COMPARING ALGORITHM FOR SENTIMENT ANALYSIS IN HEALTHCARE AND SOCIAL SECURITY AGENCY (BPJS KESEHATAN)

Asyharudin[1]; Novi Kusumawati[2]; Ulfah Maspupah[3]; Destia Sari R.F.[4]; Amir Hamzah[5]; Duwik Lukito[6]; Dedi Dwi Saputra[7]

Information Systems
Nusa Mandiri University
https://nusamandiri.ac.id/
job.adin.gmail.com[1]; novikusumawati703@gmail.com[2]; ulfahmaspupah83@gmail.com[3]; destiafadhillah19@gmail.com[4]; amirhamzah.jkt@gmail.com[5]; duwiklukito09@gmail.com[6]; dedi.eis@nusamandiri.ac.id[7]

Abstract— *Twitter is a social media that can be used to express opinions and exchange information quickly with individuals and institutions such as the Healthcare and Social Security Agency (BPJS Kesehatan). Every word that a Twitter user utters has meaning and stellar emotion. This meaning can be reached through the process of sentiment analysis. Sentiment analysis is the process of understanding and classifying emotions such as positive or negative or complaining or not complaining. This study classifies tweet data related to BPJS Health services into two classifications, namely complain and no complain. Using 1,000 data from Twitter written on the BPJS Kesehatan Twitter account. In text mining, to build a classification, the transform case, tokenize, token filter by length, stemming and stopword techniques are used. Gataframework is used to assist the preprocessing and cleansing process. Rapidminer was used to create sentiment analysis in comparing three different classification methods of the Twitter data. The method used is the Nave Bayes algorithm and the Naïve Bayes algorithm with the addition of a Synthetic Minority Over-sampling Technique (SMOTE) feature and the Naïve Bayes algorithm with an SMOTE feature that is optimized with Adaboost. The Naïve Bayes algorithm is added with the SMOTE feature which is optimized with Adaboost to get the best value with an accuracy value of 69.11%, precision 69.93%, recall 68.89% and AUC 0.770.*

Keywords: *Text Mining, Naïve Bayes, Adaboost, classification, Sentiment Analysis.*

Intisari— *Twitter salah satu media sosial yang bisa digunakan untuk menyampaikan opini dan bertukar informasi dengan cepat kepada individu maupun kepada institusi seperti Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan. Setiap kata yang diutarakan pengguna Twitter memiliki makna dan emosi tersirat. Makna tersebut bisa dipahami melalui proses sentimen analisis. Sentimen analisis merupakan proses memahami dan mengelompokkan emosi seperti positif atau negatif maupun complain atau no complain. Penelitian ini mengklasifikasikan data tweet yang berkaitan dengan layanan BPJS Kesehatan menjadi dua klasifikasi yaitu complain dan no complain. Menggunakan 1.000 data dari Twitter yang ditulis di akun Twitter BPJS Kesehatan. Pada text mining untuk membangun klasifikasi digunakan teknik transform case, tokenize, token filter by length, stemming serta stopword. Gataframework digunakan untuk membantu proses preprocessing dan cleansing. Rapidminer digunakan untuk menciptakan sentimen analisis dalam membandingkan tiga metode klasifikasi yang berbeda dari data Twitter tersebut. Metode yang digunakan adalah, algoritma Naïve Bayes dan algoritma Naïve Bayes ditambahkan feature Synthetic Minority Over-sampling Technique (SMOTE) serta algoritma Naïve bayes ditambahkan feature SMOTE yang di optimasi dengan Adaboost. Algoritma Naïve Bayes ditambahkan feature SMOTE yang di optimasi dengan Adaboost mendapatkan nilai terbaik dengan nilai accurasy 69.11%, precission 69.93%, recall 68.89% dan AUC 0,770.*

Kata Kunci: *Text Mining, Naïve Bayes, Adaboost, Klasifikasi, Sentimen Analis.*

INTRODUCTION

The Indonesian Internet Service Providers Association (APJII) conducted a survey in 2016. There are around 132.7 million internet users in Indonesia (a significant increase from 88 million

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3167

users in 2014). Of this number, 97.4% (129.2 million) are users who use the internet to access social media. The five social media with the most users are Facebook, Instagram, Youtube, Google Plus, and Twitter(Deviyanto & Wahyudi, 2018).

Twitter is a social media that can be used to express opinions and exchange information quickly to individuals and to institutions such as the Healthcare and Social Security Agency (BPJS Kesehatan). The opinion conveyed to BPJS Kesehatan is very important to improve the quality of services. Improving the quality of services at BPJS Kesehatan is very important in order to increase satisfaction with the community in obtaining good and quality health services. BPJS Kesehatan is a legal entity created to be able to organize insurance programs for health(Puspita & Widodo, 2021). Health is a state of health both physically, mentally, spiritually and socially that enables everyone to live socially and economically productive lives. while health efforts are every activity to maintain and improve health carried out by the government and or the community(Suprapto & Malik, 2019).

Every word in the opinion expressed by Twitter users has an implied meaning and emotion. This meaning can be understood through the process of sentiment analysis. Sentiment analysis is the process of understanding and classifying emotions such as positive or negative or complain or no complain.

Sentiment analysis (also known as opinion mining or emotion AI) is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, measure, and study affective states and subjective information. Sentiment analysis is widely applied to customer voice materials such as survey reviews and responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. With the advent of deep language models, such as RoBERTa, more difficult data domains can also be analyzed, for example, news texts where writers usually express their opinions/sentiments less explicitly(*Sentiment Analysis*, n.d.)

Based on the background that has been described, the authors are interested in conducting research with the title "Comparing Algorithm for sentiment analysis in Healthcare and Social Security Agency (BPJS Kesehatan)".

In order for the problem being reviewed to be more focused and achieve the predetermined targets, problem boundaries must be given, including:

a. The object used in this research is tweet data from BPJS Kesehatan twitter users in April 2022.
b. The tweet that will be used is the tweet sentence that uses Indonesian only.
c. The algorithm that will be used for classification in this research is Adaboost and NBC (Naive Bayes Classifier).
d. In this research, the stemming and stopword processes are only for Indonesian words.

MATERIALS AND METHODS

A. Data Collection Techniques

There are several ways to collect data in this research:

1) Data Analysis

Data analysis is a data processing process that aims to find useful information so that it can be used as a basis for decision making as a solution to solve a problem(Kurniasari, 2021). The data used in this study were 1,000 Indonesian-language tweets on Twitter containing the opinions of the Indonesian people on BPJS Kesehatan services. The data is selected manually, namely by selecting tweet sentences that are in Indonesian and do not contain images. The selected data is then stored in excel form. The data in this study consisted of two types, namely training data and test data. For the purposes of training data, the data that has been collected is then categorized manually to assess the sentiment in the tweet, which is included in the complain or no complaint category.

Table 1. Kind of Sentiment

Description	*Sentiment*		
	Complain	***No Complain***	***Grand Total***
Total	485	515	1.000

From the table above, there are 485 complaint data and 515 no-compliance data.

2) Text Processing Analysis

Text processing is a process of extracting, processing, organizing information by analyzing the relationship, the rules that exist in semi-structured or unstructured textual data. To be more effective in the processing process, data transformation steps are carried out into a format that is easy for user needs. This process is called text processing. Once in a more structured form with the above process, the data can be used as a data source that can be processed further. The stages for text processing consist of tokenizing, feature normalization, case folding and stopword removal(Sudiantoro et al., 2018).

B. Research Methods

The research method used is to collect tweet data using the Crawling method from Twitter. Data was taken randomly as many as 1,000 tweets in Indonesian with the keyword BPJS Kesehatan.

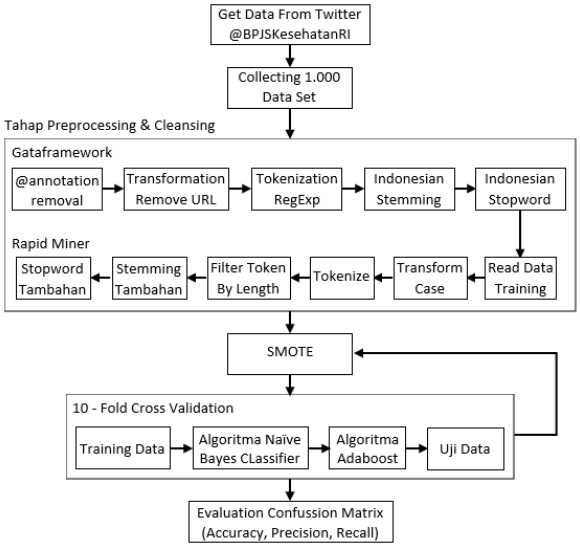

Figure 1. Research Methods

Based on the Figure 1 above, the research process begins with collecting data from Twitter using Rapid Miner, the data taken in this study is 1,000 data with the keyword BPJS Kesehatan, the data that has been collected is stored in excel format. After the data collection process is complete, the next step is to label each data complaint or no complain. After labeling the data, the next process is the preprocessing and cleansing stages. This preprocessing and cleansing stage uses two tools, Gataframewok and Rapid Miner. Gataframework is used to perform the first stage of preprocessing, in the first preprocessing stage the processes carried out are @Annotation Removal, Remove URL, Regexp, Indonesian Stemming and Indonesian stopword. In Rapid Miner, the processes carried out are Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. The last stage in this research is the process of implementing the Naive Bayes algorithm.

RESULTS AND DISCUSSION

A. Research Stages

1) Types Of Research

Sentiment analysis is used to determine the sentiment or polarity of a text whether it is Extremely positive, positive, neutral, negative, Extremely negative. Usually sentiment analysis is applied to text data of public opinion on an object, for example a review of an e-commerce product, a review of a film and comments on social media(Prima et al., 2022). The opinion is in the form of a tweet which will later become a news spread on the Twitter timeline. Each of these opinions is very important for improving the quality of BPJS Kesehatan services to the community, so that people can get good and quality services.

2) Data Collection

The data collection process was carried out using the Twitter API for the Rapid Miner application with the Query "@BPJSKesehatanRI" for the period April 2022 with 1,000 data.

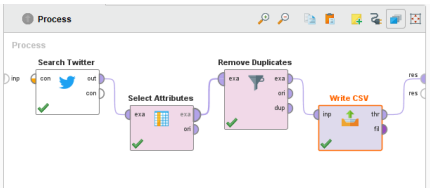

Figure 2. Twitter Data Collection Stage

Based on Figure 2 above, the Search Twitter Operator is used to connect Rapid Miner with Twitter to collect as much as 1,000 data with the keyword BPJS Health. The data collected from Twitter is only taken from the Text Column using the Select Attributes Operator, in this Text Column it contains tweets from Twitter users. To avoid duplicate tweet data, the Remove Duplicates Operator is used. The data that has been selected based on the Text Column and the duplicate data has been removed then the data saved into csv format using the Write CSV Operator.

3) Data Labeling

Data labeling is an advanced stage from the previous stage where calculations will be carried out polarity of the reviews that have been taken, so can produce two categories, namely labels(Herlinawati et al., 2020).

The data that has been collected is then given a sentiment label (Complain or No Complain) using VADER (Valence Aware Dictionary And Sentiment Reasoner). VADER is a glossary and tool for performing sentiment analysis depending on the exclusive standardized law to manifest sentiment on social media. VADER is an open source tool that is completely free. Combines word setting considerations and degree qualifications(America & States, 2021).

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3167

Table 2. Data Labeling Stage

Text	Sentiment
@BPJSKesehatanRI min, apa kartu bpjs kesehatan harua di cetak dulu untuk bisa mendapatkan pelayanan faskes? Apa boleh kita nunjukin kartu virtual di aplikasi saja?	*No_Complain*
@BPJSKesehatanRI @julio_airlangga Manfaatnys adalah pemasukan cuma^2 tanpa kewajiban mencover biaya rs bagi anggota yg kesulitan bayar tepat waktu.	*No_Complain*
@BPJSKesehatanRI Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif	*Complain*
@BPJSKesehatanRI Mau nanya kenapa BPJS nggak bisa di aktifkan lewat mobile JKN ...???	*Complain*
@BPJSKesehatanRI Apaan pandawa malah jawab mohon maaf terus	*Complain*

4) @Annotation Removal
Sometimes in a tweet a user embeds or tags another user's username using the @xxxxx notation. In this process, the stages of removing the username on each tweet are carried out.

Table 3. @Annotation Removal Stage

Before	After
@BPJSKesehatanRI Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif	Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif

5) Remove URL
In a tweet there are usually several URL links (Uniform Resource Locator) entered by the user. This URL is usually included to provide more detailed information because of the limitation of a tweet that is only 280 words. In this process, the process of removing the URL is carried out.

Table 4. Remove URL Stage

Before	After
Min ni gabisa2 Swafotonya? https://t.co/c83xRcjyif	Min ni gabisa2 Swafotonya?

6) Tokenization RegExp
In this process, the procedure for removing punctuation marks on a tweet is carried out, among others. , : " " " ' ? !, etc. In this process, every word contained in the document will be collected and then the punctuation marks, symbols or anything that is not a letter will be removed(Utami, 2018).

Table 5. Tokenization RegExp Stage

Before	After
Min ni gabisa2 Swafotonya?	Min ni gabisa Swafotonya

7) Indonesian Stemming
Stemming is a process to get the basic word from the original word in a sentence. The original word can contain affixes that are separated based on certain rules, for example the word makanan, dimakan, memakan which has the same root word, namely makan(Wardana et al., 2019).

8) Indonesian Stopword
This stage is the process of eliminating certain words in a tweet that are considered meaningless (stopwords). Basically stopword is a list of words in a language. Stopwords tend to be omitted in research related to text mining because stopwords are used repeatedly in a sentence so that stopwords are omitted so that research can focus more on words that are more important. Examples of stop words in Indonesian include yang, dan, di, dari, etc.
The essence of stopwords is to remove words that have low information value or that have no relevance to the content of the document(Hendra & Fitriyani, 2021).

9) Transform Case
In writing a tweet there are several forms of letters used by users, both uppercase and lowercase letters. At this stage all existing letters are converted to lowercase.

Table 6. Transform Case Stage

Before	After
Min ni gabisa Swafotonya	min ni gabisa swafotonya

10) Tokenize
In the tokenize process, the tokenization process is carried out in words or cutting a sentence into word for word.
The tokenization process is the process of separating a series of characters based on each word that composes them or space characters, and it is possible to delete word characters at the same time(Sari et al., 2021).

Table 7. Tokenize Stage

Before	After
Min ni gabisa Swafotonya	min ni gabisa swafotonya

11) Filter Token By Length
This filter token by length is a very interesting function, with this function we can filter tokens of a certain length, the length attribute of a

parameter needs to be specified at the minimum length and maximum length, where we can determine whether a token with the minimum length to the maximum range will stay in document or not(Kalra & Aggarwal, 2018).

In this research the minimum number of characters is 3 and the maximum number of characters is 25.

Table 8. Filter Token By Length Stage

Before	After
min	min
ni	gabisa
gabisa	swafotonya
swafotonya	

B. Implementation of the Naive Bayes Algorithm

The Naive Bayes algorithm is one of the classification techniques algorithms with probability and statistical methods proposed by British scientist Thomas Bayes, which predicts future opportunities based on past experience and is known as Bayes' theorem. The theorem is combined with Naive where it is assumed that the conditions between attributes are independent. Naive Bayes classification assumes that the presence or absence of certain characteristics of a class has nothing to do with the characteristics of other classes(Nofitri & Irawati, 2019).

The results of the model testing carried out are classifying tweets complaining and tweeting no complaints using the Maive Bayes algorithm, the Naive Bayes Algorithm is added with the Synthetic Minority Over-sampling Technique (SMOTE) feature and the Naive Bayes Algorithm is added with the Synthetic Minority Over-sampling Technique (SMOTE) feature which is optimized with Adaboost.

The Naive Bayes Classifier method is used to categorize, namely to see the opinion or tendency of opinion on a problem or object by someone, whether it tends to be in the category of complaint or no complaint. The data that has gone through the text processing process will then go through the classification stage using the Naïve Bayes Classifier to find out whether the data is in the positive category or the negative category.

1) Implementation of Naive Bayes Algorithm Only

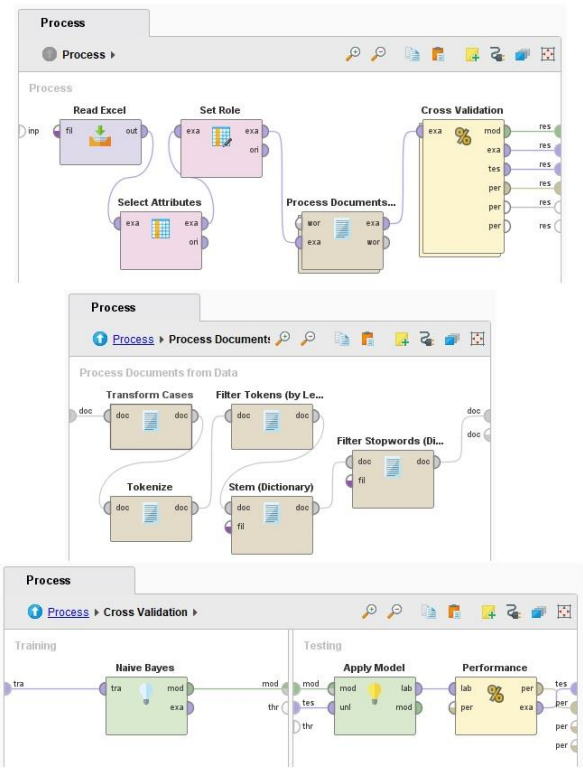

Figure 3. Nave Bayes Algorithm Implementation Only

Based on Figure 3 above, at this implementation stage the data that has gone through the preprocessing and cleansing stages are imported into Rapid Miner using the Read Excel operator. The imported data is then processed using the Process Documents operator, in the Process Documents operator there are several processes including Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. In Figure 3 above, the implementation process only uses Naive Bayes.

The implementation using only the Naive Bayes algorithm gets the result:

Table 9. Implementation results using only the Naive Bayes algorithm

Description	*Accuracy*	*Precision*	*Recall*	*AUC*
Result	71.68%	77.37%	61.17%	0.745

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3167

2) Implementation of Naïve Bayes Algorithm Plus SMOTE Features

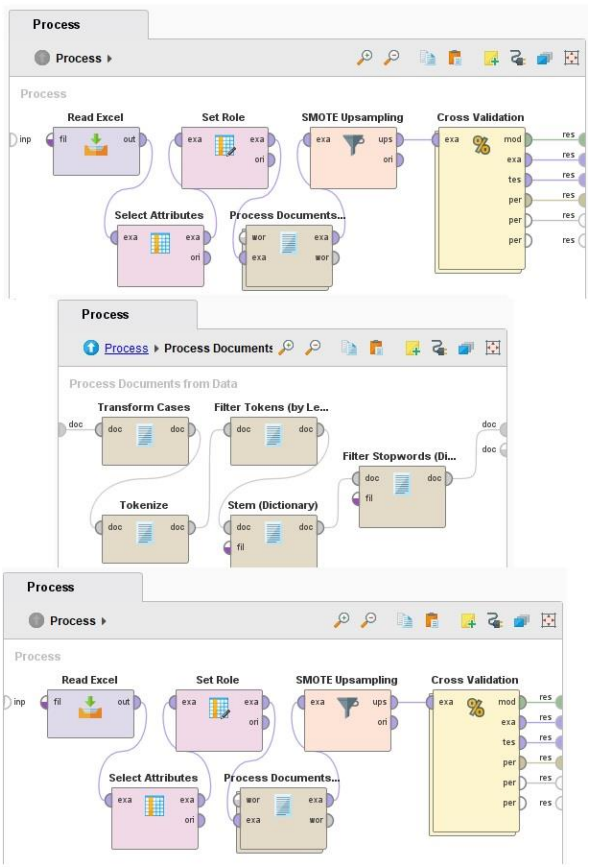

Figure 4. Implementation of Naive Bayes Algorithm Added SMOTE feature

Based on Figure 4 above, at this implementation stage the data that has gone through the preprocessing and cleansing stages are imported into Rapid Miner using the Read Excel operator. The imported data is then processed using the Process Documents operator, in the Process Documents operator there are several processes including Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. In Figure 3 above, the implementation process uses Naive Bayes added the SMOTE feature.

Implementation using the Naïve Bayes algorithm added the SMOTE feature to get the following results:

Table 10. Implementation results using the Naïve Bayes algorithm added the SMOTE feature

Description	*Accuracy*	*Precision*	*Recall*	*AUC*
Result	73.27%	80.24%	61.76%	0.755

3) Implementation of Naïve Bayes & Adaboost Algorithm Plus SMOTE Features

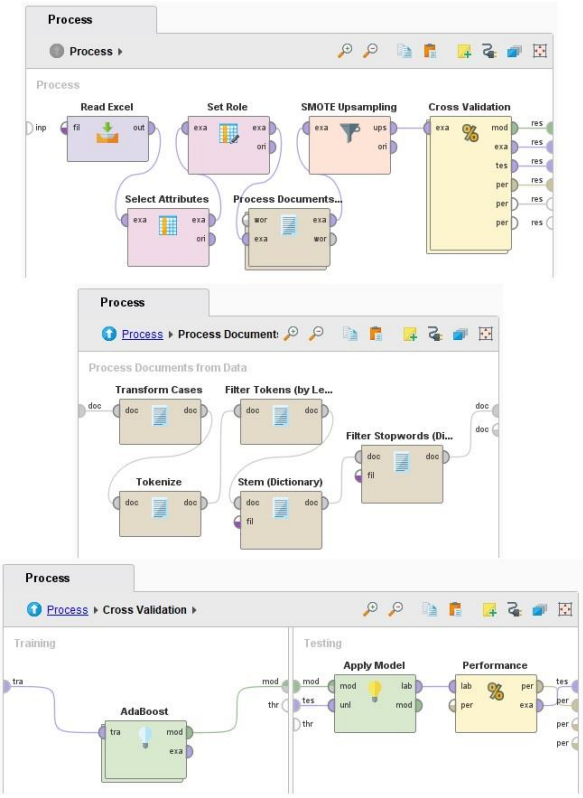

Figure 5. Implementation of Naïve Bayes & Adaboost Algorithm Plus SMOTE Features

Based on Figure 5 above, at this implementation stage the data that has gone through the preprocessing and cleansing stages are imported into Rapid Miner using the Read Excel operator. The imported data is then processed using the Process Documents operator, in the Process Documents operator there are several processes including Transform case, Tokenize, Filter Token By Length, additional Stemming and additional stopwords. In Figure 3 above, the implementation process uses Naive Bayes and Adaboost added the SMOTE feature.

Implementation using the Naïve Bayes and Adaboost algorithm added the SMOTE feature got the following results:

Table 11. Implementation Results Using Naïve Bayes Algorithm and Adaboost Plus SMOTE Features

Description	*Accuracy*	*Precision*	*Recall*	*AUC*
Result	69.11%	69.93%	68.89%	0.770

CONCLUSION

The results of this research indicate that the Naive Bayes Algorithm when added with the Synthetic Minority Over-sampling Technique (SMOTE) feature which is optimized with Adaboost

produces accuracy: 69.11%, precision: 69.93%, recall: 68.89% and AUC: 0.770. This reaserch also uses the Naïve Bayes algorithm without adding the SMOTE feature that produce accuracy: 71.68%, precision: 77.37%, recall: 61.17% and AUC: 0.745. Meanwhile, the Naïve Bayes algorithm added with the SMOTE feature produces accuracy: 73.27%, precision: 80.24%, recall: 61.76%, and AUC: 0.755. Based on the results of this research, it can be concluded that the Nave Bayes Algorithm with SMOTE features added which is optimized using Adaboost is a better classification to use than the Nave Bayes Algorithm with SMOTE features and Nave Bayes Algorithm without SMOTE features.

REFERENCE

America, N., & States, U. (2021). *Survey of Twitter Viewpoint on Application of Drugs by VADER Sentiment Analysis among Distinct Countries*. *March 2021*.

Deviyanto, A., & Wahyudi, M. D. R. (2018). Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *3*(1), 1. https://doi.org/10.14421/jiska.2018.31-01

Hendra, A., & Fitriyani, F. (2021). Analisis Sentimen Review Halodoc Menggunakan Naï¨ve Bayes Classifier. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(2), 78–89. https://doi.org/10.14421/jiska.2021.6.2.78-89

Herlinawati, N., Yuliani, Y., Faizah, S., Gata, W., & Samudi, S. (2020). Analisis Sentimen Zoom Cloud Meetings di Play Store Menggunakan Naïve Bayes dan Support Vector Machine. *CESS (Journal of Computer Engineering, System and Science)*, *5*(2), 293. https://doi.org/10.24114/cess.v5i2.18186

Kalra, V., & Aggarwal, R. (2018). Importance of Text Data Preprocessing & Implementation in RapidMiner. *Proceedings of the First International Conference on Information Technology and Knowledge Management*, *14*(January), 71–75. https://doi.org/10.15439/2017km46

Kurniasari, D. (2021). *Analisis Data Adalah: Mengenal Pengertian, Jenis, Dan Prosedur Analisis Data*. https://www.dqlab.id/analisis-data-adalah-mengenal-pengertian-jenis-dan-prosedur-analisis-data

Nofitri, R., & Irawati, N. (2019). Analisis Data Hasil Keuntungan Menggunakan Software Rapidminer. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *5*(2), 199–204. https://doi.org/10.33330/jurteksi.v5i2.365

Prima, J., Sistem, J., Komputer, I., No, V., Banjarnahor, J., Indra, E., & Sinurat, S. H. (2022). *ANALISIS PERBANDINGAN SENTIMEN CORONA VIRUS DISEASE- 2019 (COVID19) PADA TWITTER MENGGUNAKAN METODE LOGISTIC REGRESSION DAN SUPPORT VECTOR MACHINE (SVM)*. *5*(2).

Puspita, R., & Widodo, A. (2021). Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS. *Jurnal Informatika Universitas Pamulang*, *5*(4), 646. https://doi.org/10.32493/informatika.v5i4.7622

Sari, S., Khaira, U., Pradita, P., & Tri, T. S. (2021). … Beauty Shaming Di Media Sosial Twitter Menggunakan Algoritma SentiStrength: Sentiment Analysis Against Beauty Shaming Comments on Twitter Social Media …. *Indonesian Journal of …*, *1*(1), 71–78. https://journal.irpi.or.id/index.php/ijirse/article/view/55%0Ahttps://journal.irpi.or.id/index.php/ijirse/article/download/55/24

Sentiment Analysis. (n.d.). Retrieved June 30, 2022, from https://en.wikipedia.org/wiki/Sentiment_analysis

Sudiantoro, A. V., Zuliarso, E., Studi, P., Informatika, T., Informasi, F. T., Stikubank, U., & Mining, T. (2018). Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier. *Dinamika Informatika*, *10*(2), 398–401.

Suprapto, S., & Malik, A. A. (2019). Implementasi Kebijakan Diskresi Pada Pelayanan Kesehatan Badan Penyelenggara Jaminan Kesehatan (Bpjs). *Jurnal Ilmiah Kesehatan Sandi Husada*, *7*(1), 1–8. https://doi.org/10.35816/jiskh.v7i1.62

Utami, L. D. (2018). Komparasi Algoritma Klasifikasi Pada Analisis Review Hotel. *Jurnal Pilar Nusa Mandiri*, *14*(2), 261. https://doi.org/10.33480/pilar.v14i2.1023

Wardana, H. K., Swanita, I., & Yohanes, B. W. (2019). Sistem Pemeriksa Pola Kalimat Bahasa Indonesia berbasis Algoritme Left-Corner Parsing dengan Stemming. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *8*(3), 211. https://doi.org/10.22146/jnteti.v8i3.515