Sentiment analysis of student responses related to information system services using Multinomial Naïve Bayes (Case study: Telkom University)

Ainun Fauziyah Bahary, Yuliant Sibaroni, and Mohamad Syahrul Mubarok
School of Computing, Telkom University, Jl Telekomunikasi 01, Dayeuhkolot, Bandung-West, Java 40257

ainunfauziyahbahary@gmail.com, yuliant@telkomuniversity.ac.id, msyahrulmubarok@telkomuniversity.ac.id

Abstract. Current technological developments affect the types of information services provided, especially information system services that always related to the internet network. To improve existing services, user responses are needed regarding services provided, but if the response is very large, it will be difficult to know whether the services provided are good or not. By taking the right approach, the sentiment's response can be analysed quickly and automatically. Sentiment analysis is done by classifying responses into positive and negative classes, the classification method is Multinomial Naïve Bayes (MNB). Before the data is known for the sentiments, the data is manually labeled, then the data goes through the preprocessing stage, cross validation, feature extraction and then sentiment classification with the MNB classifier. Classification process using MNB method on positive and negative responses, obtaining the highest negative response results in iGracias which is a detailed response of the information system sub-service by 44.27%. The MNB method using lemma with the conditions for preprocessing has a good average accuracy of 83.24% for information system sub-services and 79.24% for internet sub-services.

1. Introduction

Most people in the era of globalization in Indonesia now have and understand how to access the internet. This affects the forms of services that exist now days, one of them is an information system service that makes it easier to find information on companies or agencies. The use of information systems in a company, especially universities is very important, and to improve the performance of these services, user response is needed.

There are many ways to collect these responses, for example by filling out questionnaires, interviews and others. From these responses it can be seen whether the services provided are good or not. To draw conclusions on the existing responses, there are still many who do it manually by checking each response. But if the number of responses is very large, it will be difficult to know the percentage of negative responses and positive responses from the service. By utilizing current technological developments, these responses can be analysed and their sentiments known automatically by doing the right approach, namely Sentiment Analysis.

Sentiment analysis is a sentiment that refers to the type of natural language processing used to understand people's moods, opinions, and sentiments regarding a particular product or film or event [1].
The sentiment analysis process is carried out by manually labelling the existing data, then passing the preprocessing process and the data went to the next step called cross validation process, then the data goes through the feature extraction and classification process. Some classification methods that can be used are Support Vector Machine, Naïve Bayes, and K-nearest Neighbor [2]. Of the three methods, Naïve Bayes is a fast method, easy to apply with simple and effective structure [3]. However, there is a specific method of Naïve Bayes namely Multinomial Naïve Bayes (MNB) which is a better method, as stated in the research of Destuardi and Sumpeno which concludes that the use for Indonesian classification, Multinomial Naïve Bayes (MNB) is better method than Naïve Bayes [4].

So in this study, the author uses the Multinomial Naïve Bayes (MNB) as a method for classification of sentiments in Indonesian-language responses related to information system services and Bag of Words (BoW) as feature extraction. This study aims to build a system, measure the accuracy performance of the MNB classification method, by testing the effect of Lemmatization on the preprocessing process, and knowing the percentage of positive and negative sentiments in the response, especially the negative response to the detail of the information system sub-service.

2. Related studies

In this study the sentiment analysis process was carried out in response using the classification method with several references to books and previous research. In the book [5] says that sentiment analysis is also called opinion mining is a field of study that analysed the opinions, sentiments, evaluations, judgments, attitudes, and emotions of people towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Opinions or user responses can be categorized into several categories or classes, namely positive or negative. In this study, classification was carried out using the Multinomial Naïve Bayes (MNB) method. MNB is a specific method of the Naïve Bayes [6] method which is a well-known method. In previous studies, sentiment analysis was carried out using the Naïve Bayes method by Gusriani, Wardhani, and Zul on the services and products of an online store. The results of the test analysis showed accuracy stability after being tested with k-Fold Cross Validation and Confusion Matrix with 93.7% accuracy where the minimum support for Frequent Itemset Mining was 0.014 [10].

Another study using the Naïve Bayes method is the use of the NB method on credit approval data, diabetes and heart disease, obtaining an average accuracy of 82.06%, 82.66%, and 81.99% [3] Taheri and Mammadov using Naïve Bayes because it is fast, easy to apply with simple and effective structure. And Naïve Bayes is also one of the top 10 algorithms for data mining [2]. Another study using the Naïve Bayes classification was also carried out by the Graphic and Campus in 1000 film review data taken from the IMDB site which showed several conclusions of test results with 65.2% and 82.06% accuracy [9] giving the great result for NB Classification.

Then in a previous study using the Multinomial Naïve Bayes (MNB) method, MNB was used in 1326 data song lyrics data to classify songs based on song topics, namely love of friendship, family, religion, nationalism and negative content. From this study, Aziz and Mubarok were able to produce good classification performance by using bag of words as a feature extraction and get an f1-measure value of 88.91% and an insurance value of 96% [9]. Then for research using MNB the other is research to deal with spam e-mail. In this study, Multinomial Naïve Bayes is used to classify e-mails that are spam e-mail and not spam e-mail. The research resulted in an accuracy of 99.308% (without mutual information) [10].

3. Research methods

3.1. System design

The design of the system in this study has the purpose of comparing the accuracy of the effect of the use of Lemmatization for preprocessing on the Multinomial Naïve Bayes classification in finding or classifying sentiments in the response. The general description of the system made in this study is shown in Figure 1. The preprocessing process, cross validation, feature extraction, classification and retrieval accuracy were made using the Visual Studio Code application with the Python programming language.
3.2. Dataset
The dataset that used in this study is student response data on Information System Services at Telkom University, the response uses Indonesian language. The data to be used is 7500 response data obtained from filling out the questionnaire in iGracias. Of all student responses taken, many responses or complaints about the information system and the handling of the information system itself are referred to as information system sub-services, with the detailed sub-services, iGracias, TUNE, blogs, e-mail and helpdesk. Then responses or complaints about accessing the internet network on campus are referred to as internet services. From the two sub-services, the sentiment will be searched or classified.

The dataset is the initial data or raw data which is then manually labeled positive or negative in each response according to the sub-service, with the label 0 for negative and 1 for positive and if the sub-service is not specified, the sentiment obtained is positive. The following is an illustration of the data compilation that will be used in this study.

| Response                                           | Sub-services |
|----------------------------------------------------|--------------|
| Layanan system informasi sangat bagus (the information system services are really good) | 1 1          |
| Sudah bagus (already good)                         | 1 1          |
| Internet kurang cepat (internet is not fast enough)| 1 0          |
| Tingkatkan (increase it)                           | 0 0          |
| Lebih ramah lagi (more friendly)                   | 0 1          |
| Hdjsafkajkhjfja                                    | 1 1          |

3.3. Preprocessing
Preprocessing is the process of processing input data or initial stages before the main stage process is carried out. This stage is carried out to convert raw data into easy data for reading in the next process [1]. The preprocessing phase that will be carried out includes Case folding, Filtering, Lemmatization, Tokenizing.

The first stage is Case folding, which is the stage of uniformization of the form of letters to lowercase or uppercase, and the elimination of numbers and punctuation. Then the Filtering process, this process is carried out using the stopword removal algorithm, to delete irrelevant words, for example,
“yang” or that, “untuk” or for, “dengan” or with, and others. The third stage is Lemmatization, this process is to find basic words and eliminate word additions. Then the last preprocessing stage is Tokenizing, to solve a set of characters in a text into a word. The stopword and lemma process uses Sastrawi which was developed by Nazief and Adriani [11].

In the preprocessing process does not address the writing of the wrong words, words that have no meaning, and words or sentences in English so that the words will still be processed in the next stage, namely the feature extraction stage. The following is an illustration of the preprocessing process with the response sentence “Yang Layanan Sistem Informasi sudah bagus, lanjutkan.”

| Process        | Result                                                                 |
|----------------|------------------------------------------------------------------------|
| Case Folding   | yang layanan sistem informasi sudah bagus lanjutkan (the information system services are good, continue) |
| Filtering      | Layanan sistem informasi bagus lanjutkan (information system services are good continue) |
| Lemmatization  | Layan sistem informasi bagus lanjut (the information system services are good, continue) |
| Tokenizing     | ‘layan’, ‘sistem’, ‘informasi’, ‘bagus’, ‘lanjut’ (‘service’, ‘system’, ‘information’, ‘good’, ‘continue’) |

In this study compare the use of Lemmatization in response, which is the process by using Lemmatization, without Lemmatization, and Lemmatization with conditions. For lemma with conditions is a stage that still uses Lemmatization but it will not eliminate the word additions in the words ‘perbaiki’ or repair or fix, ‘percepat’ or make it fast, and ‘pelayanan’ or services, because the word fix and accelerate is a negative word and if through the lemma process will produce a new word that is ‘baik’ or good and ‘cepat’ or fast which is a positive word. Whereas the word services includes an information system sub-service about the helpdesk, and if through the lemma process, it will produce a new word that is ‘layanan’ or service that has the same results as “layanan” or service which is the word for both sub-services. The following is an illustration of the effect of Lemmatization on the response sentence “perbaiki lagi untuk layanan system informasi” or “For information system services, fix it again” in this study.

| Process                  | Result                                                                 |
|--------------------------|------------------------------------------------------------------------|
| With lemmatization       | ‘baik’, ‘layan’, ‘sistem’, ‘informasi’ (‘nice’, ‘service’, ‘system’, ‘information’) |
| Without lemmatization    | ‘perbaiki’, ‘layanan’, ‘sistem’, ‘informasi’ (‘repair’, ‘service’, ‘system’, ‘information’) |
| Lemmatization with condition | ‘perbaiki’, ‘layan’, ‘sistem’, ‘informasi’ (‘repair’, ‘service’, ‘system’, ‘information’) |

3.4. Feature extraction
Feature extraction features mining from preprocessing results of student responses related to information system services at Telkom University that are ready to be used in the next process. Feature extraction that will be used for Multinomial Naïve Bayes is Bag of Words (BoW). The BoW model is a simplified representation used in natural language processing and information retrieval (IR) [12]. The process from BoW produces a dictionary containing words contained in the responses that will be used in the classification process. The following is an illustration of Bag of Words (BoW) with two responses which is “layan sistem informasi bagus lanjut” and “baik layan sistem informasi”.

Table 2. Preprocessing process.

Table 3. The effect of using lemmatization.
Table 4. Result of BoW.

| Name of Result | Result |
|----------------|--------|
| Dictionary     | {“layan”: 5, “sistem”: 6, “informasi”: 3, “bagus”: 1, “lanjut”: 4, “baik”: 2} |
|                | {“service”: 5, “system”: 6, “information”: 3, “good”: 1, “continue”: 4, “nice”: 2} |
| Vector of Response 1 | [101111] |
| Vector of Response 2 | [011011] |

3.5. Learning
This process begins with the cross validation process using the Shuffle and Split then the Multinomial Naïve Bayes (MNB) for classification. Cross validation is a process for dividing data into two parts, namely training data and testing data. In this study the type of cross validation uses as many as 10 splits with a division of 9:1 for training and testing data. The shuffle and split process is scrambling existing data and then dividing the 10% data for testing, the rest is training data. The following is an illustration of the cross validation process by using the data in Table 1 before.

Table 5. Data after cross validation.

| Data  | Response | Information system sentiment | Internet sentiment |
|-------|----------|------------------------------|-------------------|
| Train | ‘layan’, ‘sistem’, ‘informasi’, ‘bagus’ (‘service’, ‘system’, ‘information’, ‘good’) | 1                | 1                |
|       | ‘tingkat’ (‘increase’)                          | 1                | 1                |
|       | ‘lebih’, ‘ramah’ (‘more’, ‘friendly’)           | 0                | 0                |
|       | ‘internet’, ‘kurang’, ‘cepat’ (‘internet’, ‘less’, ‘fast’) | 1                | 0                |
|       | ‘hdjsafkjajkhfja’                               | 0                | 1                |
| Test  | ‘sudah’, ‘bagus’ (‘already’, ‘good’)             | 1                | 1                |

Then the classification process is one of the techniques in data mining that is used to predict classes from existing data. Classification is done because there are a number of data that is known to its class and labeled as accurately as possible. Classification consists of two steps, namely the construction of models or training data and the use of models to data testing or classification. The classification used is the Multinomial Naïve Bayes (MNB).

This method is used because MNB is a specific method of Naïve Bayes. Naïve Bayes is a fast, easy to implement method with a simple and effective structure [2]. And in previous studies, it was stated that the Multinomial Naïve Bayes was better than the Naïve Bayes method for the classification of Indonesian texts [9]. The equation for class selection or classification calculation with the Multinomial Naïve Bayes method can be defined in equation (3) by searching the results of Prior in equation (1), Conditional Probability in equation (2) and the illustration of the use of equations in table 7 using responses in table 6.

\[
P(c) = \frac{N_c}{N} \quad (1)
\]

\[
P(w|c) = \frac{\text{count}(w,c) + 1}{\text{count}(c) + |V|} \quad (2)
\]

\[
P(c|d) \propto P(c) P(w|c) \quad (3)
\]
Declaration:

\[
\begin{align*}
 w & = \text{Word} \\
 c & = \text{Class (positive or negative)} \\
 N & = \text{Sum of response} \\
 d & = \text{Response} \\
 \text{Count}(w,c) & = \text{Sum of words } w \text{ in class } c \\
 \text{Count}(c) & = \text{Sum of words in class } c \\
 |V| & = \text{Sum of vocabulary} \\
 P(c|d) & = \text{Posterior Probability, the opportunity of class } c \text{ for response } d \\
 P(w|c) & = \text{Conditional Probability (likelihood), the opportunity arises the word } w \text{ in class } c \\
 P(c) & = \text{Class Prior Probability, the opportunity arises the class } c \\
 \propto & = \text{Proportional}
\end{align*}
\]

| Information | Response | Sentiment (Class) |
|-------------|----------|-------------------|
| Data training | Layanan sudah bagus (services already good) | 1 |
|  | Bagus (good) | 1 |
|  | Jelek (bad) | 0 |
| Data testing | Sudah bagus (already good) | TBD |

| Function | Calculation | Positive | Negative |
|----------|-------------|----------|----------|
| P(c) | = 2/3 = 0,66 | = 1/3 = 0,33 |
| P(‘layan’|c) | = (1+1) / (3+4) = 2/7 = 0,29 | = (0+1) / (1+4) = 1/5 = 0,2 |
| P(‘sudah’|c) | = (1+1) / (3+4) = 2/7 = 0,29 | = (0+1) / (1+4) = 1/5 = 0,2 |
| P(‘bagus’|c) | = (2+1) / (3+4) = 3/7 = 0,43 | = (0+1) / (1+4) = 1/5 = 0,2 |
| P(‘jelek’|c) | = (0+1) / (3+4) = 1/7 = 0,14 | = (1+1) / (1+4) = 2/5 = 0,4 |
| P(c|’sudah’,’bagus’) | = 0,66 * 0,29 * 0,43 = 0,08 | = 0,33 * 0,2 * 0,2 = 0,01 |

From the results of calculations in table 7, the highest posterior results for the response to testing data “sudah bagus” or already good in table 6 is 0,08 which is obtained from the results of prior multiplication in the positive class is 0,66 and the likelihood of the word good and in the positive class of 0,29 and 0,43. And it can be concluded that “already good” is a positive response.

3.6. System evaluation

The performance evaluation process is a process carried out to measure the level of performance of a process. The level of performance is obtained from the calculation of the percentage of accuracy in each sub-service for the classification method used, following the percentage accuracy calculation formula.

\[
\text{Accuracy (\%)} = \frac{\text{Sum of correct answers}}{\text{Sum of answers}} \times 100 (4)
\]

4. Evaluation

4.1. Analysis the performance of Multimedia Naïve Bayes (MNB) and the influence of Lemmatization

System performance is measured by calculating the accuracy of the MNB method performance in classifying sentiments on responses. Before performing the classification process using MNB, testing the effect of Lemmatization on the preprocessing process was carried out. Here are the results of the
accuracy percentage of Multinomial Naïve Bayes (MNB) classification method by testing the influence of lemma on the preprocessing process.

Table 8. The results of testing the use of lemmatization (lemma).

| Iteration | Information system | Internet |
|-----------|--------------------|----------|
|           | Lemma Without lemma | Lemma With condition | Lemma Without lemma | Lemma With condition |
| 1         | 84.93% 77.73%       | 84.67% 79.87%      | 72.67% 79.87%       |
| 2         | 82.00% 77.73%       | 82.27% 79.47%      | 70.80% 79.47%       |
| 3         | 81.73% 75.47%       | 82.40% 77.73%      | 69.33% 77.47%       |
| 4         | 83.47% 77.60%       | 83.33% 78.67%      | 70.67% 78.67%       |
| 5         | 84.00% 78.13%       | 84.67% 80.13%      | 72.13% 79.87%       |
| 6         | 83.33% 78.40%       | 84.40% 79.20%      | 71.07% 78.93%       |
| 7         | 81.20% 78.13%       | 81.87% 77.73%      | 73.60% 78.00%       |
| 8         | 81.33% 75.87%       | 82.00% 79.73%      | 72.80% 79.87%       |
| 9         | 83.07% 79.33%       | 84.00% 79.20%      | 72.93% 80.53%       |
| 10        | 82.40% 78.27%       | 82.80% 79.73%      | 72.00% 79.73%       |
| Average   | 83.23% 77.67%       | 83.24% 79.24%      | 71.80% 79.24%       |
| Highest   | 84.93% 79.33%       | 84.67% 80.67%      | 73.60% 80.53%       |

From the results of the table above, we can see the difference between the highest accuracy of using lemma and lemma with conditions very small. But in Figure 2 which shows the graph comparison, it is very visible the accuracy obtained in the information system sub-service that uses lemma with more frequent conditions is above the accuracy of the information system sub-service that uses lemma without conditions with an average accuracy of 83.24% for information system sub-services and 79.24% for internet sub-services.

![Figure 2](image)

Figure 2. Comparison graph of increased accuracy on the effect of using lemma and lemma with conditions.

4.2. Detailed sentiment analysis on information system sub-services

The process of detailed sentiment analysis in the information system sub-service aims to find out the sentiments in the detailed responses that are iGacias, tune, blogs, e-mail, and helpdesk. The analysis is carried out based on the percentage of the number of negative sentiment outputs on the system so that the service provider knows the shortcomings of the services provided. The process of detailed response analysis was not carried out for internet sub-services, because internet sub-services only contained responses about accessing the internet in the campus area. The percentage of detailed negative sentiments in the information system sub-service described in Figure 3.
Based on the negative responses, most students think that iGracias needs to be improved because the server is often down, to improve the iGracias application on the play store, and the information is more updated and not only on iGracias, but on other Telkom University social media as well. Then for the helpdesk, students hope that their services will be improved, faster to respond the complaints and questions that students have, and employees are more friendly to students. And for responses about tune, many students think that tune is difficult to access, does not reach all campus areas, and hopes that tune can be used on some devices, not just one device.

5. Conclusion

In this study, the tests carried out on the influence of the lemmatization process for preprocessing which then uses the Multinomial Naïve Bayes (MNB) to classify sentiments in the existing responses. From the results of the classification carried out, the system displays the percentage output of positive responses, negative responses for each sub-service and negative responses to the detail of the information system sub-service with a negative percentage average. Which can be seen in the attachment of the percentage of positive responses, negative for each service and the percentage of responses to the details of the information system sub-service. The percentage of the three biggest negative responses sequentially is the response about iGracias by 44.27%, helpdesk by 28.38%, and tune by 26.48% which can be seen in Figure 3. So that we can know what needs to be improved and improved on Information System Services at Telkom University.

The results of the sentiment classification were obtained from the tests performed, with an average accuracy using lemmatization with conditions for the preprocessing process, namely 83.24% for information system sub-services and 79.24% for internet sub-services. For the automatic implementation of sentiment classification of information system service responses at Telkom University with user input, using the Multinomial Naïve Bayes (MNB) aspect taken from the classification results which has the best percentage of accuracy, namely the aspect of the 5th iteration for information system sub-services with accuracy of 84.67% and aspects of the 7th iteration for internet services with an accuracy of 80.95%. The MNB aspects used are priors and likelihood to calculate the posterior to find out the sentiment class of the given response.

In this study has several limitations and shortcomings, then for further research it is suggested that it can overcome the problem of words or responses that are incorrect, and have no meaning in order to obtain higher accuracy.
6. References

[1] M Govindarajan and Romina M 2013 *A Survey of Classification Methods and Applications for Sentiment Analysis* Int. J. Eng. Sci. 2319–1813

[2] X. Wu, V Kumar, J R Quinlan, J Ghosh, Q Yang, H Motoda, G J McLachlan, A Ng, B Liu, P S Yu, Z Zhou, M Steinbach, D J Hand and D Steinberg 2008 *Top 10 algorithms in data mining* vol. 14 no. 1

[3] S. Taheri and M. Mammadov 2013 *Learning the naive Bayes classifier with optimization models* Int. J. Appl. Math. Comput. Sci. vol. 23 no. 4 787–795

[4] I. Destuardi and S. Sumpeno 2009 *Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes* Semin. Nas. Pascasarj. Inst. Teknol. Sepuluh Nop.

[5] B Liu 2014 *Sentiment Analysis and Opinion Mining* vol. 70 no. 8

[6] Y Sibaroni 2008 *Analisis dan Penerapan Metode Klasifikasi untuk Pembangunan Perangkat Lunak Sistem Penerimaan Mahasiswa Baru Jalur Non-Tulis*

[7] S Gusriani, K D K Wardhani and M I Zul 2016 *Analisis Sentimen Terhadap Toko Online di Sosial Media Menggunakan Analisis Sentimen Terhadap Toko Online di Sosial Media Menggunakan Metode Klasifikasi Naive Bayes (Studi Kasus: Facebook Page BerryBenka)*

[8] N M S Hadna, P I Santosa and W W Winarno 2016 *Studi Literatur Tentang Perbandingan Metode untuk Proses Analisis Sentimen di Twitter* 57–64

[9] R A Aziz, M S Mubarok and Adiwijaya 2016 *Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes* no. August 139–148

[10] C P Susanto and E I Irawati 2015 *Algoritma Support Vector Machine Untuk Mendeteksi Sms Spam Berbahasa Indonesia* 109–116

[11] M. Adriani, J. Asian, B. Nazief, S M M Tahaghoghi and H. E. Williams 2007 *Stemming Indonesian: A Confi x-Stripping Approach* ACM Transactions on Asian Language Information Processing (TALIP) vol. 6, no. 4 1–33

[12] Soumya G K and S Joseph 2014 *Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature* vol. 16 no. 1 34–38

Acknowledgments

Thanks to Allah SWT. for giving the opportunity to get to this stage. Thank you for family and friends who have helped and encouraged the process of making this paper.