

# Multi-Label Text Classification For Automation Soft Competency Assessment Scoring

Alfan Aris Setiawan - 23522311  
Program Studi Magister Informatika  
Sekolah Teknik Elektro dan Informatika  
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung  
E-mail : alfanaris@gmail.com

**Abstract**— Assessment Center is a method which measures individual and group behavior through simulations. Several trained assessors observe and record the participants to demonstrate through the simulations. One of the simulations is called Problem Analysis. Participant asked to write down some solution for certain problem in the questions. The process of evaluating by the assessor somehow can be subjective and takes time to evaluate the whole process. Here we propose automation to grade the Soft Competency. We use multi label text classification to automate the process of evaluating the answer. But somehow as of now, the technology doesn't even near the evaluation from expert. To obtain better results, this study needs to be continued.

**Keywords**—Multi Label Text Classification, Assessment Center, Soft Competency

## I. INTRODUCTION (HEADING 1)

Assessment Center is a method which measures individual and group behavior through simulations. This evaluation is carried out by several trained assessors who observe and record the participant's behavior [1]. Although assessment materials are designed for objectivity, the assessment process tends to be subjective and requires a lot of time and human resources to analyze.

To overcome these limitations and meet the need for a more objective approach, this study is considering using multi-label classification with Machine Learning. This approach allows automation in data interpretation, reducing assessor subjectivity, as well as time and resource efficiency. In this study, a model was developed that was able to classify participants' answers according to the competencies being tested, so that the assessment process became more efficient and accurate.

This study formulates several main problems related to the use of Multi-Label Text Classification in classifying assessment participant answers, first, how to use Multi-Label Text Classification in classifying participant answers, and the second is how to use the classical algorithm in Multi-Label Text Classification to these objectives. This study aims to evaluate whether the use of Multi-Label Text Classification algorithms, can be an effective solution in the task of evaluating assessment participants' answers.

The benefit of this study is to obtain a more objective approach in analyzing data from the assessment center. This study use dataset of assessment participants' answers that have been labeled by assessors. The competency analyzed in this

study is only one with its six Key Behaviors in the Problem Analysis simulation.

This dataset is the result of an assessment center evaluation from 2022 to 2023, which involved tagging evidence of competency by assessors on participants' answers in the Problem Analysis simulation [2]. The unnecessary data will be removed. The participant's answer that has been marked by the assessor along with the key behavior fulfilled from that answer. Each participant's answer can fulfill more than one key behavior, so a special method is needed to build an accurate predictive model based on this dataset.

Text	Tagging
Lorem ipsum dolor sit amet, consectetur adipiscing elit.	KeyBe-1
Lorem ipsum dolor sit amet, consectetur adipiscing elit.	KeyBe-2
Donec vitae leo ac metus suscipit rutrum vel nec nulla.	KeyBe-4
Aliquam erat volutpat. Aliquam erat volutpat.	KeyBe-2
Aliquam erat volutpat. Aliquam erat volutpat.	KeyBe-3
Aliquam erat volutpat. Aliquam erat volutpat.	KeyBe-6

Figure 1. Dataset Example

In the multi-label classification process, it is very important to transform the data to fit a format that can be processed by machine learning algorithms. This data transformation is necessary to handle scenarios where each data instance can have more than one relevant label. In this context, the data transformation carried out aims to facilitate analysis and predictive modeling with multi-label algorithms. With this data structure, we can easily apply multi-label classification techniques because each key behavior (label) has been identified and assigned a value according to its fulfillment. This transformation also allows the model to learn from answer patterns in the context of fulfilling multiple key behaviors at once, increasing accuracy and effectiveness in multi-label predictions.

Text	KeyBe-1	KeyBe-2	KeyBe-3	KeyBe-4	KeyBe-5	KeyBe-6
Lorem ipsum dolor sit amet, consectetur adipiscing elit.	1	1	0	0	0	0
Donec vitae leo ac metus suscipit rutrum vel nec nulla.	0	0	0	1	0	0
Aliquam erat volutpat. Aliquam erat volutpat.	0	1	1	0	0	1

Figure 2. Transformed Data

The Figure 3 shows the number of texts based on the number of "Key Behaviors" tagged per text. It can be seen that the majority of texts (3074 texts) only have one "Key Behavior" tagged, followed by texts that have two "Key Behaviors" totaling 1701 texts. Meanwhile, the number of texts containing three "Key Behaviors" decreased significantly to 276 texts. Only a few texts have four "Key

Behaviors" (15 texts) and almost no texts have five "Key Behaviors" (1 text). These data indicate that most of the texts in the dataset tend to be classified by one or two key behaviors.

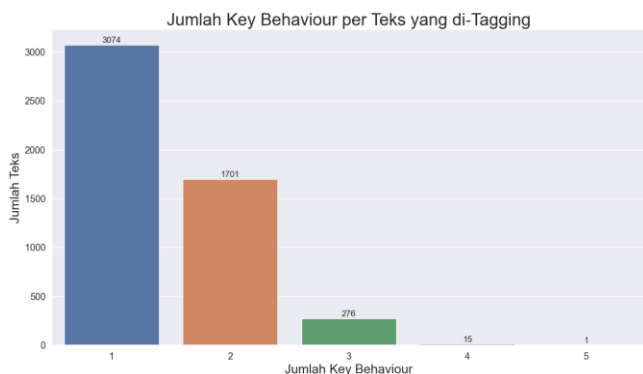


Figure 3. Count of Key Behaviour per Text

On the next step, raw text is converted into a form that can be understood and analyzed by machines [3]. This process begins with parsing, where the text is broken down into smaller structures. Then, case folding changes all letters to lowercase and tokenizing separates the text into its smallest meaningful units. Filtering removes unwanted words, and stemming reduces words to their basic forms.

The final stage is embedding with TF-IDF, where a numerical value is assigned to each word based on its frequency of occurrence and a weight reflecting its importance. The result of this process is a vector representation of text that can be used for various text processing tasks.

## II. RELATED WORK

Heider et al. [4] discusses how to improve the accuracy of HIV-1 drug resistance prediction by using multilabel classification models and cross-resistance information. The model created by training one binary classifier for each of the five drugs, turning the multi-output classification problem into five single-output problems that could be solved individually. Then, another research classifies film genres with 27 main genres using synopsis and storyline textual information. The multi-label classification approach used is binary relevance, same with Heider et al. and adds a method, namely label powerset [5].

Another research was conducted with the aim of classifying essay answers based on the STAR method using the Multi-Label K-Nearest Neighbor (MLKNN) algorithm to achieve consistent and accurate automatic scoring. The MLKNN algorithm was chosen because of its ability to handle multi-label classification, although it has several shortcomings such as low accuracy on small datasets [6].

## III. PROPOSES APPROACH

This study explores three approaches to address multilabel classification issues: Binary Relevance, Label Powerset, and Adapted Algorithm. Each approach has unique characteristics and methods that provide insights into their performance in various scenarios. Binary Relevance addresses independent classification issues by using standard classification algorithms like Logistic Regression, SVM, and Decision Trees. Label Powerset combines labels as a single label, improving the relationship between labels. Adapted Algorithm is

implemented using Multi Label k-Nearest Neighbor (MLkNN), providing a different perspective on classification.

### A. Binary Relevance

For every label, train a single single-label classifier. Every classifier forecasts if a given label will be present in a given data example or not. The total of all classifiers' predictions is the final prediction for every label. Logistic Regression, Decision Tree Classifier, and k-NN used in this study.

Text	KeyBe-1
Lorem ipsum dolor sit amet, consectetur adipiscing elit.	1
Donec vitae leo ac metus suscipit rutrum vel nec nulla.	0
Aliquam erat volutpat. Aliquam erat volutpat.	0

Text	KeyBe-2
Lorem ipsum dolor sit amet, consectetur adipiscing elit.	1
Donec vitae leo ac metus suscipit rutrum vel nec nulla.	0
Aliquam erat volutpat. Aliquam erat volutpat.	1

Text	KeyBe-3
Lorem ipsum dolor sit amet, consectetur adipiscing elit.	0
Donec vitae leo ac metus suscipit rutrum vel nec nulla.	0
Aliquam erat volutpat. Aliquam erat volutpat.	1

Figure 4. Illustration of Binary Relevance

### B. Label Powerset

Rename every unique label combination to a new one. Next, train a single-label classifier for each new label. The last prediction for each label is the classifier prediction according to the new label that results from the combination of the aforementioned labels.

Text	Alias-1	Alias-2	Alias-3
Lorem ipsum dolor sit amet, consectetur adipiscing elit.	1	0	0
Donec vitae leo ac metus suscipit rutrum vel nec nulla.	0	1	0
Aliquam erat volutpat. Aliquam erat volutpat.	0	0	1

KeyBe-1 & KeyBe 2	Alias-1
KeyBe-4	Alias-2
KeyBe-2 & KeyBe 3 & KeyBe 6	Alias-3

Figure 5. Illustration of Label Powerset

### C. Adapted Algorithm

This technique uses a single-label classification algorithm to be improved over time in order to quickly achieve multi-label classification. The examples of this approach are Multi-Label k-Nearest Neighbors (ML-kNN), Multi-Label Support Vector Machines (ML-SVM), a multi-Label Hierarchical ARAM Neural Network (MLARAM), etc. In this study we use Multi-Label k-Nearest Neighbors (ML-kNN) with 5 different k parameters.

## IV. EXPERIMENTS AND RESULTS

### A. Binary Relevance

Results obtained using the Binary Relevance approach and three algorithms used.

Pendekatan	Algoritma	Accuracy	Hamming Loss
Binary Relevance	Logistic Regression	24.98%	<b>19.20%</b>
	Decision Tree	22.81%	22.42%
	kNN (k=3)	<b>29.72%</b>	21.35%

Figure 6 Result of Binary Relevance

## B. Label Powerset

Results of the experiment using the Label Powerset pendekatan and three used algorithms.

Pendekatan	Algoritma	Accuracy	Hamming Loss
Label Powerset	Logistic Regression	29.45%	20.80%
	Decision Tree	26.50%	23.36%
	kNN (k=3)	<b>29.72%</b>	21.35%

Figure 7 Result of Label Powerset

## C. Adapted Algorithm

Results of using the Adapted Algorithm with five configurations, which are carried out on the number k, are as follows.

Pendekatan	Algoritma	Accuracy	Hamming Loss
Adopted Algorithm	MLKNN (k=2)	22.68%	24.70%
	MLKNN (k=3)	27.35%	20.66%
	MLKNN (k=5)	26.63%	20.11%
	MLKNN (k=7)	26.89%	20.16%
	MLKNN (k=10)	26.43%	19.77%

Figure 8 Result of Adapted Algorithm

## D. Analysis and Evaluation

Based on the whole sample, the relative accuracy is 22.68% for the min value and 29.72% for the highest value. Conversely, the relative tinggi of Hamming Loss is located between 24.70% and 19.20%, which is the best score. The high accuracy indicates that the model is not able to make accurate predictions in multi-label scenarios. In multi-label classification, accuracy is not always the most representative metric since every label is required for every true occurrence to be true. On the other hand, a higher Hamming loss (near 0) indicates better performance. Hamming loss indicates the presence of prediction errors, however it is more informative than the accuracy in multilabel contexts.

Even though the Hamming Loss is quite good, the model cannot be relied upon to perform well. This is further supported by a high accuracy score. One factor that contributes to the decreasing skor Hamming Loss is accuracy, which can be attributed to inaccurate data. Out of 5067 instances with 6 labels, there are 30.402 available labels. Subsequently, a more thorough analysis revealed any data imbalances. This can result in a somewhat inaccurate accuracy model; conversely, if there is a label 0 with a large number, the likelihood of Hamming Loss is reduced.

TRUE	7369	19.51%
FALSE	30402	80.49%

Figure 9 The Imbalance Data

In addition to unbalanced data, the dataset's tekst variation is also very high. There is a text with only six characters. The most abundant character count, however, is 250. Such non-seragam tagging makes the model difficult to read in the text. If six characters represent two words, it will be difficult to classify those two words based on how they appear in a large text corpus.

	data
count	5067.000000
mean	104.373199
std	51.033719
min	6.000000
25%	63.500000
50%	100.000000
75%	142.000000
max	250.000000

Figure 10 Statistic of The Data

Asesor's report is a highly contextualized report regarding the client's statement. The use of non-contextual embedding in this study may be the only factor that has to be taken into consideration. Theoretically, using contextual embedding—either word or sentence embedding—will yield better results. With a model that can "menangkap makna" from existing texts, it is expected that the model will perform better while performing classifications.

In the next section, the data must be entered in an imbang manner for each label. Label mismatch in one sample, whether it be test or train data, and even label mismatch in one sample of data also presents a challenge for the model. Utilizing "stratify" in split data will improve model performance when either oversampling or undersampling is used.

Finally, using the hyperparameter adjustment requires further investigation. Utilizing Grid Search will improve model performance by reducing function loss during training, allowing the model's performance to be improved while interpreting data.

## V. CONCLUSIONS AND FUTURE WORK

Based on the results of the study that has been completed, the following can be concluded that it is still necessary to conduct more in-depth research in order to determine a suitable analytical solution when using Multi Label Classification in automatic patient assessment. Based on the research conducted, the Binary Relevance hypothesis is the most promising hypothesis in the study of otomatising employee turnover. As of right now, the limitations of psychology as a field cannot be fully replaced by technology.

Here are some guidelines for this research as well as future research, which are use both Over Sampling and Under Sampling in subsequent research to identify labels with somewhat more difficult to predict. The number of characters in the text needs to be varied, either by pre-processing or during the assessment center's activities by the panelists, so that the number of characters that are not broken always falls between the minimum and maximum. By using contextual embedding instead of non-contextual embedding, the model may better understand the data rather than just looking at the total number of data points. Experimentation with grid search and hyperparameter adjustment is advised in further research to maximize model performance.

## REFERENCES


- [1] D. E. Rupp *et al.*, "Guidelines and Ethical Considerations for Assessment Center Operations," *J. Manage.*, vol. 41, no. 4, hal. 1244–1273, 2015, doi: 10.1177/0149206314567780.

- [2] PLN, *Direktori Kompetensi PT PLN (Persero)*, VIII. Jakarta: PLN, 2021.
- [3] J. Camacho-Collados dan M. T. Pilehvar, "On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis," *EMNLP 2018 - 2018 EMNLP Work. BlackboxNLP Anal. Interpret. Neural Networks NLP, Proc. 1st Work.*, hal. 40–46, 2018, doi: 10.18653/v1/w18-5406.
- [4] D. Heider, R. Senge, W. Cheng, dan E. Hüllermeier, "Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction," *Bioinformatics*, vol. 29, no. 16, hal. 1946–1952, 2013, doi: 10.1093/bioinformatics/btt331.
- [5] S. Kumar, N. Kumar, A. Dev, dan S. Naorem, "Movie genre classification using binary relevance, label powerset, and machine learning classifiers," *Multimed. Tools Appl.*, vol. 82, no. 1, hal. 945–968, 2023, doi: 10.1007/s11042-022-13211-5.
- [6] J. C. Larian dan Georgio Chenayan, "The Implementation of Multi Label K- Nearest Neighbor Algorithm To Classifying Essay Answers," *J. Inf. Syst. Technol. Eng.*, vol. 1, no. 3, hal. 89–94, 2023, doi: 10.61487/jiste.v1i3.38.

## PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 9 Juni 2024

  
ALFANI ARIS SETIAWAN  
23522311