customer satisfaction and natural language processing," J. Bus. Res., vol. 124, no. December 2020, pp. 264–271, 2021, doi: 10.1016/j.jbusres.2020.11.065.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 01 January 2023
DOI: 10.47191/ijcsrr/V6-i1-56, Impact Factor: 5.995
IJCSRR @ 2023
www.ijcsrr.org

11. S. Al-Otaibi et al., "Customer satisfaction measurement using sentiment analysis," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 106–117, 2018, doi: 10.14569/IJACSA.2018.090216.
12. A. Ilavendhan, S. Ranjan, and S. N. Manoharan, "An Empirical Analysis on Various Techniques Used to Detect the Polarity of Customer Satisfaction in Sentiment Analysis," pp. 4376–4386, 2021.
13. D. Kang and Y. Park, "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach," Expert Syst. Appl., vol. 41, no. 4 PART 1, pp. 1041–1050, 2014, doi: 10.1016/j.eswa.2013.07.101.
14. J. Ming, H. Quan, G. Li, and R. Law, "International Journal of Hospitality Management Understanding service attributes of robot hotels : A sentiment analysis of customer online reviews," Int. J. Hosp. Manag., vol. 98, no. July, p. 103032, 2021, doi: 10.1016/j.ijhm.2021.103032.
15. A. Andreevskaia and S. Bergler, "When Specialists and Generalists Work Together : Overcoming Domain Dependence in Sentiment Tagging," no. June, pp. 290–298, 2008.
16. M. Ghiassi and S. Lee, "A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach," Expert Syst. Appl., vol. 106, pp. 197–216, Sep. 2018, doi: 10.1016/j.eswa.2018.04.006.
17. S. Akuma, T. Lubem, and I. Terngu, "Comparing Bag of Words and TF - IDF with different models for hate speech detection from live tweets," Int. J. Inf. Technol., vol. 14, no. 7, pp. 3629–3635, 2022, doi: 10.1007/s41870-022-01096-4.
18. M. Bordoloi, "Sentiment Analysis of Product using Machine Learning Technique : A Comparison among NB, SVM and MaxEnt Sentiment Analysis of Product using Machine Learning Technique : A Comparison among NB, SVM and MaxEnt," Int. J. Pure Appl. Math., no. July, 2018.
19. J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classi fi cation methods," vol. 36, pp. 20–38, 2019, doi: 10.1016/j.ijresmar.2018.09.009.
20. S. Wang and C. D. Manning, "Baselines and Bigrams : Simple, Good Sentiment and Topic Classification," no. July, pp. 90–94, 2012.