Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024
DOI : https://doi.org/10.33395/jmp.v12i2.13416
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Terbit : 21 Januari 2024

Sentiment Analysis On Twitter Social Media Towards Minahasa Using Naive Bayes Algorithm

[1]Ayu Triana Situmorang, [2] Vivi Peggie Rantung
[1]Department of Informatics Engineering, Faculty of Engineering, Universitas Negeri Manado, Manado
[2]Academic Supervisor
[1]ayutriana2012@gmail.com, [2]vivrantung@unima.ac.id,

ABSTRACT

Sentiment analysis on social media is an effort to understand the opinions, attitudes, and emotions contained in content shared online. The study focused on sentiment analysis o f Minahasa, an area located in North Sulawesi using data obtained from the social media platform Twitter. Through the use of the Naïve Bayes algorithm, this approach aims to understand and analyze sentiment patterns associated with Minahasa. The Naïve Bayes method used in this study is a classification technique that bases its predictions on probabilities and assumptions about the independence of observed features. By utilizing text data from tweets related to Minahasa, the algorithm can classify the sentiment of each tweet into positive, negative, or neutral. This analysis allows for a deeper understanding of people's views reflected in online conversations. The data obtained from Twitter amounted to 915 tweet data which was then analyzed so as to get a variety of negative, positive and neutral sentiments.

Kata Kunci: Sentiment Analysis, Minahasa, Naïve Bayes

INTRODUCTION

Sentiment analysis is a process of analyzing text to determine the emotional tone of the author. Sentiment analysis involves the use of text analysis techniques to understand and categorize positive, negative and neutral emotions contained in writing [1].

The main purpose of sentiment analysis is to categorize the text contained in a sentence or document and then determine the opinion expressed in the sentence or document, whether it is positive, negative or neutral. Not only that, sentiment analysis can also express emotional feelings of joy, anger or sadness.

Twitter is an online networking service and microblog that allows users to send and read text-based messages to each other [2]. In 2023 Twitter has limited the number of tweets that can be read by users per day. Twitter owner Elon Musk announced that verified accounts can only read 10,000 tweets/day, while unverified accounts can read 1000 tweets, and new unverified accounts can only read 500 tweets per day. These restrictions are in place to stem the efforts of "extreme data harvesting" and data manipulation.

Minahasa Regency is one of the oldest regencies in North Sulawesi Province with its capital in Tondano. The word Minahasa itself comes from the words Minaesa, Mahasa, Minhasa which means to be one. Minahasa has eight sub- ethnic groups: Tountemboan, Tombulu, Tounsawang, Tounsea, Toulour, Pasa, Panosakan and Bantik. The most recognizable culture in Minahasa is Mapalus. Mapalus itself means the spirit of gorong royong [3].In this research, the author uses the naïve bayes algorithm to classify sentiments on the topic of Minahasa. The naïve bayes method is a data classification method based on the bayes theorem. This method is often used to predict the probability of certain classes based on existing statistical data [4].

Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13416 p-ISSN : 2089-9424

LITERATURE REVIEW

The following are some studies that are relevant to the topic:

1. Research conducted by Taufik et al., (2021) has the aim of measuring the accuracy value using Naïve bayes and support vector machine on NU and GNPH-Ulama influencers. Resulting in a better accuracy value with using support vector machine and AUC methods compared to Naïve Bayes [5].
2. Research conducted by Mahbub ah et al, (2019) focused on finding the accuracy level of the 2019 presidential election on twitter using the naïve bayes method. With a pretty good final result, where the classification of tweet data using the naïve bayes algorithm gives an accuracy of 73% [6].
3. Research conducted by Siti Rahayu, et al (2018) conducted sentiment analysis related to Shoppe online marketplace reviews with the naïve bayes method. The results obtained that Shopee online marketplace users produce an accuracy rate of 87% [7].
4. Research conducted by D. A. Muthia (2018) uses data taken from the restaurant review site www.zomato.com. From the results of data processing carried out, it is proven that the naïve bayes algorithm is superior to the support vector machine algorithm. Because the accuracy of the naïve bayes algorithm reaches 87% while the support vector machine algorithm is only 56% [8].
5. Research conducted by Eni Tri Handayani et al, (2021) conducted a search for analyzing public responses to the Covid-19 daily news using the naïve bayes method. The result of negative sentiment was 77%. The accuracy of the accuracy test results was 78% and the precision test was 92% [9].
6. Research conducted by Ahmadi, Gustian and Sembiring, (2021) has the aim of finding out the sentiments of the pros and cons of the Indonesian people about the Covid-19 vaccine with the naïve bayes method. Responses from the public are more inclined to the negative side, with 800 negative comments and 361 positive comments with an accuracy rate of 74% [10].
7. Research conducted by Ruhyana, (2019) focuses on analyzing the application of the odd/even number plate system. This research resulted in an accuracy value of 86.67%, precision of 71.43% and recall of 80.00% in positive and negative forms [11].

RESEARCH METHODS

The methods used in this research are as follows:

1. Research Data
   Twitter became the data source used to conduct sentiment analysis. Data sampling was conducted using the keyword "Minahasa" and obtained data of 915 tweets.
2. Data Crawling
   This process is a process to retrieve post data on twitter using the help of the API (Application Programming Interface) search on twitter [2]. In the context of retrieval For Minahasa-related data, we used specific keywords to search for relevant tweets.
3. Data Preprocessing
   The main problem of preprocessing sentiment analysis derived from twitter is that it is full of slang and abbreviations. Therefore, preprocessing is needed to remove irrelevant characters and reduce the quality of the model [12]. There are 5 stages that are performed:
   a. Data Cleaning
      This process removes characters that are less relevant to the topic that can reduce data quality, such as hastags, account names, and links.
   b. Tokenizing
      In this process, it breaks the text document into smaller sets of words. This stage also removes certain characters such as punctuation and filters based on text length.
   c. Normalization

Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13416 p-ISSN : 2089-9424

The normalization process aims to produce data or data tables that have an organized and structured structure, the result of this process is the formation of data sets in a form called normal form.

d. Stopword Removal
This stage removes text/words that are not related to sentiment analysis so that the text dimension will be reduced without reducing the sentiment content of the tweet.

e. Stemming Data.
This process removes suffixes and affixes on each token or word in the text. The text that has been processed becomes more standardized with words that have been systemized into their basic form.

4. Word Cloud
Word cloud is a visual image based on the frequency of how many occurrences of the same words in a collection of text, where the larger the letter size of a word, the more the number of occurrences of the word and vice versa, the smaller the letters that appear, the less the occurrence of the word.

5. Naïve Bayes Classification
Naïve Bayes is a classification method for calculating probabilities under the condition that the decision class is true. This algorithm assumes that object attributes are independent [13].
The equation of Bayes' theorem:

$$P(H|X)=(P \; X \; H \; P(H))/(P(x))$$

Where:
X : Data with unknown class
H : Data hypothesis is class specific
P(H |X): Probability of hypothesis H based on condition X
P(H) : Probability of hypothesis H
P(X|H) : Probability of X based on the condition in hypothesis H
P(X) : Probability of X
To explain the naïve bayes method, it is necessary to understand that the classification process requires clues to determine the appropriate class for the sample to be analyzed. Therefore, the naïve bayes method is customized as follows.

$$P(C)=Nc/N$$

Where:
P(C) : Probability of C
Nc : Total number of classes C
N : total number of classes

RESULTS AND DISCUSSION

1. Tweet Data Collection
The data crawling process using the Twitter API was carried out with the keyword "Minahasa". From the Twitter data crawling process, we obtained 915 tweets that have not been labeled with sentiment.

2. Data Preprocessing
After the data collection process, the data is processed by performing the text preprocessing stage. The following is the text preprocessing stage carried out by taking one tweet sample.

Tabel 3. 1 Satges of Text Preprocessing

Tahapan	Input	Output
Cleaning Data	@detiKawanua.com Danau Tondano di Sulawesi Utara akan menjadi tuan rumah gelaran Wakefest Minahasa 2023. Tempat ini dipercaya berpotensi tarik	Danau tondano di sulawesi utara akan menjadi tuan rumah gelaran wakefest minahasa

Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024
DOI : https://doi.org/10.33395/jmp.v12i2.13416
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Tahapan	Input	Output
	wisatawan. https://t.co/ouoWOykdWZ	tempat ini dipercaya berpotensi tarik wisatawan
Tokenizing	Danau tondano di sulawesi utara akan menjadi tuan rumah gelaran wakefest minahasa tempat ini dipercaya berpotensi tarik wisatawan	Danau, tondano, di, sulawesi, utara, akan, menjadi, tuan, rumah, gelaran, wakefest, minahasa, tempat, ini, dipercaya, berpotensi, tarik, wisatawan
Normalisasi	Danau, tondano, di, sulawesi, utara, akan, menjadi, tuan, rumah, gelaran, wakefest, minahasa, tempat, ini, dipercaya, berpotensi, tarik, wisatawan	Danau, tondano, di, sulawesi, utara, akan, menjadi, tuan, rumah, gelaran, wakefest, minahasa, tempat, ini, dipercaya, berpotensi, tarik, wisatawan
Remove Stopwords	Danau, tondano, di, sulawesi, utara, akan, menjadi, tuan, rumah, gelaran, wakefest, minahasa, tempat, ini, dipercaya, berpotensi, tarik, wisatawan	Danau, tondano, sulawesi', utara, tuan, rumah, gelaran, wakefest, minahasa, dipercaya, berpotensi, tarik, wisatawan
Stemming	Danau, tondano, sulawesi, utara, tuan, rumah, gelaran, wakefest, minahasa, dipercaya, berpotensi, tarik, wisatawan	Danau, tondano, sulawesi, utara, tuan, rumah, gelar, wakefest, minahasa, percaya, potensi, tarik, wisatawan

After the text preprocessing stage carried out on 915 tweet data, the net data results were obtained as much as 790 tweet data. The following is a view of the Word cloud:

Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13416 p-ISSN : 2089-9424

Figure 3. 1 Word Cloud display in Text Processing

Based on Figure 3.1, there are words that have large and small sizes. A word that has a large size means that the frequency of occurrence of the word is more in the tweet data, while a small word has a small frequency of occurrence in the tweet data used.

3. Data Labeling Stage

Data that has gone through the text preprocessing stage and grouped per word is then matched with the data dictionary to determine the polarity of the text whether it is positive, negative or neutral.

The data dictionary used is as follows:

Table 3. 2 Positive and Negative Data Dictionary

Positif		Negatif	
adaptif	dibaca	abnormal	beban
adil	dibebaskan	absurd	bejat
afinitas	diberikan	acak	bekas
afirmasi	dibersihkan	acak-acakan	bekas luka
agilely	elastis	acuh	bekas roda
berani	elite	adiktif	cengeng
berapi	emas	aib	cengking
berarti	empati	air terjun	cercaan
berbaik hati	enak	akurat	cerdik
berbakat	fleksibel	alarm	demam
berbesar hati	fleksibilitas	alasan	demoralisasi
cerdas	futuristik	alat permainan	dendam
cerdik	gagah	alergi	egois
cergas	gaib	alergik	egoisme

Table 3.2 is a data dictionary taken from the GitHub repository obtained from the link https://github.com/masdevid/ID-OpinionWords. After matching with the data dictionary, then the training set is carried out. From these results, the polarity set is obtained. A text/tweet is considered positive or negative if the word has an interest in the word dictionary, and is considered neutral if the tweet has no related words in the word dictionary.

	tweet	positive_counts	negative_counts	label
0	jajar personel polres minut laksana aman even ...	1	0	Positive
1	kapolsek belang iptu eliasr sasebohe anggota l...	0	0	Netral
2	mi teddy minahasa jual sabu blow up pimpin pol...	1	3	Negative
3	kapolres minahasa utara serta staf jajar selam...	1	1	Netral
4	mekanisme terbit sim polres minut kapolresminu...	0	0	Netral
5	target menang telak susun tim menang daerah ga...	2	0	Positive
6	personil polsek airmadidi edukasi pemuda papua...	0	0	Netral
7	personil polsek kema laksana patroli dialogis ...	0	0	Netral
8	personil polsek likupang laksana patroli malam...	0	0	Netral
9	personil polsek tan laksana patroli malam himb...	0	0	Netral

Figure 3. 2 Classification of Tweets according to sentiment class

After a training set of 790 tweets, the data is grouped into several sentiments, which can be seen in the following figure:

Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13416 p-ISSN : 2089-9424

```
In [10]: # Menghitung jumlah data dalam setiap kategori sentimen
         sentiment_counts = tweet_df['label'].value_counts()

         # Menampilkan total jumlah label sentimen
         print(sentiment_counts)

         Netral      357
         Positive    317
         Negative    116
         Name: label, dtype: int64
```

Figure 3. 3 Detail of Labeling Result

The following are the sentiment results visualized in a pie chart.

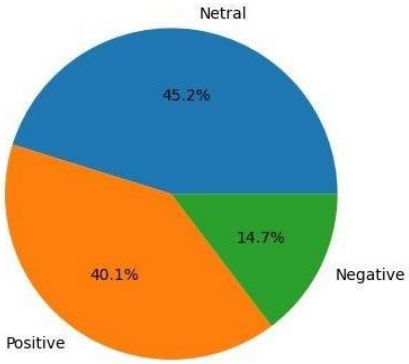

Figure 3. 4 Pie Chart of Labeling Results

From the graph above, it can be seen that tweets with positive sentiment are 40.1%, tweets with negative sentiment are 14.7%, while tweets with sentiment are 14.7% neutral as much as 45.2%. It can be interpreted that the twitter users' response to Minahasa was not too bad. Where the tweet data shows more neutral and positive sentiments compared to negative sentiments.

4. Accuracy Level in Naïve Bayes Classification

   Accuracy Level in Naïve Bayes Classification After labeling the tweet data, the naïve bayes classification stage uses the number of occurrences of words in the dataset of each class. The classification results will get the following probability:

   1. Precision is a measure that describes the extent to which positive predictions are correct out of all positive predictions made.
   2. Recall is a measure that describes the extent to which actual positive data comparisons of all positive predictions are made.
   3. F1-score is a measure that combines recall and precision to provide a more comprehensive picture of a system or model's performance in classification.
   4. Accuracy is a measure that describes the extent to which a classification system or model is correct.

   Then the classification is carried out using the naïve bayes algorithm to determine the probability of sentiment. The classification results can be seen in the following figure.

```
accuracy_score  :  0.6075949367088608
              precision    recall  f1-score   support

    Negative       0.41      0.64      0.50        22
      Netral       0.66      0.56      0.60        72
    Positive       0.67      0.66      0.66        64

    accuracy                           0.61       158
   macro avg       0.58      0.62      0.59       158
weighted avg       0.63      0.61      0.61       158
```

Figure 3. 5 Classification Results

Figure 3.5 shows that the results of the classification get 22 data with negative sentiment and 64 data with positive sentiment, and for precision gets 0.41 in negative sentiment and 0.64 in positive sentiment.

Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13416 p-ISSN : 2089-9424

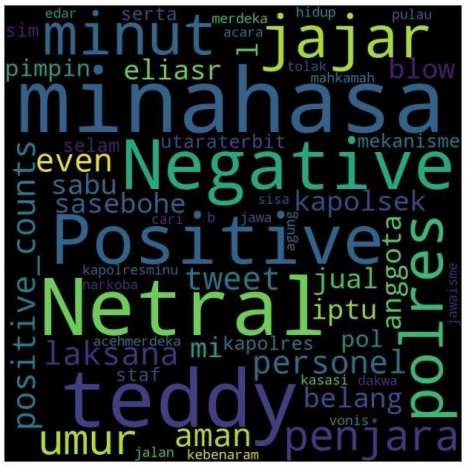

Figure 3. 6 Word Cloud display after classification

Based on the figure above, it can be seen that the word that has the highest frequency of occurrence is seen with the largest word size such as Neutral, Minahasa.

CONCLUSIONS

The analysis and test results show that the steps from data scraping to the implementation of the Naïve Bayes Classifier algorithm produce a positive conclusion on the reliability of the model in sentiment analysis of 790 processed tweet data. The classification of 790 data on Minahasa obtained a positive sentiment of 40.1%. From the accuracy results, it explains that the accuracy level tried with naive bayes modeling is quite good.

ACKNOWLEDGMENTS

I would like to thank my parents, who have supported and encouraged me to keep going against all odds. Also to my supervisor, Mrs. Vivi Piggie Rantung, and my partner, the Tourism and Culture Office of Minahasa Regency, for giving me the opportunity and providing the information needed for this research. I would also like to thank my parents who always support me every step of the way.

REFERENCES

Yamatoto, Y., & Kumatato, T. (2015). Multidimensional sentiment calculation method for Twitter based on emoticons. International Journal of Pervasive Computing and Communications, 11(2), 212–232.

A. Shiri, Introduction to Modern Information Retrieval (2nd edition), vol. 53, no. 9. In: Library Review, 2004. doi: 10.1108/00242530410565256.

Huang, Hidayat. 2021. Analisis Regresi Sederhana, Ini Penjelasannya. https://www.globalstatistik.com/analisis-regresi-sederhana-ini-penjelasannya/ Istianto, John

M. R. W. M. Baihaqi, M. Pinilih, Combination of K- Means and Support Vector Machine (SVM) to Predict Sara Elements in Tweets, vol. 7. 2020.

P. Nomleni, "Sentiment Analysis Using Support Vector Machine (Svm)," Semin. National. Technol. And Commun. 2015, 2015.

F. Taufiqurrahman, S. Al Faraby, and M. D. Purbolaksono, "Multi-label Text Classification in Indonesian Translated Hadith Using Chi Square and SVM," e-Proceeding Eng., vol. 8, no. 5.

P. Mahbubah, L. D., Zuliarso, E., Study, Twitter Sentiment Analysis in the 2019 Presidential Election Using. 2019.

Siti Rahayu, "Sentiment Analysis of Online Marketplace User Reviews Using Naïve Bayes Classifier," J. Inf. Interaktif, vol. Vol. 3, no. No. 3, 2018.

Jurnal Minfo Polgan
Volume 12, Nomor 2, Januari 2024 e-ISSN : 2797-3298
DOI : https://doi.org/10.33395/jmp.v12i2.13416 p-ISSN : 2089-9424

D. Ayu Muthia, "Integration of Genetic Algorithm and Information Gain for Feature Selection in Sentiment Analysis of Movie Reviews Using Naive Bayes Algorithm," J. Tech. Comput. AMIK BSI, vol. 4, no. No 1, pp. 186-193, 2019.

Eni Tri Handayani and Ari Sulistiyawati, "Sentiment Analysis of Public Response to Covid-19 Daily News on Twitter Ministry of Health," vol. 2, pp. 32-37, 2021.

F. Ahmadi, M. I., Gustian, D. and Sembiring, "Analysis of Public Sentiment towards the Covid-19 Case on Youtube Social Media with the Naive Bayes Method," J. Comput. Inform. (J-SAKTI), vol. 5, no. 2,
pp. 807-814, 2021.

N. Ruhyana, "Sentiment Analysis on the Implementation of the Odd/Even Number Plate System on Twitter with the Naive Bayes Classification Method," J. IKRA- ITH Inform., vol. Vol. 3, no. No 1, p. 98, 2019.

V. Kumar and A. Chadha, "Mining Association Rules in Student's Assessment Data," Int. J. Comput. Sci. Issues, vol. 9, no. 5, pp. 211-216, 2012.

Olson & Delen, "Advanced Data Mining Techniques," Springer-Verlag Berlin Heidelb., 2008.