

SENTIMENT ANALYSIS OF INTERNET SERVICE PROVIDERS USING NAÏVE BAYES BASED ON PARTICLE SWARM OPTIMIZATION

Anugrah^{1*}, Teguh Iman Hermanto², Ismi Kaniawulan³

Program Studi Teknik Informatika
STT Wastukancana
Purwakarta, Indonesia
www.stt-wastukancana.ac.id

anugrahanugrah27@wastukancana.ac.id^{1*}, teguhiman@wastukancana.ac.id², ismi@wastukancana.ac.id³
(*) Corresponding Author

Abstract

Twitter is a social media application that is widely used. Where as many as 18.45 million users in Indonesia, Twitter users can send and read messages with a maximum of 280 characters displayed. Many opinions and reviews uploaded by users via tweets on social media are experienced in everyday life. Lately, comments about internet service providers in the covid-19 pandemic have been widely reviewed by Twitter users. Problems about internet providers through words often uploaded include internet provider complaints related to network quality, package prices, user satisfaction, and others. This study aims to classify Twitter users' tweets against internet service providers in Indonesia by analyzing the sentiments of 3 internet service providers, namely with the keywords Biznet, first media, and Indihome, using the Naïve Bayes algorithm and optimization with Particle Swarm Optimization. This research is also helpful in helping to become a measure where prospective new users will see the quality of an internet service provider in Indonesia through tweets and then divide the opinion into positive and negative. The results of Biznet's research using Naïve Bayes produce an accuracy of 77.94%, and after optimization, it becomes 81.62%. First media using Naïve Bayes produces 91.39% accuracy, and after optimization, it becomes 92.88%. Indihome using Naïve Bayes produces an accuracy of 85.78%, and after optimization, it becomes 87.48%. It can be concluded that the Naïve Bayes algorithm is a good algorithm for classification, and optimization using Particle Swarm Optimization has an effect on increasing accuracy results.

Keywords: Sentiment Analysis; Internet Service Provider; Naïve Bayes; Particle Swarm Optimization

Abstrak

Twitter yaitu aplikasi media sosial yang ramai digunakan dimana sebanyak 18,45 juta jiwa penggunanya di indonesia, pengguna twitter dapat mengirim dan membaca pesan maksimal 280 karakter yang ditampilkan. Banyak opini serta review yang diunggah pengguna melalui tweet pada media sosial tersebut yang dialami dalam kehidupan sehari-hari, belakangan ini komentar tentang penyedia layanan internet pada pandemi covid-19 banyak diulas oleh para pengguna twitter. Permasalahan tentang provider internet melalui komentar yang sering diunggah diantaranya tentang keluhan provider internet yang berkaitan dengan kualitas jaringan, harga paket, kepuasan pengguna dan yang lainnya. Penelitian ini bertujuan mengklasifikasi tweet pengguna twitter terhadap penyedia layanan internet di indonesia dengan menganalisis sentimen 3 penyedia layanan internet yaitu dengan keyword biznet, first media dan indihome dengan menggunakan metode algoritma Naïve Bayes dan dilakukan optimasi dengan Particle Swarm Optimization. Penelitian ini juga bermanfaat membantu menjadi suatu ukuran dimana calon pengguna baru akan melihat kualitas suatu layanan provider internet yang ada di Indonesia melalui tweet kemudian membagi opini tersebut menjadi positif dan negatif. Hasil penelitian biznet menggunakan Naïve Bayes menghasilkan akurasi 77,94% dan setelah dioptimasi menjadi 81,62%. First media menggunakan Naïve Bayes menghasilkan akurasi 91,39% dan setelah dioptimasi menjadi 92,88%. Indihome menggunakan Naïve Bayes menghasilkan akurasi 85,78% dan setelah dioptimasi menjadi 87,48%. Dapat disimpulkan algoritma Naïve Bayes merupakan algoritma yang baik untuk klasifikasi, dan optimasi menggunakan Particle Swarm Optimization berpengaruh terhadap peningkatan hasil akurasi.

Kata kunci: Analisis Sentimen; Penyedia Layanan Internet; Naïve Bayes; Particle Swarm Optimization

INTRODUCTION

In early March 2020, the President of the Republic of Indonesia, Joko Widodo, announced that Indonesians work from home or study from home. Problems that occur when people decide to choose the best to get internet service. This policy was emphasized by the emergence of Large-Scale Social Restrictions (PSBB) instructions to suppress the spread of the Covid-19 virus (Christianto, 2020). By doing all the activities at home, we are also required to stay connected, whether doing work or learning. It also impacts increasing internet users in Indonesia during the COVID-19 pandemic.

Based on (APJII, 2019) the internet survey report of the Association of Internet Service Providers in Indonesia (APJII) 2019-2020 (Q2), internet user penetration reached 73.7%, namely 196.71 million people out of a total population of 266.91 million people in Indonesia. One of the uses of the internet is social media. Several social media that internet users in Indonesia often use are Facebook, Twitter, Instagram, LinkedIn, and others.

As a medium for communicating and seeking information, Twitter is one of the most popular social media networks, and Twitter is often used as a means of promoting products, advertisements, political campaigns, and as a means to express opinions related to criticism, suggestions, issues, and public opinion (Sudiantoro et al., 2018).

Sentiment analysis is the understanding, extracting, and processing of textual data to obtain information. There are many benefits of sentiment analysis from various points of view, including it can be used to get an overview of public perceptions of service quality, monitoring a product, sales predictions, politics, and investor decision-making (Ipmawati et al., 2017).

In 2018, sentiment analysis was carried out for optimized furniture using Particle Swarm Optimization on Naïve Bayes. The collection of data taken through the amazon website has as many as 200 reviews, divided into 100 positive and 100 negative. Experimental results can reach 93.50%, where optimization affects increasing attributes (Aulianita & Rifai, 2018).

Another research (Hayuningtyas & Sari, 2019) conducted a sentiment analysis for TMII tourism using Naïve Bayes and carried out Particle Swarm Optimization. Data collection through travel websites with 100 reviews, 50 positive and 50 negatives. The results show that the accuracy using only Naïve Bayes is 70%, and after optimization using PSO, it increases to 94.02%

The review on the traffic application using the algorithm used is Naïve Bayes with the use of Particle Swarm Optimization. The results of the first experiment on the test using Naïve Bayes got an accuracy of 69.50% and an AUC graph of 0.488. Then the second test after the Naïve Bayes test was optimized using PSO, where the results increased by 76.24%, and the AUC graph became 0.636 (ER & Solecha, 2021).

A review of distance learning on Twitter was conducted by sentiment analysis in 2021 using the naive Bayes algorithm. They were collecting data through Twitter tweets using the Twitter API as much as 600 data. The accuracy using 3-fold Cross Validation results reached 93% (Khairina & Fitriati, 2021).

Therefore, this study aims to implement particle swarm optimization for sentiment analysis of internet service providers' tweets using Naïve Bayes.

RESEARCH METHODS

The method uses the Naïve Bayes algorithm and Particle Swarm Optimization to improve optimization to analyze Twitter users' tweets against internet service providers. The design for the plot in this research is shown through the research framework in Figure 1. below:

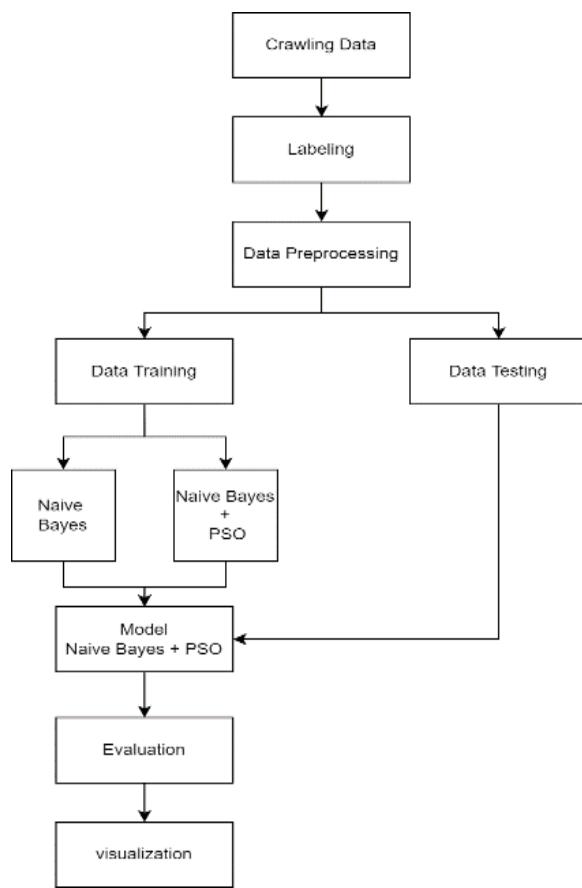


Figure 1. Research Framework

Crawling Data

Data crawling is the stage that aims to collect or download data from a database. Data collection from this analysis is taken through the Twitter server with the help of the Twitter API (Ramadhan & Setiawan, 2019). The data crawling uses the orange tool and is carried out three times because it will look for three tweet documents from the internet provider or ISP, carried out on 04-18 April 2022.

Labeling

Adding sentiment labeling to each tweet document is obtained into two parts: the positive sentiment for Twitter users' tweet documents with a positive tone and the negative sentiment for negative Twitter users' tweet documents. Because it will be divided into positive and negative only, tweet documents that are neutral, which are not opinions, and from official accounts related to keywords will be deleted.

Data Preprocessing

Preprocessing is used for helpful sentiment analysis, making data management easy.

Preprocessing is also carried out before the classification process. This process is needed to clean or remove unnecessary characters or words. This stage aims to optimize the data in the following procedure: classification (Muttaqien, 2016).

a. Transform Cases

Because it is inconsistent in words in the tweeted document, do a lowercase which aims to change all text to lowercase, because the text data or content contained from the crawling data has different letters, namely capital and lowercase letters (Harjanta, 2015).

b. Cleansing

Tweets found that several things will interfere with the analysis process, so in this Cleansing stage, some deletions are carried out using the replace operator in RapidMiner tools, such as removing duplicates to remove the same tweet document so that the tweeted document is neater before the labeling process is carried out. Furthermore, several deletions are carried out, such as removing URL, removing RT (retweet), removing hashtag (#), removing mention (@), and regexp or regular expression. This stage is done to remove certain characters or words with regular expressions set to remove punctuation marks (Rustiana & Rahayu, 2017).

c. Tokenization

Document tweets will be tokenized or split the text or separate word by word, or it can be called sentence fragmentation. Another goal of the tokenization process is to make each word's weight value easier. This process uses the tokenize operator in the document from the data process in the RapidMiner tools (Harjanta, 2015).

d. Filtering

At this stage, the filtering process is carried out to remove words that are not used, namely the stopwords process, to remove or eliminate connecting words such as "and," "which," "in," "is," and others (Rustiana & Rahayu, 2017).

Naïve Bayes

Naïve Bayes is one of the algorithms used for classification. The inventor is a scientist from Great Britain or England named Thomas Bayes, and the classification assumes that one attribute or class has no characteristics with other features (Bustami et al., 2013). The method used is probability and statistics.

Naïve Bayes will carry out a classification that aims to study data patterns so that they can make decisions for data that have not been previously processed and generate sentiment for tweet documents that have not been labeled with a sentiment class, which is positive or negative. In this

process, Naive Bayes will learn from training data using a formula to find the highest probability or opportunity from previous experience or tweet documents. In this case, it is test data.

Particle Swarm Optimization

Particle Swarm Optimization (PSO) was first formulated in 1995 by R. Eberhart and J. Kennedy and is an algorithm of feature selection that functions to solve an optimization problem. Described or the thought process behind this algorithm is inspired by the social behavior of animals, such as flocks of birds or herds of fish in search of food or survival (Haupt & Haupt, 2004).

Optimization for increasing attribute weight uses Particle Swarm Optimization (PSO) using the Optimize Weight operator on the RapidMiner tools, where previously the Naive Bayes algorithm was implemented and then added the PSO process to find a more optimal value. The application of Naive Bayes and PSO is for several attributes of inertia weight (inertial weight), local best weight (best local weight), and global best weight (best global weight). The number of particles determines the population size (number of population) in PSO.

Evaluation

The evaluation in this study uses a confusion matrix on cross-validation for the calculation of accuracy, precision & recall. The analysis starts from the initial stage of using NB and then compares it with the addition of optimized weight or NB - PSO from the three internet provider keywords.

A confusion Matrix is a matrix in the form of two times two that can display results in the form of a percentage of the classification performance, which is carried out with several formulas for measuring the performance (Andika et al., 2019).

Visualization

At the visualization stage, researchers use Microsoft Power BI tools to visualize data with several charts, namely Stacked Column Chart, Pie Chart, Clustered Column Chart, and Wordcloud. Plus, an explanation for the results of the analysis carried out.

Power BI is software from Microsoft to make it easier to visualize the results of processing data with a dashboard. The resulting report can be displayed via a web or mobile device, this software provides many charts, and you can get templates that support the software through the store by downloading it for free (Hanifah, 2020). This software also includes business intelligent software,

which can connect many data with several extensions that support it.

RESULTS AND DISCUSSION

Crawling Data

The data obtained is carried out through a data crawling process on the Twitter social media page with the help of the Twitter API (Application Programming Interface), which is obtained through a request using a Twitter account on Twitter Developer. Figure 2 below shows the Crawling Data process carried out using orange tools.

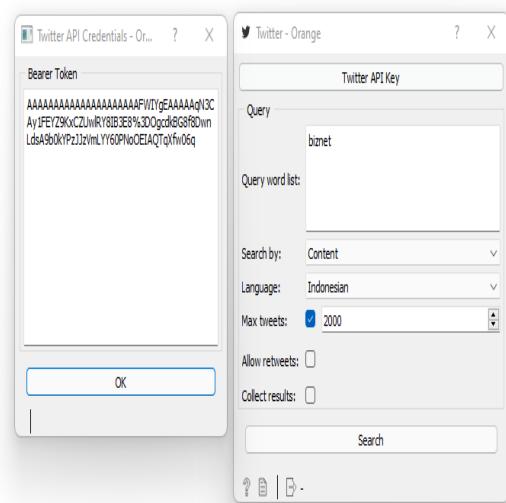


Figure 2. Crawling Data

This crawling process is carried out three times because it will search for three tweet data of internet service providers (ISP) in Indonesia with the keywords "Biznet," "first media," and "Indihome." The total tweet documents obtained from this crawling process are 1,050 tweets. The amount of data crawling from each searched keyword is shown in table 1.

Table 1. Amount of Data Crawling

Keyword	Amount of Data
Biznet	136
First Media	267
Indihome	647

Labeling

The next stage is labeling. At this stage, the researcher marks the excel file obtained through the crawling process by adding a sentiment column. In the excel file, the researcher separates several columns that are only needed, namely the text or tweet data content and the sentiment column. The

labeling process is done manually, with the given classes being positive and negative. Table 2 displays the amount of data labeled as a consequence of each keyword search and an example of labeling data that has been assigned a sentiment.

Table 2. Labeling Example

Text	Sentiment
keren filmnya keren berlangganan dengan menggunakan indihome	positif
marah sama indihome jelek	negatif

The amount of data labeling results from each searched keyword is shown in table 3 below:

Table 3. Amount of Data Labeling Results

Keyword	Amount of Data	Positive Sentiment	Negative Sentiment
Biznet	136	40	96
First	267	22	245
Media			
Indihome	647	113	534

Data Preprocessing

a. Transform Cases

Convert all letters in the tweeted document to lowercase. An example of the transform cases in a tweet document can be seen in table 4 below:

Table 4. Transformation Process

Before Transform Cases	After Transform Cases
Kecewa dengan IndiHome yang Mempersulit PDA (Perpindahan Alamat) - https://t.co/j7ZkXS9B0	kecewa dengan indihome yang mempersulit pda (perpindahan alamat) - @indihome @IndiHomeCare #IndiHome">https://t.co/j7zkxs9b0tcc @indihome @IndiHomeCare #IndiHome

b. Cleansing

Several deletions are carried out at this stage, such as punctuation, URL, rt, #, and @. An example of the cleansing process in a tweet document can be seen in table 5 below:

Table 5. Cleansing Process

Before Cleansing	After Cleansing
------------------	-----------------

kecewa dengan indihome yang mempersulit pda (perpindahan alamat) - @indihome @indihomecare #indihome">https://t.co/j7Zkxs9b0tcc @indihome @indihomecare #indihome	kecewa dengan indihome yang mempersulit pda perpindahan alamat
---	---

c. Tokenization

Break up text or words, or it can be called sentence beheading. An example of the tokenization process in a tweet document can be seen in table 6 below:

Table 6. Tokenization Process

Before Tokenization
kecewa dengan indihome yang mempersulit pda perpindahan alamat
After Tokenization
kecewa dengan indihome mempersulit pda perpindahan alamat

d. Filtering

At this stage, the conjunction is removed or commonly called stopwords removal. An example of the filtering process in a tweet document can be seen in table 7 below:

Table 7. Filtering Process

Before Filtering	After Filtering
kecewa dengan indihome yang mempersulit pda perpindahan alamat	kecewa indihome mempersulit pda perpindahan alamat

Naïve Bayes

After data preprocessing is done, at this stage, several steps are carried out with several operators used, for the algorithm validation stage uses a cross-validation operator where the operator is divided into two parts of the process, namely training and testing. For the training process, a classification algorithm is used, namely naive Bayes, while in the testing process, a model is applied that is connected to performance and to run the naive Bayes algorithm or operator in the previous training process. The application process in RapidMiner can be seen in Figure 3. as follows:

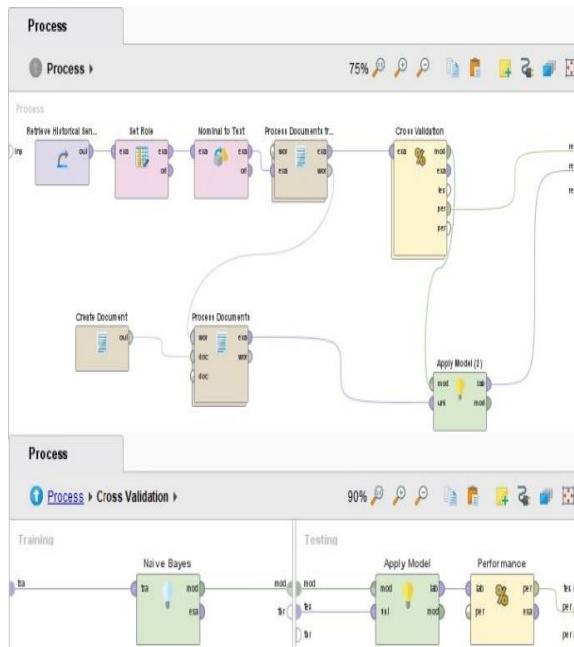


Figure 3. Naïve Bayes Process in RapidMiner

Furthermore, the connecting process begins with creating a document, processing the paper, and applying the model to predict the testing data.

Particle Swarm Optimization

In the previous stage, a classification algorithm using naive Bayes was carried out at this stage adding an optimized weights operator for Particle Swarm Optimization to optimize the increase in attributes.

There is a cross-validation operator for naive Bayes in the optimize weights operator because it will optimize the increase in the optimization algorithm. The process in RapidMiner can be seen in Figure 4. below:

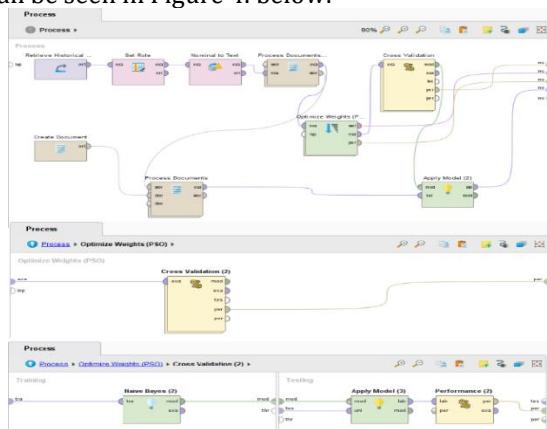


Figure 4. Naïve Bayes and Particle Swarm Optimization Process in RapidMiner

Evaluation

In the previous classification results, researchers will analyze and calculate the performance of the nave Bayes algorithm and naive Bayes - PSO using a confusion matrix to calculate accuracy, precision, and recall using RapidMiner tools.

a. Biznet

The results of several measurements for evaluating Biznet keywords can be seen in table 8 below:

Table 8. Biznet Evaluation Results

Measurement	NB	NB + PSO
Accuracy	77,94%	81,62%
Precision	62,50%	65,31%
Recall	62,50%	80,00%

b. First Media

The results of several measurements for evaluating First Media keywords can be seen in table 9 below:

Table 9. First Media Evaluation Results

Measurement	NB	NB + PSO
Accuracy	91,39%	92,88%
Precision	93,70%	95,20%
Recall	97,14%	97,14%

c. Indihome

The results of several measurements for evaluating Indihome keywords can be seen in table 10 below:

Table 10. Indihome Evaluation Results

Pengukuran	NB	NB + PSO
Accuracy	85,78%	87,48%
Precision	93,29%	95,55%
Recall	89,08%	88,89%

Visualization

At this stage, visualization is carried out to display a chart for the data results using Microsoft Power BI: Stacked Column Chart, Pie Chart, Clustered Column Chart, and Wordcloud.

a. Stacked Column Chart

Figure 5 shows a stacked column chart used to visualize the count of tweet documents for each internet service provider using a bar chart.

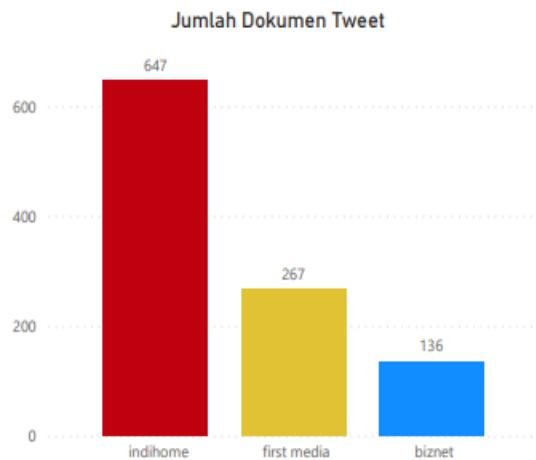


Figure 5. Tweet Document Count Chart

b. Pie Chart

Figure 6 shows the pie chart used to visualize the count of sentiment attributes across all tweet documents.

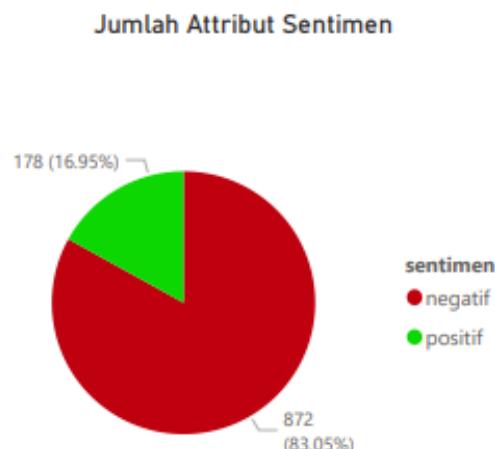


Figure 6. Sentiment Attribute Count Chart

c. Clustered Column Chart

Figure 7 shows a chart used to visualize the number of sentiment attributes for each ISP.

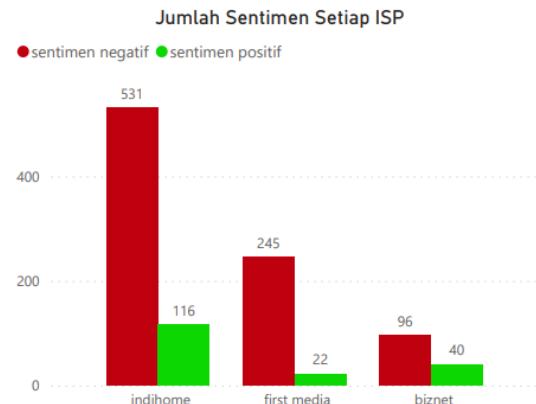


Figure 7. Amount Sentiment Attribute Chart of Each ISP

Figure 8 shows the confusion-matrix evaluation results chart for each ISP.

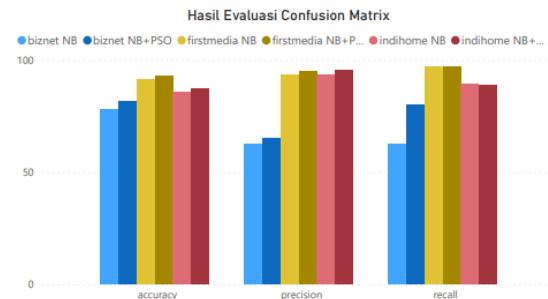


Figure 8. Confusion-Matrix Evaluation Results Chart

d. Wordcloud

Figure 9 shows a list of words that often appear.

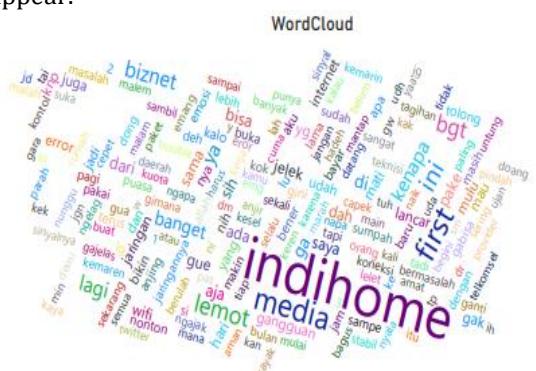


Figure 8. Wordcloud Visualization

The accuracy value as a classification result is used as a benchmark in determining the sentiment of a comment or tweet carried out in the confusion matrix test, coupled with optimization using particle swarm optimization to improve accuracy. But in some cases, other classification algorithms are

better at testing. So it is necessary to make classification algorithm comparisons to get better accuracy values, such as naive Bayes and support vector machines (Fikri et al., 2020).

CONCLUSIONS AND SUGGESTIONS

Conclusion

It is concluded from the research that has been done comparing the results of the Naive Bayes algorithm and optimization using particle swarm optimization on three internet service providers in Indonesia. The results of the Naive Bayes classification accuracy are 77.94%, and after PSO optimization has increased by 3.68% to 81.62%. The results of the Naive Bayes classification accuracy are 91.39%, and after PSO optimization has risen by 1.49% to 92.88%. The first keyword is Biznet. The second keyword is the first media. The third keyword is Indihome, the results of the naive Bayes classification accuracy are 85.78%, and after PSO optimization is done, it increases by 1.70% to 87.48%. From the results of the three keywords above, it can be concluded that Naive Bayes is a reasonable classification algorithm, and optimization using Particle Swarm Optimization has an effect on increasing accuracy results.

Suggestion

This research only classifies three internet service providers in Indonesia. Further research is expected to classify all internet service providers in Indonesia and compare them with other classification algorithms to get better accuracy values.

REFERENCES

Andika, L. A., Azizah, P. A. N., & Respatiwulan, R. (2019). Analisis Sentimen Masyarakat terhadap Hasil Quick Count Pemilihan Presiden Indonesia 2019 pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier. *Indonesian Journal of Applied Statistics*, 2(1), 34–41. <https://doi.org/10.13057/ijas.v2i1.29998>

APJII. (2019). Survei. Diambil kembali dari Asosiasi Penyelenggara Jasa Internet Indonesia. In apjii.or.id. <https://apjii.or.id/survei2019x>

Aulianita, R., & Rifai, A. (2018). Optimasi Particle Swarm Optimization Pada Naive Bayes Untuk Sentiment Analysis Furniture. *Information Management for Educators and Professionals*, 3(1), 31–40.

Bustami, 2012, Teknik, D. I., & Bayes, N. (2013). "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah." *TECHSI-Jurnal Teknik Informatika*, 146(Klasifikasi), 128–146.

Christianto, H. (2020). Penggunaan Media Internet dalam Pemenuhan Hak atas Pendidikan di Masa Pandemi Covid-19: Perspektif Hak Asasi Manusia dan Hukum Pidana. *Jurnal Ham*, 11(2), 239–253. <http://repository.ubaya.ac.id/38038/>

ER, Y., & Solecha, K. (2021). *Implementasi Particle Swarm Optimization (PSO) pada Analysis Sentiment Review Aplikasi Trafi menggunakan Algoritma Naive Bayes (NB)*. 7(1), 25–29. <https://doi.org/10.31294/jtk.v4i2>

Fikri, M. I., Sabrina, T. S., & Azhar, Y. (2020). Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter. *Smatika Jurnal*, 10(02), 71–76. <https://doi.org/10.32664/smatika.v10i02.455>

Hanifah, A. (2020). *Pembangunan Business Intelligence Pada Toserba Koperasi Karyawan Semen Padang (KKSP) Berbasis Dashboard System* [Universitas Andalas]. <http://scholar.unand.ac.id/56876/>

Harjanta, A. T. J. (2015). Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining. *Informatika UPGRIS*, 1(1), 1–9. <http://journal.upgris.ac.id/index.php/JIU/article/view/804>

Haupt, R. L., & Haupt, S. E. (2004). *Practical Genetic Algorithms* (2nd ed.). John Wiley & Sons.

Hayuningtyas, R. Y., & Sari, R. (2019). Analisis Sentimen Opini Publik Bahasa Indonesia Terhadap Wisata TMII Menggunakan Naïve Bayes dan PSO. *Jurnal Techno Nusa Mandiri*, 16(1), 37–42. <https://doi.org/10.33480/techno.v16i1.115>

Ipmawati, J., Kusrini, & Taufiq Luthfi, E. (2017). Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen. *Indonesian Journal on Networking and Security*, 6(1), 28–36. <http://www.ijns.org/journal/index.php/ijns/article/view/1444>

Khairina, P. R., & Fitriati, D. (2021). Sentiment Analysis of Twitter Data on Remote Learning Using Naïve Bayes Algorithm. *Jurnal Riset Informatika*, 3(3), 203–210. <https://doi.org/10.34288/jri.v3i3.187>

Muttaqien, F. (2016). *Implementasi Text Mining Pada Aplikasi Pengawasan Penggunaan Internet Anak*. 53(9), 8–24.

Ramadhan, D. A., & Setiawan, E. B. (2019). Analisis Sentimen Program Acara di SCTV Pada Twitter Menggunakan Metode Naive Bayes Dan



Support Vector Machine. *Seminar Nasional Teknologi Fakultas Teknik Universitas Krisnadwipayana*, 6(2), 9736–9743. <https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/10708>

Rustiana, D., & Rahayu, N. (2017). Analisis Sentimen Pasar Otomotif Mobil: Tweet Twitter Menggunakan Naïve Bayes. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, 8(1), 113–120. <https://doi.org/10.24176/simet.v8i1.841>

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.

