Telematics Work Field Review Text Classification Using the Naïve Bayes Method

Aries Maesya
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
aries.maesya@binus.ac.id

Yaya Heryadi
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
yayaheryadi@binus.edu

Yulyani Arifin
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
yarifin@binus.edu

Lukas
Cognitive Engineering Reseach Group (CERG) Faculty of Engineering, Universitas Katolik Indonesia Atma Jaya, Indonesia
lukas@atmajaya.ac.id

Wayan Suparta
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
drwaynesparta@gmail.com

Abstract— **Indonesia will benefit from a demographic boom in 2030 with a higher labor supply than in earlier decades. Then in industrial revolution 4.0 robotics and artificial intelligence will take the place of low-skilled or menial employment that don't require specialized expertise (AI). To aim research is Telematics Work Field Review Text Classification Using the Naïve Bayes Method. The method using Multinomial Naïve Bayes model which is trained to learn from patterns in training data set without being programmed explicitly. Then, based on the Term Frequency - Inverse Document Frequency, consider the weighting of the word used (TF-IDF). The text classification stage is then carried out using the multinominal nave bayes classification method with evaluation using the confusion matrix, following the acquisition of the TF-IDF value. In the study it took data with web crawling techniques on social media sites twitter. The data collected was 936 data consisting of 7,8% negative sentiments, 26,4% positive sentiments, and 65,8% neutrals. The results of accuracy testing using the Confusion Matrix. And from the results of such tests resulted in an accuracy of 66%, precision 73%, and recall 85%.**

Keywords—**Telematics, Naïve Bayes, Confusion Matrix, TF-IDF**

I. INTRODUCTION

Indonesia will benefit from an increase in its population, which is a phenomena known as the Industrial Revolution 4.0. Demographic benefit in that there will be a significantly greater supply of labor than in past years in Indonesia. Then came the argument that the Fourth Industrial Revolution will automate many jobs, especially those requiring little or no training and menial labor, and that these jobs would be replaced by machines and artificial intelligence (AI)[1]. The work environment is the entire tooling tool and material faced by the surrounding environment where a person works, his work methods and work arrangements both as an individual and as a group. The term "The New Hybrid Technology" also refers to the combination of computing and communication principles known as telematics, sometimes known as ICT (Information and Communication Technology). ICT has been proliferated in many disciplines, for example in economic[2], business[3] and agriculture[4]. It will also be a significant challenge for the domestic workforce to increase industrial expertise, certification, and competency. Technology 4.0. There are many news-on-news sites or public responses on other social media to employment in industry 4.0. The response contained positive, negative, and neutral sentiment towards employment, especially in the field of Indonesian Telematics. This classification text can be used to obtain an overview of people's perceptions of the growth and empowerment of Indonesian telematics jobs both tending to positive, negative, and neutral sentiments.

The previous research explained that all sentiment analysis activities used the Naïve bayes classification, getting a classification accuracy value of 77% for the dataset before the IPO took place and 76% for the dataset taken after the IPO. Although the accuracy is not very high, it is good enough to produce accurate prediction values based on the dataset used. Based on the predictions made, the negative sentiment on the Bukalapak application decreased after the IPO and the increase in positive sentiment[5]. This research aims to build a document classification system of Naïve Bayes. Naïve Bayes was chosen because it can obtain high accuracy only with lower training data. The results obtained showed that the performance of Naïve Bayes. In this research, data mining can be done using text mining and machine learning methods[6]. One of the text mining algorithms is the extraction of TF-IDF features with the Naïve Bayes method and the programming language used is python. The first thing that must be emphasized in conducting this research is by using TF-IDF weighting by first doing the preprocessing stage, namely case folding, tokenizing, stopwords, and stemming. Then the classification stage uses the Naive Bayes Classifier method and for data retrieval using web crawling.

II. MATERIALS AND METHODS

This study uses the Knowledge Discovery and Data Mining method which is a structured analysis process to obtain correct, new, useful information and find patterns from large and complex data. Data mining (DM) is at the core of the KDD process, namely by using certain algorithms to explore

data, build models and find unknown patterns[7]. The research method comprises of several stages including data selection, pre-processing, data transformation, data mining, and data evaluation.

A. Data Input

The data source for this study was taken using a crawling method based on the keyword "Telematics". The data in the dataset is also often not all used, therefore only the data corresponding for analysis will be taken from the dataset[8]. The data is retrieved using python programming language with crawling technique. Data crawling is a stage in research that aims to collect or download data from a dataset.

B. Data Preprocessing

Preprocessing step in this study aims to prepare the text into data for further processing[10]. The stages in preprocessing are: (i) Case Folding to convert uppercase letters to lowercase[11], (ii) Tokenizing to convert sentence sets into single words[12], (iii) Filtering for the disposal of words not used in data grouping[13], and (iv) Stemming for changes in words that are experiencing growth to base words[14].

C. Data Transformation

Data transformation in this study aims to transform or combine data into an appropriate format for processing in data mining. Some data mining methods require special data formats before they can be implemented[15]. In this study, we used the weighting of the word TF-IDF as a stage of transforming data into vector forms. Weighting the feature by combining term frequency and inverse document frequency (TF-IDF) will result in a composite weight for each word term in each document[16].

D. Naïve Bayes Modeling

The modeling step in this study aims to address classification problem using Multinomial Naïve Bayes model which is trained to learn from patterns in training data set without being programmed explicitly. In this stage, the model is trained to learn interesting patterns that can be predicted from large-scale data models. The model training algorithm assumes that features are taken from multinomial distributions. Having been trained, the Naive Bayes model predicts future events based on previous experiences based on Bayes' Theorem[17].

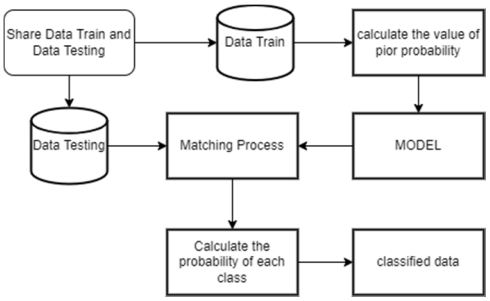


Figure 1: Naïve Bayes Modelling

The Naïve Bayes is a probabilistic model whose algorithm consists of several stages as follows :

1. Counts the numberof classes/labels.
2. Counts the same number of cases of the same class.
3. Multiplies all class variables.
4. Compare the results of all variavles.

In general, the calculation formula of naïve bayes can be formulated using the following formula:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Where:
X = Data with unknown classes.
Y = The X data hypothesis which is a specific class of a particular class.
P(H|X) = Hypothesis probability H based on a specific condition X.
P(H) = Hypothesis probability H (prior probability).
P(X|H) = Probability X based on hypothesis conditions H.
P(X) = Probability X.

E. Model Performance Validation and Evaluation

This process aims to measure or calculate the value of whether the system that has been created is in accordance with expectations or not[18]. In this study, an scoring system was held based on the accuracy of predictions of positive sentiment or negative sentiment using a confusion matrix, so that result of the scoring can get an accuracy value. Accuracy is the total validity of the model that calculated as the sum of true classifications divided by the total number of classifications. The formula to obtain accuracy is described below:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{(TP+TN+FP+FN)}} \quad (2)$$

III. RESULT AND DICUSSION

Based on the data that has been obtained, then the analysis is carried out as a problem solving. The analysis was carried out using the crawling on social media Twitter. When crawling data, the tool used is tweepy, tweepy is a library of the python programming language that is useful for pulling data through Twitter based on the desired keywords[19]. Obtained the results of review data of 936 data, with a negative sentiment of 73, a sentiment positive of 247, and a neutral sentiment of 616. the results of the comparison between polarity can be seen in the following figure.

Table 1: Data Polarity

Tweet Polarity	Quantity
Positive	247
Negative	73
Neutral	616

After the analysis process, it can be seen that the comparison between neutral sentiment is more than positive

and negative sentiment. From this figure explains that neutral sentiment is more than positive and negative sentiment.

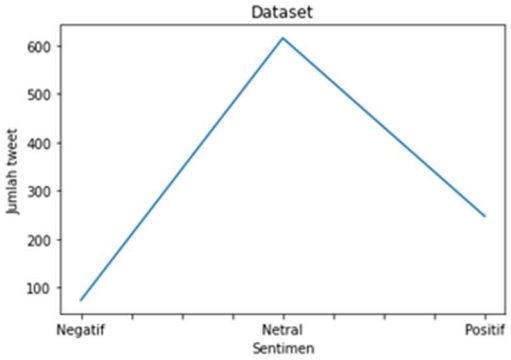


Figure 2: Comparison between class

Before classifying data on documents, the stage that needs to be done first is the data preprocessing process. At the data preprocessing stage, raw tweet data is first carried out the process of case folding, tokenizing, filtering and stemming. This research proposes several steps in text pre-processing, such as :

1. Removing URLs task handles the URLs in tweet
2. Removing special characters of Twitter such as #hastags, @username, and RT (retweet).
3. Removing symbols or numbers (e.g.!, #, $, *, 1234, etc.)
4. Normalize lengthening words, for example the word will be normalized which means spirit.
5. Tokenization separates a stream of text into parts called tokens.
6. The public figure name which appears in the tweet will be omitted in this step.
7. Case folding transforms words into similar form (lowercase or uppercase).
8. Change slang words into standard word based on dictionary.
9. Stemming is used to reduce the affixes and suffixes in the word.
10. Removing stopwords task removes the stopword

Stemming is a stage for the process of solving the variants of a word into basic words. Stem (root) is the part of the root that remains after the suffix is removed which consists of prefixes (prefix), Suffixes (insert), and confixes (prefix and suffix combination). Then an example of the stemming process can be seen in the following table.

Table 2: Result of Stemming

Before	After
Training	Train
Hoping	Hope
Conducting	Conduct
Disturbing	Disturb

The next process of data that has passed the preprocessing process enters into the data transformation process using TF-IDF with the aim of converting words into a valuable vector. At this stage, it is the processing stage of giving the weight of a word term to the weight calculation document which is carried out with two concepts, namely Term Frequency (TF) is the frequency of occurrence of the word (t) in the sentence (d). Document frequency (DF) is the number of sentences in which a word (t) occurs. the following is an example of the calculation of the tf-idf weighting shown in this Table 3.

Table 3: Examples of review weighting cases

Document	The term that represents the document		Category
	ENGLISH	INDONESIA	
Q	how footwear industry	gimana industri alas kaki	?
D1	how absorb labor	gimana serap tenaga kerja	Negative
D2	Oversupply labor.	Tenaga kerja oversupply	Positive
D3	Quite disturb	Cukup ganggu	Negative
D4	nice satisfied	Bagus puas	Positive

After going through the weight calculation stage using the tf-idf algorithm, the test data and training data will enter the vector length calculation stage between the training data document and the test data document where the level of similarity between the training data and test data will be calculated as shown in this Table 4.

Table 4: Calculation of TF-IDF Weight

Term		TF					DF	D\DF	IDF	Weight				
IND	ENG	Q1	D1	D2	D3	D4				Q	D1	D2	D3	D4
gimana	how	1	1				4	4	0.60206	0.602026	0.60206			
industry	Industry	1					1	4	0.60206	0.602026				
Alas	Footwear	1					1	4	0.60206	0.602026				
Kaki		1					1	4	0.60206	0.602026				
Serap	Absorb		1				1	4	0.60206		0.602026			
tenaga	labor		1				1	4	0.60206		0.602026			
kerja			1				1	4	0.60206		0.602026			
Advertising	Advertise			1			1	4	0.60206			0.60206		
oversupply	Oversupply			1			1	4	0.60206			0.60206		
cukup	Quite				1		1	4	0.60206				0.60206	
ganggu	Disturb				1		1	4	0.60206				0.60206	
bagus	nice					1	1	4	0.60206				0.60206	0.60206
puas	Satisfied					1	1	1	0.60206					0.60206
Max Value										**0.01191**	**0.03285**	**0.03285**	**0.13139**	**0.03285**

After getting the word weighting on the TF-IDF process. The next step is with the implementation of the model using

the Multinominal Naïve Bayes method. At this classification stage, naïve Bayes is divided into two processes, namely the training process and the testing process. Where the training process will be carried out first for training, then the testing process will be carried out by referring to the probability of the results of the training dataset.

1. *Data Training*

At this stage, the data that has been obtained weight values are then used as training data to be used as a reference for the formation of a classification model. At this stage, a prior probability calculation process and laplace smoothing calculation will be carried out. Calculating the Pior Probabilities P(c) of each class is to calculate the chances of a document appearing in a certain category, with a formula like the following:

$$P(c) = \frac{Nc}{N} \tag{3}$$

2. *Data Testing*

The stage after the training process. The flow in the testing process is almost the same as the flow of the training process, which distinguishes at the end of the testing process, the calculation of the final probability value will be carried out. At this stage, testing is carried out with test data into a model that has been formed at the training stage. the results of the preprocessing data in table 3 are then carried out the matching process, namely the process of finding the same term in the training data. the results of the example can be seen in his below.

$$\begin{aligned} &P(Pos|D1) = 1*0.03285= 0.03285 \\ &P(Net|D1) = 0.5*0.03285 = 0.16425 \\ &P(Neg|D1) = 0*0.\ 03285 = 0 \end{aligned} \tag{4}$$

From the results of the probability calculation, it is obtained that the probability of a positive class against D1 has the highest value of 0.16425, so it can be said that D1 belongs to the Neutral class (the test result is correct) because the most class in the dataset is Neutral Class.

Table 5: Classification Result

1	2	3	4
D1	D2	D3	D4
0, 16425	0, 16425	0, 21823	0
Neutral	Neutral	Positive	
Classification Result			
Neutral			

In validation and evaluation, a pattern evaluation is carried out which aims to evaluate whether the model is good or not, namely by looking at how the results are validated using a confusion matrix. the results of the Confusion matrix can be seen in the following figure.

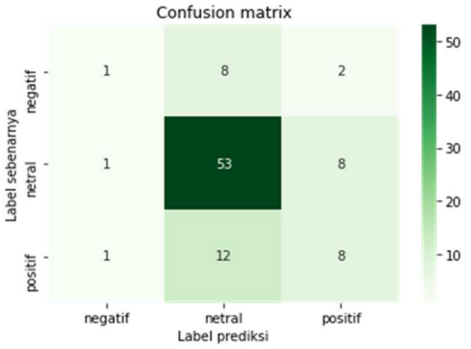
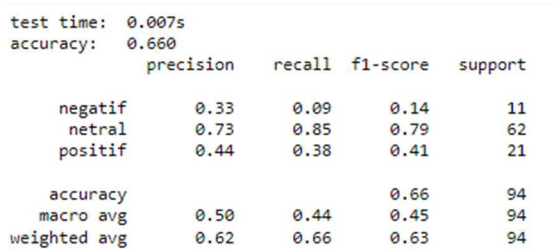


Figure 3: Confusion Matrix Result

In research using naive bayes, it did not get such an accuracy result of about 66%, because the data used had unstable in determining the negative positive neutral class. this result can be seen in the following figure.



test time: 0.007s
accuracy: 0.660

	precision	recall	f1-score	support
negatif	0.33	0.09	0.14	11
netral	0.73	0.85	0.79	62
positif	0.44	0.38	0.41	21
accuracy			0.66	94
macro avg	0.50	0.44	0.45	94
weighted avg	0.62	0.66	0.63	94

Figure 4: Accuracy Naive Bayes

From the test results using the confusion matrix using training data 70:30 test data from a total of 936 data obtained on average with a precision value of 73%, recall 85%, and accuracy of 66% of the scale of 0% - 100%, so it can be seen that precision is lower than the recall value, the effectiveness of the sentiment analysis system for Telematics work reviews using Naive bayes is said to be effective, this system is said to be effective if the precision and recall value above 50% are said to be ineffective if the value is below 50%.

Next process as a result of Wordcloud visualization, this Wordcloud process displays words based on the frequency of occurrence of the word, the larger the word displayed, the more frequent the appearance of the word[20]. The results of this wordcloud can be seen in the following fig. 5.

Figure 5: Wordcloud Result

IV. CONCLUSION

In the study entitled Telematics Work Review Classification Text Using Naïve Bayes Method, it took data with web crawling techniques on social media sites twitter. The data collected was 936 data consisting of 7,8% negative sentiments, 26,4% positive sentiments, and 65,8% neutrals. Then the data is entered in the preprocessing process and this stage aims to eliminate unused words and clean the data from noise. After the data is clean the next process is by weighting words using TF-IDF where the words are converted into vectors so that the data can be processed during data modeling using Naïve Bayes. Then, after the word weighting process, an implementation process or data classification model is carried out using Multinominal Naïve Bayes and for the results of accuracy testing using the Confusion Matrix. And from the results of such tests resulted in an accuracy of 66%, precision 73%, and recall 85%. the effectiveness of the sentiment analysis system for Telematics work reviews using Naive bayes is said to be effective, this system is said to be effective if the precision and recall value above 50% are said to be ineffective if the value is below 50%.

ACKNOLEDGEMENT

The research described in this paper was corporated with Laboratory of the Computer Science Binus University and Laboratory of The Computer Science Pakuan University.

REFERENCES

[1] Huang, Ming-Hui, and Roland T. Rust,. 2018. "Artificial intelligence in service." Journal of Service Research 21.2: 155-172.

[2] Coase, Ronald H. "Economics and contiguous disciplines. 2019. " The organization and retrieval of economic knowledge. Routledge, 481-495.

[3] Haseeb, Muhammad, et al. 2019. "Industry 4.0: A solution towards technology challenges of sustainable business performance." Social Sciences 8.5 : 154.

[4] Zhai, Zhaoyu, et al. 2020. "Decision support systems for agriculture 4.0: Survey and challenges." Computers and Electronics in Agriculture 170 : 105256.

[5] Somantri, O., 2017. "Text mining untuk klasifikasi kategori cerita pendek menggunakan naïve bayes (nb)." Jurnal Telematika, 12(1), pp.7-12.

[6] Yanuargi, B., 2022. "Analisis Sentemen Terhadap Aplikasi Bukalapak Sebelum IPO dan Sesudah IPO Menggunakan Algoritma Naive Bayes." JNANALOKA, pp.17-25.

[7] Zanuardi, A. and Suprayitno, H., 2018. "Analisa karakteristik kecelakaan lalu lintas di jalan ahmad yani surabaya melalui pendekatan knowledge discovery in database." Jurnal Manajemen Aset Infrastruktur & Fasilitas, 2(1).

[8] Viloria, Amelec, et al. 2018. "Methodology for the design of a student pattern recognition tool to facilitate the teaching-learning process through knowledge data discovery (big data)." International conference on data mining and big data. Springer, Cham.

[9] Ramadhan, Dery Anjas, and Erwin Budi Setiawan. 2019. "Analisis Sentimen Program Acara di SCTV pada Twitter Menggunakan Metode Naive Bayes dan Support Vector Machine." eProceedings of Engineering 6.2.

[10] Nata, G. N. M. and Yudiastra, P. P., 2017. "Preprocessing Text Mining pada email box berbahasa Indonesia." E-Proceedings KNS&I STIKOM Bali, pp.479-483.

[11] Filcha, A. and Hayaty, M., 2019. "Implementasi Algoritma Rabin-Karp untuk Pendeteksi Plagiarisme pada Dokumen Tugas Mahasiswa." JUITA: Jurnal Informatika, 7(1), pp.25-32.

[12] Fikri, Mujaddid Izzul, Trifebi Shina Sabrila, and Yufis Azhar. 2020. "Perbandingan Metode Naïve Bayes dan Support Vector Machine pada Analisis Sentimen Twitter." Smatika Jurnal 10.02 : 71-76.

[13] Wibisono, A. D., Rizkiono, S. D. and Wantoro, A., 2020. "Filtering Spam Email Menggunakan Metode Naive Bayes." Telefortech: Journal Of Telematics And Information Technology, 1(1), pp.9-17.

[14] Simarangjir, M.S.H., 2017. "Studi Perbandingan Algoritma-Algoritma Stemming Untuk Dokumen Teks Bahasa Indonesia." Jurnal Inkofar, 1(1).

[15] Y. D. Pramudita, S. S. Putro, and N. Makhmud, 2018. "Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer," J. Teknol. Inf. dan Ilmu Komput., vol. 5, no. 3, p. 269, doi: 10.25126/jtiik.201853810.

[16] Gunawan, B., Sastypratiwi, H. and Pratama, E. E., 2018. "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes." JEPIN (Jurnal Edukasi dan Penelitian Informatika), 4(2), pp.113-118

[17] Abbas, Muhammad, et al. 2019. "Multinomial Naive Bayes classification model for sentiment analysis." IJCSNS Int. J. Comput. Sci. Netw. Secur 19.3 : 62.

[18] Mustafa, M. Syukri, Muh Rizky Ramadhan, and Angelina P. Thenata., 2018. "Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier." Creative Information Technology Journal 4.2: 151-162..

[19] Razaq, E. R. M., Jacob, D. W. & Hamami, F., 2021. "Analisis Sentimen Kepuasan Mahasiswa Terhadap Pembelajaran Online Selama Pandemi Covid-19 Pada Media Sosial Twitter Menggunakan Perbandingan Algoritma Klasifikasi." eProceedings of Engineering, 8(5).

[20] Putri, W. T. H., & Hendrowati, R. 2018. "Penggalian Teks Dengan Model Bag of Words Terhadap Data Twitter." Jurnal Muara Sains, Teknologi, Kedokteran dan Ilmu Kesehatan, 2(1), 129-138.

[21] Pamungkas, F. S., & Kharisudin, I. 2021. "Analisis Sentimen dengan SVM, NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter." In PRISMA, Prosiding Seminar Nasional Matematika (Vol. 4, pp. 628-63).