Plagiarism Checker X Originality Report

Similarity Found: 10%

Date: Thursday, November 26, 2020
Statistics: 266 words Plagiarized / 2717 Total words
Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

The 7th International Conference on DV-Xa Method IOP Conf. Series: Materials Science and Engineering 835 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/835/1/012057 1 Text Mining Pre-Processing Using Gata Framework and RapidMiner for Indonesian Sentiment Analysis S Kurniawan 1*, W Gata 1, D A Puspitawati 1, I K S Parthama 2, H Setiawan 3 and S Hartini 1 1Sekolah Tinggi Manajemen Informatika dan Komputer N usa Mandiri, Indonesia 2Universitas Pramita Indonesia, Indonesia 3Sekolah Tinggi Manajemen Informatika dan Komputer B ani Saleh, Indonesia * kurniawan.sgt@gmail.com Abstract.

Research in the field of Text Mining in general sti ll 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 wo rds, emoticons, and Indonesian stemming.

The GATA Framework Text Mining provided is one of t he 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, Evaluati on, 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 T ext Mining in Indonesian, the need for pre-processing in Indonesian is very important. GATA Fr amework is an option for

pre-processing devices that can be combined with Repidminer device s, as seen from the results of the excellent FUPRS. 1. Introduction Research in the field of Text Mining in general sti ll uses text in English, Arabic, China or others language, while for texts in Indonesian is still ve ry limited, so it requires good tools to help Indon esian researchers to conduct research in the field of tex t mining in Indonesian.

Pre-processing needed for t ext mining processes such as deleting '@' notation, 'ht tp' removal, Indonesian stopwords, normalize acronym words, slang words, emoticons, and Indonesi an 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 min ing 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 class ification algorithm commonly used for text mining such as Support Vector Machine (SVM) and Naïve Baye s (NB).

Although the RapidMiner application has provided many menus, for the use of Indonesian Acronym, Indonesian stemming, Indonesian slang The 7th International Conference on DV-Xa Method IOP Conf. Series: Materials Science and Engineering 835 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/835/1/012057 2 words, and others related to the Indonesian text pr e-processing, but is still very limited and require s innovation.

Text mining or Natural Language Processing research , especially Indonesian Language Stemming, already exists in an application called Sastrawi us ing 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 a nd there are still limitations if must be integrated w ith other software.

One of the studies using the Sa strawi application is research conducted by Agastya and Ar tha 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 rese archers in making research in the field of text-min ing of Indonesian language.

The application is built us ing 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 progra m 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 Frame work in 2012 and changed its name to the GATA framework in 2017. GATA Framework has been abl e to overcome various external problems, namely usability, capability, response, security, e xistence, and reliability, as well as internal fact ors, namely ease of syntax or code that is easy to use a nd 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. Exce ll 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 softw are, this study uses the FURPS quality model method which consists of, Functionality, Use, Reali ty, Performance, and Support in the form of descriptive statistics. FURPS (Functionality, Usabi lity, 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 pro cessing data, which consists of several stages, namely Bussiness Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment with a combination of text pre-processin g using the GATA Framework pre-processing application combined with RapidMiner applications t o process words in Indonesian, as shown in figure 1. The 7th International Conference on DV-Xa Method IOP Conf.

Series: Materials Science and Engineering 835 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/835/1/012057 3 Figure 1 . Combination of GATA Framework – Pre-processing an d RapidMiner 2. Business Understanding The use of data in the form of text comments or twe ets 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 a lso known as opinion mining, studies people's sentiments towards certain entities. The Internet i s a resourceful place with respect to sentiment information. From a user's perspective, people are able to post their own content through various soci al media, such as forums, micro-blogs, or online socia l networking

sites [6].

Research related to sentiment analysis that has use d text pre-processing, among others, as has been do ne by Nia Kusuma Wardhani for sentiment analysis in on line news articles related to the coordinator of th e maritime ministry using the Naïve Bayes classificat ion algorithm and Support Vector Machine which was optimized with Particle Swarm Optimization by u sing Rapidminer.

The results of the study obtained the Naive Bayes algorithm which was optimized with PSO having a high accuracy value and being a solution to the problem of sentiment analysis in on line news articles [7]. Other research for sentiment analysis as conducted by Siswanto on research related to the classificati on of comment analysis on social media related to Moto GP, research uses the Naive Bayes classification algorithm and Support Vector Machine to classify co mments on social media and then classify them as analysis of positive or negative sentiment with Rap idminer. The study resulted in accuracy for sentime nt prediction of 95.50% for the Support Vector Machine algorithm [8].

There is also another study conducted by Tirana Noo r Fatyanosa related to the comparison of the classification methods for sentiment analysis in In donesian social media, in addition to using the pre - processing text also uses several classification me thods compared to one another, the classification methods include; Summation, Average on Tweet, Avera ge on Tweet with the objective score, Weighted Average, and Naïve Bayes method.

The results of the se experiments found that Naïve Bayes produced high precision, recall and accuracy values for neut ral and positive sentiments, but it was not good fo r negative sentiment [9]. The overall research carrie d out related to sentiment analysis begins with the The 7th International Conference on DV-Xa Method IOP Conf.

Series: Materials Science and Engineering 835 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/835/1/012057 4 pre-processing phase of the text which then continu es to use the classification method to obtain posit ive or negative results from the dataset used. 3. Data Understanding When using Twitter social media, there are signs su ch as # (hashtag), @ (user), HTTP, and other signs that need to be removed so that text can be used be tter.

While the use of pre-processing methods in Indonesian that can be used is the Indonesian acron ym method, Indonesian stopwords, Indonesian Stemming, and others. Retrieving data from the Twit ter 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 di ctionary facility to change the acronym, and stopwords, but it is still limited to English, Chin ese, and Arabic, while for Indonesian it is still n ot available. The Indonesian text processing stage for the removal of a hashtag, @, HTTP, Indonesian acronym, Indonesian Stopwords, and Indonesian stemm ing can use the GATA Framework application as shown in Figure 3.

The technique commonly used t o pre-processing Indonesian text is @annotation Removal, Remove URL, Tokenization: REGEXP, Transfor mation Not (Negative), Indonesian Stemming, and Indonesian Stopwords Removal. An expl anation 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: htttp://www.gataframework.com/textmining The 7th International Conference on DV-Xa Method IOP Conf.

Series: Materials Science and Engineering 835 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/835/1/012057 5 Table 1. Description techniques on Text mining – Pre-proces sing - GATA Framework No Techniques Description 1 @ Annotation Removal Remove the @ sign and descri ption 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", e tc. 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 ne eded, 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 ne xt step is to do modeling using RapidMiner.

In RapidMiner Pre-processing stages can be done and us e sentiment algorithms such as SVM and NB algorithms. Some pre-processing features commonly u sed on RapidMiner, namely: Tokenize, Filter Stopwords (unnecessary words using a dictionary), F ilter 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 cro ss-validation and T-Test. Figure 4.

Pre-processing Text Using RapidMiner The next step is to make the overall model of all o perators 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 Source: RepidMiner 9.2 The 7th International Conference on DV-Xa Method IOP Conf.

Series: Materials Science and Engineering 835 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/835/1/012057 6 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. F igure 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. Table 2.

Results of the FURPS questionnaire No Aspek Score Result Value/max * 100 1 2 3 4 5 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 The 7th International Conference on DV-Xa Method IOP Conf.

Series: Materials Science and Engineering 835 (2020) 012057 IOP Publishing doi:10.1088/1757-899X/835/1/012057 7 Stopwords Removal. In addition, these tools can als o 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 & App lications (IJSEA) , 9(2), 41-50. [3] M. Hofmann, R. K. (2014). RapidMiner: Data Mining Use Cases and Business Anal ytics Applications . Boca Ranton: Chapman and Hall/CRC. [5] Nia Kusuma Wardhani, S. K. (2018).

Sentiment An alysis Article News Coordinator Minister of Maritime Affairs Using Algorithm Naive Bayes and Su pport Vector Machine with Particle Swarm Optimization. Journal of Theoretical and Applied Information Tech nology , 96(24), 8365- 8378. [6] Nia Kusuma Wardhani, W. G. (2017). IMPLEMENTASI FRAMEWORK MTG DALAM PENGEMBANGAN CRUD, PEMBUATAN LAPORAN, 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 Soci al Media Sentiment Analysis . Paper presented at International Conference on Su stainable 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 (I CAITI). [9] Umair Shafique, H. Q. (2014). A Comparative Stu dy 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 Us ing Product Review Data. Journal of

Big Data , 2(1), 2-14.

INTERNET SOURCES:

<1% - https://www.researchgate.net/post/What_are_the_steps_one_should_follow_to_prepare_a_gold_standard_dataset
<1% - https://ejournal.bsi.ac.id/ejurnal/index.php/paradigma/search/authors?searchInitial=&authorsPage=17
<1% - https://www.researchgate.net/publication/252067921_The_Effect_of_Stemming_on_Arabic_Text_Classification_An_Empirical_Study
<1% - https://www.researchgate.net/publication/4329399_Intelligent_heart_disease_prediction_system_using_data_mining_techniques
<1% - https://www.javatpoint.com/text-data-mining
<1% - https://medium.com/sciforce/text-preprocessing-for-nlp-and-machine-learning-tasks-3e077aa4946e
<1% - https://www.prlog.org/10282839-kdnuggets-poll-rapidminer-again-the-number-1-open-source-data-mining-tool.html
<1% - https://www.encyclopedia.com/places/asia/indonesian-political-geography/indonesia
<1% - http://ufdc.ufl.edu/UFE0041851/00001
<1% - https://es.scribd.com/document/43545809/Application-of-Data-Mining-in-EBusiness-Finance
<1% - https://elibrary.judiciary.gov.ph/thebookshelf/showdocs/1/56650
1% - https://projectstation.co.in/sentiment-analysis-product-rating/
1% - https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2
<1% - https://core.ac.uk/download/pdf/327322709.pdf
<1% - https://rafalab.github.io/dsbook/examples-of-algorithms.html
<1% - https://www.researchgate.net/publication/344059311_Sentiment_Analysis_of_the_Body-Shaming_Beauty_Vlog_Comments
<1% - https://www.researchgate.net/scientific-contributions/2055074793_Haseeb_Qaiser
1% -

https://www.journaltocs.ac.uk/index.php?action=browse&subAction=pub&publisherID=456&journalID=8450&pageb=1
<1% - https://zombiedoc.com/ic-csod.html
1% - https://flylib.com/books/en/2.809.1.46/1/
<1% - https://publik.tuwien.ac.at/publist.php?lang=1&Fak=8&inst=1100&func=0&s..
<1% - http://www.altcancer.net/news/coronavirus2Aug20.htm
<1% - https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/
<1% - https://stackoverflow.com/questions/56308116/should-feature-selection-be-done-before-train-test-split-or-after
<1% - http://jatit.org/volumes/ninetysix24.php
1% - https://sinta.ristekbrin.go.id/journals/detail?id=4532
1% - https://www.semanticscholar.org/author/Fahad-Salmeen-Al-Obthani/1423559392
<1% - https://flyccs.com/jounals/IJSEA/Home.html
1% - https://repository.widyatama.ac.id/xmlui/bitstream/handle/123456789/7865/PAPER%20ARI%20PURNO%20%281%29.pdf?sequence=6
<1% - https://towardsdatascience.com/social-media-sentiment-analysis-part-ii-bcacca5aaa39
<1% - https://www.researchgate.net/publication/326552878_Deep_Leaning_Architectures_and_its_Applications_A_Survey
1% - http://repository.uinsu.ac.id/8706/1/Prosiding%20ICAITI%202018.pdf
1% - https://www.researchgate.net/publication/336522122_Evaluation_of_the_effect_of_learning_disabilities_and_accommodations_on_the_prediction_of_the_stability_of_academic_behaviour_of_undergraduate_engineering_students_using_decision_trees