INFORMATION EXTRACTION ON INCOMING LETTER USING NAIVE BAYES METHOD

Chandra Ratiwi[1], Nelly Indriani[2]

Program Studi Teknik Informatika
Fakultas Teknik dan Ilmu Komputer Universitas Komputer Indonesia
Jl. Dipati Ukur 114 Bandung
E-mail : chandrartw@gmail.com[1], indiwidi@gmail.com[2]

ABSTRACT

Information extraction is a method of text mining that aims to find, retrieve or process information which contained in an object. Based on the results of information extraction previous research, document has a deficiency that rule-based method must have a structured and formatted document. Based on the existing problem, in this study the Naive Bayes method will be used as a machine learning algorithm to extract information by classifying sentence by sentence which contained in the incoming letter, because incoming mail is not always identical to the existing format. Before the letter is classified by Naive Bayes the letter will go through to text preprocessing which consists of filtering, tagging, and tokenization. After preprocessing the data is given a weight value for each feature based on sentence text which found in the letter for the probability calculation of Naive Bayes. Based on the results of the confusion matrix test, the results are really good and it can be concluded that the Naive Bayes algorithm can be applied to information extraction systems on incoming letters with very high accuracy results of 96.96%.
Kata Kunci : Information Extraction, Supervised Learning, Classification, *Naive Bayes*, Incoming Letter

1. INTRODUCTION

Letters are one of the tools used to communicate between one party and another to convey written information. Each letter has a certain value of information such as the name of the sending agency, address, date, subject matter and so on. To obtain this information, the need arises to extract information automatically from the data in the form of text in the letter so that it can be stored in the form of data [1].

Information extraction is an automatic search for structured information such as entities, relationships between entities, and attributes that describe entities from unstructured sources [2]. The search for information extraction is a system that functions to find specific data in Natural Language Text. In a form or table template and the text components, the data is taken to be extracted into data that has usefulness [3]. Information extraction on patterned text was previously studied by Dimas Mustaqwa with the title "Implementation of Information Extraction in Thesis Documents Using Ruled Based Methods" but even though the accuracy level generated by the method is 100% but the method must have exactly the same pattern as the dataset so does not match text that does not always have the same pattern [4]. Based on the problem based on the rule, according to Marina Dorash, the solution that can be done is to use machine learning [5]. One method of machine learning is Naïve Bayes. Naïve Bayes has several advantages and advantages of other algorithms, namely simple, fast and has a high degree of accuracy [6]. Information extraction was carried out by Oman Sumantri in his research that Naïve Bayes had an accuracy rate of 14% better than SVM [7]. This is confirmed that based on research proven by Xhemali, Hinde and Stone in his research said that the Naïve Bayes Classifier has a better level of accuracy than Decision Tree and Neural Network [8].

2. INFORMATION EXTRACTION

Information extraction is an automatic search for structured information such as entities, relationships between entities, and attributes that describe entities from unstructured sources [2]. The search for information extraction is a system that functions to find specific data in Natural Language Text. In a form or table template and the text components, the data is taken to be extracted into data that has usefulness [3]. Information extraction acts as a system that will recognize unstructured data that has information that has not been categorized and has no specific meaning. Unstructured does not mean that data is confusing or unclear in the sense that the

information contained in it cannot be directly translated by a computer so that it requires a system called information extraction..

3. RESEARCH METHORD

3.1 Letter Analysis

Data sources for incoming letters are officially obtained physically and the internet with the rules of an official letter from the service. In the input letter data has restrictions, namely :

1. Document has .pdf format
2. Logo instansi, cap, dan tanda tangan tidak termasuk dalam ekstraksi informasi pada penelitian..
3. Agency logo, stamp, and signature are not included in the extraction of information in the study..

The following are the things contained in the letter:

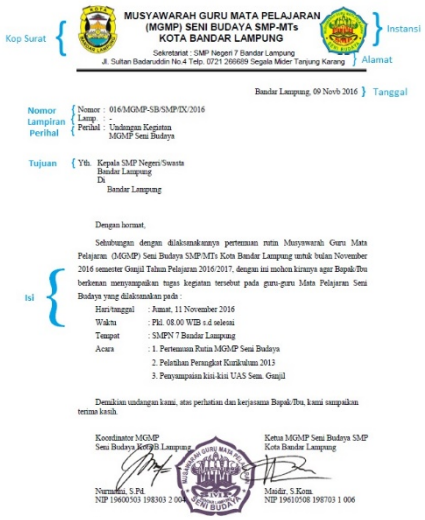


Figure 1. Letter

3.2 System Analysis

System analysis can be defined as the decomposition of a complete system into its component parts with a view to identifying and evaluating existing problems and obstacles so that improvements can be proposed in assisting the system design process. System Analysis Implementation of Information Extraction in the Entry Letter with the Naïve Bayes Method includes several parts, namely :

1. Convering format data.
2. Text *Preprocessing* analysis (*Filtering*, *Tagging*, dan *Tokenization*).
3. Weighting Analysis.
4. *Naïve Bayes* algorithm analysis.
5. Information Extraction Analysis.

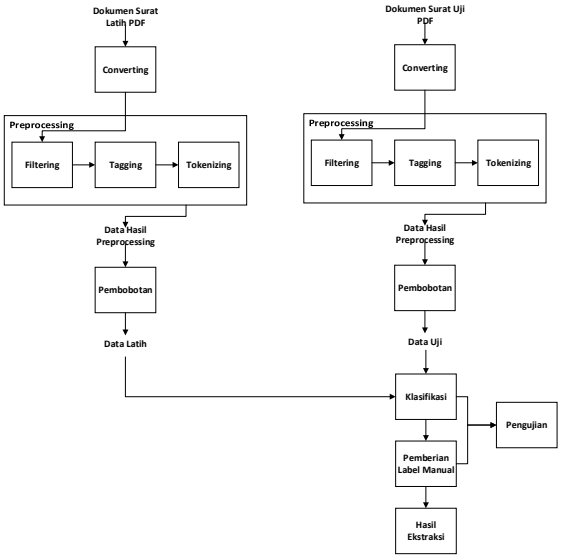


Figure 2. Information Extraction System Description

3.3 Converting Process

The converting process is the process of converting the data format into the data format needed to go to the preprocessing process. Conversions made in this study are * .pdf file formats to * .xml. The reason for choosing converting is that based on conversion experiments carried out into various formats such as docx, csv or txt, it has constraints on the results obtained and xml is chosen because the resulting text has an html tag, from the start of the conversion and has a tag per line so as to maintain the shape of the letter and suitable for letters that have sentence data per line for dates, numbers, attachments and destinations so that it is easier to process text preprocessing. The converting process can be seen in Figure 3, and the results in Figure 4 with a partial example of the results of the conversion process.

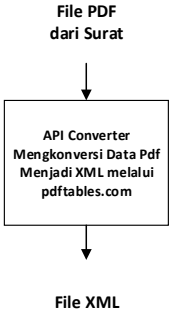


Figure 3. Converting Process



Figure 4. Converting Result

3.4 Preprocessing Analysis

Preprocessing is a process for preparing data to be processed in the main process using Naive Bayes. There are three processes in preprocessing in this study, namely filtering, tagging, tokenization. Preprocessing depictions can be seen in Figure 5.

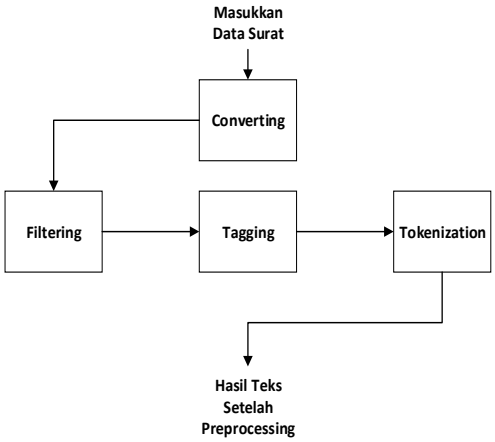


Figure 5. Preprocessing

3.4.1 Filtering

The filtering process can also be called stopwords. The motive that stopwords must be removed from a text is that they make the text look heavier and less important for research. Removing stopword reduces the dimensions of the term space. These words are treated as stop words. Stopwords are deleted from documents because they are not measured as keywords in text mining applications [9]. Characters to be filtered can be seen in table 2 and the process can be seen in Figure 6 and the results in Figure 7.

Table 1. Filtering Actions

Condition	Action
<td>,<tr>,</tr>,</td>	Replace by <>
Colspan, Style text	Colspan, Style text
Words /tag in HTML	Erased by strip_tags function
Except on top condition	No Action

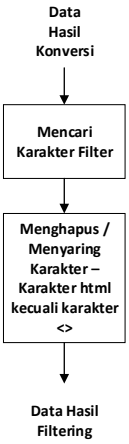


Figure 6. Filtering Process

<document page-count="1">
<page number="1">
<table data-filename="php2E3E.tmp" data-page="1" data-table="1">< colspan="5">MUSYAWARAH GURU MATA PELAJARAN < colspan="4">(MGMP) SENI BUDAYA SMP-MTs KOTA BANDAR LAMPUNG <>Sekretariat : SMP Negeri 7 Bandar Lampung <><>< colspan="7"><>Jl. Sultan Badaruddin No.4 Telp. 0721 266689 Segala Mider Tanjung Karang < colspan="5">Bandar Lampung, 09 Novb 2016<>Nomor :016/MGMP-SB/SMP/IX/2016 Lamp. :< style="text-align: right">- <>Perihal :Undangan Kegiatan <><>MGMP Seni Budaya <>YthKepala SMP Negeri/Swasta <>Bandar Lampung Di <><>Bandar Lampung <><>Dengan hormat, <><>Sehubungan dengan<>dilaksanakannya<>pertemuan<>rutin<>Musyawarah<>Guru<>Mata<>Pelajaran< colspan="7">(MGMP) Seni Budaya SMP/MTs Kota Bandar Lampung untuk bulan November<>< colspan="9">2016 semester Ganjil Tahun Pelajaran 2016/2017, dengan ini mohon kiranya agar Bapak/Ibu<>berkenanm enyampaikan tugas kegiatan tersebut<>pada<>guru-guru<>Mata<>Pelajaran<>Seni<>< colspan="4">Budaya yang dilaksanakan pada : Hari/tanggal : Jumat, 11 November 2016 <>Waktu<>: Pkl. 08.00 WIB s.d selesai <>Tempat<>: SMPN 7 Bandar Lampung <>Acara: 1. Pertemuan Rutin Acara: 1. Pertemuan Rutin MGMP Seni Budaya <>2. Pelatihan Perangkat Kurikulum 2013 <>3. Penyampaian kisi-kisi UAS Sem. Ganjil <><>< colspan="8">Demikian undangan kami, atas perhatian dan kerjasama Bapak/Ibu, kami sampaikan<>terima kasih. <><>Koordinator MGMP<><>< colspan="4">Ketua MGMP Seni Budaya SMP<><>Seni Budaya Kota B.Lampung<>Kota Bandar Lampung GURU MATA<>< colspan="4"><>Nurmaini, S.Pd.<>S<>< colspan="4">Maidir, S.Kom.<>NIP 19600503 198303 2 004<>E N<>A< colspan="4">NIP 19610508 198703 1 006<><>I<>AY< colspan="4"><>< colspan="9">B U D</table></page>
</document>

Figure 7. Filtering Result

3.4.2 Tagging

Tagging is a conditional application that is used to provide conditional tags for certain words or phrases in a sentence depending on the keywords contained in the sentence identified by sed command [10].

Tagging is a step to mark the key words that will be the beginning of the sentence and as a barrier or stemming that will be extracted. But in letters often there are differences in writing or additional characters so that the character adjustment is needed to add tags. Keywords in tagging in this process can be seen in table 3 and the word adjustments in table 4.

Table 2. Tagging Keywords

Kata Kunci	
Nomor	Perihal
No	H a l
Hal	Lampiran
Lamp	Dengan Hormat
Lamp.	Lamp:
Kepada	Assalam
Yth.	Dalam rangka
Jalan	Tempat
Tag	
<>	

Table 3. Adjustment Words

Kata Kunci	Adaption Words Adjustment
Lamp:	Lamp : atau _Lamp:
Dengan Hormat	Dengan hormat
No	No. / _No / No_
Perihal	_Perihal
Jalan	Jln / Jln. / Jl / Jl. / jl / jl.
Yth.	_Yth. / yth. / YTH. / Yth
Nomor	Nomor: / _Nomor:

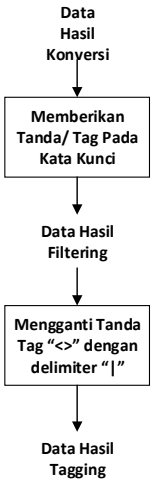


Figure 8. Tagging Process

|MUSYAWARAH GURU MATA PELAJARAN (MGMP) SENI BUDAYA SMP-MTs KOTA BANDAR LAMPUNG Sekretariat : SMP Negeri 7 Bandar Lampung|Jl. Sultan Badaruddin No.4 Telp. 0721 266689 Segala Mider Tanjung Karang Bandar|Lampung, 09 Nov 2016|Nomor :016/MGMP-SB/SMP/IX/2016|Lamp. : -|Perihal :Undangan Kegiatan MGMP Seni Budaya|Yth.Kepala SMP Negeri/Swasta Bandar Lampung Di Bandar Lampung|Dengan Hormat, Sehubungan dengan dilaksanakannya pertemuan rutin Musyawarah Guru Mata Pelajaran(MGMP) Seni Budaya SMP/MTs Kota Bandar Lampung untuk bulan November 2016 semester Ganjil Tahun Pelajaran 2016/2017, dengan ini mohon kiranya agar Bapak/Ibu berkenanmenyampaikan tugas kegiatan tersebut pada guru-guru Mata Pelajaran Seni Budaya yang dilaksanakan pada : Hari/tanggal : Jumat, 11 November 2016 Waktu : Pkl. 08.00 WIB s.d selesai Tempat : SMPN 7 Bandar Lampung Acara: 1. Pertemuan Rutin MGMP Seni Budaya 2. Pelatihan Perangkat Kurikulum 2013 3. Penyampaian kisi-kisi UAS Sem. Ganjil Demikian undangan kami, atas perhatian dan kerjasama Bapak/Ibu, kami sampaikan terima kasih. Koordinator MGMP Ketua MGMP Seni Budaya SMP Seni Budaya Kota B.Lampung Kota Bandar Lampung GURU MATA Nurmaini, S.Pd. S Maidir, S.Kom. NIP 19600503 198303 2 004 E N ANIP 19610508 198703 1 006 I AY B U D|

Figure 9. Tagging Result

3.4.3 Tokenization

Tokenization is the process of solving the flow of textual content into words, terms, symbols, or some other meaningful elements called tokens [9]. The Tokenization process starts from reading the delimiter tag then the system will separate sentences per sentence based on delimiter "|" and the sentence will be entered into the database. The tokenization process can be seen in Figure 10 and the results in Figure 11.



Figure 10. Tokenization Process

No	Kalimat
1	MUSYAWARAH GURU MATA PELAJARAN (MGMP) SENI BUDAYA SMP-M
2	Jl. Sultan Badaruddin No.4 Telp. 0721 266689 Segala Mider Tanjung Karang Bandar
3	Lampung, 09 Novb 2016
4	Nomor :016/MGMP-SB/SMP/IX/2016
5	Lamp. : -
6	Perihal :Undangan Kegiatan MGMP Seni Budaya
7	Yth.Kepala SMP Negeri/Swasta Bandar Lampung Di Bandar Lampung
8	Dengan Hormat, Sehubungan dengan dilaksanakannya pertemuan rutin Musyawara

Figure 11. Tokenization Result

3.5 Weighting

Weighting is the process after the incoming mail data has gone through all stages of preprocessing. Weighting for feature extraction is using the Rule-Based algorithm by giving a boolean value to the feature. Some local features are adjusted based on Peng's research [3] and some of the development features are 11 parameters. The following is the feature table used in table 4 and the weighting process in figure 12.

Table 4. Feature Extraction

Local Feature	
EMAIL	Text has email
ALLCAPS	All text characters are uppercase
CONTAINDIGIT	Text has at least 1 number character
CONTAINDASH	Text has at least 1 strip
CONTAINMORESLASH	Text has at least 1 slash
INITIALKEYWORD	Text has at least 1 keyword
WORD	Text has word itself
CONTAINCOLON	Text has at least 1 colon
STRINGLENGTH	Text length more than 270 characters
CONTAINCOMMA	Text has at least 1 comma
CONTAINDATE	Text has date format

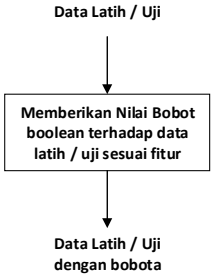


Figure 12. Weighting Process

3.6 Naive Bayes Classification

In the Naive Bayes Algorithm, training data is needed for the learning phase by calculating each probability in the training data from each feature extraction which contained in the training data before the algorithm classifies each testing. Suppose there is data with weighting results as shown in Figure 13. Then the data is calculated by multiplying each probability feature in each class, the data will have results and compare the opportunities in each class. Each of the highest probability results will be classified as a result of the highest probability class as

shown in Figure 14. The classification process can be seen in Figure 15.

	Kalimat	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	PT ROCKET MANA	0	1	0	0	0	0	0	0	0	0	0

Figure 13. Data Testing

DATA UJI	Fitur	Nilai Fitur	DI KELAS							
			1	2	3	4	5	6	7	8
1	x1	0	1	2/5	1	1	1	1	1	1
	x2	1	4/5	0	0	0	0	0	0	0
	x3	0	2/5	0	0	0	3/5	4/5	3/5	0
	x4	0	1	1	1	1	3/5	1	1	2/5
	x5	0	1	1	1	0	1	4/5	4/5	4/5
	x6	0	1	0	1	0	0	0	0	2/5
	x7	0	1	1	1	1	1	0	1	2/5
	x8	0	1	4/5	1	0	0	1/5	1	0
	x9	0	1	1	1	1	1	1	1	0
	x10	0	1	1/5	0	1	1	1	4/5	0
	x11	0	1	1	0	1	1	1	4/5	1/5
Peluang Pada Setiap Kelas			8/25	0	0	0	0	0	0	0
Persentase Peluang			100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00

Figure 14. Classification Calculation

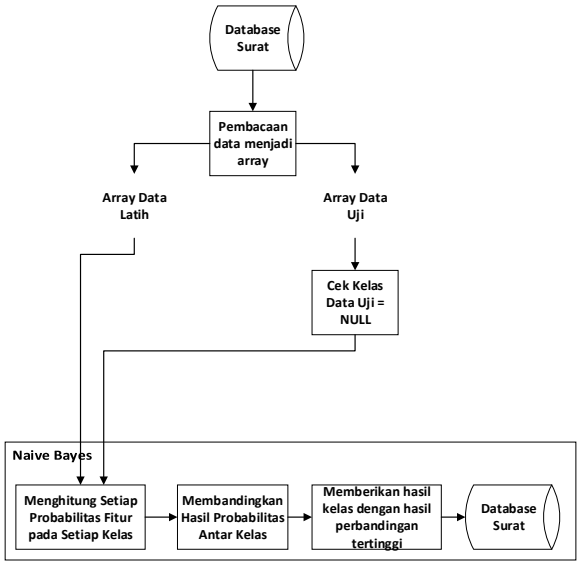


Figure 15. Classification Process

4. TESTING AND STUDY

The testing of this research was conducted to determine the level of accuracy of the Naive Bayes classification in extracting information. Testing is done with confusion matrix. There is a main test with data that is balanced with 27 data letters that each letter has 8 data with a total of 216 training data. This test was also carried out with an unbalanced training data scenario to compare the results obtained. Testing of balance data can be seen in Figure 16

		PREDIKSI							
		1	2	3	4	5	6	7	8
AKTUAL	1	9	0	0	0	0	0	1	0
	2	0	9	0	0	0	0	0	0
	3	0	0	7	0	0	0	0	0
	4	0	0	0	9	1	0	0	0
	5	0	0	0	0	7	0	0	0
	6	0	0	0	0	0	10	0	0
	7	0	4	0	0	0	0	6	0
	8	3	0	0	0	0	0	0	7

Figure 16. Confusion Matrix

$$\text{Accuration} = \frac{\frac{9+60}{9+60+3+1} + \cdots + \frac{6+62}{6+62+1+4} + \frac{7+63}{7+63+0+3}}{8} * 100\%$$

$$= \frac{0.94+0.94+1+0.98+0.98+1+0.93+0.96}{8} * 100\%$$

$$= 96{,}96\%$$

$$\textit{Precision} = \frac{9+9+7+9+7+10+6+7}{(3+9)+(4+9)+(0+7)+(0+9)+(1+7)+(0+10)+(1+6)+(0+7)}$$

$$= \frac{64}{73} * 100\%$$

$$= 87{,}67\%$$

$$\textit{Recall} = \frac{9+9+7+9+7+10+6+7}{(1+9)+(0+9)+(0+7)+(1+9)+(0+7)+(0+10)+(4+6)+(3+7)}$$

$$= \frac{64}{73} * 100\%$$

$$= 87{,}67\%$$

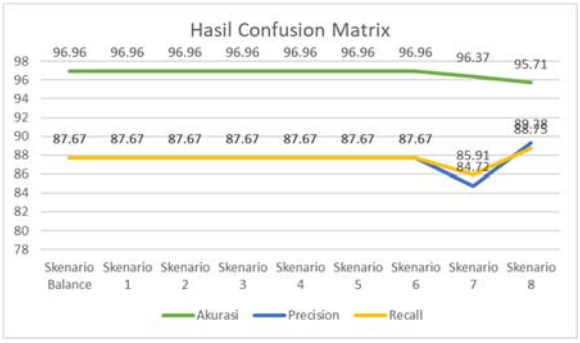
Based on the result, 96.96% accuracy obtained with the same Precision and Recall which is 87.67%. Then compare the accuracy values obtained with the data imbalance scenario with training data that has the composition as in table 5 and the results of the confusion matrix of each scenario are in table 6 and the graph in Figure 17:

Table 5. Scenario Composition

Skenario	Kelas							
	1	2	3	4	5	6	7	8
1	40	25	30	15	30	25	19	22
2	20	40	30	20	30	25	19	22
3	20	30	40	20	30	25	20	21
4	23	23	24	40	30	25	24	27
5	30	23	20	25	40	27	25	26
6	25	27	27	29	20	40	23	25
7	27	14	32	27	22	25	40	25
8	25	27	20	24	30	25	25	40

Table 6. Accuration Result By Scenario

No	Skenario	Precision	Recall	Accuration
1	Data *Balance*	87,67	87,67	96,96%
Skenario *Imbalance* Data				
2	Scenario 1	87,67%	87,67%	96,96%
3	Scenario 2	87,67%	87,67%	96,96%
4	Scenario 3	87,67%	87,67%	96,96%
5	Scenario 4	87,67%	87,67%	96,96%
6	Scenario 5	87,67%	87,67%	96,96%
7	Scenario 6	87,67%	87,67%	96,96%
8	Scenario 7	84,72%	85,91%	96,37%
9	Scenario 8	89,28%	88,75%	95,71%



Gambar 1. Grafik Hasil Confusion Matrix

4.1 Testing Conclusion

Based on the results of the tests with the calculations performed for each scenario, the results obtained with the balance training data are 96.96% accuracy with precision 87.67% and 87.67% recall and the same results obtained with scenarios 1 to 6, but in the scenario 7 and 8 have a difference that in scenario 7 has decreased accuracy 96.37% and even in scenario 8 it decreased to 95.71%. Precision and Recall in scenario 7 also decreased to 84.72 and 85.71. In scenario 8 Precision and Recall goes up to 89.28% and 88.75%. This is due to data imbalances in grades 7 and 8 which greatly affect the classification results so that they experience errors in classification. In scenario 7, there are 4 data changes that have the actual class is class 7 but are classified as class 5. Decreasing precision occurs to classify class 2, so that the true positive data decreases. Then in scenario 8 for class 8 increases 1 which becomes true positive and classifying class 7 which becomes false negative data or class 2 classed data by the system increases. Another reason is that in the test data it is obtained class 7 with the x3 feature, that is class 7 data that contains numbers and in class 8 containing e-mail in it, causing classified data as class 1. Whereas in the training data, it has insufficient data for destination or class 7 which has an x3 or number feature, and contents or class 8 that has an x1 feature or that contains e-mail. So it can be concluded that the training data is very influential in classifying naive bayes and the balance of the data.

5. CONCLUSION AND SUGGESTION

5.1 Conclusion

Pada penelitian yang telah dilakukan ini dapat ditarik kesimpulan bahwa :

1. Naive bayes which is an algorithm for information extraction has a very high level of accuracy and naive bayes is suitable for information extraction systems in incoming mail.

5.2 Suggestion

Based on this research, there is a problem in the conversion process through the API where absurd symbols are found which cause the information extraction process to be unsuccessful. The suggestions for further study are as follows :

1. A better conversion API is needed in recognizing the writing in the letter so that it can finalize the errors in the conversion.
2. Data must has good image quality so that the conversion process runs smoothly without problems.

REFERENCES

[1] F. Peng, “Accurate Information Extraction from Research Papers Using Conditional Random Fields,” *Inf. Process. Manag. an Int. J.*, vol. 42, 2004.

[2] S. Sarawagi, “Information extraction,” *Commun. ACM*, vol. 39, 2008.

[3] A. Riaddy, “6 Ekstraksi Informasi pada Makalah Ilmiah dengan Pendekatan Supervised Learning,” *e-Proceeding Eng.*, vol. 3, no. 1, 2016.

[4] D. Mustaqwa, “Implementasi Ekstraksi Informasi Pada Dokumen Teks Skripsi Menggunakan Metode Rule Based,” Universitas Komputer Indonesia, 2017.

[5] M. Dorash, “Machine Learning vs. Rule Based Systems in NLP.” unpublished.

[6] A. Syarifah, “Pemanfaatan Naive Bayes Untuk Merespon,” vol. 4, 2015.

[7] O. Somantri, “Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB),” *J. Telemat.*, vol. 12, 2017.

[8] D. Xhemali, C. J. Hinde, and R. G. Stone, “Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages,” *Int. J. Comput. Sci.*, vol. 4, no. 1, 2009.

[9] S. Vijayarani, J. Ilamathi, and M. Nithya, “Preprocessing Techniques for Text Mining - An Overview,” *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, 2015.

[10] A. Kao and S. R. Poteet, *Natural language processing and text mining*. New York: Springer, 2007.