IATMI22-057

Artificial Intelligence (AI) Natural Language Processing (NLP) Applications For Identifying Suspended Job Root Cause

Dr. Joko Nugroho PHW*[1], Faried Efendi[2], Baruna Satria[3], Yustinawati[4], Ramdhan Ari Wibawa[5], and Muhammad Shafwan Faturrahman[6]
[1, 2, 3,4,5,6] PERTAMINA
* Email: joko.phw@pertamina.com

One of the characteristics of Big Data is data variety. This applied to work-over rig operations, which deliver reports to inform daily and summary rig program and activities, known as tour-report. Tour-report contains unstructured data, which requires extensive, time-consuming effort to read, digest, and extract meaningful information for further handling and analysis. The root causes of suspended jobs are critical to determine remediation strategy. In some cases, suspended jobs with severe problems are forced to be idle wells or long-term closed (LTC). Here are a few examples of suspended job causes, i.e., killing problem or high casing pressure, wellbore assembly overpull or stuck, fishing left in the wellbore, and well mechanical completion issue. This information can be taken by digesting unstructured data available in tour-report to come up with a remediation program, i.e., rework with rig capable of handling high casing pressure, fishing job, and wellbore repair. To gain insight and knowledge discovery of tour report, PT Pertamina Hulu Rokan (PHR) Heavy Oil Asset Optimization Team (HO AOT) and Integrated Optimization Decision Support Center (IODSC) are collaborating to deliver pilot of Artificial Intelligence (AI) Natural Language Processing (NLP) approach to identify suspended job root causes from idle wells (~800 wells in Heavy Oil fields) to come up with a remediation strategy, mainly in supporting the reactivation of LTC wells in the Rokan field program. This study shows the impact of fuzzy string-matching in handling data pre-processing. In this study, several algorithms are combined with the fuzzy string-matching method. Based on this study, fuzzy string-matching can improve the performance of classification by more than 100% in most all algorithms used.

Keyword(s): Tour-report, Text mining, Fuzzy string-matching.

1 Introduction

Unstructured data in the form of text is commonly found in oil industry. In regular activities such as well intervention and drilling jobs, they are typed on a tour-report explaining the detailed information in chronological order. In addition, they contain crucial information about the condition of the well based on rig crew’s findings during the job, including the results of well integrity test, casing collapse, casing parted, and the problem they encountered during the job, such as stuck and killing problem. This information is vital to the engineer, for example, in determining well candidates for heavy well repair programs such as plug abandoned, sidetrack, or replacement well. Moreover, the quantitative result that can be delivered from the tour report would be beneficial in preparing the long-term strategy for the field.

This study examines the preliminary step, which is data pre-processing as a fundamental step in any data-driven projects such as text mining, Machine Learning (ML), or Artificial Intelligence (AI), including in the field of Natural Language Processing (NLP). As a tour-report is prepared by humans, typing errors and misspelling words are inevitable yet lexical meaning can still be understood. In the context of petroleum industry, standard abbreviations are commonly used on tour-reports to express certain repetitive activities to simplify report content. Moreover, many words might have different meanings in the oil industry compared to general usage. Therefore, the structure of the sentences would be helpful for the computer to understand the context and the meaning of sentences. For this purpose, NLP comes to the rescue to extract the information from human language by analyzing text and grammatical syntax since it is equipped with a certain mechanism to accomplish that such as “attention mechanism” in the transformer-based model. However, without proper data pre-processing, the modeling process in data-driven project would be challenging as explained in (Wardana, 2019). As described by this paper, by applying better data pre-processing, the vector's quality significantly increases, which is indicated by accuracy improvement.

2 Theory

2.1 Text Mining

According to (Noshi, 2019), text classification or document classification is also classified as text mining. Text mining is a method that can retrieve information from text-based data and then process it in the next step such as classification with machine learning. There are also several phases of text mining described in (Noshi, 2019), such as collecting unstructured data, solving the incongruities data, transforming it into structured data, and examining the trend or classifying the data to help the people in the decision-making process.

2.2 Text Preprocessing

Text preprocessing is a process that converts text data into a form of data that can be processed with machine learning. According to (Kozhevnikov, 2020) and (Safaei, 2018), text preprocessing consists of tokenizing, filtering, and stemming tasks. Tokenization is defined as a fragmentation process of a sentence or sequence of characters into words called tokens (Noshi, 2019), and it is known as the first phase of text preprocessing (Webster, 1992). Filtering such as removing stop words, is one of the text preprocessing methods that can improve the performance of a classification (Saif, 2014). Stemming is a method that can reduce morphological variants of words into their root form (Hull, 1996).

2.3 Fuzzy String-Matching

Data interpretation of text-based data is very crucial. Especially for a text-based data that have an unstructured characteristic. The incorrect data caused by spelling mistakes can make a piece of misleading information (Onifade, 2011). Fuzzy String Matching uses the Levenshtein distance method to measure the difference between two words or phrases in handling misleading data interpretation based on its similarity (Zhang, 2017). In this study, the fuzzy string-matching will be combined with Decision Tree (TR), Random Forest (FR), XGBoost (XGB), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), and Logistic Regression (LR).

3 Methodology

3.1 *Tour Reports Data*

The dataset consists of 144 tour reports that contain well name and job remarks that contain root causes information of Long Term Closed (LTC) wells. All the reports have unique well names. The job remark is an unstructured text report which was made by the field engineer of the drilling or well intervention team. The LTC root causes analysis consists of 47 suspended jobs due to stuck and overpull, 46 suspended jobs due to killing problems, 18 suspended jobs due to fish left in the wellbore, 12 suspended jobs due to sand, clay, or congeal problems, 11 suspended jobs due to mechanical issues, and 10 suspended jobs due to shallow tag. In this study, the job remark will be the input data and the LTC root cause analysis will be the label or the output of the model. The example of tour reports shows in table 1.

Table 1. Sample Data of Tour Reports

Well Name	Job Remark	LTC Cause Criteria
3RXX-01	KILLED WELL ON TOTAL 1920 BKW, BLEED OF PRESSURE TO FBT GOT HOT WATER ON RETURN. FINAL PRESSURE SITP 20 PSI, SICP 20 PSI. RDMO. WELL PENDING @ 11:30 HRS	Suspended job due to killing problem
5UXX-01	MIRU.CWSI.KW.RECIPROCATED STACK TBG PUMP W/40000 LBS UNSUCCES AND DECISION PENDING WELL.REPORTED TO MR. SITORUS OPERATOR AREA 1/7.	Suspended job due to stuck & overpull
8MXX-01	R/D Rig, MIRU to 8M-XXX. Cont'd MIRU 100%, CDW, TOH Plgr, ND WHC, NU BOP, Shell test BOP, RIH GPPT Tool inside string got tagged at 304 ft, R/D WPF, N/D BOP, TIH Plunger @ 493 ft, Test Connection Tree. Job Suspend.	Suspended job due to shallow tag

The challenge of this dataset is that the job remark was written by many different people, which makes it have no word type characteristic. People often describe something in different words or phrases. Moreover, it has no specific structure to make the job remarks, so the randomness of the tour reports is very high. Besides, the amount of LTC root cause analysis is different. It can make the model biased to major root causes. Because of that, the model must be very agile in handling those problems.

3.2 *Model Development*

The data needs to be processed before it can be used to build the models. Data pre-processing that was used in this research consist of several processes. The first step is tokenizing to split all the words from the job remark. Then remove all the English stop words such as ‘and’, ‘the’, ‘a’, etc. The punctuation and numbers were also removed in this process. The selected words which are considered as an unimportant words will also be removed in this process such as ‘psi’, ‘lbm’, ‘lbs’, etc. After that, the text will be processed by the method named stemming. As explained before, stemming can handle some variations of words that have the same root word.

Vectorization combined with the fuzzy string-matching method is the next step after data pre-processing is completed. The results of this process are pairs of 2 words and pairs of 3 words which were ordered continuously based on the data. After that, the vector will be processed by fuzzy string-matching. Fuzzy string-matching will classify some different vector with high similarity into a vector. This method can handle the typo word that was made by human error in the process of making tour reports. As a comparison to the fuzzy string-matching method, the original vector or the vector without a fuzzy string-matching process will be used to build the models.

Figure 1 below shows the word cloud of 2 pair word that has been filtered with the fuzzy string-matching method. The size of the word correlated with the frequency of the appearance of the words in the tour reports. The bigger display of the words means that the frequency of the words is high. We can see in the figure 1. that "cool well" and "kill well" are relatively big than others even though it is not the biggest size. Those two words are correlated with the number of "Suspended jobs due to killing problem" in the LTC Cause Criteria that have 46 samples of data. The other words in the word cloud are the common words that appear in the tour reports.

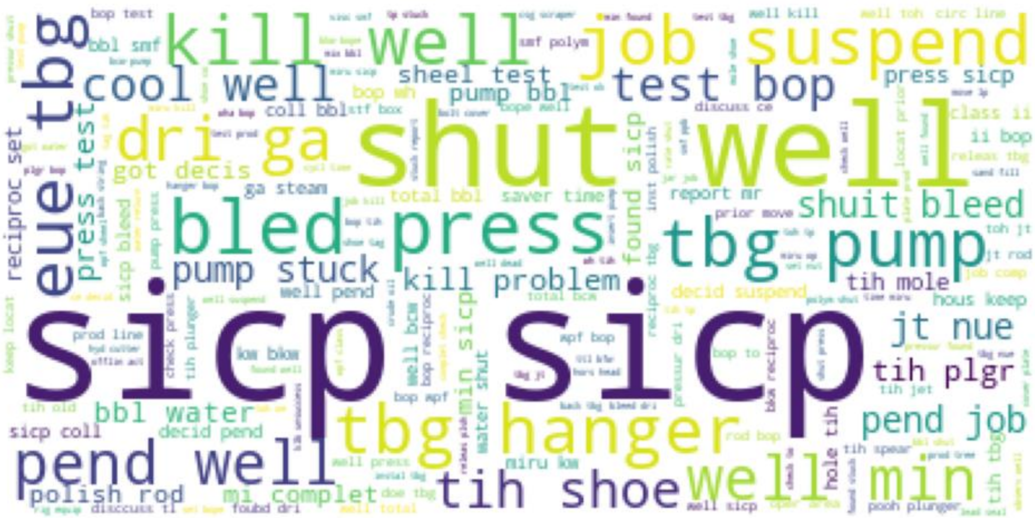

Figure 1. Word Cloud of 2 Pair Word in The Tour Reports

5-fold cross-validation was used to build the model for both methods. This model selection type was used to check the robustness of the model. By using this method, all of the data will be parted into the training set and also the validation set. The illustration of 5-fold cross-validation is shown in figure 2.

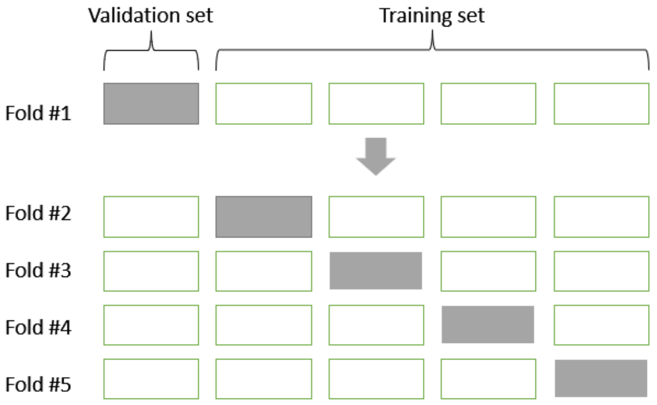

Figure 2. Illustration of 5-Fold Cross-Validation [8] (Berrar, 2018)

The metrics that was used in this study is accuracy. Accuracy can be determined by the formula 1 as follows (Kozhevnikov, 2020).

$$accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \qquad (1)$$

True positive means the number of cases that are recognized correctly in true class, true negative means the number of cases that are recognized correctly in false class, false positive means the number of cases that are recognized incorrectly and belong to false class, false negative means the number of cases that are recognized incorrectly and belong to true class. Where in multi-classification cases, true class means the class that was same as the label or output and false class means the class that was different with the label or output.

4 Results and Discussion

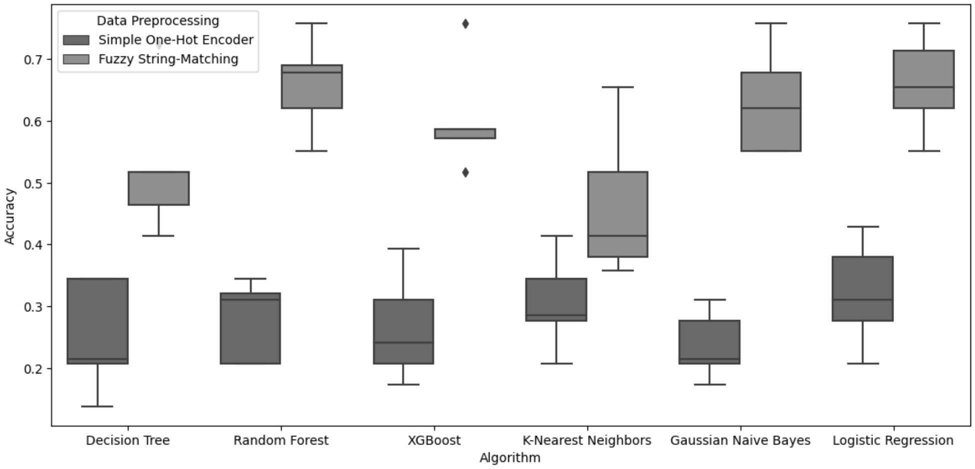

Figure 3. Accuracy Distribution Comparisons of All Algorithms

Figure 3 shows the accuracy distribution comparisons of all the algorithms. The random forest, gaussian naïve Bayes, and logistic regression combined with fuzzy string-matching have been shown as the best model because their accuracy is relatively higher than others. Even though the XGBoost has the lowest variance, the top 3 model stated before has higher accuracy mean than XGBoost. The lowest performance combined with fuzzy string-matching of all the algorithms shows by KNN. KNN has the lowest performance compared to the others. The fuzzy string-matching has improved the KNN model, but it still has the lowest accuracy mean. In general, in figure 3 we can see that the fuzzy string-matching method gives better performance than the simple one-hot encoder for all algorithms.

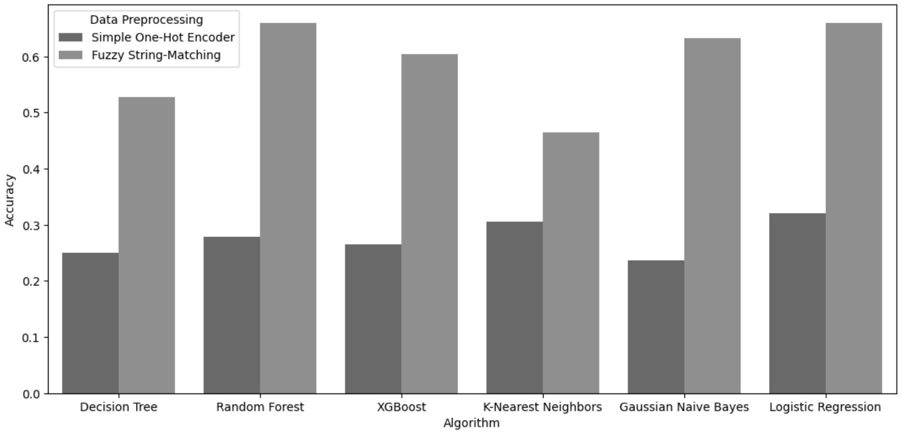

Figure 4. Mean Accuracy Comparisons of All Algorithms

In addition, figure 4 shows the mean accuracy comparisons of all algorithms. It has shown that the fuzzy string-matching method has enabled the model to learn better. The model combined with fuzzy string-matching performs better than the simple one-hot encoder because the fuzzy can handle the typo word caused by human error. Moreover, the fuzzy also can handle the short words made by the report maker. The variation of short words is correlated with the number of people who make the reports. In figure 4, we can see that the fuzzy string-matching method can improve performance by more than 100% in almost all algorithms. The highest improvement of accuracy mean is 168% when fuzzy was combined with the gaussian naïve Bayes algorithm and the lowest improvement was shown by the KNN algorithm by 50%.

5 Conclusions

The classification algorithm that can analyze tour reports can help engineers to work faster. There are so many aspects that must be considered when we are going to analyze the tour reports, and it could take so much time if hundreds of reports must be classified. This research delivers a solution for engineers who want to minimize the effort and maximize the time used for work. The algorithm is quite simple as explained before. However, the value of this algorithm can minimize more than 90% of man work hours to analyze the tour reports.

The method of fuzzy string matching can extract information better than the simple one-hot encoder. The study shows that the fuzzy string-matching method can make the classification have a better performance compared to the simple one-hot encoder, the fuzzy can improve more than 100% in almost all of the algorithms that have been tested in this study.

For the future, there are many improvements opportunity. The most important aspect in building a classification model is the amount of training data. The labeled data is one of the most challenging in building the model. It requires many resources to label the data. Moreover, we also need an expert engineer to label the data, so it can minimize the error that can deliver a piece of misleading information.

Try the other machine learning algorithms that also can improve the possibilities to improve the model's performance. Besides, we can add another feature extraction to the data or use an optimization method to define the hyperparameter of the algorithm. We also can use deep learning to classify tour reports better. Despite all, fuzzy string matching still can be a suitable method for further study because it can significantly improve the model's performance based on this study.

6 References

[1] Kozhevnikov, V., Pankratova, E. 2020. Research of the Text Data Vectorization and Classification Algorithms of Machine Learning. Theoretical & Applied Science. 85. 10.15863/TAS.2020.05.85.106.

[2] Noshi, C., Schubert, J. 2019. A Brief Survey of Text Mining Applications for the Oil and Gas Industry. 10.2523/19382-MS.

[3] Safaei, Saeid, et al. 2018. A brief survey of text mining: Classification, clustering and extraction techniques.

[4] Onifade, O., Onifade, O., Akomolafe, P. 2011. A Fuzzy Search Model for Dealing with Retrieval Issues in Some Classes of Dirty Data. ICIQ 2011 - Proceedings of the 16th International Conference on Information Quality.

[5] Wardana, R., Afriyanto, R., Nasution, D. 2019. Machine Learning Approach in Identifying Wellbore Integrity Issue from Drilling Reports in Mahandini Field. Proceeedings, Indonesian Petroleum Association Forty-Third Annual Convention & Exhibition. IPA19-E-318

[6] Berrar, D. 2018. Cross-Validation. 10.1016/B978-0-12-809633-8.20349-X.

[7] Hull, D. A. 1996. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for* Information Science **47** (1): 70 – 84.

[8] Webster, J., Kit, C. 1992. Tokenization as the initial phase in NLP. 1106-1110. 10.3115/992424.992434.

[9] Saif, H., Fernandez, M., Alani, H. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14). 810-817.

[10] Zhang, S., Hu, Y., Bian, G. 2017. Research on string similarity algorithm based on Levenshtein Distance. 2247-2251. 10.1109/IAEAC.2017.8054419.