The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.4264

Question Similarity Detection in Indonesian Language Consumer Health Forums with Feature-based Binary Classification Approach

Eka Putri Irianti[1], Alfan Farizki Wicaksono[2]
eka.putri13@ui.ac.id[1], alfan@cs.ui.ac.id[2]
[1,2] Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia

Article Information

Received : 7 Jul 2024
Revised : 12 Jul 2024
Accepted : 8 Aug 2024

Keywords

Question similarity detection, Binary classification, Ensemble boosting, ADASYN, SMOTE

Abstract

Two questions are considered similar if the same response can be given to both. Due to the increase in users of consumer health forums, a growing number of similar questions are not being adequately answered. Identifying duplicate questions in online medical Question Answering (QA) forums offers several advantages for users and medical professionals. Therefore, it is crucial for online medical QA forums to identify similar questions to provide relevant and useful answers. This study examines a feature-based binary classification method for detecting similar questions in the Indonesian consumer health domain. The results indicate that the feature-based classification approach using the CatBoost model yields the best performance. The research also explores techniques to address class imbalance in the dataset, finding that imbalanced learning technique such as ADASYN and SMOTE results in improved classification performance. This study also analyzes discriminative features for identifying semantic similarity between question pairs, concluding that a combination of distance, medical, and encoding features produce the best results.

A. Introduction

In online consumer health forums, users can post detailed inquiries to receive precise answers. However, as the user base of these medical Q&A platforms expands, the volume of queries significantly surpasses the capacity of qualified health professionals—specifically, doctors—to provide answers [1]. Furthermore, the proliferation of duplicate questions results in numerous inquiries remaining inadequately addressed. Consequently, the detection of similar questions emerges as a critical issue in online health Q&A forums.

Identifying similar questions in these forums is essential not only for users seeking information but also for the health experts providing answers. One of the primary advantages of detecting duplicate questions is the reduction in user search time [2]. Additionally, recognizing similar questions allows medical professionals to leverage previous answers to address new inquiries effectively [3].

The identification of duplicate questions can be approached as a binary classification problem, involving comparing and determining whether two questions are semantically similar [4],[5]. This binary classification can be implemented using feature-based techniques with conventional machine learning models [2],[6],[7] or through end-to-end deep learning models [5], [8]. Such methodologies are applicable for detecting similar questions within the health domain.

This research aims to explore a feature-based binary classification approach, examining various features that can be extracted from annotated question pair datasets to identify similar questions in Indonesian-language consumer health forums. This study employs a boosting ensemble model which has been proven to surpass the performance of traditional models [9] and reduce variance while producing more stable and accurate predictions [10]. However, the dataset used in this study exhibits class imbalance, with fewer instances of similar question pairs compared to dissimilar ones. This imbalance presents challenges for evaluating the classifier's performance. Therefore, this study will also investigate imbalance learning techniques to address class imbalance in the dataset.

B. Related Works

The identification of semantically similar questions through a feature-based approach involves designing various textual attributes, including topic similarity, lexical similarity, and syntactic features [5]. Prior research proposed the calculation of cosine similarity across four discrete components—title, description, topic, and tags—resulting in a composite score derived from these weighted components [11]. Subsequent studies have enhanced the methodology of [11] by integrating information from titles, bodies, and tags to capture comprehensive textual data from questions, thereby mitigating the risk of overlooking synonymous terms distributed across different sections of question pairs. This research extracted features such as cosine similarity, term overlaps, entity overlaps, entity type overlaps, and WordNet similarity for training classification models [12].

Jabbar et al. [2] implemented binary classification using Gradient Tree Boosting (GTB), an ensemble learning technique, to detect similar questions. Their research utilized 40 selected features that encapsulated both semantic and structural similarities between question pairs. These features included traditional and non-traditional distance measures such as TF-IDF distance, Word Mover's Distance (WMD), graph-based

structural similarity measures, and embeddings-based distances like Word2Vec and Doc2Vec.

Conversely, Ansari & Sharma [6] employed GloVe word embeddings and leveraged 300-dimensional vectors from Google News to extract various distance and text-based features. Although 28 features were initially extracted, only the 20 most effective features were retained for detecting similar questions. This study utilized ensemble boosting algorithms, such as XGBoost, and a deep learning model, specifically LSTM.

In machine learning, ensemble models can outperform traditional models [9] and also deep learning models on tabular data [10]. Besides better performance, ensemble boosting models also require minimal hyperparameter tuning. Additionally, ensemble methods can reduce variance, resulting in more stable and accurate predictions [10]. They also offer advantages in terms of statistical robustness, computational efficiency, and representational capacity [13].

C. Research Method

Dataset

In this research, the dataset utilized comprises annotated question pairs without incorporating any supplementary information (e.g., user search history). The dataset originates from [14], which provides a test dataset in Indonesian for the health domain. This data set includes queries paired with several relevant (positive) questions and several irrelevant (negative) questions. The dataset construction was undertaken by [14] and was derived from Indonesian health Q&A websites.

The dataset consists of 2437 annotated question pairs employed as both training and testing data across all experimental phases of this research. The training and testing split follows a 70:30 ratio, with 1729 pairs allocated to training and 708 pairs to testing. The class distribution within the training and test sets, with non-similar to similar question pairs, is 88:12 and 75:25 respectively, highlighting a class imbalance in the dataset.

For the task of similar question detection using a binary classification approach, the annotation labels of the question pairs were adjusted accordingly. Non-relevant annotations were assigned a label of 0 (zero) for non-similar, whereas partially relevant and highly relevant annotations were labeled as 1 (one) to denote similar question pairs. Additionally, the question text, which includes the title and content, was preprocessed and concatenated into a single text string with a `[SEP]` separator between the title and content. Following preprocessing, the dataset underwent feature engineering for the purpose of feature extraction.

Feature Engineering

During the feature engineering process, the question text dataset will be vectorized using TF-IDF and several pretrained language encoding models to compute similarity features. A total of 38 features were successfully extracted and will be tested in this study. This research will be analyzed 16 out of 28 features proposed by [6]. These selected features are as follows:

Table 1. Features Used in this Study from [6]

No.	Feature	Feature Type
1	Difference in the length of questions	Basic
2	Qratio	Fuzzy
3	Wratio	Fuzzy
4	Partial ratio	Fuzzy
5	Token set ratio	Fuzzy
6	Token sort ratio	Fuzzy
7	Partial token set ratio	Fuzzy
8	Partial token sort ratio	Fuzzy
9	Word mover's distance (WMD)	Distance
10	Cosine distance	Distance
11	Minkowski distance:	Distance
12	Cityblock distance	Distance
13	Euclidean distance	Distance
14	Jaccard distance	Distance
15	Canberra distance	Distance
16	Braycurtis distance	Distance

This study also includes BM25 score as a feature, and proposes additional textual features, namely:

The ratio of length disparity to total sentence length, calculated by the formula:

$$d(Q_1, Q_2) = \frac{|Q_1 - Q_2|}{Q_1 + Q_2} \quad (1.1)$$

with Q_1 and Q_2 being the number of words in question sentence 1 and question sentence 2.

The ratio of the length dispasrity to the average sentence length of question pair, calculated as follows:

$$d(Q_1, Q_2) = \frac{|Q_1 - Q_2|}{\left(\frac{Q_1 + Q_2}{2}\right)} \quad (1.2)$$

In addition to the above features, this study also proposes new features specific to the healthcare domain, based on the presence of medical terms in the question sentences. Firstly, researchers extract medical terms from questions by referring to a list of medical terms[1]. Next, the extracted medical terms from each question are

[1] https://univindonesia-my.sharepoint.com/:x:/g/personal/raniah_nur_office_ui_ac_id/EbZeMDTuRHVEprQ_azswu8EBNqEyQJP0CzHK_vPR7o0Lw?e=DPu4hy

combined (joined) into one sentence. Finally, the combined medical terms from each pair of questions will be included to obtain the features below.

Intersection of medical terms sets, calculated by:

$$d(MQ_1, MQ_2) = MQ_1 \cap MQ_2 \quad (1.3)$$

with MQ_1 and MQ_2 are sets of medical terms on question sentence 1 and question sentence 2.

Jaccard score of intersection of medical terms sets, calculated by the formula:

$$d(MQ_1, MQ_2) = \frac{MQ_1 \cap MQ_2}{MQ_1 \cup MQ_2} \quad (1.4)$$

The ratio of intersection of medical terms sets to the union of all terms in a question pair, that is:

$$d(MQ_1, MQ_2) = \frac{MQ_1 \cap MQ_2}{Q_1 \cup Q_2} \quad (1.5)$$

with MQ_1 and MQ_2 are sets of medical terms on question sentence 1 and question sentence 2 and Q_1 and Q_2 are sets of terms on question 1 and question 2.

This study also proposes additional features involving encoding from a pretrained language model. From this encoding, Euclidean distance and Cosine Similarity are computed, resulting in features on 0.

Table 2. Features Extracted from Encoding of Pre-trained Language Model

Model	Features
firqaaa/indo-sentence-bert-base[2]	Euclidean distance and Cosine Similarity
indolem/indobertweet-base-uncased[3]	Euclidean distance and Cosine Similarity
stevenwh/indobert-base-p2-finetuned-mer-80k[4]	Euclidean distance and Cosine Similarity
cahya/distilbert-base-indonesian[5]	Euclidean distance and Cosine Similarity
thonyyy/pegasus_indonesian_base-finetune[6]	Euclidean distance and Cosine Similarity
panggi/t5-base-indonesian-summarization-cased[7]	Euclidean distance and Cosine Similarity

This study also proposes RAKE [15] for extracting keyphrases to generate new features. Keyphrase extraction involves using custom stopwords composed of words or phrases with low IDF values, as well as extracting keyphrases without any additional stopwords. The extracted keyphrases are concatenated and then vectorized using TF-IDF. Finally, Euclidean and Cosine Similarity distances are computed based on these word vectors, resulting in the features listed in 0**Error! Reference source not found.**.

[2] https://huggingface.co/firqaaa/indo-sentence-bert-base
[3] https://huggingface.co/indolem/indobertweet-base-uncased
[4] https://huggingface.co/stevenwh/indobert-base-p2-finetuned-mer-80k
[5] https://huggingface.co/cahya/distilbert-base-indonesian
[6] https://huggingface.co/thonyyy/pegasus_indonesian_base-finetune
[7] https://huggingface.co/panggi/t5-base-indonesian-summarization-cased

Table 3. Keyphrases Features Extracted from RAKE

Keyphrases	Features
With Custom Stopwords	Euclidean distance and Cosine Similarity
Without Custom Stopwords (original)	Euclidean distance and Cosine Similarity

Feature Importance

To understand how selected features influence classification performance and to identify discriminative features in question similarity detection, this research will analyze performance of these feature combinations:

Table 4. Feature Combinations Analyzed in Feature Importance

No.	Feature Combination	Description
1	Combination of All Features	The combination includes all features that have been extracted.
2	Combination of Textual Features	Eleven textual features such as difference in the length of question pairs, ratio of length disparity to total sentence length, Jaccard score, ratio of length disparity to average sentence length of question pair, and fuzzy features (Qratio, Wratio, Partial Ratio, Token Set Ratio, Token Sort Ratio, Partial Token Set Ratio, Partial Token Sort Ratio) obtained using the fuzzywuzzy[8] library.
3	Combination of Distance-based Features	Eight distance features and similarity scores encompassing BM25, WMD (Word Mover's Distance) and distance features computed from TF-IDF vectorization of question sentences, i.e. Cosine, Euclidean, Minkowski, Canberra, Braycurtis, and Manhattan distance.
4	Combination of Medical Features	Three medical features: intersection of medical terms sets, Jaccard score of intersection of medical terms sets, and the ratio of intersection of medical terms sets to the union of all terms in a question pair, utilized in machine learning classification models.
5	Combination of *Encoding* Features	Cosine Similarity and Euclidean distance features are derived from encoding of several pretrained language models detailed in Table 2.
6	Combination of Keyphrases Features	Cosine Similarity features and Euclidean distance extracted from keyphrases vectorization using TF-IDF.
7	Ablation Study	Each feature combination's impact on classification performance is assessed by systematically removing them from the analysis one by one.

Feature-based Binary Classification

This study explores a binary classification approach to identify similar questions. Formally, similarity detection between questions Q_i and Q_j is framed as estimating the binary random variable $P(Y = 1|Q_i, Q_j, \theta)$ where θ denotes the model parameters learned from data. When presented with a new question Q_i (Query Question), the system computes $P(Y = 1|Q_i, Q_j, \theta)$ for all Q_j (Candidate Questions) in the dataset.

Based on prior research, this study experiments with various ensemble boosting models. In ensemble boosting, base models are sequentially trained, with each subsequent model aimed at correcting errors from previous models until performance plateau. During each iteration, training sample weights are adjusted.

[8] https://pypi.org/project/fuzzywuzzy/

Final predictions aggregate weighted votes from base models. Specifically, this research investigates three ensemble boosting models: AdaBoost, a traditional method, and XGBoost and CatBoost, both gradient boosting machines [16].

AdaBoost: AdaBoost maintains a weighted distribution over training data [17]. Initially, all weights are uniform, with subsequent iterations increasing weights of misclassified samples. This mechanism compels weak learners to focus on challenging instances. Each weak learner seeks a suitable weak hypothesis under the current data distribution. AdaBoost adjusts sample weights in proportion to prediction accuracy and indirectly proportional to classification errors [16]. Moreover, final predictions are weighted based on classifier accuracy during training.

XGBoost: XGBoost is a scalable machine learning system tailored for boosting trees. Its scalability features include sparsity-aware split discovery, weighted quantile sketching for approximate weight computation, and cache-aware data block processing for efficiency with large datasets [18]. To mitigate overfitting, XGBoost employs regularization, shrinkage, and feature subsampling strategies. It adopts an exact greedy algorithm for optimal tree splitting.

CatBoost: CatBoost implements gradient boosting using binary decision trees [19]. These trees recursively partition feature spaces into nodes based on splitting attributes, with leaf nodes providing class label predictions. CatBoost leverages Target Statistics (TS) for categorical feature treatment and ordered boosting. It integrates permutation-based strategies at different gradient boosting stages to address prediction shifts stemming from target leakage. These innovations elevate CatBoost's performance beyond that of other boosting models.

Imbalanced Learning Techniques

The quantity of samples per class significantly impacts model performance [20]. Increasing sample size diminishes the marginal gains in model performance and reduces the sensitivity of error estimation to intra-class bias-variance. Conversely, errors in approximating inter-class bias-variance tend to escalate, potentially leading to classification errors in trained models.

In addressing class imbalance, [21] introduced Synthetic Minority Over-Sampling Technique (SMOTE), a method that generates synthetic samples for the minority class by applying operations in the feature space rather than directly in the data space. SMOTE enhances the minority class by creating synthetic samples alongside random K-nearest neighbors. This study explores several oversampling techniques available in the imbalanced-learn API [22], including SMOTE, RandomOverSampler, ADASYN, and variants of SMOTE like BorderlineSMOTE, SVMSMOTE, and KMeansSMOTE.

Evaluation Metrics

The choice of metrics plays a pivotal role in evaluating our models, shaping how we measure their performance compared to each other and to baseline standards.

Accuracy: These metrics are foundational for evaluating text classification models. Accuracy measures the proportion of correctly predicted samples without distinguishing between positive and negative predictions.

Precision, Recall, and F1 Score: Precision evaluates the correctness of positive predictions, while recall assesses the coverage of actual positive samples. F1 Score is the harmonic mean of precision and recall. These metrics are crucial for evaluating classification on imbalanced test data, where the majority of samples belong to one specific class [23], [24]. In imbalanced datasets, precision reflects the accuracy of the minority class, while recall measures its coverage. A classifier performs optimally when precision, recall, and F1 score reach 1, and poorly when all three metrics are 0.

D. Result and Discussion

Feature-based Binary Classification Analysis

The experimental results are presented in Table 5. Precision, recall, and F1-Score values reported in the table are for the 'similar' label, with the best results highlighted in bold. Overall, using all extracted features, CatBoost model achieved the highest classification performance with an accuracy of 0.79. Not only superior in accuracy, but the CatBoost model also achieved the highest precision, recall, and F1-score values, reaching 0.75, 0.21, and 0.32, respectively. These findings are consistent with recommendation from [25] to incorporate CatBoost as a predictive model. On the other hand, the XGBoost model exhibited the lowest classification performance with accuracy, precision, recall, and F1-score of 0.77, 0.66, 0.13, and 0.22, respectively. This aligns with findings from [26] that CatBoost outperforms XGBoost.

Table 5. Performance of Feature-based Classification in Detecting Similar Questions

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.77	0.66	0.13	0.22
AdaBoost	0.78	**0.75**	0.14	0.23
CatBoost	**0.79**	**0.75**	**0.21**	**0.32**

The experimental results indicate that feature-based classification using the CatBoost model achieved the best performance compared to other models, albeit not yet optimal. This could be attributed to class imbalance issues. Therefore, we explore imbalanced learning techniques application on the training set. Furthermore, we conduct an ablation study to assess the impact of selected features on performance and identify discriminative features crucial for detecting similar questions.

Implementation of Imbalanced Learning Technique Analysis

The next step involved experimenting with the application of imbalanced learning techniques. In this study, imbalanced learning techniques were applied exclusively to the training set for feature-based classification experiments using ensemble boosting models across all extracted features. The experiments with imbalanced learning techniques were conducted using bootstrapping with 20 resampling iterations for each technique. Subsequently, each experiment was evaluated on the same test set. The experimental results with imbalanced learning techniques presented here represent the averages obtained from 20 iterations for each technique and ensemble boosting model.

The application of imbalanced learning techniques can alter the sample size and class distribution within a dataset. These changes vary across different imbalanced learning methods. For instance, oversampling with ADASYN shifts the distribution of similar question pairs in the dataset, increasing the class proportion from 37% to 49%, as illustrated in Figure 1. Specifically, ADASYN adds 1265 samples to the training set, expanding it from an initial 1729 to 2994 samples. Similarly, KMeanSMOTE oversampling adds 1310 similar question pairs, resulting in a dataset size of 3039 samples. On the other hand, RandomOverSampler, SMOTE, BorderlineSMOTE, and SVMSMOTE each add 1305 annotated similar samples to the dataset. The total dataset sizes after applying imbalanced learning techniques are summarized in Table 6.

Figure 1. Class Distribution in the Dataset Before and After Applying Imbalanced Learning Techniques

Table 6. Dataset Distribution Before and After Applying Imbalanced Learning Techniques

Dataset	All	Similar	Not similar
Test	708	175	533
Training	1729	212	1517
ADASYN	2994	1477	1517
RandomOverSampler	3034	1517	1517
SMOTE	3034	1517	1517
BorderlineSMOTE	3034	1517	1517
SVMSMOTE	3034	1517	1517
KMeansSMOTE	3039	1522	1517

The feature-based classification results with imbalanced learning techniques in Table 7 indicate that, overall, imbalanced learning methods tend to decrease accuracy, except in experiments involving the CatBoost model with ADASYN, BorderlineSMOTE, and SMOTE techniques. Additionally, these techniques generally reduce precision values. However, imbalanced learning techniques typically enhance recall and F1-score across all experiments. Exceptions in recall and F1-score values were observed only in experiments using CatBoost and XGBoost with SVMSMOTE.

Table 8 demonstrates that the best classification performance is achieved when the training set undergoes oversampling using ADASYN and SMOTE techniques. Specifically, CatBoost model performance with these imbalanced learning techniques reached an accuracy and F1-score of 0.81 and 0.60, respectively, in feature-based binary classification for detecting similar question pairs in an Indonesian health forum dataset. Conversely, applying SVMSMOTE to the training set for classification with the AdaBoost model showed the lowest performance, with an accuracy of only 0.51 and F1 score of 0.24.

These results also indicate that the applied imbalanced learning techniques can significantly impact the classification model's performance. Imbalanced learning techniques have the potential to introduce noise or lead to overfitting, which can affect classification performance [27]. This is evident in the comparison of additional sample counts generated by ADASYN and SVMSMOTE as shown in Table 6. Despite SVMSMOTE generating more additional samples, its performance significantly underperforms compared to ADASYN.

These results indicate that applying imbalanced learning techniques can significantly impact the performance of classification models. Incorrect choices of these techniques have the potential to introduce noise or lead to overfitting, which can adversely affect classification performance [27]. This is demonstrated by comparing the additional sample counts generated by ADASYN and SVMSMOTE, as shown in Table 6. Despite SVMSMOTE generating more additional samples, its performance significantly underperforms compared to ADASYN."

Table 7. Feature-Based Classification Performance by Model and Imbalanced Learning Technique

Imbalanced Learning Technique	CatBoost				XGBoost				AdaBoost			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
None	0.79	0.75	0.21	0.32	0.77	0.66	0.13	0.22	0.78	0.75	0.14	0.23
ADASYN	**0.81**	0.61	**0.59**	**0.60**	0.74	0.48	0.80	0.60	0.70	0.44	0.88	0.59
BorderlineSMOTE	0.80	0.62	0.51	0.56	0.76	0.51	0.77	0.61	0.75	0.50	0.83	0.62
KMeansSMOTE	0.77	0.57	0.22	0.31	0.72	0.37	0.20	0.26	0.71	0.38	0.27	0.31
RandomOverSampler	0.79	0.61	0.39	0.48	0.72	0.46	0.74	0.57	0.71	0.45	0.86	0.59
SMOTE	**0.81**	**0.62**	0.57	**0.60**	0.74	0.49	0.80	0.61	0.71	0.46	0.88	0.60
SVMSMOTE	0.75	0.00	0.00	0.00	0.71	0.22	0.08	0.08	0.51	0.20	0.43	0.24

Overall, the CatBoost model exhibits superior accuracy in feature-based binary classification compared to other ensemble boosting models, both with and without the application of imbalanced learning techniques. However, the F1 score of the CatBoost model does not exhibit that level of superiority as its accuracy. Whereas, the highest F1 score is achieved by the AdaBoost model when applying the BorderlineSMOTE technique. Figure 2 and Figure 3 depicts a comparative graph of accuracy and F1 score based on models and imbalanced learning techniques.

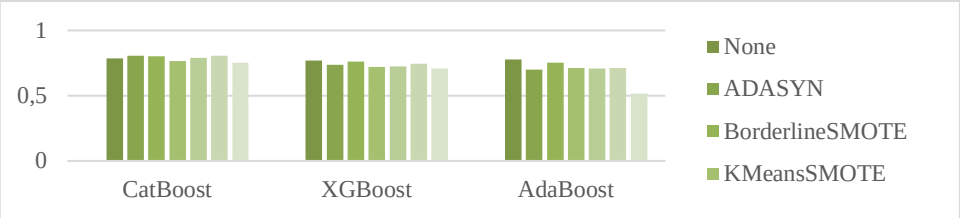

Figure 2. Accuracy of Feature-Based Binary Classification Using Ensemble Boosting Models and Imbalanced Learning Techniques

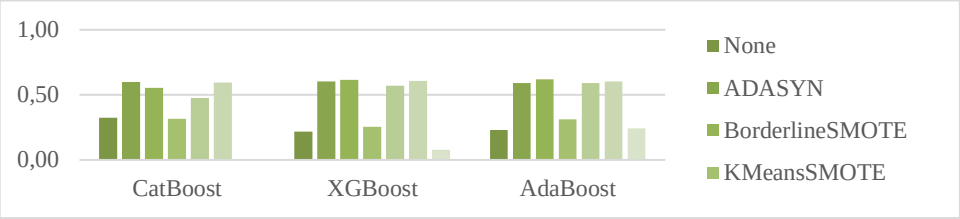

Figure 3. F1 Score of Feature-Based Binary Classification Using Ensemble Boosting Models and Imbalanced Learning Techniques

Table 8 presents a comparison of best classification results before and after applying imbalanced learning techniques, particularly using the CatBoost model. The table shows that implementing imbalanced learning techniques achieves the best classification performance with accuracy of 0.81 and F1 score of 0.60. The recall value also reaches higher result compare to the scenario where these techniques are not applied. However, the precision shows a decline in performance. Nevertheless, the precision and recall values from applying imbalanced learning techniques are reasonably balanced.

Table 8. Best Classification Performance from Each Experiment With and Without Imbalanced Learning Techniques

Model	Accuracy	Precision	Recall	F1 Score
CatBoost	0.79	0.75	0.21	0.32
CatBoost with ADASYN	**0.81**	0.61	0.59	**0.60**
CatBoost with SMOTE	**0.81**	0.62	0.57	**0.60**

Ablation Study

We have conducted feature-based binary classification experiments using ensemble boosting models with all selected features described in subsection Feature Engineering. The experiments revealed that the CatBoost model achieved the best performance using the entire set of extracted features. However, when compared to the best-performing model, other models exhibited relatively lower performance, indicating that not all features were effective. Therefore, the author conducted an ablation study to identify discriminative feature combinations for identifying similar questions. 0 presents the feature combinations and the features used in the ablation study in this research.

Table 9. Feature Combinations in Feature-Based Classification Experiments

Name of Feature Combination	Features Used
Distance	Distance-based feature combination as mentioned in Table 4
Encoding	Encoding feature combination as mentioned in Table 4
Keyphrases	Keyphrases feature combination as mentioned in Table 4
Medical	Medical feature combination as mentioned in Table 4
Textual	Textual-based feature combination as mentioned in Table 4
All	Distance+Keyphrases+Textual+Medical+Encoding
NonEncoding	Distance+Keyphrases+Textual+Medical
NonKeyphrases	Distance+Encoding+Textual+Medical
NonTextual	Distance+Encoding+Keyphrases+Medical
NonMedical	Distance+Encoding+Keyphrases+Textual
NonDistance	Encoding+Keyphrases+Textual+Medical
Feature Set 1	Distance+Textual+Medical
Feature Set 2	Keyphrases+Distance+Medical
Feature Set 3	Keyphrases+Distance+Textual
Feature Set 4	Keyphrases+Textual+Medical
Feature Set 5	Distance+Medical+Encoding
Feature Set 6	Distance+Textual+Encoding
Feature Set 7	Keyphrases+Distance+Encoding
Feature Set 8	Keyphrases+Medical+Encoding
Feature Set 9	Keyphrases+Textual+Encoding
Feature Set 10	Textual+Medical+Encoding
Feature Set 11	Distance+Medical
Feature Set 12	Distance+Textual
Feature Set 13	Keyphrases+Distance
Feature Set 14	Keyphrases+Medical
Feature Set 15	Keyphrases+Textual
Feature Set 16	Textual+Medical
Feature Set 17	Distance+Encoding
Feature Set 18	Medical+Encoding
Feature Set 19	Keyphrases+Encoding
Feature Set 20	Textual+Encoding

We further analyzed the impact of selected features in detecting similar questions in the Indonesian health forum dataset. Figure 4 and Figure 5 illustrate the comparison of accuracy and F1 score before and after removing feature combinations using the ADASYN and SMOTE imbalanced learning techniques. Generally, the average accuracy and F1 score from applying both imbalanced learning techniques on the selected feature set tend to be similar. In both techniques, the highest average accuracy and F1 score were achieved when classifying using the Nonkeyphrases feature set, meaning when the combination of Keyphrases features was eliminated from the selected features. Consistent with these results, when only the Keyphrases feature

combination was used for classification, the average accuracy was the lowest compared to other feature combinations. Meanwhile, consistently across ADASYN and SMOTE applications, the lowest average F1 score was obtained when classifying using only the medical feature combination.

Overall, it is observed that the average accuracy and F1 score exhibit similar patterns in both ADASYN and SMOTE applications. This pattern shows an increase in average accuracy and F1 score when a feature combination is eliminated from the classification. The largest differences in average accuracy and F1 score occur between using the Keyphrases combination and the NonKeyphrases combination. Therefore, feature analysis should be expanded to include other feature combinations, such as Feature Set 1 through Feature Set 20 in 0.

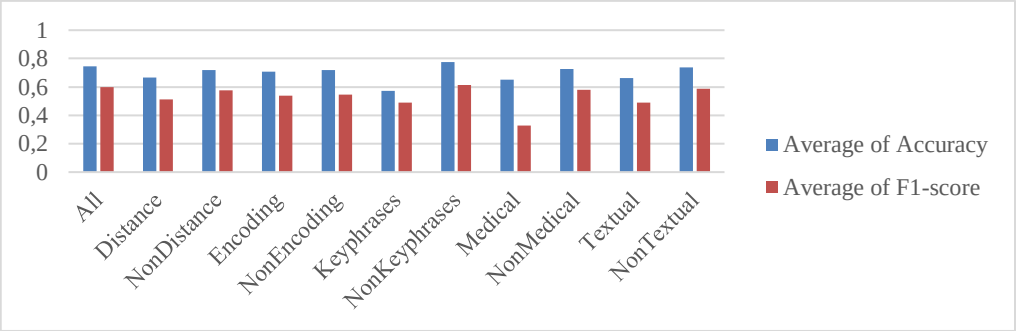

Figure 4. Average Accuracy and Average F1-Score from Experiments using ADASYN Imbalanced Learning Technique on Selected Features

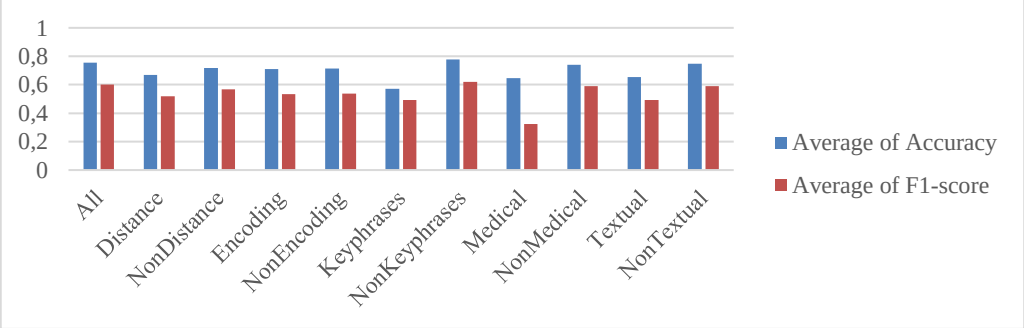

Figure 5. Average Accuracy and Average F1-Score from Experiments using SMOTE Imbalanced Learning Technique on Selected Features

In this section, the author will compare the top five classification performances among selected features and then contrast these performances with those achieved using all features. Table 10 presents the top five classification performances with selected features. From these results, it is evident that the highest F1 score is achieved when the classification involves only Feature Set 5 with the ADASYN technique using the CatBoost model, achieving an F1 score of 0.63 with a standard deviation of 0.01. Feature Set 5 combines distance, medical, and encoding features, representing the most discriminative feature combination in detecting similar and dissimilar question pairs. This finding aligns with the feature analysis results in the preceding subsection, which indicated that eliminating keyphrase features can enhance performance. Meanwhile, the highest accuracy is achieved using Feature Set 5 with the SMOTE oversampling technique, Feature Set 11 with BorderlineSMOTE, and NonKeyphrases with SMOTE, reaching 0.82 with a standard deviation of 0.01. These results

demonstrate that specific feature sets can impact classification performance positively or negatively.

Table 10. Top 5 Best Classification Performance with Selected Features and Implementation of Imbalanced Learning Techniques

Feature Combination	Model	Imbalanced Learning Technique	Accuracy	Precision	Recall	F1-score
Feature set 5	CatBoost	ADASYN	0.81 ± 0.01	0.62 ± 0.02	0.64 ± 0.02	0.63 ± 0.01
Feature set 5	CatBoost	SMOTE	0.82 ± 0.01	0.63 ± 0.01	0.62 ± 0.02	0.62 ± 0.01
Feature set 11	CatBoost	BorderlineSMOTE	0.82 ± 0.01	0.64 ± 0.02	0.60 ± 0.03	0.62 ± 0.02
Non-Keyphrases	CatBoost	SMOTE	0.82 ± 0.01	0.65 ± 0.01	0.59 ± 0.02	0.62 ± 0.02
Non-Keyphrases	CatBoost	ADASYN	0.81 ± 0.01	0.63 ± 0.02	0.59 ± 0.02	0.61 ± 0.02

Table 11 illustrates the performance changes in classification when feature elimination is performed. These results demonstrate the necessity of feature selection to achieve optimal classification performance. This can be seen from the increased precision, recall, and F1 score in the CatBoost model with ADASYN and SMOTE when text-based and keyphrase features are removed. Specifically, precision increases by 0.01 and recall by 0.05 points with both imbalanced learning techniques, while F1-score increases by 0.03 points with ADASYN and 0.02 points with SMOTE. Additionally, there is a 0.01-point increase in accuracy observed in the application of SMOTE when using Feature Set 5.

Table 11. Feature-Based Classification Results with Imbalanced Learning Techniques Before and After Feature Elimination

Feature Combination	Model	Accuracy	Precision	Recall	F1 Score
All Features	CatBoost with ADASYN	0.81	0.61	0.59	0.60
	CatBoost with SMOTE	0.81	0.62	0.57	0.60
Feature Set 5	CatBoost with ADASYN	0.81 ± 0.01	0.62 ± 0.02	**0.64 ± 0.02**	**0.63 ± 0.01**
	CatBoost with SMOTE	**0.82 ± 0.01**	**0.63** ± 0.01	0.62 ± 0.02	0.62 ± 0.01

E. Conclusion

This study explored feature-based binary classification approaches to detect similar questions in the domain of consumer health in Indonesian language. The research findings demonstrate that the identification of similar questions using the CatBoost model significantly outperforms other ensemble boosting models. However, the dataset used in this research exhibits imbalanced class proportions, which necessitated the application of techniques to address imbalanced datasets in feature-based binary classification research. Experimental results indicate that the most effective techniques for addressing imbalanced dataset are ADASYN and SMOTE

applied to the training set, followed by classification using the CatBoost model, achieving an accuracy and F1-score of 0.81 and 0.60, respectively.

Additionally, ablation study experiments were conducted to investigate the impact of text similarity-based features in question similarity detection and to identify discriminative features in identifying similar questions. From these experiments, it was concluded that a combination of distance features, medical features, and encoding features are discriminative in detecting pairs of similar questions in Indonesian consumer health forums. With only these discriminative features, the classification performance reached the highest average accuracy of 0.82 and highest F1-score of 0.63, each with a standard deviation of 0.01.

E. Acknowledgment

We would like to thank Statistics Indonesia (BPS) for financial support in conducting this research. Mrs. Irianti also wants to thank BPS for supporting her during her studies at Universitas Indonesia.

F. References

[1] C. McCreery, N. Katariya, A. Kannan, M. Chablani e X. Amatriain, "Domain-relevant embeddings for medical question similarity," *arXiv preprint arXiv:1910.04192,* 2019.

[2] M. S. M. Jabbar, L. Kumar, H. Samuel, M.-Y. Kim, S. Prabhakar, R. Goebel e O. Zaïane, "On generality and knowledge transferability in cross-domain duplicate question detection for heterogeneous community question answering," *arXiv preprint arXiv:1811.06596,* 2018.

[3] H. Naderi, S. Madani, B. Kiani e K. Etminani, "Similarity of medical concepts in question and answering of health communities," *Health informatics journal,* vol. 26, p. 1443–1454, 2020.

[4] X. Zhang, X. Sun e H. Wang, "Duplicate question identification by integrating framenet with neural networks," em *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[5] Z. Xu e H. Yuan, "Forum duplicate question detection by domain adaptive semantic matching," *IEEE Access,* vol. 8, p. 56029–56038, 2020.

[6] N. Ansari e R. Sharma, "Identifying semantically duplicate questions using data science approach: A quora case study," *arXiv preprint arXiv:2004.11694,* 2020.

[7] H. H. Alfartosy e H. K. Khafaji, "A New Feature Extraction, Reduction, and Classification Method for Documents Based on Fourier Transformation.," *International Journal of Intelligent Engineering & Systems,* vol. 16, 2023.

[8] D. J. Shah, T. Lei, A. Moschitti, S. Romeo e P. Nakov, "Adversarial domain adaptation for duplicate question detection," *arXiv preprint arXiv:1809.02255,* 2018.

[9] A. Mohammed e R. Kora, "Computer and Information Sciences," *Journal of King Saud University–Computer and Information Sciences,* vol. 35, p. 757–774, 2023.

[10] R. Shwartz-Ziv e A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion,* vol. 81, p. 84–90, 2022.

[11] Y. Zhang, D. Lo, X. Xia e J.-L. Sun, "Multi-factor duplicate question detection in stack overflow," *Journal of Computer Science and Technology,* vol. 30, p. 981–997, 2015.

[12] M. Ahasanuzzaman, M. Asaduzzaman, C. K. Roy e K. A. Schneider, "Mining duplicate questions in stack overflow," em *Proceedings of the 13th International Conference on Mining Software Repositories*, 2016.

[13] T. Ngo-Yel e A. Dutt, "A Study on Efficacy of Ensamble Methods for Classification Learning," *ICIS 2009 Proceedings,* p. 69, 2009.

[14] S. Nurhayati, Pencarian Pertanyaan Serupa pada Forum Konsultasi Kesehatan Online dengan Pendekatan Perolehan Informasi, Depok: Universitas Indonesia, 2019.

[15] A. N. Hakim, R. Mahendra, M. Adriani e A. S. Ekakristi, "Corpus development for indonesian consumer-health question answering system," em *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017.

[16] S. Rose, D. Engel, N. Cramer e W. Cowley, "Automatic keyword extraction from individual documents," *Text mining: applications and theory,* p. 1–20, 2010.

[17] S. González, S. García, J. Del Ser, L. Rokach e F. Herrera, "A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities," *Information Fusion,* vol. 64, p. 205–237, 2020.

[18] Y. Freund, R. Schapire e N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence,* vol. 14, p. 1612, 1999.

[19] T. Chen e C. Guestrin, "Xgboost: A scalable tree boosting system," em *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush e A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems,* vol. 31, 2018.

[21] Y. Song, D. Yang, W. Wu, X. Zhang, J. Zhou, Z. Tian, C. Wang e Y. Song, "Evaluating landslide susceptibility using sampling methodology and multiple machine learning models," *ISPRS International Journal of Geo-Information,* vol. 12, p. 197, 2023.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, p. 321–357, 2002.

[23] G. LemaÃŽtre, F. Nogueira e C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of machine learning research,* vol. 18, p. 1–5, 2017.

[24] C. D. Manning, Introduction to information retrieval, Syngress Publishing,, 2008.

[25] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu e J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM computing surveys (CSUR),* vol. 54, p. 1–40, 2021.

[26] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz e G. A. Saheed, "Comparison of the CatBoost classifier with other machine learning methods,"

International Journal of Advanced Computer Science and Applications, vol. 11, 2020.

[27] S. Jhaveri, I. Khedkar, Y. Kantharia e S. Jaswal, "Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns," em *2019 3rd international conference on computing methodologies and communication (ICCMC)*, 2019.

[28] S. Rahmadani, A. Subekti e M. Haris, "Improving Classification Performance on Imbalanced Medical Data using Generative Adversarial Network," *Jurnal Ilmu Komputer dan Informasi,* vol. 17, p. 9–17, 2024.

[29] S. Zhuang e G. Zuccon, "TILDE: Term independent likelihood moDEl for passage re-ranking," em *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[30] G. Zhou, Y. Liu, F. Liu, D. Zeng e J. Zhao, "Improving question retrieval in community question answering using world knowledge," em *Twenty-third international joint conference on artificial intelligence*, 2013.

[31] Y. Zhao e J. Zhang, "Consumer health information seeking in social media: a literature review," *Health Information & Libraries Journal,* vol. 34, p. 268–283, 2017.

[32] X. Zhao e J. X. Huang, "Bert-QAnet: BERT-encoded hierarchical question-answer cross-attention network for duplicate question detection," *Neurocomputing,* vol. 509, p. 68–74, 2022.