Classification of Public Complaint Data in SMS Complaint Using Naive Bayes Multinomial Method

Lucan Yance Nanlohy
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
lucan.18071@mhs.its.ac.id

Eko Mulyanto Yuniarno
Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekomulyanto@ee.its.ac.id

Supeno Mardi Susiki Nugroho
Department. of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mardi@its.ac.id

***Abstract*—SMS Complaint is an electronic public complaint tool for reporting issues on government performance. Text mining classification utilized to determine the value of each complaint category. The SMS data in this study sourced from the SMS Complaint Service of Ambon City Government. There were 6 categories of classification, namely Public Service, Infrastructure, Bureaucracy, Health, Education, and Social. The classification performed to measure levels of accuracy of the Stemming process and non-Stemming process represented in Matrix with values of recall, precision, and f1 score. The methods used in the measurement were Naıve Bayes Multinomial. With the naıve Bayes method, an accuracy level with stemming of 91.38% obtained, and while the accuracy level without stemming was 90.73%. The result showed that the naıve Bayes method could be used effectively to predict complaint data through stemming.**

Keywords—Naive Bayes, Text Mining, SMS Complaint, Classification, Multinomial

I. Introduction

Along with the rapid advancement of technology, information has undoubtedly become a primary need since its distribution is the key to stay up-to-date with current events. One communication media to elicit information is SMS Complaint Service, which the government uses to bridge the communication between communities and government agencies. SMS complaint service is a medium that accommodates public aspirations and is used to evaluate government performance electronically. Through this service, public complaints have been steadily increasing, thereby calling for quick, transparent, and accountable handling. Public complaints are received in the form of text to facilitate information search for ongoing issues. Text Mining is a periodic information search process where users are faced with a set of documents of data mining analysis tools [1].

Based on problems received through SMS Complaint, there is a need to analyze public complaint data collected through SMS complaint by classifying complaints into several categories. Classification is a method for identifying models that distinguish data concepts and classes [2]. Research conducted Amelia Rahman about online new classification using Naive Bayes to categorize Indonesian news automatically using multinomial using TF-IDF weighted test results average accuracy of 86.62% while Multinomial DF-thresholding reached 86.28%[3]. High level of accuracy with simple calculations. The Dataset used in this study was 113 reporting. Data is classified into 3 groups namely Information, Kamtibnas, and Criminal acts. Research conducted to produce high accuracy average, namely recall 93%, precision 90%, and f-measure 92% [4]. A comparison of some types of multinomial naïve Bayes in the document classifies. The datasets used in this study were 20 newsgroups, industry sectors, Web KB, and Reuters-21578. These datasets are frequently used datasets in the study of text classification. The research proves that modifications to the transformed weight normalized complement naïve Bayes (TWNBC) are not required to obtain optimal results for some datasets. However, the use of TFIDF in Word weighted is proven to significantly increase the accuracy of most datasets. the use of document length normalization can reduce the accuracy of weighing with TF-IDF [5]. Therefore, this study uses the Naive Bayes Classifier method to predict SMS complaint data in specific categories and measurements based on stemming and without stemming and uses feature extraction to calculate the value weight of the frequency of occurrence of words. The Naïve Bayes method proves to provide quite satisfactory results when used for text classification [6]. One of the models of Naïve Bayes that is often used in the classification of text is multinomial Naïve Bayes. [7].

II. Material and Method

In this study, the algorithm and method used were naıve Bayes with 4 stages, namely: SMS Complaint dataset, Feature Extraction, Preprocessing, and Classification. The following process flowchart explained in Fig. 1 and an explanation of the data token in Ambon City complaints SMS service database.

A. Data Preparation

Illustrates the process flow that will be carried out with the first stage data sourced from the Ambon City Government's SMS complaint database. The dataset will be divided into six class categories namely Public Services, Bureaucracy, Infrastructure, Education, Health, and Social Affairs.

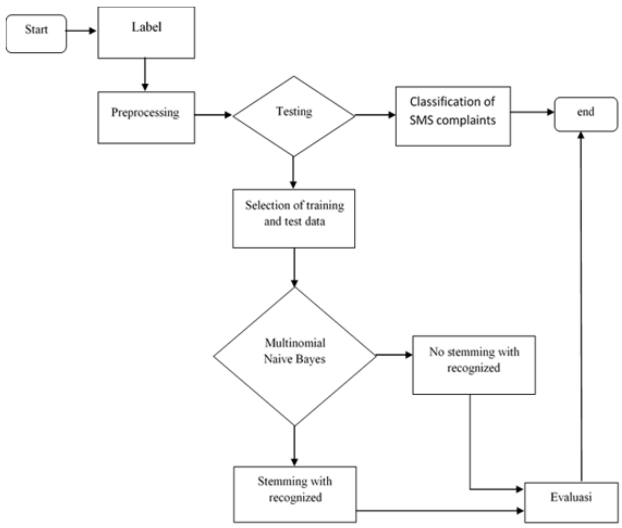

Fig. 1. System Flowchart Diagram

978-1-7281-8131-8/20/$31.00 ©2020 IEEE

The number of datasets used was 846 out of 1038 complaints. The first test with 189 training data will be added up to reach 846 then will be tested through the process of stemming and without stemming with vocabulary taken randomly from each document text test data. Classification of public complaints using the multinomial naive Bayes method consists of several stages with an explanation that can be seen in the Fig. 1.

B. *Features Extraction NLP*

1) *Number of Word*

The first stage extracted the number of words in each complaint to determine, in general, words that contained positive and negative sentences in each complaint.

2) *Number of Characters.*

In stage second stage, the process carried out to count the numbers of characters in each complaint.

3) *Average word.*

This stage then calculates the average word length of each complaint by dividing the number of words by the total length of the complaint. this process also carried out to help in obtaining information that will later be used to improve modeling during classification.

4) *Numbers of word*

At this stage, the stop removal process is applying for words that have meaning that is not important.

C. *Preprocessing*

They are preprocessing performed to transform unstructured textual data into a structured model. A structured model was needed so that the data could be processed and analyzed. Preprocessing, at every stage, could be performed manually; however, due to the sheer amount of data, an automated method would be beneficial. The following is the flow of the preprocessing system. The processing stage consists of case folding, stopword removal, tokenization, and stemming.

- Case Folding, letters in a document into one form such as changing capital letters into lowercase.
- Stop Removal, It was a process where important terms from previous stages taken then remove unimportant terms.
- Word Removal Frequency, In NLP (Natural Language Processing), stop words are neglected words during the processing. The words would later be stored in a stop list. The main criteria of stop word. selection were frequency of occurrence, which was usually for conjunctions such as “and,” “or,” “but,” “will,” and others. There were no exact rules in determining stop words to be used.
- Spilling Correct, It was a process of filtering data generated by previous stages by correcting words to be used in the next process.
- Tokenization, This process performed to break down text documents into a separate set of words or, in other words, the process of butchering words from sentences.
- Stemming, Stemming aimed to transform affixed words into basic words by removing prefixes and suffixes in every word.

D. *Term Frequency-Inverse document frequency*

Weight calculation with term frequency-inverse document frequency (TF-IDF) uses the combination of two values, term frequency and inverse frequency document obtained by dividing the number of overall documents with the number of documents where the word frequently occurred. The following is the formula to determine the weight of TF-IDF [8].

$$Tf\,Idf\,(W) = tf\;x\;log\;\frac{n}{df(w)} \qquad (1)$$

Where :

$tf\;idf$	: Frequency of t word from the key in d document.
dft	: The number of documents containing t-word from a word
$W\,d.t$	: d document weight on keywords
D	: The number of all documents in the d to t key.

E. *Naive Bayes Multinomial*

Naive Bayes Classifier is a machine learning method that uses the concept of probability [9]. In the naive Bayes method, text classification can be represented with term groups where each term in a document is assumed to be independent of each other [6]. The Naive Bayes method proves to provide quite satisfactory results when used for text classification.[7]. One of the models of Naïve Bayes that is often used in the classification of text is multinomial Naïve Bayes. The following is the equation:

$$P(c\mid d) \propto P(c) \prod_{k=1}^{n} P(t_k \| c) \qquad (2)$$

where :

$P(c\|d)$	: Probabilities document D is in Class C
$P(c)$	: Prior probability a document to be in Class C
$\{t_1, t_1, t_1, \ldots, t_n\}$	:Tokens in document d which are part of the vocabulary with the amount of n.
$P(t_k\|c)$	: tk conditional probability is in a document in Class c

Classification aims to determine the best class in a document. In determining the best class naive Bayes look for the maximum value of a posteriori (MAP) class C with the equation.

$$c_{map} = argmax\;\hat{P}(c) \prod_{k=1}^{n} \hat{P}(t_k|c) \qquad (3)$$

P is written with $\hat{P}$ because the true value of $P(c|d)$ dan $P(t_k|c)$ is not yet known and will be counted during the training process [7].

In Equation 3 there is a probability value that can be multiplied. Therefore the summation process is on the logarithm of probability. The class with the algorithm of the highest probability is a class with probability in a document ; $\log(xy) = \log(x) + \log(y)$. Equation 3 that uses the algorithm of probability can be expressed in equation 4

$$c_{map} = \underset{c \in C}{\arg\max} \left[\log \hat{P}(c) + \sum_{1 \leq k \leq n} \log \hat{P}(t_k|c) \right] \quad (4)$$

$\hat{P}$ and $\hat{P}(t_k|c)$ Can be generated by calculating the maximum likelihood which is the relative frequency of the parameters. Prior is expressed in equation (5).

$$\hat{P}(c) = \frac{N_c}{N} \quad (5)$$

where :

$\hat{P}(c)$: Prior probability a document is in Class C

N_c : Number of Class C documents

N : Number of whole documents

$\hat{P}(t|c)$ is the probability of the relative term T in the document to be in Class C, expressed by equation (6).

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \quad (6)$$

where :

$\hat{P}(t|c)$: The conditional term T probability resides in a document in Class C

T_{ct} : Number of term T occurrences in a document with category C

$\sum_{t' \in V} T_{ct'}$: The total frequency of all term in class C

In the calculation of maximum likelihood, each vocabulary in the document will be added 1 to avoid the value 0 on each category. Equations can be expressed as follows:

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'})+N} \quad (7)$$

where N : Total number of vocabulary terms

III. Experiment and Analysis

The data used in this study are SMS complaints, derived from the Ambon City Government database with a total of 1038 SMS complaints consisting of 6 categories: Public Services, Infrastructure, Bureaucracy, Social Affairs, Education, Health.

A. Testing through stemming and without stemming

Testing at this stage to get the result of the test data in theform of confusion matrix through stemming and without stemming. Table 1 explains that the prediction results in each class category with training data of 189 with vocabulary taken randomly as many as 1747 stemming and without stemming 2034 vocabulary.

TABLE I. Table Categories From Training Data 189

No.	Catagories	Number of Values
1	Public Service	67
2	Bureaucracy	34
3	Infrastructure	48
4	Education	14
5	Health	20
6	Social	4

TABLE II. Confusion Matriks Table Test Data 189

	recall	precision	f1_score	accuracy
Stemming	0,94	0,97	0.96	92,11%
Without Stemming	0,89	1,0	0,94	89,47%

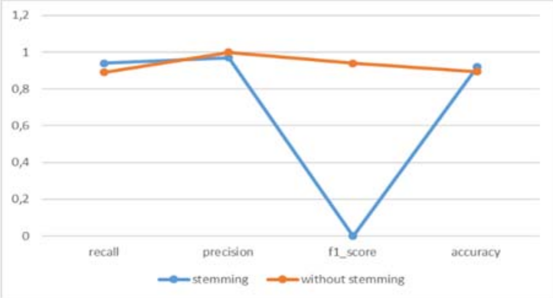

Fig. 2. Matrix Confusion Graph

Table 2 explains that the test result using data 189 with vocabulary chosen randomly from the extraction of the overall feature of data as much as 1747/2034 with accuracy stemming 92,11% and without stemming 89,47%.

Fig. 2 showed the results of confusion matrix testing, where the accuracy value and recall, precision, and f1 score showed a similar value of 92,11%.

Table 3 explains that the prediction results in each class category with training data of 472 with vocabulary randomly selected as many as 3083 vocabularies.

Table 4 explains that the test result using data 472 with vocabulary taken randomly from the extraction of the overall feature of data as much as 3083/3756 with the result of accuracy stemming 92% and without stemming 89,47%.

Fig. 3 showed the test result of 472 data with the highest and the lowest recall value of 0.91 and 0.90, respectively. The highest precision was 0.97, and the lowest was 0.96. The highest accuracy was 89.47%.

TABLE III. Table Categories From Training Data 472

No.	Catagories	Number of Values
1	Public Service	157
2	Bureaucracy	102
3	Infrastructure	115
4	Education	22
5	Health	59
6	Social	12

TABLE IV. Confusion Matriks Table Test Data 472

	Recall	precision	f1_score	accuracy
Stemming	0,91	0,97	0.94	89,47%
without Stemming	0,90	0,96	0,93	87,37%

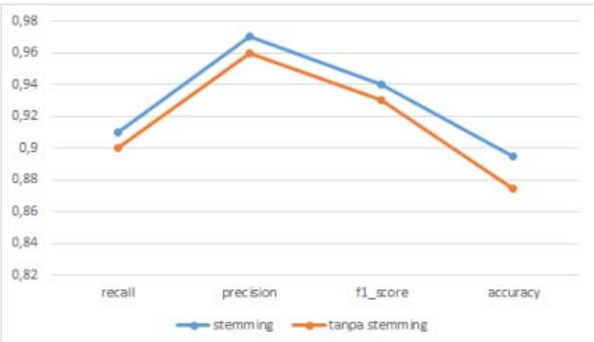

Fig. 3. Matrix Confusion Graph

TABLE V. TABLE CATEGORIES FROM TRAINING DATA 660

No.	Catagories	Number of Values
1	Public Service	213
2	Bureaucracy	148
3	Infrastructure	145
4	Education	40
5	Health	88
6	Social	16

TABLE VI. CONFUSION MATRIKS TABLE TEST DATA 660

	recall	precision	f1_score	accuracy
Stemming	0,91	0,97	0.94	89,47%
Without Stemming	0,90	0,96	0,93	87,37%

Table 5 explains that the prediction results in each class category with training data of 660 with vocabulary taken randomly as many as 3685 vocabularies.

Table 6 explains that the test result using data 660 with vocabulary taken randomly from the extraction of the overall feature of data as much as 3685/4547 with accuracy stemming 90,91% and without stemming 87,88%.

Fig. 4 showed the test result of 660 data with the highest and the lowest recall value of (0.91) and (0.90), respectively. The highest precision was (0.97), and the lowest was (0.96). The highest accuracy was 90.91%.

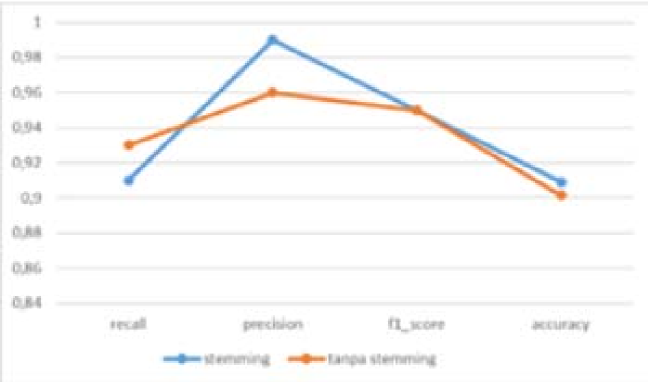

Fig. 4. Matrix Confusion Graph

Table Categories From Training Data 754

No.	Catagories	Number of Values
1	Public Service	243
2	Bureaucracy	180
3	Infrastructure	166
4	Education	44
5	Health	95
6	Social	16

TABLE VII. CONFUSION MATRIKS TABLE TEST DATA 754

	recall	Precision	f1_score	accuracy
Stemming	0,93	0,97	0.95	91,39%
Without Stemming	0,93	0,96	0,95	90,73%

Table 7 explains that the prediction results in each class category with training data of 754 with vocabulary taken randomly as many as 4013/4943 without stemming vocabularies.

Table 8 explains that the test result using data 754 with vocabulary taken randomly from the extraction of the overall feature of data as much as 4013/4943 with the result of accuracy stemming 91,39% and without stemming 90,73%.

Fig. 5 showed the test result of 754 data with the highest and the lowest recall value of (0.91) and (0.91), respectively. The highest precision was (0.97), and the lowest was (0.96). The highest accuracy was 91.39%.

Table 9 explains that the prediction results in each class category with training data of 846 with vocabulary taken randomly as many as 4023/5013 without stemming vocabularies.

Fig. 5. Matrix Confusion Graph

TABLE VIII. TABLE CATEGORIES FROM TRAINING DATA 846

No.	Catagories	Number of Values
1	Public Service	261
2	Bureaucracy	200
3	Infrastructure	184
4	Education	62
5	Health	107
6	Social	16

TABLE IX. CONFUSION MATRIKS TABLE TEST DATA 846

	recall	precision	f1_score	accuracy
Stemming	0,93	0,97	0.95	88,82%
without Stemming	0,93	0,96	0,94	88,24%

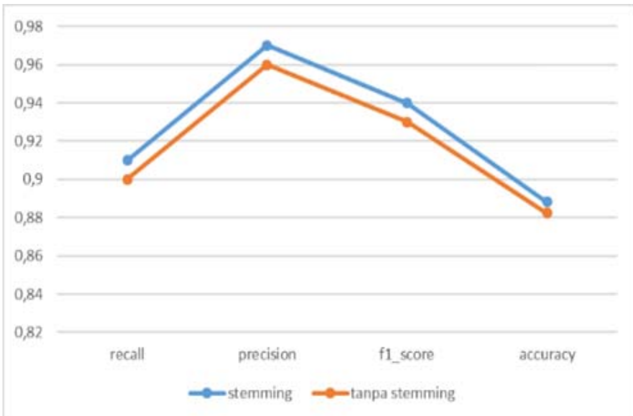

Fig. 6. Matrix Confusion Graph

TABLE X. TABLE OF TERM FREQUNCY-INVERSE DOCUMENT FREQUENCY

Categories	Recall	Precision	F1_score
Public Service	1,0	0,4	0,57
Bureaucracy	1,0	0,4	0,57
Infrastructure	0	0	0
Education	1,0	0,4	0,57
Health	0,54	0,7	0,61
Social	0,54	0,7	0,61

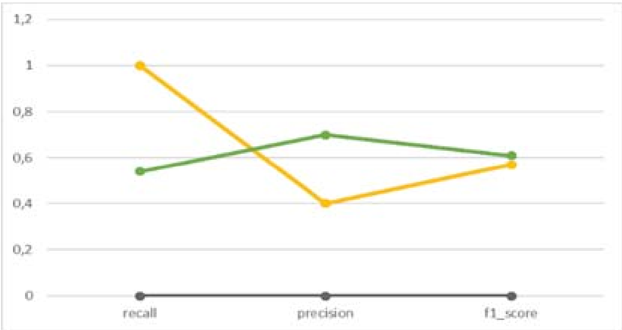

Fig. 7. The term frequency graph - Inverse document frequency

TABLE XI. TABLE OF TERM FREQUNCY-INVERSE DOCUMENT FREQUENCY

Categories	Recall	Precision	F1_score
Public Service	0,93	0,50	0,65
Bureaucracy	0,33	1,0	0,50
Infrastructure	0,93	0,50	0,65
Education	0,93	0,50	0,65
Health	0,64	0,70	0,67
Social	0,64	0,70	0,67

Table 10 explains that the test result using data 846 with vocabulary taken randomly from the extraction of the overall feature of data as much as 4023/5013 with the result of accuracy stemming 88,82% and without stemming 88,24%

Fig. 6 showed the test result of 846 data with the highest and the lowest recall value of (0.90) and (0.93), respectively. The highest precision was (0.93), and the lowest was (0.96). The highest accuracy was 88.82%.

Table 11 the test results using training data 189 and 383, in which output is obtained as a ratio value. The value of the ratio of showed good results in the category of public service bureaucracy and Education. precision results show a ratio of 0.4 to 0.7.

Fig. 7. illustrates the value of the recall ratio, showing good results in the category of public service and bureaucracy, while the lowest value is in infrastructure. The highest value of the rate of precision is in the social and health categories.

Table 12 the test results using training data 472 and 825, which output obtained as a ratio value. the value of the ratio showed good results in the category of public service, Infrastructure, and dan education with the value of 0,93.

Fig 8. illustrates the value of the recall ratio, showing good results in the category of public service and bureaucracy, while the lowest value is in Infrastructure. The highest value of the ratio of precision is in the bureaucracy and health categories.

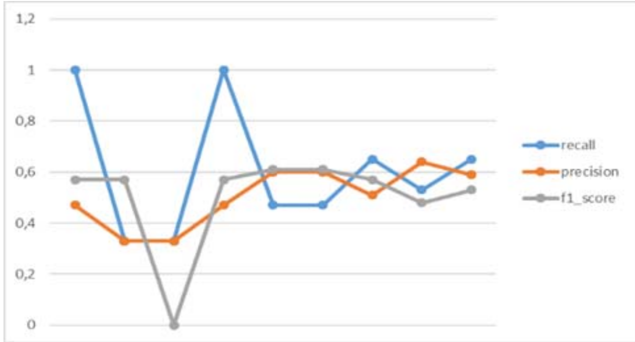

Fig. 8. The term frequency graph - Inverse document frequency

TABLE XII. TABLE OF TERM FREQUNCY-INVERSE DOCUMENT FREQUENCY

Categories	Recall	Precision	F1_score
Public Service	0,92	0,50	0,65
Bureaucracy	0,92	1,0	0,50
Infrastructure	0,06	0,50	0,65
Education	0,92	0,50	0,65
Health	0,92	0,70	0,67
Social	0,58	0,70	0,67

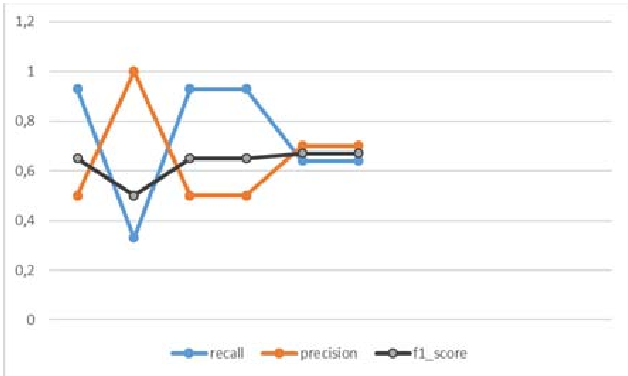

Fig. 9. The term frequency graph - Inverse document frequency

Table 13 the test results using training data 660 and 1076, which output obtained as a ratio value. The value of the ratio showed good results in the category of public service, bureaucracy, education, and health, with a value of 0,92.

Fig. 9. illustrates the value of the recall ratio, showing good results in the category of public service, bureaucracy, education, and health, while the lowest value is in Infrastructure. The highest value of the ratio of precision is in the Infrastructure.

IV. DISCUSSION

This study shows that the Naive Bayes Multinomial method is good at classifying and predicting text documents, as evidenced by the confusion matrix's accuracy and results. The frequency test in each category shows a recall result of (189) 0.54% on health and social. 1.0% value is in public services, bureaucracy, and education, test with 472 data results of 0.92% indicated in the four lowest value categories are in infrastructure. Data (660): The highest results are in 3 categories (0.93%), and bureaucracy is at the lowest value. While the precision results show values of 0.4% to 0.7% (189) in each category, in data (472), the value of precision is 0.44% public services, bureaucracy, education, health, social, and infrastructure by 1.0%. data (660) a precision value of

0.50% consisting of the categories of public services, infrastructure, education, bureaucracy by (1.0%), and health, social by 0.70%.

V. CONCLUSION

Based on the study results, it can conclude that Electronic-based SMS complaint service is a two-way communication that can be used by two parties to communicate. The Naıve Bayes Multinomial method used proves to be capable of classifying, as evidenced by the accuracy of training data that goes up to 91.39% from 80% with training data that continues to increase. Stop word extraction feature process is helpful for the process stemming. The stemmer performed with the Indonesian language proves to be good, as seen from the accuracy obtained.

REFERENCES

[1] R. Feldman and J. Sanger, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007.

[2] Y. Permadi, "Text categorization using N-Gram for Indonesian language documents," 2008.

[3] A. Rahman, W. Wiranto, and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART J. Teknol. dan Inf.*, vol. 6, no. 1, pp.32–38,2017,doi: 10.20961/ITSMART.V6I1.11310.

[4] F. Handayani and S. Pribadi, "Implementation of the Naive Bayes Classifier algorithm in the classification of automatic text complaints and reporting community through Call Center Service 110," vol. 7, no. 1, 2015.

[5] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited," pp. 488–489, 2004.

[6] T. Jo, *Concepts, Implementation, and Big Data Challenge*. Springer International Publishing AG, 2019.

[7] C. D and P. R. H. S, *Manning, Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.

[8] Y. A. Sari and E. Y. Puspaningrum, "Semantic News document Search uses Essential Dimension of Latent Semantic Indexing by using Document Frequency and Information Gain Thresholding feature reduction," *Semin. Nas. Teknol. Inf. dan Multimed.*, no. July, pp. 27–32, 2015.

[9] D. Susandi and U. Sholahudin, "Utilization of Vector Space Model on the application of the algorithm Nazief Adriani, KNN and Similarity Cosine function for the use of IDF and WIDF in the prototype of the Indonesian text classification system," *J. ProTekInfo*, vol. 3, no. 1, pp. 22–29, 2016.