Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis

S Kurniawan1*, W Gata1, D A Puspitawati1, I K S Parthama2, H Setiawan3 and S Hartini1

1Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Indonesia
2Universitas Pramita Indonesia, Indonesia
3Sekolah Tinggi Manajemen Informatika dan Komputer Bani Saleh, Indonesia

*kurniawan.sgt@gmail.com

Abstract. Research in the field of Text Mining in general still uses text in English, Arabic, China or others language, while for text in Indonesian is still very limited, so it requires good tools to help Indonesian researchers to conduct research in the field of text mining in Indonesian. Pre-processing is needed for text mining processes such as deleting notation '@', 'http' removal, Indonesian stopwords, normalizing acronym, slang words, emoticons, and Indonesian stemming. The GATA Framework Text Mining provided is one of the options for conducting text mining research in Indonesian and has been used by several researchers. There are several known data mining processing methods, including KKD, CRISP-DM, and SEMMA, all three of which are quite reliable methods. CRISP-DM which consists of; Bussiness Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment is a method that is quite widely used in research in the field of text mining which can be combined with text pre-processing. With so much research in the field of Text Mining in Indonesian, the need for pre-processing in Indonesian is very important. GATA Framework is an option for pre-processing devices that can be combined with Rapidminer devices, as seen from the results of the excellent FUPRS.

1. Introduction
Research in the field of Text Mining in general still uses text in English, Arabic, China or others language, while for texts in Indonesian is still very limited, so it requires good tools to help Indonesian researchers to conduct research in the field of text mining in Indonesian. Pre-processing needed for text mining processes such as deleting ' @' notation, ' http' removal, Indonesian stopwords, normalize acronym words, slang words, emoticons, and Indonesian stemming.

Many software can be used to do text mining, one of which is RapidMiner software. RapidMiner is one of the most widely used worldwide open source data mining solutions [1]. The use of RapidMiner in various studies related to data mining and text mining has been very good. In RapidMiner many menus are available and can be used for text mining, such as retrieving text data from Twitter, Facebook and pre-processing, before pre-processing using a classification algorithm commonly used for text mining such as Support Vector Machine (SVM) and Naïve Bayes (NB). Although the RapidMiner application has provided many menus, for the use of Indonesian Acronym, Indonesian stemming, Indonesian slang
words, and others related to the Indonesian text pre-processing, but is still very limited and requires innovation.

Text mining or Natural Language Processing research, especially Indonesian Language Stemming, already exists in an application called Sastrawi using the PHP and Python programming languages that were built in 2016. In general, this device can be used but needs special expertise in installing it and there are still limitations if must be integrated with other software. One of the studies using the Sastrawi application is research conducted by Agastya and Artha which discusses the influence of Indonesian Language Stemmer [2].

In 2018, a personal named Windu Gata developed web-based applications for pre-processing such as the elimination of Indonesian stopwords, Indonesian stemming, Indonesian Acronym, Indonesian Slang and others intended to help students and other researchers in making research in the field of text-mining of Indonesian language. The application is built using a framework called the GATA Framework (http://www.gataframework.com). The application is an alternative in Indonesian pre-processing text, the application also provides an application program interface (API) feature for sending data from external applications. While the GATA framework is a framework based on the PHP programming language that was developed with the name MTG Framework in 2012 and changed its name to the GATA framework in 2017. GATA Framework has been able to overcome various external problems, namely usability, capability, response, security, existence, and reliability, as well as internal factors, namely ease of syntax or code that is easy to use and has used the Model View Controller (MVC) programming pattern [3]. Currently, the application device can process pre-processing in the form of a single form, or upload data in the form of Ms. Excell files with templates and web services.

In this study, the focus is on how the development and use of pre-processing text with GATA Framework Text Mining and RapidMiner in processing sentiment analysis in Indonesian. In addition, to evaluate the quality of the GATA Framework text mining software, this study uses the FURPS quality model method which consists of, Functionality, Use, Reality, Performance, and Support in the form of descriptive statistics. FURPS (Functionality, Usability, Reliability, Performance, Supportability) Quality Model is a model introduced by Rober Grady, where the development was carried out by IBM in Rational Software [3].

This study choose to use the CRISP-DM method in processing data, which consists of several stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment with a combination of text pre-processing using the GATA Framework pre-processing application combined with RapidMiner applications to process words in Indonesian, as shown in figure 1.
2. Business Understanding

The use of data in the form of text comments or tweets from microblogging twitter is often used to process sentiments from each comment or tweet. The sentiment is an attitude, thought, or judgment prompted by feeling. Sentiment analysis, which is also known as opinion mining, studies people’s sentiments towards certain entities. The Internet is a resourceful place with respect to sentiment information. From a user’s perspective, people are able to post their own content through various social media, such as forums, micro-blogs, or online social networking sites [6].

Research related to sentiment analysis that has used text pre-processing, among others, as has been done by Nia Kusuma Wardhani for sentiment analysis in online news articles related to the coordinator of the maritime ministry using the Naïve Bayes classification algorithm and Support Vector Machine which was optimized with Particle Swarm Optimization by using Rapidminer. The results of the study obtained the Naïve Bayes algorithm which was optimized with PSO having a high accuracy value and being a solution to the problem of sentiment analysis in online news articles [7].

Other research for sentiment analysis as conducted by Siswanto on research related to the classification of comment analysis on social media related to MotoGP, research uses the Naïve Bayes classification algorithm and Support Vector Machine to classify comments on social media and then classify them as analysis of positive or negative sentiment with Rapidminer. The study resulted in accuracy for sentiment prediction of 95.50% for the Support Vector Machine algorithm [8].

There is also another study conducted by Tirana Noor Fatyanosa related to the comparison of the classification methods for sentiment analysis in Indonesian social media, in addition to using the pre-processing text also uses several classification methods compared to one another, the classification methods include; Summation, Average on Tweet, Average on Tweet with the objective score, Weighted Average, and Naïve Bayes method. The results of these experiments found that Naïve Bayes produced high precision, recall and accuracy values for neutral and positive sentiments, but it was not good for negative sentiment [9]. The overall research carried out related to sentiment analysis begins with the
pre-processing phase of the text which then continues to use the classification method to obtain positive or negative results from the dataset used.

3. Data Understanding

When using Twitter social media, there are signs such as # (hashtag), @ (user), HTTP, and other signs that need to be removed so that text can be used better. While the use of pre-processing methods in Indonesian that can be used is the Indonesian acronym method, Indonesian stopwords, Indonesian Stemming, and others. Retrieving data from the Twitter microsite on RapidMiner can use the Search Twit operator as shown in Figure 2.

![Figure 2. Search Twitter and Write Excell Using RapidMiner](Source: RapidMiner 9.2)

In the RapidMiner application there is already a dictionary facility to change the acronym, and stopwords, but it is still limited to English, Chinese, and Arabic, while for Indonesian it is still not available. The Indonesian text processing stage for the removal of a hashtag, @, HTTP, Indonesian acronym, Indonesian Stopwords, and Indonesian stemming can use the GATA Framework application as shown in Figure 3. The technique commonly used to pre-processing Indonesian text is @annotation Removal, Remove URL, Tokenization: REGEXP, Transformation Not (Negative), Indonesian Stemming, and Indonesian Stopwords Removal. An explanation of each technique option in the GATA Framework text mining can be seen in table 1.

![Figure 3. Indonesian Pre-processing Using GATA Framework](Source: http://www.gataframework.com/textmining)
**Table 1.** Description techniques on Text mining – Pre-processing - GATA Framework

| No | Techniques                        | Description                                                                 |
|----|-----------------------------------|-----------------------------------------------------------------------------|
| 1  | @ Annotation Removal              | Remove the @ sign and description that is often used to greet or mark other accounts. |
| 2  | Transformation: Remove URL        | Remove the URL from the tweet used.                                         |
| 3  | Tokenization: Regexp              | Remove marks other than letters and also eliminate numbers.                 |
| 4  | Transformation NOT                | Connect words that have inverse meanings like "no", "no", etc.             |
| 5  | Indonesian Stemming               | Returning a word becomes a basic word.                                      |
| 6  | Indonesian Stopword removal       | Remove words that have no meaning.                                         |

4. **Data Preparing**

At this stage, each data used will be labeled as needed, if you use the model sentiment then each comment can be labeled "Yes" or "No". The amount of data labeled "Yes" and "No" is recommended to have the same amount. If you have an unequal number, you can use the balancing method, namely SMOTE.

5. **Modeling**

After completing the data preparation stage, the next step is to do modeling using RapidMiner. In RapidMiner Pre-processing stages can be done and use sentiment algorithms such as SVM and NB algorithms. Some pre-processing features commonly used on RapidMiner, namely: Tokenize, Filter Stopwords (unnecessary words using a dictionary), Filter Token By Length, and Generate n-Grams (relationship of words to other words), as shown in figure 4 While the next step is the use of SVM and NB algorithms which are validated using 10-fold cross-validation and T-Test.

![Figure 4. Pre-processing Text Using RapidMiner](image)

The next step is to make the overall model of all operators use RapidMiner, namely: ReadExcel, SetRole, Nominal to Text, Process Document (figure 4), SMOTE, Weight by GINI Index, Multiply, Cross Validation for SVM, Cross Validation for NB, and T-Test, as shown in figure 5.

![Figure 5. Example of Modeling Using RapidMiner With SVM and NB Algorithm](image)

Source: RepidMiner 9.2
In the cross-validation stage, there are two columns, namely training and testing. The training column uses the SVM algorithm and NB algorithm can be seen as in Figure 6.

![Figure 6. Example of Cross-Validation SVM and NB](source: RapidMiner 9.2)

6. Evaluation
At this stage, what must be done is to choose the best algorithm by looking at the value of accuracy produced by the RapidMiner application device. If you have got the best accuracy value from the algorithm model used and consider the accuracy to be good enough, then it will continue with the deployment stage.

In this study, the use of FURPS was used in the form of descriptive statistics to assess Text Mining - Pre-processing applications on the GATA framework. A total of 21 application users or researchers were given a questionnaire by giving questions about FURPS. The results of the questionnaire are explained in table 6, where the highest value of the questioner is 5 multiplied by the number of evaluators of 21, equal to 105. The scoring formula is the maximum number of divided values * 100.

| No | Aspek     | Score | Result Value/max * 100 |
|----|-----------|-------|------------------------|
| 1  | Functionality | 13 8  | 87.62                  |
| 2  | Usability   | 15 6  | 85.71                  |
| 3  | Reliability | 1 13 7| 82.86                  |
| 4  | Performance | 1 3 11 6 | 70.48            |
| 5  | Supportability | 1 13 7 | 82.86                  |
|    | Average     |       | **81.90**              |

From the results obtained the highest value is seen from the Functional parameter aspect which has a value of 87.62, while the lowest value is in the Performance parameter aspect. The average value of all aspects is 81.90, which is greater than 80 which means that the application GATA Framework text mining is very good to use.

7. Deployment
At this stage, we can use two models, use RapidMiner Server, or develop applications by choosing the programming language that suits your needs. In some studies that have been done, there are those who use VB.net and PHP at the deployment stage.

8. Conclusion
The amount of research in the field of text mining in Indonesian, the need for pre-processing in Indonesian is very important. GATA Framework is one of the recommended pre-processing device options, seen from the results of the excellent FUPRS. The device supports Data Understanding on the CRISP-DM method or other methods, namely ‘@’ annotation Removal, Remove URL or ‘http’, Tokenization: REGEXP, Transformation Not (Negative), Indonesian Stemming, and Indonesian
Stopwords Removal. In addition, these tools can also be combined with RapidMiner tools, and have online process features and web services (API) that can be linked to other applications.

References

[1] Agastya, I. M. (2018). PENGARUH STEMMER BAHASA INDONESIA TERHADAP PEFORMA ANALISIS SENTIMEN TERJEMAHAN ULASAN FILM. Jurnal Tekno Kompak, 12(1), 18-23.
[2] Fahad Salmeen Al-Obthani, A. A. (2018). TOWARDS CUSTOMIZED SMART GOVERNMENT QUALITY MODEL. International Journal of Software Engineering & Applications (IJSEA), 9(2), 41-50.
[3] M. Hofmann, R. K. (2014). RapidMiner: Data Mining Use Cases and Business Analytics Applications. Boca Ranton: Chapman and Hall/CRC.
[5] Nia Kusuma Wardhani, S. K. (2018). Sentiment Analysis Article News Coordinator Minister of Maritime Affairs Using Algorithm Naive Bayes and Support Vector Machine with Particle Swarm Optimization. Journal of Theoretical and Applied Information Technology, 96(24), 8365-8378.
[6] Nia Kusuma Wardhani, W. G. (2017). IMPLEMENTASI FRAMEWORK MTG DALAM PENGEKSPORAN LAPIOR, GRAFIK MODEL DAN EKSPORT BERKAS MENGGUNAKAN BAHASA PEMROGRAMAN PHP. International Journal of Human Capital Management (IJHCM), 1(02), 81-94.
[7] Tirana Noor Fatyanosa, F. A. (2017). Classification Method Comparison on Indonesian Social Media Sentiment Analysis. Paper presented at International Conference on Sustainable Information Engineering and Technology (SIET).
[8] Siswanto, Y. P. (2018). Classification Analysis of MotoGP Comments on Media Social Twitter Using Algorithm Support Vector Machine and Naive Bayes. Bali, Indonesia: International Conference on Applied Information Technology and Innovation (ICAITI).
[9] Umair Shafique, H. Q. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.
[10] Xing Fang, J. Z. (2015). Sentiment Analysis Using Product Review Data. Journal of Big Data, 2(1), 2-14.