Available online at http://iorajournal.org/indx.php/orics

Operations Research: International Conference Series

e-ISSN: 2722-0974
p-ISSN: 2723-1739

Vol. 3, No. 2, pp. 41-51, 2022

Analysis of Employment Sentiment in the Indonesian Telematics Field Use Multinomial Naive Bayes and Vector Space Model

Tomi Herdiawan[1], Eneng Tita Tosida[2], Aries Maesya[3*]

[1,2,3] *Computer Science Study Program, Faculty of Mathematics and Natural Sciences, Pakuan Bogor University*

**Corresponding author email: a.maesya@unpak.ac.id*

Abstract

Indonesia in 2030 experienced a demographic bonus in the sense that Indonesia would have far more labor supply than in previous years. Then there is a discourse that this 4.0 industrial revolution will replace a lot of work, especially low-skilled work or does not require special skills and rough jobs replaced by machinery and artificial intelligent (AI). To obtain the value of the percentage of positive, negative and neutral sentiments from the public regarding the impact of the industrial revolution 4 against labor and employment on online news media sites and social media Twitter, the authors conducted a study "analysis of employment sentiment in Indonesian telematics using multinomial naïve bayes. " The author uses the preprocessing stages including the case folding, tokenizing, stopword, and stemming. Then weighting with Term Frequency - Invers Document Frequency (TF-IDF). After that the classification stage was done using the multinomial Naive Bayes Classifier method and compare it with the Vector Space model classification. The evaluation used is the Confusion Matrix evaluation method. This study produced an evaluation value in the multinomial method of Naïve Bayes for news data to produce an accuracy of 81.75%, average precision 82.77%, and the average recall of 78.15%. Whereas with the Vector Space model method for news data produces an accuracy of 67.88%, average precision 65.59%, and the average recall of 70.56%. On Twitter data with the Multinomial Naïve Bayes method resulted in an accuracy of 88.80%, average precision 93.75%, and the average recall of 74.44%. On Twitter data with the Vector Space Model method resulted in 85.60% accuracy, average precision 76.44% and average recall of 86.07%.

Keywords: Mathematics, instructions for authors, manuscript template

1. Introduction

The phenomenon of the appearance of the Industrial Revolution 4.0 coincided with the demographic bonus that will be experienced by Indonesia in 2030 (Warsito, 2019). The demographic bonus in the sense that Indonesia will have a far more labor supply than in previous years. Then there is a discourse that this 4.0 industrial revolution will replace a lot of work, especially low-skilled work or does not require special skills and rough jobs replaced by machinery and artificial intelligent (AI). The question is whether Indonesia is ready to face the 4.0 industrial revolution, and any positive and negative impact that is brought through the 4.0 industrial revolution.

Industrial Revolution 4.0 is the convergence of information technology into the world of industry. Through the Internet of Things (IoT) and Big Data, technology can collect previous data to then predict what to do a technology to work efficiently, which we usually call Machine Learning.

The effects of the Industrial Revolution 4.0 are not replacing workers but replacing their jobs, so need new skills, for example the database administrator position will be replaced by the data analyst and it requires new skills. So, what must be done is the government and the company prepares a curriculum and educational structure that is able to overcome the change in skills and demand from the industry. The government must be serious in its efforts to overcome changes, if not companies will revoke their investment and leave Indonesia (Cuaresma et al., 2014).

Telematics or commonly called ICT (Information and Communication Technology) is a combination of computing and communication concepts also known as "The New Hybrid Technology". It will also be the main challenge for domestic employment in increasing competency, certification, and mastery of industrial technology 4.0. Many news on website sites or community responses on social media regarding employment in the industrial era 4.0. this. The response - the response contains positive sentiment, negative sentiment, and neutral social sentiment to employment, especially in the field of Telematics of Indonesia. Sentiment analysis can be used in obtaining a general description of community perceptions in the growth and empowerment of Indonesian telematics employment whether it tends to be positive, negative, or neutral sentiment. The focus of research sentiment analysis is to analyze the opinion of a document in the form of text.

Some of the previous studies related to the research conducted by Naz et.al (2018) concerning Twitter user sentiment analysis using the Support Vector Machine method. Porter algorithm is used in the process of stemming for feature extraction and term frequency method for weighting. The software is built using the PHP programming language for the server side that runs on the Windows Azure and Java cloud platform for the client side that runs on the Android platform. From the results of the study with 1,400 tweets on the dataset and 200 test data obtained accuracy of 79.5%.

Research conducted by Chen and Fu (2018) concerning the implementation of multinomial Naïve Bayes Classifier to classify letters out so that it can determine the mail number automatically. The classification system is supported by Confix-Stripping Stemmer to determine the basic word and TF-IDF for the word weighting. Testing is measured using Confusion Matrix. From the test results show the system has the level of Accuracy, Precision, Recall, and F-Measure in a row of 89.58%, 79.17%, 78.72%, 77.05%.

Research conducted by Huq et al (2017) concerning the application of sentiment analysis on Twitter users using the KNN method. This research uses the PHP programming language, the data used is data tweet and data retrieval using scraper. Classifications in two classes, namely positive sentiment and negative sentiment the classification method using the K-Nearest Neighbor and TF-IDF weighting. From the test results it is known that the biggest accuracy value is 67.2% when K = 5.

Research conducted by Wongkar and Angdresey (2019 concerning sentiment analysis using the Naive Bayes Classifier method on student questionnaires with PHP programming languages, and data retrieval manually from the questionnaire which is divided into two parts, namely training data and test data based on classification research This method results in accuracy 80%.

Research conducted by Zuraiyah et.al (2018) with the title Application of Vector Space Model in Aggregator Online Job Vacancies Results obtained using job vacancies from job vacancies such as JobsDB, Jobs, Monster, and JobStreet This system can find job openings The right, fast and accurate according to the query input by the user.

Based on previous research, the author would conduct an analysis of the Indonesian Telematics Employment Sentiment analysis using the TF-IDF weighting by first the preprocessing stage was carried out, namely the case of folding, tokenizing, stopword, and stemming. Then the classification stage used the multinomial Naive Bayes Classifier method and compare it with the Vector Space model classification using the PHP programming language and data retrieval with a web scraping. The advantage of the multinomial method of Naïve Bayes which has a decimal number input and the frequency of the word is very influential on the results obtained, compared to ordinary Naive Bayes in general only has input 1 and 0. Other advantages, namely to obtain a high probability value used Laplace Smoothing so that the value of each Each probability is not equal to 0. If the value of the word probability is 0 then the training and testing data will never be enough to represent the frequency when there are rare events. Naïve Bayes can work well even with the presence of features that have strong depedencies in DataSet (Domingos et.al, 1997). One variant of the Naïve Bayes method to handle the multinomial data used in the text classification is the Multinomial Naïve Bayes method. The multinomial model produces better accuracy than the multivariate model of Bernoui for the classification of text on data with large quantities of vocabulary while mutivariate Bernouuli reverse (McCallum et.al, 1998). While the main advantages of the Consine Similarity method which is part of the Vector Space Model equation is not affected by the short length of a document. In addition, consine similarity has high accuracy when faced with classifications with many types of classes (Bhattacharjee, 2015) and labor competence in line with the industrial revolution in Indonesia. It is expected that by making a quantitative aggregation of sentiments on social media social media and news media results from this study will show how sentiment polarization of the workforce conditions of job factors and labor competencies in Indonesia.

2. Data and Research Methods

2.1 Data Selection

The data selection stage is the stage to usethe data. The data in the *database* is also often not all used, therefore only the appropriate data for analysis will be retrieved from the *database*.

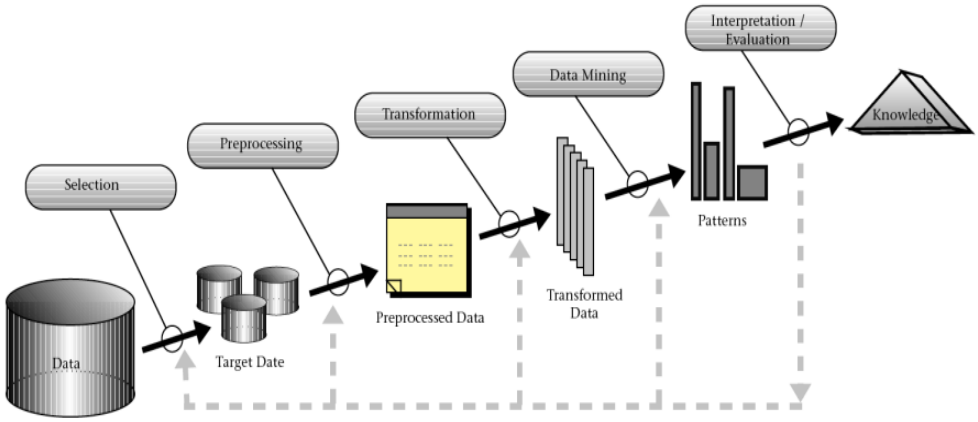


Figure 1. Metode *Knowledge Discovery and Data Mining* (KDD)

2.2 Preprocessing

Preprocessing is to prepare text into data that will undergo further processing. The stages in *preprocessing* are *Case Folding* converting uppercase letters to lowercase letters, *Tokenizing* changing a collection of sentences into single words, *filtering* the disposal of words that are not used in data grouping, *stemming* the change of words that have a growth into the base word.

2.3 Transformation

Data transformation is the process by which data is changed or merged into the appropriate format for processing in data *mining.* Some *data mining methods* require a special data format before they can be applied.

2.4 Data Mining

The *data mining* stage is at the core of data analysis, the process of finding insights, interesting patterns, as well as descriptive, understandable and predictive of large-scale data models. By looking at the fundamental nature of data modeled as a data matrix, which emphasizes geometric and algebraic views as well as probabilistic interpretations of data. *Data mining* is the process of finding interesting patterns or information in selected data that are interrelated by using certain techniques or methods. Techniques, methods or algorithms in *data mining* vary widely. The selection of the right method or algorithm depends largely on the overall purpose and process of KDD. At this stage using *the text mining* process where the initial data used is in the form of text and then processed using *data mining* techniques, this is interrelated. To find such knowledge patterns the methods used were *multinomial Naïve Bayes* and Vector *Space Model.*

2.5 Interpretation / Evaluation

The pattern of information generated from *the data mining* process needs to be displayed in a form that is easily understood by interested parties. This stage is part of a KDD process called *interpretation.* This stage includes examining whether the patterns or information found contradict previously existing facts or hypotheses.

3. Results And Discussions

3.1 Result

In this chapter will explain about the results of website-based applications that have previously been designed. This research will display the results of the classification of sentiment analysis regarding Indonesian telematics employment using *the Multinomial Naïve Bayes* method and *the Vector Space Model* Method. Then there will be evaluation calculations with *confusion matrix* so that we know how much accuracy the accuracy of classification in each method in the application that has been seized.

The main page is the page that was first visited at the time the application was opened by the admin after going through *the login* process. On this page there is a category menu, source menu, news and twitter *training* data menu, news and twitter data *testing* menu, graphs, profiles and admin logout. This page explains briefly about the research to be done as well as the location of the research. The main page view can be seen in Figure 2.

Figure 2. Home Page

The source page is a page that contains the source id description and the name of the source where the data was retrieved. While the category page contains the category id and category label. There is a search menu, export basic, export all, export selected, hide / show pagination, refresh, toggle, columns, and export file type. The source and category page views can be viewed in Figure 3.

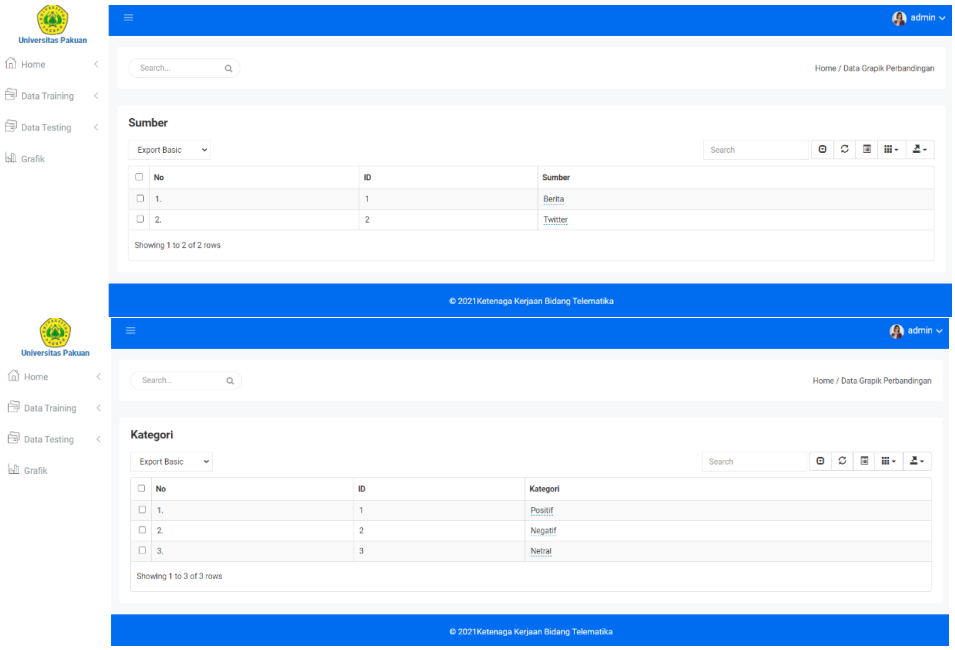
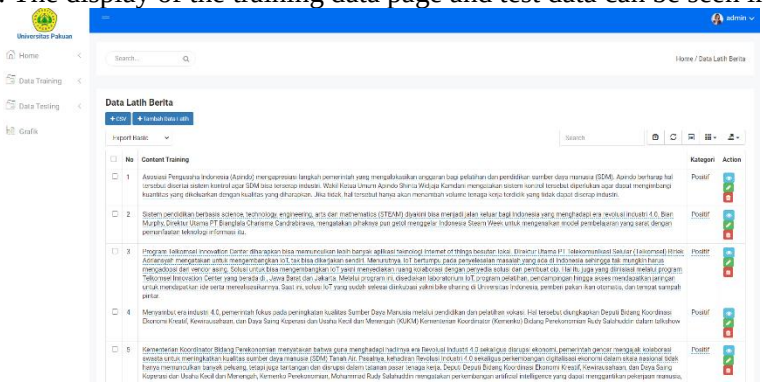


Figure 3. Source and Category page

A training data page is a page that contains information about the content of each training or test data document. Training or test data pages are divided into two, namely news data pages and twitter data pages. There are search menus, export basic, export all, export selected, hide / show pagination, refresh, toggle, columns, and export file types, add data through csv files and add data through forms. There is a process action, change and delete on each training and test document. The display of the training data page and test data can be seen in Figure 4.



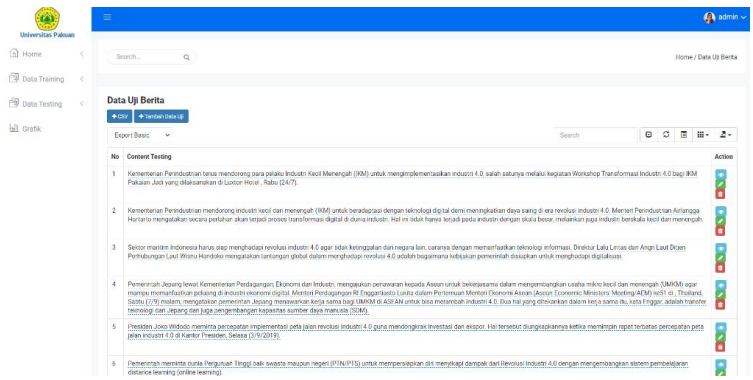


Figure 4. Training And Test Data Page

This page is used to add CSV files that contain a lot of data or training documents for news and twitter is then labeled a positive, negative, or neutral category on each file that has been grouped in the training data. While on the menu add CSV file in the test data there is no selection of category labels.

This page is a page used to add a news sentence or twitter review that is then labeled a positive, negative, or neutral category on each sentence in the training data. While on the menu add review files in the test data there is no selection of category labels. After the sentence is stored, the data will increase one data or document in the database. This page is used to visualize data in the form of graphs of all sentiment analysis results that have been processed based on categories of positive, negative, neutral and data that have not been tested. There are 4 charts, namely the graph of classification results using the *Multinomial Naïve Bayes* method and the Vector Space *Model* method on news and twitter.

3.2 Discussion

This application was built with the aim of analyzing Indonesian people's sentiments regarding telematics employment in the era of the industrial revolution 4.0 based on employment factors and labor competence by doing quantitative aggregation. Data is obtained using *web scraping* techniques using google *chrome extension*, namely *web scraper*. *Scraping* is a technique used to extract large amounts of data from websites where extracted data is stored to a local file on a computer or to a *database* in table or *spreadsheet* format. The website page is generally built using a markup language such as HTML or XHTML which is then processed to retrieve important data around the information we need related to the employment of Indonesian telematics field.

The process of retrieving data with *web scraping* technique is carried out on several sites including social media sites twitter, online news sites such as kompas.com, bisnis.com, kabar24.id, tribunnews.com, infonawacita.com, republika.co.id, akurat.com, antaranews.com, and kominfo.go.id. From the *web scraping* process produces 1,323 data divided into 2 *training* data and data *testing* then stored in CSV format. The classification process is carried out through 2 stages, including *the training* and *testing stages.* The stage of learning or training is the process of extracting documents that have been known categories. This process is done by the establishment and formation of a bag of *words* on each *training* document. Each stored word is analyzed *weighting* values against all documents in each classification so that it can calculate the probability of the word that will be used at the *testing* stage (Kalokasari et.al, 2017). For more details, you can see it in Figure 5.

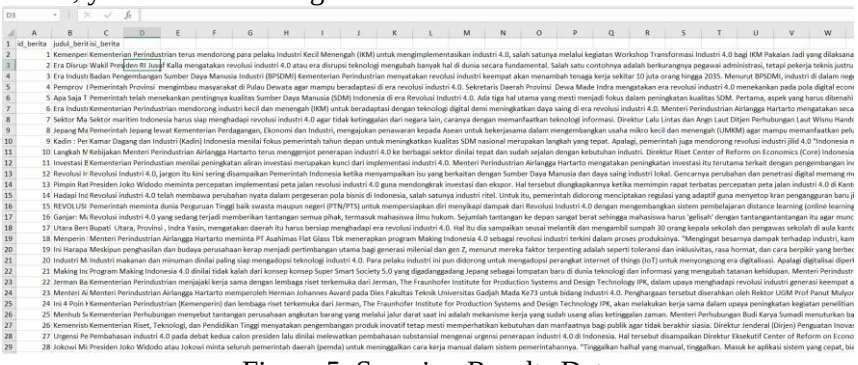


Figure 5. Scraping Results Data

Further analysis needs to be done to ensure that the topics taken in accordance with what is needed in this study, after this process is done the process of cleaning the data manually, then the data is entered into the database either through *importing data* at *localhost* or importing CSV files in the application we create. Data that has gone through the cleaning stage and imported into the *database* can be seen in Figure 6.

	id_ulasan	id_sumber	ulasan	id_kategori
Edit Copy Delete	1	1	Asosiasi Pengusaha Indonesia (Apindo) mengapresias...	1
Edit Copy Delete	2	1	Sistem pendidikan berbasis science, technology, en...	1
Edit Copy Delete	3	1	Program Telkomsel Innovation Center diharapkan bis...	1
Edit Copy Delete	4	1	Menyambut era industri 4.0, pemerintah fokus pada ...	1
Edit Copy Delete	5	1	Kementerian Koordinator Bidang Perekonomian menyat...	1
Edit Copy Delete	6	1	Indonesia siap memberikan dukungan negaranegara be...	1
Edit Copy Delete	7	1	Pertumbuhan bisnis logistik seiring dengan peningk...	1
Edit Copy Delete	8	1	Perguruan tinggi diminta aktif menyiapkan SDM yang...	1
Edit Copy Delete	9	1	Kementerian Perindustrian terus mendorong pengemba...	1
Edit Copy Delete	10	1	Presiden Joko Widodo atau Jokowi menyebut kondisi ...	1
Edit Copy Delete	11	1	Teknologi cloud computing atau komputasi awan dapa...	1
Edit Copy Delete	12	1	Dunia telah masuk ke era baru, yakni disrupsi yang...	1
Edit Copy Delete	13	1	PT Pan Brothers Tbk gencar melakukan otomatisasi d...	1
Edit Copy Delete	14	1	Surabaya, Gatra.com - Kementerian Komunikasi dan ...	1
Edit Copy Delete	15	1	Jakarta, Kominfo - Komputasi awan bisa menjadi so...	1

Figure 6. Data After Cleaning and Imported to Database

Data that has entered the database needs to be processed again by the applications that we have created by doing *preprocessing* stages including *case folding*, *tokenizing*, *stopword removal* or *filtering* and *stemming.* For more details, you can see it in Figure 7.

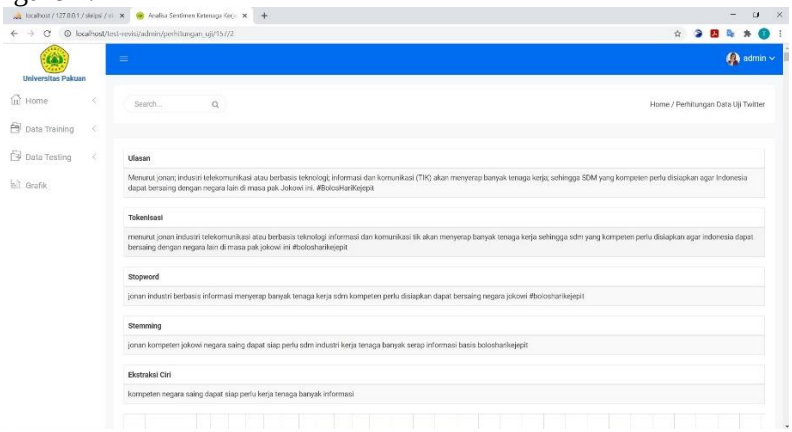
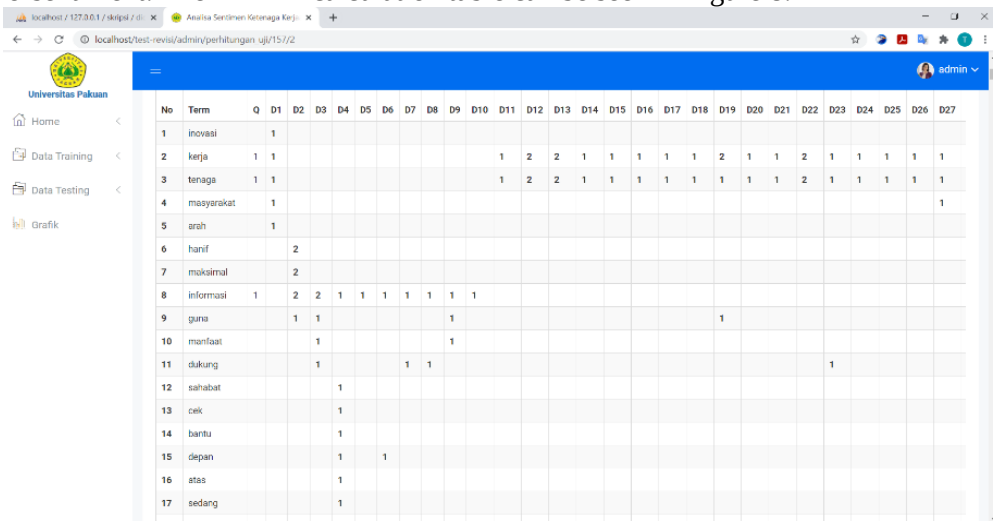


Figure 7. *Preprocessing*

Furthermore, the data that goes through the *preprocessing* stage will be calculated the weight of each word using the TF-IDF method. To make it easier to find the aggregate value of this sentiment analysis TF-IDF will multiply the value of TF (Term*Frequency)*by DF (Document*Frequency)*so that the weight of each word can be used to calculate the classification ofsentiment. The TF-IDF calculation table can be seen in Figure 8.



No	Term	Q	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20	D21	D22	D23	D24	D25	D26	D27
1	inovasi		1																										
2	kerja	1	1										1	2	2	1	1	1	1	1	2	1	1	2	1	1	1	1	1
3	tenaga	1	1										1	2	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1
4	masyarakat		1																										1
5	arah		1																										
6	hanif			2																									
7	maksimal			2																									
8	informasi	1		2	2	1	1	1	1	1	1	1																	
9	guna			1	1						1										1								
10	manfaat				1						1																		
11	dukung				1				1	1															1				
12	sahabat					1																							
13	cek					1																							
14	bantu					1																							
15	depan					1		1																					
16	atas					1																							
17	sedang					1																							

Figure 8. TF-IDF Calculation Table

Data after weighting with TF-IDF then calculated the value of its sentiment classification using *the Multinomial Naïve Bayes* method and *the Vector Space Model* Method. The classification table with *the Multinomial Method of Naïve Bayes* can be seen in Figure 9 and the classification table with the Vector Space *Model* method can be seen in Figure 10.

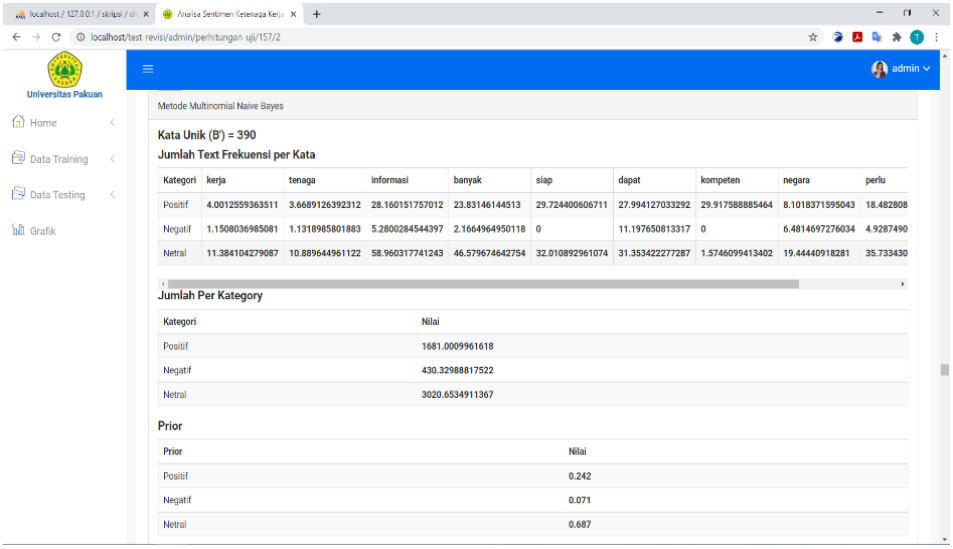


Figure 9. Classification Table with MNB Method

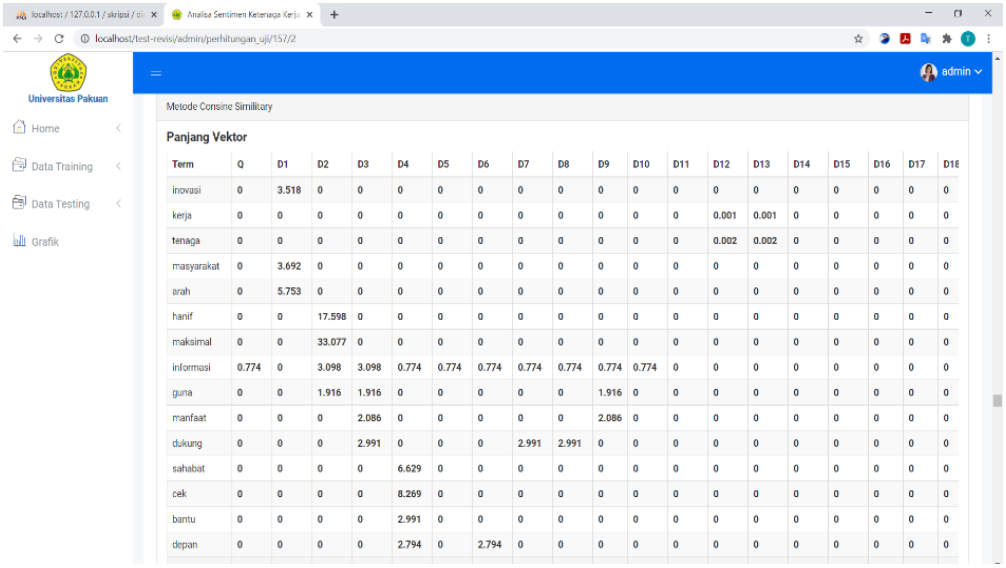


Figure 10. Classification Table with VSM Method

After getting all the sentiment category values from the test data then the data will be represented values into graph form. Using 1,323 data divided into two, namely training data as much as 936 training data and test data as much as 387 data in this study the Indonesian telematics field employment sentiment analysis system produced the highest positive value on news data using *the Multinomial Naïve Bayes* method with a percentage of 48.17%. While in news data using the Vector Space Model method *the* system produces the highest neutral value with a percentage of 51.09%. On twitter data using the *Multinomial Naïve Bayes* method the system generated the highest neutral value with a percentage of 77.2%. While on twitter data using the Vector Space Model method *the* system produces the highest neutral value with a percentage of 60.4%. In this graph shows data more towards neutral towards positive this is evidenced in a news that the world is currently faced with job disruption in several industrial lines. The demand for individuals with mastery of new skills also becomes an inevitability for every company. For example, in the world of technology. In the past, maybe this type of work such as big data specialists, artificial intelligence (AI) specialists, or data analysis has not needed its role. However, from the last few years to decades to the next, demand for workers in this sector is predicted to skyrocket. The government applies science, technology, engineering, arts, mathematics (STEAM) learning methods to improve the quality of education (Kompas.com). The impact of major changes in industry 4.0 has been researched and studied by the McKinsey Global Institute with the conclusion: Industry 4.0 will have a wide impact in the industrial world because of the complete utilization of the digitalization process until 2030 the industrial revolution process can be a threat in Indonesia today has a large workforce but at the same time has a high unemployment rate as well. The importance of understanding the technological environment turned into a demand to prepare for it, almost all professions are currently affected either directly or indirectly. The graph of employment sentiment analysis in Indonesian telematics can be seen in Figure 11.

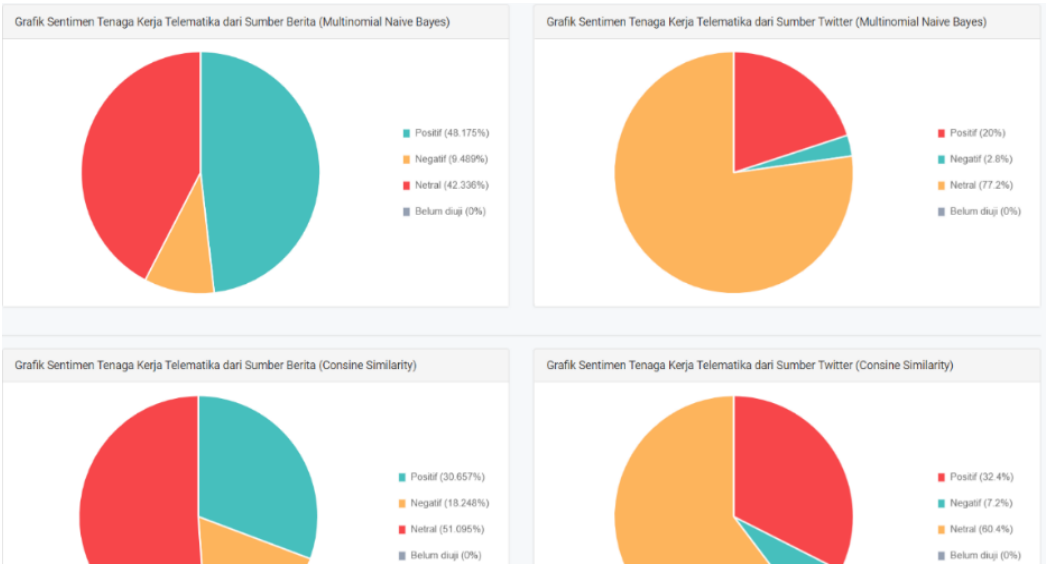


Figure 11. Graph of Employment Sentiment Analysis in Telematics

This stage is a testing phase of the system that has been built. Tests are conducted to find out the shortcomings of the system created. In addition, it can be known whether the system is functioning properly as desired or not.

Structural trials aim to test whether the applications that have been made run according to the design structure or not. From the results of the trial it can be known that the application is in accordance with the design structure. The results of the trial can be seen in Table 1.

This stage is done to test whether the function, button or form that has been made to function according to its function, the test results can be seen in Table 2.

Validation trials are conducted to determine the system's ability to provide relevant search results. The trial was conducted using *a confusion matrix* by calculating the accuracy, *precision*, and *recall* of the results of sentiment categories by the system with actual sentiment categories. Here is a table of comparison results between *actual* and *prediction* news data. As well as calculating accuracy, *precision*, and *recall* values with *Naïve Bayes' Multinomial* method for news data can be seen in Table 1.

Table 1. *Confusion Matrix* Table with *Naïve Bayes Multinomial* Method for News Data

		Predictions Class		
		Positive (Class A)	Negative (Class B)	Neutral (Class C)
Actual Class	Positive (Class A)	50	2	6
	Negative (Class B)	4	11	1
	Neutral (Class C)	12	0	51

$$\text{Accuracy} = \frac{50+11+51}{51+2+6+3+11+1+12+0+51} * 100\% = \frac{112}{137} * 100\% = 81.75\%$$

Precision :

1. Positive $= \frac{50}{50+4+12} * 100\% = \frac{50}{66} * 100\% = 75.75\%$
2. Negative $= \frac{11}{11+2+0} * 100\% = \frac{11}{13} * 100\% = 84.62\%$
3. Neutral $= \frac{51}{51+6+1} * 100\% = \frac{51}{58} * 100\% = 87.93\%$

Average Precision $= \frac{75.75 + 84.62 + 87.93}{3} = \frac{248,3}{3} = 82.77\%$

Recall :

1. Positive $= \frac{50}{51+2+6} * 100\% = \frac{50}{59} * 100\% = 84.75\%$
2. Negative $= \frac{11}{11+4+1} * 100\% = \frac{11}{16} * 100\% = 68.75\%$
3. Neutral $= \frac{51}{51+12+0} * 100\% = \frac{51}{63} * 100\% = 80,95\%$

Average Recall $= \frac{84.75+68.75+80.95}{3} = \frac{234.45}{3} = 78.15\%$

Table of comparison results between *actual* and *prediction* data twitter. As well as calculating accuracy, *precision*, and *recall* values with *Naïve Bayes' Multinomial* method for twitter data can be seen in Table 2.

Table 2. *Confusion Matrix* Table with *Naïve Bayes Multinomial* Method for Twitter Data

		Predictions Class		
		Positive (*Class* A)	Negative*(Class* B)	Neutral*(Class* C)
Actual Class	Positive (*Class* A)	47	0	19
	Negative*(Class* B)	0	7	6
	Neutral*(Class* C)	3	0	168

$$\text{Accuracy} = \frac{47+7+168}{47+0+19+0+7+7+3+0+168} * 100\% = \frac{222}{250} * 100\% = 88.80\%$$

Precision :

1. Positive = $\frac{47}{47+0+3} * 100\% = \frac{47}{50} * 100\% = 94\%$
2. Negative = $\frac{7}{7+0+0} * 100\% = \frac{7}{7} * 100\% = 100\%$
3. Neutral = $\frac{168}{168+19+6} * 100\% = \frac{168}{193} * 100\% = 87.05\%$

Average Precision = $\frac{94+100+87.05}{3} = \frac{281.25}{3} = 93.75\%$

Recall :

1. Positive = $\frac{47}{47+0+19} * 100\% = \frac{47}{66} * 100\% = 71.21\%$
2. Negative = $\frac{7}{7+0+6} * 100\% = \frac{7}{13} * 100\% = 53.85\%$
3. Neutral = $\frac{168}{168+3+0} * 100\% = \frac{168}{171} * 100\% = 98.25\%$

Average Recall = $\frac{71.21+53.85+98.25}{3} = \frac{223.31}{3} = 74.44\%$

Table 3. Accuracy, *Precision,* and *Recall* Results from News and Twitter Data Using *Naïve Bayes' Multinomial* Method

No	Data Source	Category	*Precision*	*Recall*	Accuracy
1.	**News**	**Positive**	75,75%	84,75%	81,75%
		Negative	84,62%	68,75%	
		Neutral	87,93%	80,95%	
		Average	82,77%	78,15 %	
2.	**Twitter**	**Positive**	94%	71,21%	88,80%
		Negative	100%	53,85%	
		Neutral	87,05%	98,25%	
		Average	93,75%	74,44%	

From the results of the evaluation of confusion matrix news and twitter data using *the Multinomial Method Naïve Bayes* can be concluded that the average accuracy of news data produces 81.75%, precision 82.77%, recall 78.15%. While on twitter data produces an average accuracy of 88.80%, precision 93.75%, recall 74.44%.

The following is a table of comparison results between *actual* and *prediction* news data. As well as calculation of accuracy, *precision,* and *recall* values with the Vector Space *Model* method for news data can be seen in Table 4.

Table 4. *Confusion Matrix* Table with Vector *Space Model* Method for News Data

		Predictions Class		
		Positive (*Class* A)	Negative*(Class* B)	Neutral*(Class* C)
Actual Class	Positive (*Class* A)	32	5	21
	Negative*(Class* B)	2	13	1
	Neutral*(Class* C)	8	7	48

$$\text{Accuracy} = \frac{32+13+48}{32+6+21+2+12+1+8+7+48} * 100\% = \frac{93}{137} * 100\% = 67.88\%$$

Precision :

1. Positive = $\frac{32}{32+2+8} * 100\% = \frac{32}{42} * 100\% = 76.19\%$
2. Negative = $\frac{13}{13+5+7} * 100\% = \frac{13}{25} * 100\% = 52\%$
3. Neutral = $\frac{48}{48+21+1} * 100\% = \frac{48}{70} * 100\% = 68.57\%$

Recall :

1. Positive = $\frac{32}{32+5+21} * 100\% = \frac{32}{58} * 100\% = 54.24\%$
2. Negative = $\frac{13}{13+2+1} * 100\% = \frac{13}{16} * 100\% = 81.25\%$
3. Neutral = $\frac{48}{48+8+7} * 100\% = \frac{48}{63} * 100\% = 76.19\%$

Average Recall = $\frac{54.24+81.25+76.19}{3} = \frac{211.68}{3} =$

Average Precision = $\frac{76{,}19+52+68.57}{3} = \frac{196.76}{3} =$ 65.59%	70.56%

Table of comparison results between *actual* and *prediction* data twitter. The *Confusion Matrix* table and the calculation of accuracy, *precision*, and *recall* values with the Vector Space *Model* Method for Twitter Data can be seen in Table 5.

Table 5. *Confusion Matrix* Table with Vector *Space Model* Method for Twitter Data

		Predictions Class		
		Positive (*Class* A)	Negative*(Class* B)	Neutral*(Class* C)
Actual Class	Positive (*Class* A)	59	1	6
	Negative*(Class* B)	1	11	1
	Neutral*(Class* C)	21	6	144

Accuracy = $\frac{59+11+144}{59+1+6+1+12+1+21+5+144} * 100\% = \frac{214}{250} * 100\% = 85.60\%$

Precision : 1. Positive = $\frac{59}{59+1+21} * 100\% = \frac{59}{81} * 100\% =$ 72.84% 2. Negative = $\frac{11}{11+1+6} * 100\% = \frac{11}{18} * 100\% =$ 61.11% 3. Neutral = $\frac{144}{144+6+1} * 100\% = \frac{144}{151} * 100\% =$ 95.36% *Average Precision* = $\frac{72.84+61.11+95.36}{3} = \frac{229.31}{3} =$ 76.44%	*Recall* : 1. Positive = $\frac{59}{59+1+6} * 100\% = \frac{59}{66} * 100\% =$ 89.39% 2. Negative = $\frac{11}{11+1+1} * 100\% = \frac{11}{13} * 100\% =$ 84.62% 3. Neutral = $\frac{144}{144+21+6} * 100\% = \frac{144}{171} * 100\% =$ 84.21% *Average Recall* = $\frac{89.39+84.62+84.21}{3} = \frac{258.22}{3} =$ 86.07%

Table 6. Accuracy, *Precision*, and *Recall* Results from News and Twitter Data Using Vector *Space Model* Method

No	**Data Source**	**Category**	***Precision***	***Recall***	***Accuracy***
1.	**News**	**Positive**	76.19%	54.24%	67.88%
		Negative	52%	81.25%	
		Neutral	68.57%	76.19%	
		Average	65.59%	70.56%	
2.	**Twitter**	**Positive**	72.84%	89.39%	85.60%
		Negative	61.11%	84.62%	
		Neutral	95.36%	84.21%	
		Average	76.44%	86.07%	

From the results of the evaluation of confusion matrix news and twitter data using the *Vector Space Model* method it can be concluded that the average accuracy of news data produces 67.88%, precision 65.59%, recall 70.56%. While on data Twitter produces an average accuracy of 85.60%, precision 76.44%, recall 86.07%.

Based on the results of this study the system showed in the testing of news data and twitter the accuracy and *precision* of *the Multinomial Naïve Bayes* method is higher than the Vector Space Model method. *Naïve Bayes' Multinomial* method *recall* value in news data is higher than the Vector Space Model's *recall* value in news data. But the *recall* value *of Naïve Bayes' Multinomial* method on twitter data is lower than the *recall* value of the Vector *Space Model* method on twitter data. It can be concluded that this system is sufficiently feasible to be used in analyzing community sentiment based on labor factors and employment so that it can be an evaluation material by increasing labor competence and expanding employment.

4. Conclusion

In the study entitled *Employment Sentiment Analysis in Indonesian Telematics Using Multinomial Naïve Bayes*, it took data with *scraping* techniques on twitter.com social media sites and on online news sites such as kompas.com, bisnis.com, kabar24.id, tribunnews.com, gatra.com, infonawacita.com, republika.co.id, akurat.co, antaranews.com,

and kominfo.go.id. The data collected as much as 1,323 data. It was then divided into 185 news training data, 137 news test data, 751 twitter training data, and 250 twitter test data which was then saved to the database.

After going through the scraping stages, *the preprocessing* and word-weighting with TF-IDF, as well as comparing the classification results with the *Multinomial Method Naïve Bayes* with the addition of *laplace smoothing* to avoid the probability of 0 and the Vector Space *Model* method ending with *consine similarity* to look for similarity of documents.

This application has gone through several stages of trials including structural trials, functional trials, validation trials. At the validation trial stage is done using *confusion matrix. Naïve Bayes' Multinomial* Method for news data yielded an accuracy value of 81.75%, an average *precision* of 82.77%, and an average *recall* of 78.15%. While with the *Vector Space Model* method for news data produces an accuracy value of 67.88%, average *precision* 65.59%, and average *recall* 70.56%. On twitter data with the Multinomial method *Naïve Bayes* produced an accuracy of 88.80%, average *precision* 93.75%, and average *recall* 74.44%. On twitter data with the Vector Space *Model* method produces an accuracy of 85.60%, average *precision* 76.44% and average *recall* 86.07%. Based on the results of this study the system showed in the testing of news data and twitter the accuracy and *precision* of *the Multinomial Naïve Bayes* method is higher than the Vector *Space Model* method. *Naïve Bayes' Multinomial* method recall value in news data is higher than the Vector Space Model's *recall* value in news data. But the *recall* value *of Naïve Bayes' Multinomial* method on twitter data is lower than the *recall* value of the Vector *Space Model* method on twitter data. It can be concluded that this system is sufficiently feasible to be used in analyzing community sentiment based on labor factors and employment so that it can be an evaluation material by increasing labor competence and expanding employment.

References

Bhattacharjee, J. 2015. Constructivist Approach to Learning– An Effective Approach of Teaching Learning. *International Research Journal of Interdisciplinary & Multidisciplinary Studies (IRJIMS), 1*(4), 65-74.

Chen H. and Fu D. 2018. An Improved Naïve Bayes Classifier for Large Scale Text. *Advances in Intelligent Systems Research, 146*, 33-36.

Cuaresma, J. C., Lutz, W., and Sanderson, W. 2014. Is the Demographic Dividend an Education Dividend? *Demography*, 299-315.

Domingos. 1997. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. *Proceedings of ICML*.

Huq, R. M., Ali, A., and Rahman, A. 2017. Sentiment Analysis on Twitter Data using KNN and SVM. *International Journal of Advanced Computer Science and Applications*, *8*(6), 19-25.

McCallum. 1998. A Comparation of Event Models for Naïve Bayes Text Classification. *Proceedings of AAAI. Pennsylvania*.

Naz, S., Sharan, A. and Malik, N. 2018. Sentiment Classification on Twitter Data Using Support Vector Machine. *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 676-679.

Warsito, T. 2019. Attaining The Demographic Bonus in Indonesia. *Jurnal Pajak dan Keuangan Negara, 1*(1), 134-139.

Wongkar, M. and Angdresey, A. 2019. Sentiment Analysis Using Naive Bayes Algorithm of The Data Crawler: Twitter. *Conference: Fourth International Conference on Informatics and Computing (ICIC)*. 1-5.

Zuraiyah, T. A., Wihartiko, F. D., amd Effendi, E. 2018. Implementation of Vector Space Model in Online Jobs Vacancy Aggregator. *International Journal of Engineering & Technology*. *7*(3). 385-388.