

# Oversampling vs. undersampling in TF-IDF variations for imbalanced Indonesian short texts classification

I Nyoman Prayana Trisna, Ni Wayan Emmy Rosiana Dewi, Muhammad Alam Pasirulloh

Department of Information Technology, Faculty of Engineering, Udayana University, Bali, Indonesia

## Article Info

### Article history:

Received Jul 29, 2024

Revised Oct 25, 2024

Accepted Dec 26, 2024

### Keywords:

Bahasa Indonesia

Imbalanced dataset

Oversampling method

Short-text classification

Term frequency inverted

document frequency

Undersampling method

## ABSTRACT

Even though it is considered a more traditional method compared to more modern algorithms, term frequency inverted document frequency (TF-IDF) nevertheless produces good results in a range of text mining tasks. This study assesses the effectiveness of several TF-IDF modifications for short text classification. Imbalanced datasets are another issue that is addressed in this research. To rectify the imbalanced issue, we integrate standard, log-scaled, and boolean TF-IDF in short text classification with undersampling and oversampling methods. Precision, recall, and f-measure metrics are used to evaluate each experiment. The best result is obtained when applying boolean TF-IDF with the oversampling method. Oversampling methods outperform the undersampling methods in every experiment, although there are some cases where experiments with undersampling methods are considerable. Additionally, our conducted study reveals that employing modified TF-IDF, such as boolean or log-scaled versions, provides greater advantages to classification performance, particularly in handling imbalanced datasets, when compared to solely relying on the standard TF-IDF approach.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

I Nyoman Prayana Trisna

Department of Information Technology, Faculty of Engineering, Udayana University

St. Kampus Bukit Jimbaran, Kuta Selatan, Badung, Bali, Indonesia

Email: prayana.trisna@unud.ac.id

## 1. INTRODUCTION

For years, many methodologies were conducted for the best result in text mining for text document, specifically in machine learning approach. As unstructured data, document needs to be vectorized to undergo into machine learning methodologies. The combination of term frequency (TF) along with inverse document frequency (IDF) is still prominent vectorization for text mining. This method is also known as term frequency inverted document frequency (TF-IDF). Kowsari *et al.* [1] states although TF-IDF – or TF alone failed to capture syntactic and semantic characteristic of text, the computation of vectorization is quite simple and manageable with low-resource computing. Compared to state-of-the-art document vectorization such as word-embedding [2], fastText by Facebook [3], or even transformer-based language model like BERT [4] TF-IDF still yields decent performance for multiple text mining task. This is also proven by prior research by Marcinczuk *et al.* [5], who attempts to compare modern approaches such as word2vec and BERT with TF-IDF as “classical one” in four types of datasets. The experiment shows that TF-IDF ranks as 1<sup>st</sup> place in 2 datasets and 2<sup>nd</sup> in 1 dataset with significantly faster computation time.

Document classification is one of popular text mining task. This is done by grouping the documents into multiple class with label. Kowsari *et al.* [1] specifies there are four types of document scopes in text classification: document level, paragraph level, sentence level, and sub-sentence level, although most of document classification is focused on sentence and/or document level [6]. This level of scope also defines

length of the classified text. Compared to document level scope like news articles [7] or paragraph level document like abstract [8], sentence level documents have significant differences. The main difference is the number of texts in the document, where document or paragraph level document have longer text than sentence level document. Shorter document means the resource for document tasks are scarcer and noisier compared to longer text. The scarcity of text resource in sentence level document classification will be seen as sparse vector [9].

Alongside the sparse vector which is resulted from shorter text, the common problem for document classification is the imbalanced distribution of classes. It is usually rephrased as imbalanced dataset. Imbalanced dataset could lead to inadequate performance of the classifier model. This is because majority of classification models require balanced classes to obtain optimal performance [10]. The problem of imbalanced dataset becomes more uncertain and problematic if the dataset is imbalanced to the extreme. The extreme imbalanced dataset means there are majority and minority classes, where the presence of the minority classes is only represented with little to nearly none instances compared to majority classes [11].

However, for the past years methodologies for solving imbalanced dataset are developed. Experimental review done by Tanha *et al.* [12] reveals four methods to handle imbalanced dataset: data-level method, algorithm-level method, hybrid-method, and boosting-method. The data-level methods work by resampling the number of the instances in the dataset [13]. The resampling process is called undersampling when the instances with majority class are sampled down so that those instances are balanced with minority instances. On the contrary, when the minority instances are synthesized so the minority classes have the same distribution as majority classes, it is called oversampling [13]. Both of undersampling and oversampling technique are vastly applied in recent researchers. Furthermore, these techniques are furtherly developed into several algorithms for multiple cases, like adaptive synthetic (ADASYN) [14], synthetic minority oversampling technique (SMOTE) [15], random-based undersampling [16], and neighbor-based undersampling [17].

Our research experiments the data-level imbalanced handling method by comparing oversampling technique and undersampling technique for short text classification. The experiments are done specifically in Bahasa Indonesia's dataset. Although it is spoken roughly by hundreds of million speakers worldwide [18], the resources for Bahasa Indonesia text mining task are limited [19]. Furthermore, our research employs traditional TF-IDF vectorizer even so because its simplicity and the beneficial to the computational time.

## 2. RELATED WORK

The following paragraphs will discuss about the related works that inspire the research. The related works focus on the vectorizer method, imbalanced dataset handlings, machine learning methods, and the text classification problems. Zhu *et al.* [20] utilizes TF-IDF method for hot topic detection in news articles. This research refines TF-IDF vectorizer to adapt to time-distributed information and user attention. The refined vector is then clustered with clustering method to extract the hot topics of the news network. Similar subject of hot topic detection is also conducted by Bok *et al.* [21] who modifies TF-IDF to carry out the temporal information of document frequencies. In addition to modified document frequency, Bok *et al.* [21] scales the term frequency of the words into logarithmic scale. The logarithmic scaling of term frequency is also done in the comparative research by Piskorski and Jacquet [22]. The comparison is conducted between log-scaled TF-IDF character N-grams and word embedding for fine-grained classification task shows that log-scaled TF-IDF approach outperform word embedding approach in most tasks.

Imbalanced dataset handlings are done in several previous researches. Ishaq *et al.* [15] combine oversampling technique with several data mining techniques to improve the prediction of heart failure case. This research employs SMOTE to oversample the minority class which is the mortality case. The conducted research also shows that random forest classifier yields the most promising results based on several evaluations. In network attack, Zuech *et al.* [16] explores the sampling methods by undersampling the majority class. This research also shows that random forest classifier outperforms most of the classifiers. Another research by Oskouei and Bigham [23] experiments oversampling and undersampling techniques in extremely imbalanced dataset. The explored datasets consist of 13 standard real datasets from open-source repository. The research shows that in imbalanced problem resampling method is crucial, and is more preferred than exploring the influence of the classifier. The research also concludes that oversampling methods outperform in all cases compared to undersampling methods.

As stated in previous paragraph, random forest classifier yields sufficiently well performance in classification, including the problem with imbalanced dataset [15], [16]. Triayudi and Fitri [24] explores various method of classifications in educational data mining. Even though it is not the perfect result, random forest classifier performs well in majority task, especially in modelling without any feature selections. In another imbalanced case, Mohammed *et al.* [25] experiments several models of classification in transactions data. The data contains immense number of columns and rows. Random forest classifier outperforms all other models in oversampling technique, and falls into 2<sup>nd</sup> position in undersampling

technique. These prior researches [15], [16], [25] concludes that random forest classifier is befitting for classification problem with imbalanced problem.

There have been several studies done on short text mining. Bernard *et al.* [26] explore the clustering method for tracking news stories in short messaging in Covid-19 area. This research utilizes the sparse TF-IDF combined with Transformer as the vectorization methods. Previous method by Miranda *et al.* [27] was used in this research [26], which uses supervised clustering from monolingual and crosslingual approaches. Besides that, unsupervised K-means was also utilized and combined with prior research [27]. The result shows that TF-IDF is still robust for doing multiple short text mining, even when is combined with other type of vectorization. In another research by Marivate and Sefara [28] conducted text classification in multiple tasks. The research utilizes global augmentation method which uses synonym augmentation, semantic similarity augmentation, and round-trip translation. Although the loss is reduced when the global augmentation is employed, the result shows that the reduction of loss with global augmentation is not significant. Moreover, this research explores the method in English which has large resources and corpora. The proposed method of this research may not necessarily be applicable in other languages.

In Bahasa Indonesia text, Setiabudi *et al.* [29] explores the effect misspelled word in Bahasa Indonesia's text classification. Levenshtein distance is employed to fix the misspelled word. The misspelled correction itself is conducted before the model performs as preprocessing. The result shows that with the misspelled correction with the Naïve Bayes model outperform the baseline model by 8.2%. However, this research also shows that the addition of this preprocessing adds the complexity and elapsed time of the model. Santoso *et al.* [30] work with sentiment analysis and hoax classification in Bahasa Indonesia. The study suggests using particle swarm optimization (PSO) to increase Naïve Bayes' accuracy. Both researches by Setiabudi *et al.* [29] and Santoso *et al.* [30] further prove that Bahasa Indonesia's resources for text mining and classification are lacking.

### 3. METHOD

This section is divided into several subsections: research methodology, dataset, scenario of experiment, and evaluation metrics. Each subsection will be explained further.

#### 3.1. Research methodology

The flowchart in Figure 1 briefly describes how this research is accomplished. The research begins with the collected dataset. The dataset will be explained further in subsection 3.2. Before undergo any process, the dataset is then preprocessed using usual standard preprocess for text mining, which are tokenization, case-folding, and stemming [31]. The data then is splitted into two parts; training data and testing data. The training data will be the base of the TF-IDF vectorization, and results in bag-of-words. The bag-of-words is used as vectorizer to transform both training data and testing data. Once all the data is transformed into vector using bag-of-words, the training data is applied into the model. Inspired by prior researches [15], [16], [25], the random forest algorithm is employed to yield the better result. The testing process is done after the random forest model is build, engaging the testing data as the benchmark for the prediction result.

The evaluation metrics will be used to compare the performance of each experiment. In Figure 1, the bolded blocks are the processes that will be experimented with multiple scenario. The detail of experiment scenarios and the evaluation metrics are respectively elucidated in subsections 3.3 and 3.4.

#### 3.2. Dataset

The dataset used is the title of the final assignment of Information Technology students at Universitas Udayana. The classified class of each assignment is the topic of corresponding assignment. This dataset is described simply in Table 1.

This dataset is used because of the following reasons:

- The dataset is presented in formal Indonesian language.
- The dataset title of each assignment, contains a relatively short number of words compared to paragraphs or abstracts of the assignments. This supports the sentence level classification.
- The dataset has many classes, but there are some classes that have a very small number of instances compared to other classes.

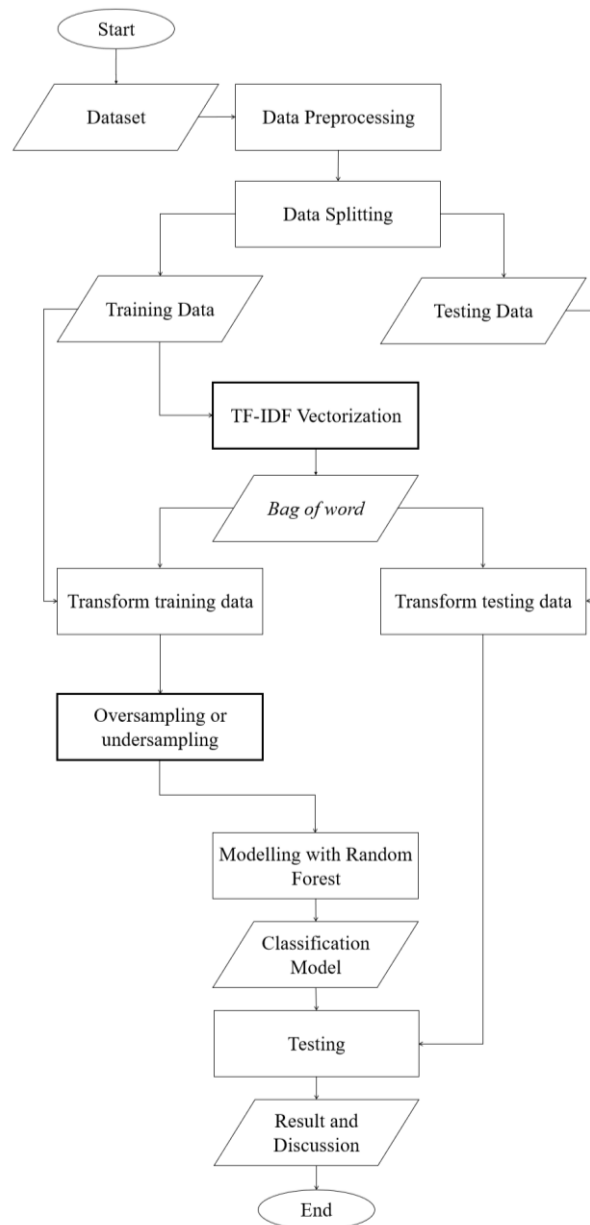


Figure 1. The research method explained in flowchart

Table 1. Summary of the dataset

Topic	Code	Number of assignments
Digital Bali tourism	DBT	10
Digital imaging system	DIS	12
Data science	DS	52
Digital economy	ED	14
Internet of thing	IOT	18
Information system	IS	74
Network and cloud computing	NCC	11
IT governance	TKTI	9
Total		200

### 3.3. Scenario of experiment

As pictured in Figure 1, the bolded blocks are the process which will be experimented in this research. Thus, the TF-IDF vectorization methods and the sampling methods are the parameters of the experiment. Table 2 explains the parameters and the how those parameters contribute in each experiment.

Table 2. Experimented scenarios

Scenario	TF-IDF used	Sampling method
Scenario 0	Standard TF-IDF	No sampling
Scenario 1	Standard TF-IDF	Oversampling
Scenario 2	Log-scaled TF-IDF	Oversampling
Scenario 3	Boolean TF-IDF	Oversampling
Scenario 4	Standard TF-IDF	Undersampling
Scenario 5	Log-scaled TF-IDF	Undersampling
Scenario 6	Boolean TF-IDF	Undersampling

Scenario 0 in Table 2 is used as baseline for the model. In these scenarios, log-scaled and boolean modification of TF-IDF are introduced [32]. The standard TF-IDF is defined in (1), where the log-scaled TF-IDF is formulated in (2), and boolean TF-IDF is in (3).

$$tfidf_{t,d} = tf_{t,d} \times \log\left(\frac{N}{df_t}\right) \quad (1)$$

$$\log - tfidf_{t,d} = \left(1 + \log(tf_{t,d})\right) \times \log\left(\frac{N}{df_t}\right) \quad (2)$$

$$boolean - tfidf_{t,d} = \begin{cases} \log\left(\frac{N}{df_t}\right), & \text{if } tf_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$tf_{t,d}$  itself represents the frequency of term  $t$  in document  $d$ , and  $df_t$  represents number of documents that contain term  $t$ . The total document in the collection is symbolized with  $N$ . In such,  $tfidf_{t,d}$  represents the TF-IDF value of term  $t$  in document  $d$ .

Combined with TF-IDF modifications, sampling methods are also experimented. Both of oversampling and undersampling are conducted in different scenarios. The oversampling method is carried out by adding new instances to minority classes so that those classes have the same number of instances with the majority class. In contrast, undersampling method cuts the number of instances in majority classes, resulting the majority and minority classes have the same total of instances [33].

### 3.4. Evaluation metrics

The evaluation will be concluded in each scenario in Table 2. Precision, recall, and f-measure are used to evaluate the experiments. Precision and recall are more favorable in imbalanced dataset for their abilities to elaborate the model performance in specific class, rather than overall dataset with all classes. This is the opposite of accuracy measure, which evaluate the overall performance of model. Accuracy tends to measure the performance of the model by the majority class [34].

Precision is defined as ratio of correct prediction with total prediction, whereas recall is ratio of correct prediction with total of actual classes. Both precision and recall are measured in specific classes, f-measure combines both of precision and recall, and is used to measure in specific classes as well. They are different with accuracy which measures in overall classes. In (4) to (6) show the formula of precision, recall, and f-measure respectively, where  $c$  represents specific class in the case.

$$p_c = \frac{\text{number of correct prediction for class } c}{\text{number of prediction for class } c} \quad (4)$$

$$r_c = \frac{\text{number of correct prediction for class } c}{\text{number of actual class } c} \quad (5)$$

$$f1 - measure_c = \frac{2 \times p_c \times r_c}{p_c + r_c} \quad (6)$$

All of the metrics in (4) to (6) will be summarized in weighted average. The weighted average accounts the number of classes in the testing data. In (7) explains the calculation of weighted average further.

$$WA(m) = \frac{\sum_{i=1}^{n_c} m_c \times s_c}{\sum_{i=1}^{n_c} s_c} \quad (7)$$

$WA(m)$  represents the weighted average of metric  $m$ . Metric  $m$  can be either precision, recall, or f-measure.  $m_c$  represents the measurement of metric  $m$  in class  $c$ , and  $s_c$  represents total member of class  $c$  in testing data.

#### 4. RESULT AND DISCUSSION

Our experiments with various parameters as in Table 2 yield complying results. Table 3 illustrates in brief about the results of our experiments. The numbers in Table 3 are the weighted average as explained in (7). The detailed result of each experiment projected in the confusion matrix. Tables 4 to 6 show the confusion matrix for each scenario.

From the Table 3, it can be concluded that scenario 3 with oversampled log-scaled TF-IDF results the best scores of all metrics. The baseline model in scenario 0 –using standard TF-IDF with no sampling, already shows excellent performance. Its average precision, recall, and f-measure are exceeding 80%. All of the scenarios with the oversampling method (scenario 1-3) provide better outcomes than baseline scenario. On the other hand, scenario 4-6, which employ the undersampling strategy, considerably underperform the baseline model in all criteria.

Table 3. Result of each scenarios in summary

Scenario	Avg. precision (%)	Avg. recall (%)	Avg. f-measure (%)
Scenario 0	80.965	85.000	81.900
Scenario 1	81.943	87.500	83.955
Scenario 2	82.500	87.500	84.152
Scenario 3	<b>91.373</b>	<b>90.000</b>	<b>88.764</b>
Scenario 4	67.905	57.500	53.583
Scenario 5	69.750	60.000	56.455
Scenario 6	66.766	57.500	53.021

Table 4. Confusion matrix for baseline scenario

		Predicted label							
		DBT	DIS	DS	ED	IOT	IS	NCC	TKTI
True label (baseline scenario)	DBT	2	0	0	0	0	0	0	0
	DIS	0	2	0	0	0	0	0	0
	DS	0	0	9	0	0	1	0	0
	ED	0	0	0	0	0	3	0	0
	IOT	0	0	0	0	4	0	0	0
	IS	1	0	0	0	0	14	0	0
	NCC	0	0	0	0	0	0	2	0
	TKTI	0	0	0	0	0	1	0	1

Table 5. Confusion matrixes for scenarios with oversampling

		Predicted label							
		DBT	DIS	DS	ED	IOT	IS	NCC	TKTI
True label (scenario 1)	DBT	2	0	0	0	0	0	0	0
	DIS	0	2	0	0	0	0	0	0
	DS	0	0	10	0	0	0	0	0
	ED	0	0	1	0	0	2	0	0
	IOT	0	0	0	0	4	0	0	0
	IS	1	0	0	0	0	14	0	0
	NCC	0	0	0	0	0	0	2	0
	TKTI	0	0	0	0	0	1	0	1
True label (scenario 2)	DBT	2	0	0	0	0	0	0	0
	DIS	0	2	0	0	0	0	0	0
	DS	0	0	10	0	0	0	0	0
	ED	0	0	0	0	0	3	0	0
	IOT	0	0	0	0	4	0	0	0
	IS	1	0	0	0	0	14	0	0
	NCC	0	0	0	0	0	0	2	0
	TKTI	0	0	0	0	0	1	0	1
True label (scenario 3)	DBT	2	0	0	0	0	0	0	0
	DIS	0	2	0	0	0	0	0	0
	DS	0	0	10	0	0	0	0	0
	ED	0	0	1	1	0	1	0	0
	IOT	0	0	0	0	4	0	0	0
	IS	1	0	0	0	0	14	0	0
	NCC	0	0	0	0	0	0	2	0
	TKTI	0	0	0	0	0	1	0	1

Table 6. Confusion matrixes for scenarios with undersampling

		Predicted label							
		DBT	DIS	DS	ED	IOT	IS	NCC	TKTI
True label (scenario 4)	DBT	2	0	0	0	0	0	0	0
	DIS	0	2	0	0	0	0	0	0
	DS	0	0	9	0	0	1	0	0
	ED	1	0	0	2	0	0	0	0
	IOT	0	0	0	0	4	0	0	0
	IS	6	1	4	1	0	2	0	1
	NCC	0	0	1	0	0	0	1	0
	TKTI	1	0	0	0	0	1	0	1
True label (scenario 5)	DBT	2	0	0	0	0	0	0	0
	DIS	0	2	0	0	0	0	0	0
	DS	0	0	9	0	0	1	0	0
	ED	1	0	0	2	0	0	0	0
	IOT	0	0	0	0	4	0	0	0
	IS	6	1	3	1	0	2	0	2
	NCC	0	0	0	0	0	0	2	0
	TKTI	1	0	0	0	0	1	0	1
True label (scenario 6)	DBT	2	0	0	0	0	0	0	0
	DIS	0	2	0	0	0	0	0	0
	DS	0	0	9	0	0	1	0	0
	ED	1	0	0	2	0	0	0	0
	IOT	0	0	0	0	4	0	0	0
	IS	6	1	4	1	0	2	0	1
	NCC	0	0	1	0	0	0	1	0
	TKTI	1	0	0	1	0	1	0	1

The outcome of resolving the unbalanced dataset issue is displayed in Table 3. Based on our completed situations, the oversampling methods greatly outperform the undersampling methods in short text classification. Comparing short text classification to long text classification, the completed bag-of-words has significantly fewer text resources.

According to De Boom *et al.* [9], sparse vectors are produced when text resources for short text classification are limited. Undersampling makes the already limited resources even more so. Scarce resources produce a jumble of words that are unable to distinguish between classification classes. On the other hand, the classifier model may more easily identify the classes by creating a synthesis vector with oversampling since the bag-of-words vector has larger dimensions. With the limited dataset in this research, the undersampling approach is more likely to produce sparse vector, thus resulting poor result compared to the oversampling approach with synthetic vector.

Despite producing poor performance, this research demonstrates that undersampling can beat baseline models, even oversampling methods, especially in minority classes. Class ED will be used as an illustration. Class ED is consistently misspredicted in the baseline model from Table 4, and no additional classes are projected to be ED, thus making the precision and recall results for ED are zero. This issue persists even in cases of oversampling, with only one case of class ED is accurately predicted by scenario 3 (best scenario). The undersampling approach prevents this. Based on Table 6, two of the three instances of class ED in the testing data are correctly classified by the scenario using the undersampling method. Better recall for class ED in the undersampling approach is the outcome of this.

Class ED has the fifth-lowest number of instances; the other classes with less instances than class ED are DBT, DIS, NCC, and TKTI. Recalculating the f1-measure, precision, and recall for these five classes yields the following results for each scenario: Table 7. Based on Table 7, scenarios 4-6 with undersampling demonstrates the same, if not superior, recall for minority class classification, even though scenario 3 still performs best in average precision and f-measure for minority classes.

Table 7. Result of each scenarios (DBT, DIS, ED, NCC, and TKTI only)

Scenario	Avg. precision (%)	Avg. recall (%)	Avg. f-measure (%)
Scenario 0	66.667	63.636	63.030
Scenario 1	66.667	63.636	63.030
Scenario 2	66.667	63.636	63.030
Scenario 3	<b>93.939</b>	72.727	<b>76.667</b>
Scenario 4	61.212	72.727	60.000
Scenario 5	58.182	<b>81.818</b>	64.242
Scenario 6	57.071	72.727	57.954

Accounting only minority classes as shown in Table 7 indicates that the undersampling method can still perform well in minority classes despite overall performance results that are dropping, particularly when we consider coverage of true prediction as the primary factor. However, the trade-off of this approach with undersampling method is the decline performance of majority classes. As stated in Table 6, with the escalation in the performance of minority classes such as class ED, the undersampling method tends to neglect the majority such as class IS. As indicated in Table 3, this results in a decline in the overall outcomes for undersampling method.

The experiment results in Tables 3 and 7 yield further question: which TF-IDF modification is the best for either oversampling and undersampling method? We now simply pay attention to the oversampling method's output, which is shown in Table 5. Scenarios 1-3's confusion matrices are essentially the same. Scenarios 1-3 are identical but for the class ED. Class ED is only accurately classified in Scenario 3, and that is only in one out of three instances. For this reason, even though there isn't much of a difference between scenarios 1-3, scenario 3 with Boolean TF-IDF is the best scenario out of all the TF-IDF modifications in the oversampling method.

In short text document like title, one word typically occurs in few occurrence, sometimes even almost once. As a result, the regular TF-IDF (as in scenario 1) will produce a vector that is almost exclusively 1 and 0 if the terms do not occur. A vector that contains only 1 and 0 is referred to be boolean vector. Scenario 3 is executed using the word's boolean feature; if the term exists, its occurrence will be considered as True (or 1) and vice versa. This is why the scenarios 1-3 have similar result as stated in Table 5, with scenario 3 has slight superiority. Despite Boolean TF-IDF vectorization with undersampling method shows dominance in performance, the best undersampling method doesn't share the similar vectorization. According to Tables 3 and 7, scenario 5 which makes use of log-scaled TF-IDF, produces the best results when using the undersampling method.

In undersampling method, the bag-of-words dimension is much less than in the non-sampling method. In comparison to standard bag-of-words from non-sampling approach, the smaller bag-of-words from undersampling method returns even more sparse vector. Log-scaled TF-IDF can be used to solve the sparse vector problem. This is due to the fact that log-scaled TF-IDF always yields non-zero TF based on (2). The occurring terms will be valued greater than 1 –depending on the log value of the frequency, while the non-occurring terms will be valued at 1 (not 0). The absence of 0 in the vector results the less sparse vector. Compared to standard TF-IDF or boolean TF-IDF, where non-occurring terms are valued at 0, this is significantly different because they would create the more sparse vector. Apparently, by making limited resources even less with undersampling method, the less sparsed vector is needed.

From the prior discussion, the experiment also find that class ED is the most misclassified class in most scenarios, where class ED is primarily misclassified as class IS. Meanwhile in undersampling approach most classes are misclassified as IS. This can be shown from scenario 0 as in Table 4, scenarios with an oversampling method as in Table 5, and scenarios with undersampling as in Table 6.

The last assignment topic, digital economy (ED) in Table 1 has numerous intersections with other topics particularly information system (IS). This is due to the fact that titles pertaining to the digital economy typically use the words “*bangun*” (building), “*implementasi*” (implementing), or “*rancang*” (designing), even “*sistem informasi*” (information system), which are terms that can be used to refer to information systems. As shown in Table 8 (in Appendix), we can observe that several phrases from class ED are also widely used in class IS by selecting five examples from the entire dataset for each ED and IS.

## 5. CONCLUSION

TF-IDF, however regarded as a traditional approach in comparison to contemporary algorithms, continues to yield excellent results in a variety of text mining tasks. In this study, the use of several TF-IDF modification for short text categorisation is evaluated. Another problem is imbalanced datasets are a common issue in text mining jobs. In order to address the imbalanced problem, we combine either oversampling and undersampling methods with standard, log-scaled, and boolean TF-IDF in short text classification. Each experiment is assessed using measurements of precision, recall, and f-measure.

According to the results, we find that the undersampling method performs badly when compared to the standard approach, whereas the oversampling method performs significantly better than the standard approach in several TF-IDF modification. On the other hand, the undersampling technique covers true prediction better than the standard and oversampling method if only minority classes are measured, leading to a better recall measurement. Our experiment also find that boolean TF-IDF is slightly better utilized than the standard TF-IDF if combined with oversampling method. Despite of poor performance for undersampling method, our experiment shows that log-scaled TF-IDF is better suited, because its ability to handle sparse vector. With these findings, we believe that utilizing oversampling approach combined with boolean TF-IDF vectorization is best suited for imbalanced short text classification, especially in Indonesian language which



has limited resources. Additional research on TF-IDF modification can be conducted in the future, particularly when compared to the most advanced word2vec and large-language modelling (LLM) techniques. It is also feasible to explore in future research using the hybrid technique by adjusting the oversampling and undersampling.

## APPENDIX

Table 8. Sample titles for class ED and IS

Sample titles labelled as ED	Sample titles labelled as IS
<i>Analisis bisnis proses dan implementasi enterprise resource planning (ERP) pada Batik Srikandi Banyuwangi</i> (English: Business Process Analysis and Implementation of ERP in Batik Srikandi Banyuwangi)	<i>Perancangan sistem informasi pengembangan karir untuk mahasiswa teknologi informasi berbasis website</i> (English: Web-based Information System of Career Development for Student of Information Technology Department)
<i>Rancang bangun customer relationship manajemen pada sistem informasi penjualan jasa wahana watersport berbasis website</i> (English: Designing and Implementation of Web-based CRM on Wahana Watersport Service)	<i>Rancang bangun aplikasi pengenalan biota laut melalui buku bergambar dengan augmented reality</i> (English: Designing and Implementation of Marine Biota Identification through Augmented Reality-based Drawing Book)
<i>Rancang knowledge base system dan CRM pada toko batik nyoman berbasis website</i> (English: Designing Web-based Knowledge Base System and CRM of Nyoman Batik Shop)	<i>Analisis sistem menggunakan metode system usability scale (sus) dan concurrent think aloud (cta) terhadap kepuasan pengguna</i> (English: System Analysis on Customer Satisfaction Using System Usability Scale and Concurrent Think Aloud)
<i>Penerapan electronic-customer relationship management (E-CRM) berbasis website pada PT Panca Niaga Bali</i> (English: Implementation of Web-based Electronic CRM in Panca Niaga Bali Ltd.)	<i>Penerapan business intelligence untuk menentukan strategi marketing pada krisna oleh – oleh bali menggunakan microsoft power BI</i> (English: Implementation of Business Intelligence for Marketing Strategy on Krisna Oleh-Oleh Bali Using Microsoft Power BI)
<i>Implementasi enterprise resource planning (ERP) pada CV. Cipta Anugerah Bakti Mandiri</i> (Web-based ERP Implementation on CV. Cipta Anugerah Bakti Mandiri)	<i>Rancang bangun sistem informasi manajemen dan pengawasan proyek konstruksi kolam renang berbasis website pada mimba pool</i> (English: Designing and Implementation of Information System for Web-based Management and Monitoring Pool Construction in Mimba Pool Company)

## ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Udayana University for the generous financial support provided through the Study Program Preferred Research Grant with contract number B/255.30/UN14.4.A/PT.01.03/2024. This research would not have been possible without their commitment.





## REFERENCES

- [1] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, pp. 1–68, Apr. 2019, doi: 10.3390/info10040150.
- [2] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Advances in Neural Information Processing Systems*, vol. 26, pp. 3111–3119, 2013.
- [3] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint*, arXiv:1607.01759, 2016.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*, arXiv:1810.04805, 2018.
- [5] M. Marcinczuk, M. Gniewkowski, T. Walkowiak, and M. Bedkowski, "Text document clustering: Wordnet vs. TF-IDF vs. word embeddings," *Proceedings of the 11th Global Wordnet Conference*, pp. 207–214, 2021.
- [6] O. Arandjelović, "Targeted Adaptable Sample for Accurate and Efficient Quantile Estimation in Non-Stationary Data Streams," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 848–870, Jul. 2019, doi: 10.3390/make1030049.
- [7] G. Xiaoning, T. De Zhern, S. W. King, T. Y. Fei, and L. H. Shuan, "News reliability evaluation using Latent Semantic Analysis," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 16, no. 4, pp. 1704–1711, 2018, doi: 10.12928/TELKOMNIKA.v16i4.9062.
- [8] I. N. P. Trisna and A. Nurwidyantoro, "Single document keywords extraction in Bahasa Indonesia using phrase chunking," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 4, pp. 1917–1925, 2020, doi: 10.12928/TELKOMNIKA.V18I4.14389.
- [9] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognition Letters*, vol. 80, pp. 150–156, 2016, doi: 10.1016/j.patrec.2016.06.012.
- [10] H. S. Al-Ash, M. F. Putri, P. Mursanto, and A. Bustamam, "Ensemble Learning Approach on Indonesian Fake News Classification," in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, 2019, pp. 1–6, doi:




- 10.1109/ICICoS48119.2019.8982409.
- [11] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
  - [12] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting methods for multi-class imbalanced data classification: an experimental review," *Journal of Big Data*, vol. 7, no. 1, pp. 1–47, 2020, doi: 10.1186/s40537-020-00349-y.
  - [13] G. Lema, F. Nogueira, and 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, no. 17, pp. 1–5, 2017.
  - [14] A. Alhudaif, "A Novel Multi-class Imbalanced EEG Signals Classification Based on the Adaptive Synthetic Sampling (ADASYN) approach," *PeerJ Computer Science*, vol. 7, pp. 1–15, 2021, doi: 10.7717/PEERJ-CS.523.
  - [15] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.
  - [16] R. Zuech, J. Hancock, and T. M. Khoshgoftaar, "Detecting web attacks using random undersampling and ensemble learners," *Journal of Big Data*, vol. 8, no. 1, pp. 1–20, 2021, doi: 10.1186/s40537-021-00460-8.
  - [17] P. Vuttipittayamongkol and E. Elyan, "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data," *Information Sciences*, vol. 509, pp. 47–70, 2020, doi: 10.1016/j.ins.2019.08.062.
  - [18] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," in *2019 7th International Conference on Information and Communication Technology, ICoICT 2019*, 2019, pp. 1–6, doi: 10.1109/ICoICT.2019.8835370.
  - [19] D. Munandar, A. F. Rozie, and A. Arisal, "A multi domains short message sentiment classification using hybrid neural network architecture," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2181–2191, 2021, doi: 10.11591/EEI.V10I4.2790.
  - [20] Z. Zhu, J. Liang, D. Li, H. Yu, and G. Liu, "Hot Topic Detection Based on a Refined TF-IDF Algorithm," *IEEE Access*, vol. 7, pp. 26996–27007, 2019, doi: 10.1109/ACCESS.2019.2893980.
  - [21] K. Bok, Y. Noh, J. Lim, and J. Yoo, "Hot topic prediction considering influence and expertise in social media," *Electronic Commerce Research*, vol. 21, no. 3, pp. 671–687, 2021, doi: 10.1007/s10660-018-09327-2.
  - [22] J. Piskorski and G. Jacquet, "TF-IDF Character N-grams versus Word Embedding-based Models for Fine-grained Event Classification: A Preliminary Study," in *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, 2020, pp. 26–34.
  - [23] R. J. Oskouei and B. S. Bigham, "Over-sampling via under-sampling in strongly imbalanced data," *International Journal of Advanced Intelligence Paradigms*, vol. 9, no. 1, pp. 58–66, 2017, doi: 10.1504/ijaip.2017.10002026.
  - [24] A. Triayudi and I. Fitri, "Comparison Of The Feature Selection Algorithm In Educational Data Mining," *Telkonnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 6, pp. 1865–1871, 2021, doi: 10.12928/TELKOMNIKA.v19i6.21594.
  - [25] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," in *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 2020, pp. 243–248, doi: 10.1109/ICICS49469.2020.239556.
  - [26] G. Bernard, C. Suire, C. Faucher, A. Doucet, and P. Rosso, "Tracking News Stories in Short Messages in the Era of Infodemic," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 18–32, 2022, doi: 10.1007/978-3-031-13643-6\_2.
  - [27] S. Miranda, A. Znotinš, S. B. Cohen, and G. Barzdins, "Multilingual clustering of streaming news," *arXiv preprint*, arXiv:1809.00540, 2018.
  - [28] V. Marivate and T. Sefara, "Improving Short Text Classification Through Global Augmentation Methods," in *CD-MAKE 2020: Machine Learning and Knowledge Extraction*, pp. 385–399, 2020, doi: 10.48550/arXiv.1907.03752.
  - [29] R. Setiabudi, N. M. S. Iswari, and A. Rusli, "Enhancing text classification performance by preprocessing misspelled words in Indonesian language," *Telkonnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 4, pp. 1234–1241, 2021, doi: 10.12928/TELKOMNIKA.v19i4.20369.
  - [30] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *Telkonnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 799–806, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14744.
  - [31] R. A. Ramadani, I. K. G. D. Putra, M. Sudarma, and I. A. D. Giriantari, "A new technology on translating Indonesian spoken language into Indonesian sign language system," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3338–3346, 2021, doi: 10.11591/ijece.v11i4.pp3338-3346.
  - [32] C. D. Manning, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/cbo9780511809071.
  - [33] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information*, vol. 14, no. 1, pp. 1–15, 2023, doi: 10.3390/info14010054.
  - [34] R. Soleymani, E. Granger, and G. Fumera, "F-measure curves: A tool to visualize classifier performance under imbalance," *Pattern Recognition*, vol. 100, 2020, doi: 10.1016/j.patcog.2019.107146.

## BIOGRAPHIES OF AUTHORS






**I Nyoman Prayana Trisna**     received the Bachelor's and Master's degree in Computer Science and Electronics from the Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2017 and 2020 respectively. He is currently an Assistant Professor in the Information Technology Study Program, Faculty of Engineering, Udayana University, Bali, Indonesia. His current research interests include machine learning, evolutionary computation, natural language modelling, and text mining. He can be contacted at email: prayana.trisna@unud.ac.id.



**Ni Wayan Emmy Rosiana Dewi**    received her M.Kom. degree in Computer Science from Ganesha University of Education in Singaraja, Bali, with a thesis titled “Detection of Class Regularity with Support Vector Machine methods.” Since 2022, she has served as a permanent lecturer in the Information Technology Department at Udayana University. In her role as an educator, she teaches several courses and actively contributes to self-development through article writing and participation in community service activities, demonstrating her commitment to education and the community. She can be contacted at email: [emmyrosiana@unud.ac.id](mailto:emmyrosiana@unud.ac.id).



**Muhammad Alam Pasirulloh**    received the Master degree in Information Technology from the Indonesia University, Jakarta, Indonesia, with the Dissertation “*Analisis Critical Knowledge Penerbangan Dan Antariksa: Studi Kasus Lembaga Penerbangan dan Antariksa (LAPAN)*”. He is a lecturer in the Bachelor’s Degree of Information Technology, Udayana University, Bali, Indonesia since 2022. His research interests are in information technology, IT governance, knowledge management, system simulation, IT audit, and decision support system. He can be contacted at email: [muhammad.alam@unud.ac.id](mailto:muhammad.alam@unud.ac.id).