

Aspect based sentiment analysis for Modelling Internet Service Provider Product During Covid-19 Pandemic

Abstract

The use of the internet is growing faster. Internet Service Provider is a company or service provider that provides services in making the internet. Various providers have been present in Indonesia such as *Telkomsel*, *XL*, *Indosat* and 3. Providers continue to try to make developments in the services they provide by adjusting technological advances and consumer needs. One of the innovations made by several providers is to provide an all-digital provider in the services they provide. There are 2 providers that released this innovation, namely *by.U* and *MPWR*. To find out consumer preferences for *by.U* and *MPWR* providers, it is needed from the customer's opinion. Twitter social media can be a place to get opinions about consumer opinion. Consumer opinion are grouped into 4 feature categories, namely service features, price, network and product categories. The process of categorizing consumer opinion in the form of tweets uses the *cosine similarity* algorithm and in the sentiment analysis uses the *Multinomial Naïve Bayes method*. Testing was carried out with 2 datasets, namely a dataset for provider *by.U* totaling 100 tweets and the dataset for *MPWR* totaling 100 tweets. The accuracy obtained is 75% for *by.U* dataset and 50% for the *MPWR* dataset from the total test data of 20 test data.

Keywords: *Cosine Similarity, Multinomial Naïve Bayes, by.U, MPWR*

1. INTRODUCTION

Internet is needed in order to be able to access information in our time by using a provider. Provider is a service provider in using the internet. Service providers or providers continue to innovate in the services they make to meet consumer needs [2]. *by.U* and *MPWR* are one of the new providers or service providers as the first full digital service providers in Indonesia. To verify the quality of the products that the provider makes, the provider needs the consumer preferences of *by.U* and *MPWR* users, namely customer opinions. Because customers are so important to business, service companies should research customer satisfaction information to find ways to improve their services. Customer feedback can be collected through social media such as Twitter.

Twitter is a social media that can disseminate information quickly in the form of comments, opinions, criticisms and suggestions, both positive and negative [1]. Aspects that increase satisfaction include product quality, price, service, and network [2].

The data in the form of tweets will later be analyzed by sentiment using the Multinomial Naïve Bayes method with positive and negative categories and the Cosine Similarity algorithm to classify tweets by calculating their similarity based on queries. Based on the background explained about consumer preferences for *by.U* and *MPWR* products, the author will conduct a study entitled "Analysis of Sentiment Against *MPWR* and *by.U*

Providers Based on Multinomial Naïve Bayes Using Cosine Similarity Method.

2. RELATE RESEARCH

Sentiment analysis is the process of processing attitudes, feelings, data about problems and events to obtain information. The task of sentiment analysis is to classify the polarity value of the text in the document, sentence or aspect feature that can be positive, negative or neutral.

One of the related studies conducted by Putri Dinda Utami, this study describes sentiment analysis of Indonesian cosmetic reviews using the *naïve Bayes classifier* algorithm. The test data used are 100 data and the preprocessing text used consists of case folding, tokenizing, stopword and filtering. The results showed the accuracy of getting a value of 80%.

One of the related studies was conducted by Haqqizar and Larasyanti. This study describes sentiment analysis on *Telkomsel* service providers with data regarding the keywords *#Telkomsel* or *@Telkomsel*. The data used are 151 tweet data. The results of the research show that the public gives "negative sentiment" to *Telkomsel*'s services, with the highest accuracy value of 70.21%, precision of 70.11% and recall of 70.33% taken from 151 tweet data.

Other related research was conducted by Rizki Tri Wahyuni, Dhidik Prastiyanto, and Eko Suprptono. This research consists of 50 thesis

documents. This study discusses how to automatically classify documents into different folders in the database in order to better manage existing documents and the accuracy rate is 98%. Classification is done using the Cosine Similarity algorithm with TF-IDF calculations.

The last related research was conducted by Rakhman Halim Satrio, Mochammad Ali Fauzi and Indriati. This study discusses the classification of tweets on twitter using the K-Nearest Neighbor (K-NN) method. The number of data tested is 140, with descriptions of 100 data as training data & 40 data as testing data. In this study, the classification consists of several substances such as football, volleyball, and tennis. The results obtained from this study with $k = 1$, an accuracy of 75.0%;

3. METHODOLOGY

3.1. Research design

The research design that will be used in this study includes collecting data in the form of tweet data, text preprocessing which consists of several stages such as case folding, tokenizing, normalization, stopwords removal and stemming, cosine similarity, TF-IDF word weighting, sentiment classification using multinomial naive bayes, validation and evaluation

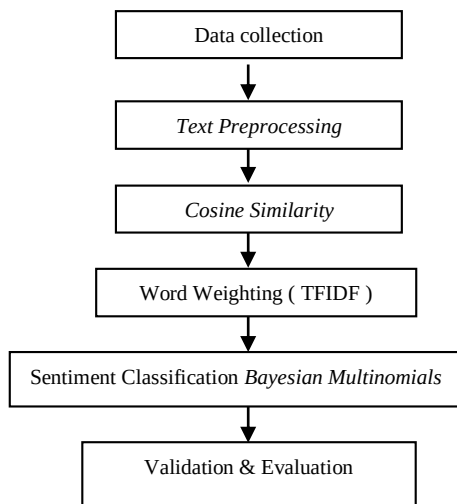


Figure 1. Research flow design

3.2. Data collection

Data collection is done by crawling the data using API Keys on Twitter. The data obtained comes from searches by.U and MPWR. Furthermore, the data obtained will go through the Text Preprocessing stage.

3.3. Text Preprocessing

Text preprocessing This is done to filter or prepare words in the narrative data so that they are clean from noise before weighting. The processes of text processing include case folding, tokenizing, normalization, stopwords removal and stemming.

3.4. Cosine Similarity

Cosine Similarity is the calculation of the similarity between the tweet and query data. The similarity between query and document can be calculated using query vector and tweet vector. (Sha'bani, 2018)

Cosine Similarity Formula (1):

$$\text{Cosine a} = (1) \frac{A.B}{|A||B|} \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2 \times \sum_{i=1}^n (B_i)^2}}$$

3.5. TF-IDF Word Weighting

The weighting technique is used to calculate the weight value of a word in the document using the TF-IDF algorithm. The first step in finding the weight value of a word is to calculate the term frequency for each narrative data. The second step is to calculate the inverse document frequency (IDF). The last step is to calculate the TF-IDF weight value by multiplying the TF value by IDF [5].

3.6. Classification

The classification process used in this research process uses the Multinomial Naïve Bayes (MNB) method. Multinomial Naive Bayes is a condition where the probability is calculated without taking into account a sequence in each word and information that already exists in the document or sentence. The algorithm also counts the number of words that appear in the document data [4].

3.5. Validation and Evaluation

The validation process is carried out to get the most optimal accuracy results, in this study K-Fold Cross Validation as the basis for validation. In this study, it was determined that the K fold value used was 2, 4, 5 and 10. In the selection of the K-Fold value there were no formal rules that became the basis for determining the value.

Furthermore, the evaluation in this study can be done after the validation stage. The most optimal evaluation results are used in the evaluation as the basis for the calculation value. For the value of K-Fold Validation in the evaluation, the most optimal was chosen at the stage of the validation process.

4. RESULTS AND DISCUSSION

4.1. Data

Data collection was taken with a total of 100 data with the keyword #by.U 100 data and 100 data with the keyword #MPWR.

Table 1. Data of tweets by.U

No	Tweet	Label
1	@Telkomsel min kalo trasnfer pulsa dari simpati loop ke by.u bisa nggak ya? terima kasih	Positive
2	by u knp deh tumben lemot ??	Negative
3	Mana kalo udh mati listrik by.u sinyalnya bolot banget	Negative
4	Sinyalmu kok blendre :(	Negative
5	Pake by.u nder, disitu ada kuota khusus zen	Positive
6	@fingerpieckk yh ok, by.u juga murah	Positive
7	@garutfess By.u atau ga indosat, sinyal sama kayak telkom. Smartfren kalo kuota banyak	Positive
8	,ini sinyal by.u dari tadi cuma satu-dua bar mulu dah provider digital yg enak	Negative
9	kayaknya cuma by.u ga ada masa aktif, tp mpwr sama liveon juga bagus promonya	Positive
10	tumben jam segini by.u buat download drama korea lemot.... :(	Negative

Table 2. MPWR tweet tweet data

No	Tweet	Label
1	Indosat Resmi Luncurkan Provider Digital MPWR	Positive
2	Curiga sender nya mpwr staff :(((	Negative
3	Gak jelas, gak ada keterangan yg ngejelasin apa itu mpwr, cara pakenya gimana, dll...	Negative
4	@MPWR_ID Yang dipikirkan nambah New User aja. User lama gak diperhatikan. Siap2 jadi kartu sekali pake buang Tandingan By.U dari	Negative
5	Indosat MPWR. Lagi promo simcard serebu doang cuss @ajifachmi @MPWR_ID	Positive
6	Baru juga aktivasi ngab, mayan serebu dapet 10gb free nich	Positive
7	@MPWR_ID Kok ga sampey ya kartunya	Negative
8	buat kalian yg lagi fakir kuota, bisa install mpwr,	Positive

lumayan promo 1rb 10gb 30hr	
barusan dah beli simcard mpwr cuma seribu, padahal udah ada indihome cuma seruuu aja kalo dapet yg diskonan gitu HAHHAHA	Positive
@mpwr_id nomor saya ga aktif lagi (ga dapet sinyal) siapa yang bisa bantu?	Negative

4.2. Query

In the tweet data there are 4 types of categories. The category of each tweet is determined from the cosine similarity value between the tweet and the query that has been prepared as the basis for each query. The query is contained in table 3 as follows:

Table 3. Categories and queries

Category	Query
Service	bad, good, register, response, register, application, cs, send, login receipt, order, registration
Price	expensive, cheap, price, buy, pay, pulse
Network	slow, loud, smooth, net, signal
Product	promo, active, activation, unlimited, code, referral, package, quota, card, provider

4.3. Cosine Similarity Results

The results of the cosine similarity calculation are obtained between the calculation of the vector distance between the tweet and the query. The highest value will determine the category of the tweet. The greater the value of the distance between the tweet and the query gives the conclusion that the tweet is a tweet with the category with the largest value. The results of cosine similarity on the by.U dataset can be seen in table 4.

Table 4. Results of Category Classification based on Cosine Similarity dataset by.U

Category	Amount of data
Service	9
Price	20
Network	32
Product	39

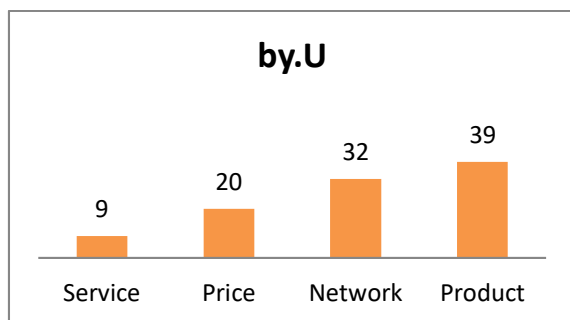


Figure 1. Results of Category Classification based on Cosine Similarity by.U dataset with chart

The results obtained from the calculation of cosine similarity on the dataset by.U get the results with the number of data in the service category totaling 9 data, the price totaling 20 data, network totaling 32 data and product totaling 39 data.

Table 5. Results of Category Classification based on Cosine Similarity MPWR dataset

Category	Amount of data
Service	17
Price	19
Network	11
Product	53

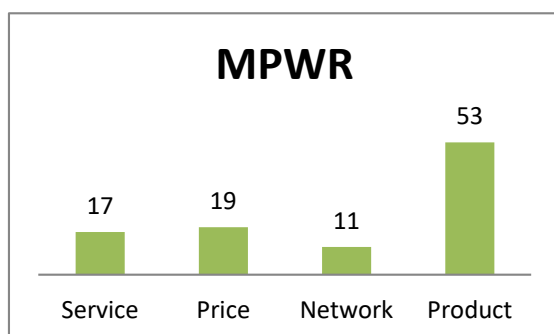


Figure 2. Results of Category Classification based on Cosine Similarity MPWR dataset with chart

The results obtained from the calculation of cosine similarity on the dataset MPWR get the results with the number of data in the service category totaling 17 data, the price totaling 19 data, network totaling 11 data and product totaling 53 data.

4.4. Validation

Validation is used to determine the best K-Fold to be used as a basis for evaluation. The K-Fold values used are 2, 4, 5 and 10. This stage is carried out on the training data of each dataset. The results of each K-Fold as the highest value are shown in Tables 6 and 7.

Table 6. Results of accuracy, precision, recall with data training on by.U dataset

K-Fold	Stage-	Accuracy	Precision	Recall
2	2	62%	52.17%	70.59%
4	3	85%	85.71%	75%
5	3	93.75%	100%	85.71%
10	6	87.50%	100%	75%

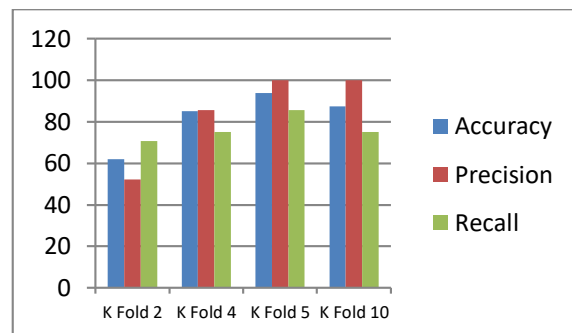


Figure 3. Results of accuracy, precision, recall with data training on by.U dataset with chart

The results obtained from the calculation of accuracy, precision and recall from the K-Fold with values of 2,4,5 and 10, namely for K-Fold 2, the results obtained are 62% on accuracy, 52.17% precision and 70.59% recall. For K-Fold 4, it gets 85% accuracy, 85.71% and 75% recall. For K-Fold 5, the results are 93.75% accuracy, 100% and 85.71% recall. For K-Fold 10, the results are 87.50% accuracy, 100% and 75% recall.

Table 7. Results of accuracy, precision, recall with data training on MPWR dataset

K-Fold	Stage-	Accuracy	Precision	Recall
2	1	80%	75%	75%
4	3	85%	100%	75%
5	4	93.75%	87.50%	100%
10	5	100%	100%	100%

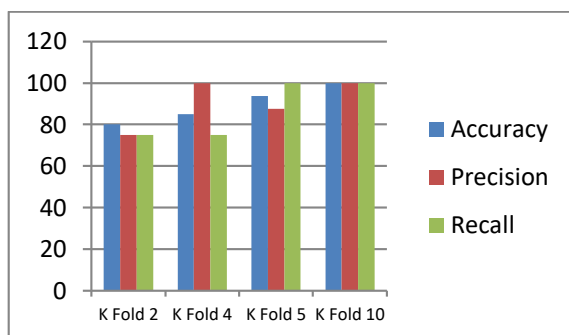


Figure 4. Results of accuracy, precision, recall with data training on MPWR dataset with chart

The results obtained from the calculation of accuracy, precision and recall from the K-Fold with values of 2,4,5 and 10, namely for K-Fold 2, the results obtained are 80% on accuracy, 75% precision and 75% recall. For K-Fold 4, it gets 85% accuracy, 100% and 75% recall. For K-Fold 5, the results are 93.75% accuracy, 87.50% and 100% recall. For K-Fold 10, the results are 100% accuracy, 100% and 100% recall.

4.5. Evaluation

Validation is used to determine accuracy, precision, and recall based on the most optimal K-Fold value, this value is obtained from the previous validation stage. Calculations at the evaluation stage use data testing on each dataset. The results of the evaluation phase are shown in tables 8 and 9.

Table 8. Results of accuracy, precision, recall in the by.U . dataset

K-Fold	Stage-	Accuracy	Precision	Recall
5	3	75%	0%	0%

Table 9. Results of accuracy, precision, recall on the MPWR . dataset

K-Fold	Stage-	Accuracy	Precision	Recall
10	5	50%	0%	0%

Based on the results of calculations on the test data with the amount of test data used amounted to 20 test data. For the by.U dataset, the results are 75% accuracy, 0% precision and 0% recall. As for the MPWR dataset, the results are 50% accuracy, 0% precision and 0% recall.

5. CONCLUSIONS AND SUGGESTIONS

Based on the tests carried out, the following conclusions can be drawn:

1. The results of the Cosine Similarity algorithm on the provider by.U dataset produce 9 tweets including the service category, 20 tweets including the price category, 32 tweets

including the network category and 39 tweets including the product category.

2. The results of the Cosine Similarity algorithm on the MPWR provider dataset produce 17 tweets including the service category, 19 tweets including the price category, 11 tweets including the network category and 53 tweets including the product category.
3. The results of sentiment analysis using the Multinomial Naïve Bayes method on the by.U dataset give 75% accuracy, 0% precision and 0% recall, while the MPWR dataset gives 50% accuracy, 0% precision and 0% recall.
4. The difference in accuracy results was obtained due to differences in datasets and also too few test data, amounting to 20.

Based on the tests carried out, the following conclusions can be drawn:

1. Using other methods in determining category classification and sentiment analysis

REFERENCES

- [1] Haqqizar, N., & Larasyanti, TN (2019). Sentiment Analysis of Telkomsel Telecommunication Service Providers on Twitter Using the Naïve Bayes Method. Proceedings of TAU SNAR-TEK 2019 National Seminar on Engineering and Technology, 10(2), 1–15.
- [2] Wibisono, LE (2011). The Influence of Product Quality, Service Quality and Price Perception on Customer Satisfaction of 4G XL in Yogyakarta. Journal of Physics A: Mathematical and Theoretical, 44(8), 51.
- [3] Sya'bani, MM, & Umilasari, R. (2018). Application of the Cosine Similarity Method and TF / IDF Weighting on the Book Synopsis Classification System at the Jember District Attorney's Library. Justindo (Indonesian Journal of Systems & Technology), 3(1), 31–42.
- [4] Shofiya, Feni(2020) . Comparison of Support Vector Machine (Svm) and Multinomial Naive Bayes (Mnb) Algorithms in Final Project Abstract Classification (Case Study: Faculty of Engineering, University of Muhammadiyah Jember). Undergraduate thesis, University of Muhammadiyah Jember
- [5] Putra Nuansa, E. (2017). Sentiment Analysis of Twitter Users Against the Election of the Governor of DKI Jakarta Using Naïve Bayesian Classification and Support Vector Machine Methods. Sepuluh Nopember Institute of Technology Surabaya, 1–101.
- [6] Zhang, S. (2014). International Journal of Information and Communication Technology Education. SIGI Global Platform, 10(4), 17. <https://www.igi->

global.com/gateway/article/120616#pnlRecommendationForm

- [7] Rintyarna, BS, Sarno, R., & Fatichah, C. (2019). Semantic features for optimizing supervised approach of sentiment analysis on product reviews. *Computers*, 8(3). <https://doi.org/10.3390/computers8030055>
- [8] Rintyarna, BS, Sarno, R., & Fatichah, C. (2019). Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks. *Journal of Big Data*, 6(1). <https://doi.org/10.1186/s40537-019-0246-8>