ISSN 0125-9008 (Print); ISSN 2301-8593 (Online)
DOI: https://dx.doi.org/10.14203/j.baca.v41i2.563
SK Dirjen Risbang -Kemristekdikti No 21/E/KPT/2018 (Peringkat 2 SINTA)

ANALYZING THE IMPACT OF RESAMPLING METHOD FOR IMBALANCED DATA TEXT IN INDONESIAN SCIENTIFIC ARTICLES CATEGORIZATION

Ariani Indrawati[1*], Hendro Subagyo[2], Andre Sihombing[3], Wagiyah[4], Sjaeful Afandi[5]
[1,2,3,4,5]Indonesian Institute of Science

[*]Correspondence: indrawati.ariani@gmail.com

Submission: 02-06-2020; **Review:** 30-08-2020; **Accepted:** 07-09-2020; **Revised:** 30-10-2020

ABSTRACT
The extremely skewed data in artificial intelligence, machine learning, and data mining cases are often given misleading results. It is caused because machine learning algorithms are designated to work best with balanced data. However, we often meet with imbalanced data in the real situation. To handling imbalanced data issues, the most popular technique is resampling the dataset to modify the number of instances in the majority and minority classes into a standard balanced data. Many resampling techniques, oversampling, undersampling, or combined both of them, have been proposed and continue until now. Resampling techniques may increase or decrease the classifier performance. Comparative research on resampling methods in structured data has been widely carried out, but studies that compare resampling methods with unstructured data are very rarely conducted. That raises many questions, one of which is whether this method is applied to unstructured data such as text that has large dimensions and very diverse characters. To understand how different resampling techniques will affect the learning of classifiers for imbalanced data text, we perform an experimental analysis using various resampling methods with several classification algorithms to classify articles at the Indonesian Scientific Journal Database (ISJD). From this experiment, it is known resampling techniques on imbalanced data text generally to improve the classifier performance but they are doesn't give significant result because data text has very diverse and large dimensions.

ABSTRAK
Dataset yang tidak seimbang jika digunakan pada kecerdasan buatan, *machine learning*, dan *data mining* sering kali memberikan hasil yang keliru. Hal tersebut dikarenakan algoritma *machine learning* dirancang untuk berkerja secara optimal dengan data yang seimbang. Namun, sering kali kita diharuskan untuk melakukan proses analisis data menggunakan dataset yang tidak seimbang. Cara yang paling umum digunakan untuk menangani permasalahan ketidakseimbangan data adalah dengan melakukan *resampling* untuk mengubah jumlah data pada kelas mayoritas atau minoritas sehingga membentuk dataset yang seimbang. Beberapa teknik *resampling* telah diajukan, baik *oversampling*, *undersampling*, maupun kombinasi dari keduanya. Teknik *resampling* ini memungkinkan untuk meningkatkan atau menurunkan performa dari model klasifikasi. Teknik *resampling* dengan data terstruktur sudah banyak diterapkan pada beberapa penelitian, namun penerapan *resampling* pada data tidak terstruktur belum banyak dilakukan. Hal tersebut menimbulkan pertanyaan apakah teknik *resampling* dapat diterapkan pada tidak terstruktur seperti teks yang memiliki dimensi yang banyak dan karakter yang sangat beragam. Pada penelitian ini kami mencoba menerapkan teknik *resampling* pada dataset artikel *Indonesian Scientific Journal Database* (ISJD) untuk memahami bagaimana pengaruhnya terhadap beberapa model klasifikasi. Dari hasil eksperimen diketahui bahwa secara umum teknik *resampling* ini dapat meningkatkan performa dari model klasifikasi, namun tidak memberikan hasil yang signifikan.

Keywords: Imbalanced data; Resampling techniques; Machine learning; Classification; Journal; ISJD

1. INTRODUCTION
The problem of imbalanced data has got more and more hot topics in recent years. Imbalance data is the condition where the number of instances in one class is significantly lower than the other classes. Imbalance data is a challenging problem in artificial intelligence, machine learning, and data mining topic. Most machine learning algorithms are designated to work best with balanced data that the target classes have similar prior probabilities. However, the real situation is often the ratios of prior probabilities between classes are extremely skewed in the high dimensionality and extremely sparse.

Typically, in the imbalanced dataset problem, it is more difficult to classify members of the minority class than members of the majority class. This happens because machine learning algorithms do not consider the class distribution, they are usually designed to improve accuracy by reducing the error. Many researches have reported data mining with an imbalanced data distribution often give misleading result, such as diagnostic of rare diseases, fraud detection, network intrusion detection, detection of oil spills from radar images, text classification, marketing, etc.

The most popular technique for handling imbalanced data is resampling a training dataset in order to balance the class distribution before the data used as input to the machine learning process. Resampling is a process that modifies the number of instances in the majority and minority classes into a standard balanced data. It will be much easier for machine learning to process the balanced data.

There are 3 approaches resampling, under-sampling by reducing some samples from the majority class, over-sampling by adding more samples to the minority class, or a combination of both under-sampling and over-sampling. Many resampling methods have been proposed, Random Over Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE) (Chawla, 2002), Borderline SMOTE (Han, 2005), kMeans SMOTE (Last, 2017), Support Vector Machine SMOTE (SVM-SMOTE) (Zhang, 2018), Adaptive Synthetic (ADASYN) (He, 2008), Random Under Sampling (RUS), TomekLinks (Tomek, 1976) Edited Nearest Neighbors (ENN) (Wilson, 1972), etc.

Some previous studies have been implemented those resampling method to their cases. Batista, Prati, and Monard analyze the behavior of several over-sampling and under-sampling methods to deal with the problem of learning from imbalanced in thirteen UCI data sets which each dataset has been collapse in 2 classes (positive and negative), they use C4.5 as classifier method (Batista, 2004). Xie, Hao, Liu, and Lin in 2019 have been proposed the fused case-control screening to balancing the p53 mutant dataset before detecting the transcriptional activity (active or inactive) (Xie, 2019). Padurariu and Breaban also dealing the imbalanced data text with oversampling methods (Padurariu, 2019). Al-Azani and El-Alfy use SMOTE to highly imbalanced data sentiment analysis in short Arabic text (Al-Azani, 2017). Suh, Kim, Song, Leegu, Yu, and Mo comparing of oversampling methods on imbalanced topic classification of Korean news articles (Suh, 2017). Fernandez, del Rio, Chawla, and Herra compared RUS, ROS, and SMOTE using MapReduce with two subsets of the Evolutionary Computation for Big Data and Big Learning (ECBDL'14) dataset (Fernández, 2017).

Loyola-González, Martínez-Trinidad, Carrasco-Ochoa, García-Borroto, have studied the use of resampling methods combined with contrast pattern based classifiers for data mining and classification tasks on imbalanced databases (Loyola-González, 2016). Krawczyk analyzed different aspects of imbalanced learning s uch as classification, clustering, regression, data mining and big data analytics (Krawczyk, 2016). Blagus and Lusa use SMOTE to balancing three breast cancer gene expression data sets and classify each of them with kNN (Blagus, 2013). Li, Sun, and Zhu study on data imbalance problem in text classification on several form such as text distribution, class size, and overlapping class (Li, 2010). Yanminsun, Wong, and Kamel provides a review of the classification of imbalanced data regarding: the appli- cation domains; the nature of the problem; the learning difficulties with standard clas- sifier learning algorithms; the learning objectives and evaluation measures; the reported research solutions; and the class imbalance problem in the presence of multiple classes (Yanminsum, 2011).

In this research, we investigated the impact from 5 oversampling techniques are ROS, SMOTE, Borderline SMOTE, KMeans SMOTE, SVM-SMOTE, and ADASYN, 3 undersampling methods are RUS, Tomek, and ENN, However, oversampling and undersampling method has some flaws. Oversampling can lead to model overfitting, since it will duplicate instances from minority

class, while undersampling can end up leaving out important instances that provide important differences in the majority class. We also tried combined oversampling and undersampling methods are SMOTEENN and SMOTETomek, to Gaussian Naïve Bayes, Multinomial Naïve Bayes, SVM with linear kernel, SVM with RBF kernel and k-NN for handle highly imbalanced data text in Indonesian scientific articles categorization.

2. LITERATURE REVIEW

In this section, we briefly describe the basic idea to understand how each resampling methods works to balancing the imbalanced data. Illustration before and after resampling data can be seen in Figure 1.

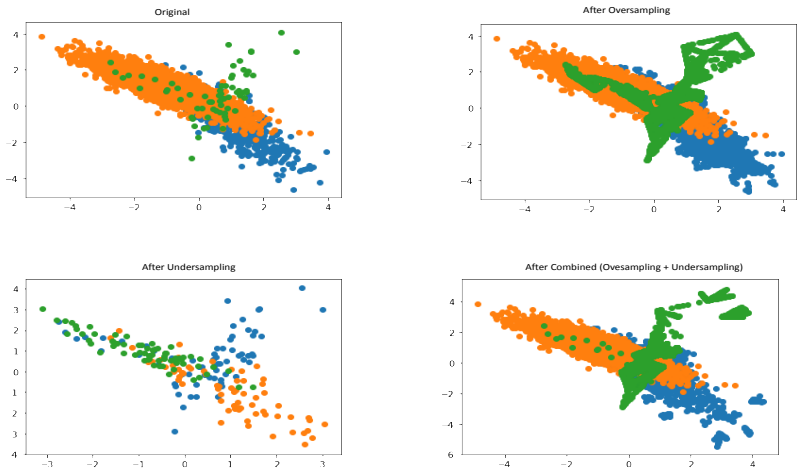

Figure 1. Illustration before and after resampling

- Random Over Sampling (ROS). ROS is simply duplicating the data samples in minority classes and adding them to the training datasets. ROS increases the size of the training data set through repetition of the original samples until the class distribution is balance.
- Synthetic Minority Oversampling Technique (SMOTE). SMOTE was introduced by Chawla in 2002. Similar to ROS, SMOTE is also increase the size of the training dataset and its variety by generating artificial samples in the training dataset by interpolating between existing data points of the minority class that are closer to each other. SMOTE algorithm is described in (Chawla, 2002).
- Borderline SMOTE. Borderline SMOTE is variant of the original SMOTE, proposed by Han, Wen-Yuan, and Bing-Huan in 2005. Borderline-SMOTE generate their synthetic samples along the borderline of minority and majority classes. Figure 3 illustrates before and after resampling with Borderline SMOTE, before resampling class 0 have 100 data but class 1 only have 10 data, after resampling both class 0 and class 1 have 100 data.
- KMeans SMOTE. Felix Last, Georgius Douzas, Fernando Bacao tried to apply KMeans to SMOTE in their research (Last, 2017). KMeans SMOTE generating minority class samples in safe and crucial areas of the input space.
- SVM-SMOTE. This algorithm is a variant of SMOTE which use SVM to locate the decision boundary defined by the support vectors and examples in the minority class that close to the support vectors become the focus for generating synthetic examples (Zhang, 2018).
- Adaptive Synthetic (ADASYN). ADASYN proposed by Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li in 2008. The essential idea of ADASYN reducing the bias and adaptively learning to generate some synthetic data samples for the minority classes based on dynamic adjustment of weights and an adaptive learning procedure according to data distributions. Algorithm ADAYSN is described in (He, 2008).

- Random Under Sampling (RUS). RUS does the opposite from ROS, it removes some samples from the majority class to balanced it with minority class.
- Tomek Links. Tomek (1976) proposed an algorithm for resampling dataset, named Tomek Links. This algorithm detects pairs of instances from the nearest opposite classes to determine borderline between majority and minority classes.
- Edited Nearest Neighbors (ENN). This method proposed by Zhang in 2008 which uses the edited nearest neighbor algorithm to select some samples to be removed to balanced it with minority class.
- SMOTEENN. SMOTEENN is a combined method between oversampling method using SMOTE and undersampling using ENN.
- SMOTETomek. SMOTETomek is a combined method between oversampling method using SMOTE and undersampling using Tomek.

3. METHOD

Figure 2 illustrates 3 important stages in this research, are text processing, resampling and categorization, and evaluation.

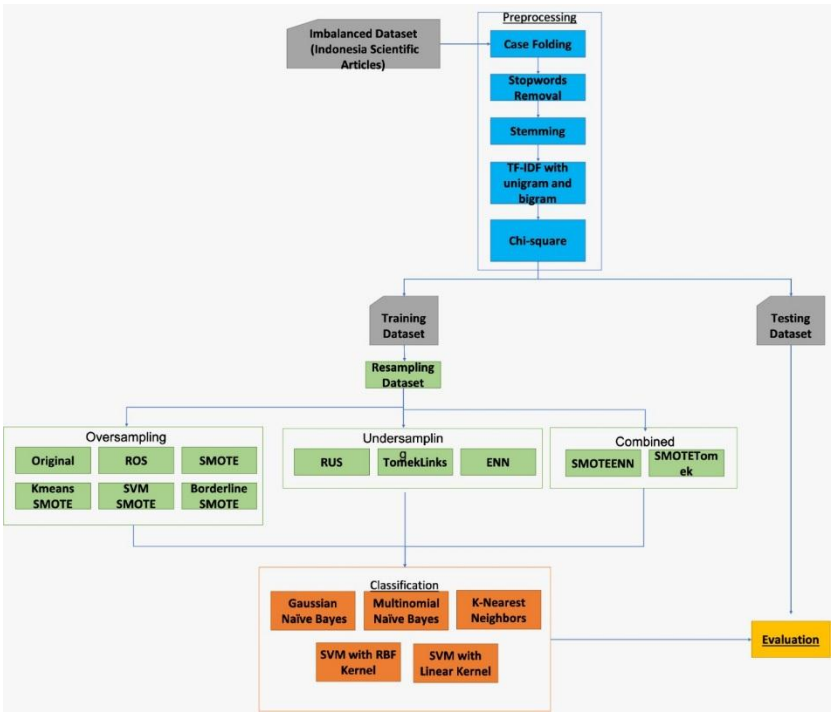

Figure 2. Methodology

a) Text Processing. In this stage, we have 2 tasks, which are text pre-processing and feature weighting.

Text Pre-Processing

The raw textual data is mostly unstructured. So, before carrying out the categorization process, it is necessary to processing the abstract to make it in a structured form and can enhance the classifier's performance significantly (Haddi, 2013). There are a few steps to take in the text pre-processing phase, are:

- Case folding: the entire text in the abstract will be converted to lowercase letters.
- Stopwords removal: the process to remove stopwords so that only words that are considered important will be used. Words such as conjunctions will be removed.
- Stemming: change a word to the basic word form that builds it. In Indonesian texts all the words added both suffixes and prefixes are also omitted

Feature Extraction

When dealing with text, we should represent each document to a vector of word frequencies. At this stage we are using Term Frequency - Inverse Document Frequency (TF-IDF) with unigram and bigram. We also use *Chi-Square* for feature selection to select the important features in each class.

b) Resampling and Categorization

In this research, we use 3 approaches in the resampling, are oversampling, undersampling, and combined.

- Oversampling: ROS, SMOTE, Borderline SMOTE, KMeans SMOTE, SVM SMOTE, and ADASYN.
- Undersampling: RUS, TomekLinks, and ENN.
- Combined: SMOTEENN and SMOTETomek.

Resampling result from each method will be used to generated classification model. For classification machines, we use SVM with Linear and RBF Kernel, Naïve Bayes both Gaussian and Multinomial, K-NN.

c) Evaluation

We use precision, recall, and f-1 measure.

- Precision
- Recall
- F1 Score

4. RESULTS AND DISCUSSIONS

Data in this research retrieved from the Indonesian Scientific Journal Database (ISJD) from 2013 until 2019. ISJD is a database containing journals published by journal publishers in Indonesia. We retrieve abstract in Indonesian language and label category from each article. After cleansing the data, we have 26708 data. The data is not well balanced as the distribution of the dataset can be seen in Figure 3.

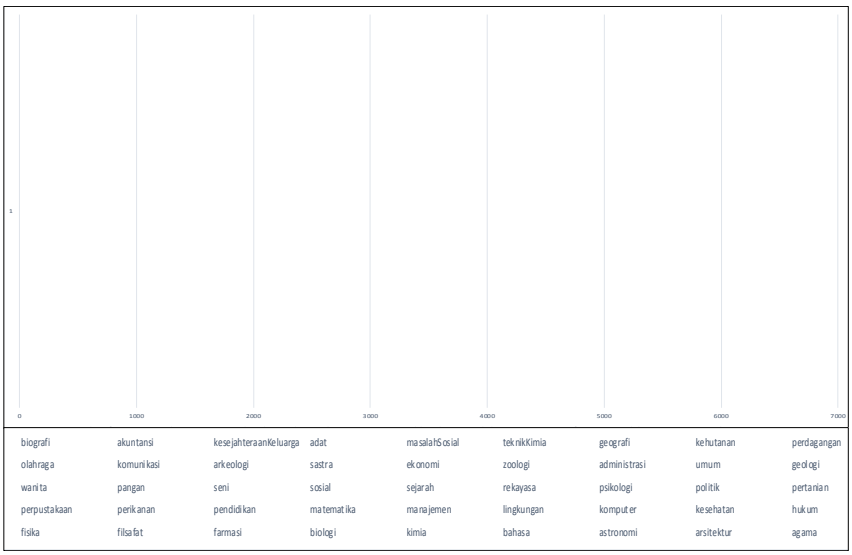

Figure 3. Data distribution

In this research, we investigated the over-sampling and under-sampling techniques to imbalance data text. However, oversampling and undersampling method has some flaws. Oversampling can

lead to model overfitting, since it will duplicate instances from minority class, while undersampling can end up leaving out important instances that provide important differences in the majority class. So, we also tried combined oversampling and undersampling method to see how this combined method affect the classifier performance.

a) Oversampling

The precision, recall, and f-measure can be seen in Figure 4, we can see the highest F-Measure from oversampling techniques result is SMOTE with SVM-Linear classifier. It is known some oversampling techniques can improve classifier performance, ROS, SMOTE, SVM SMOTE, Borderline SMOTE, KMeans SMOTE outperformed the original model, only ADASYN has decreased the classifier performance. In general, using SMOTE technique will improve the result compare to using ROS technique. SMOTE technique and other modifications and extension not only increase the size of the minority class but on the other hand, it also increases the variety of your data. Variation in the training data set, help machine to not learning too much specific from only a few examples. However, sometimes we need to careful with the result, whether the variation of the data that has been generated by SMOTE is valid. From all classification methods we used, MNB is the most affected from these oversampling techniques, MNB F-Measure increase by 0.19 to 0.21. Interesting things here are that the oversampling method techniques give a negative impact on kNN, especially with SMOTE, SVM SMOTE, Borderline SMOTE, KMeans SMOTE. It happens because SMOTE generates their synthetic samples by interpolating between existing data points of the minority class that are closer to each other. It is very possible that the data generated is so close to other classes, so that it is difficult for KNN to classify new data from resampling results.

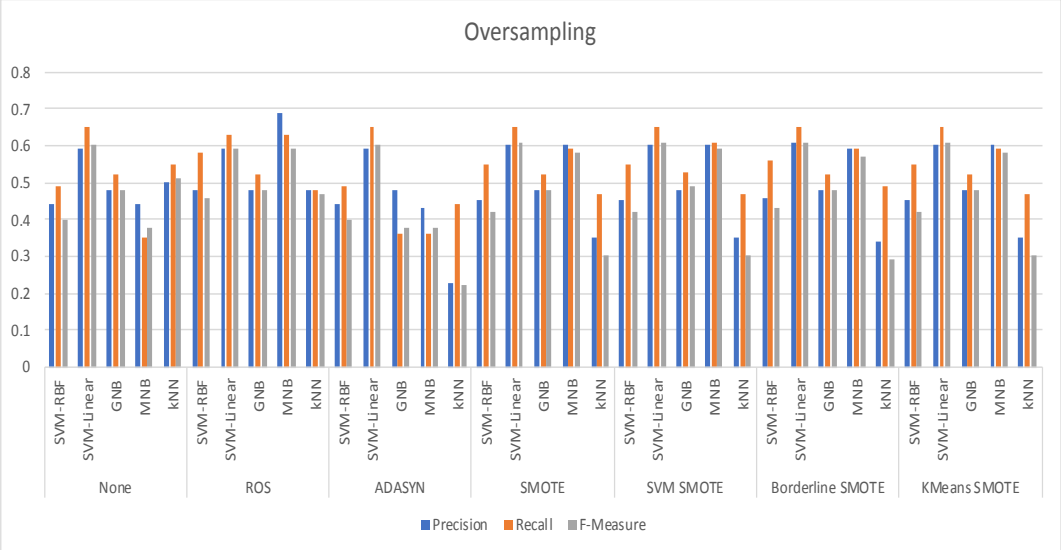

Figure 4. Precision, recall, and F-Measure oversampling techniques result

b) Undersampling

The precision, recall, and f-measure can be seen in figure 5. We can see the highest F-Measure from undersampling techniques result is TomekLinks with SVM-Linear classifier. In contrast to oversampling, undersampling techniques mostly decrease the classifier performance, except for RUS in SVM-RBF and MNB, and TomekLinks in SVM-RBF which undersampling techniques can increase the classifier performance by 0.01 to 0.09. The most affected classifier is SVM-Linear. This technique has the advantage in terms of times and memory complexity compare to oversampling because we are decreasing the size of the data. While in the process we may remove some potential data that could be important for the learning process. Another undersampling technique we use in this research has overcome the problem by removing data that has been identified as redundant or get a high score in similarity.

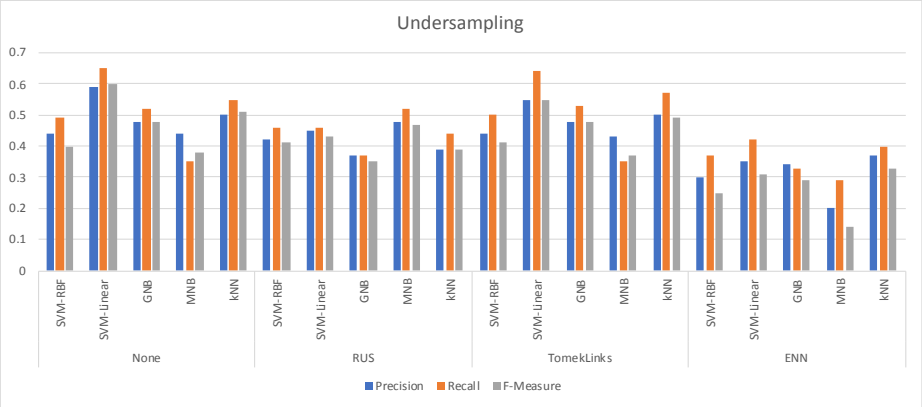

Figure 5. Precision, recall, and F-Measure undersampling techniques result

c) Combined

In this section, we combined the two techniques described before. it is typically the better approach in the resampling method compare to used only oversampling or undersampling individually. First, we could remove some redundant data in the majority class so it will decrease the size with the hope to improve the times and memory complexity, in the other hand for the minority class we increase the data using appropriate oversampling techniques until all the classes in data set is balance. The precision, recall, and f-measure can be seen in figure 6. We can see the highest F-Measure from the combined methods between oversampling and undersampling techniques result is SMOTENN with SVM-Linear classifer. The combination resampling method increases the classifier performance by 0.01 to 0.21, except for kNN decrease the classifer by 0.18 to 0.23. Same as the oversampling method, the most affected classifier is MNB.

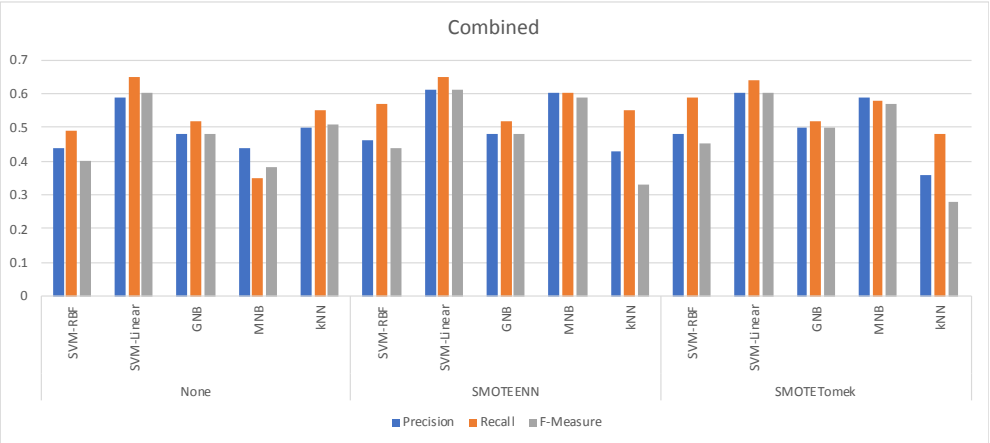

Figure 6. Precision, recall, and F-Measure combined techniques result

Results and discussion should be arranged in separate sub-headings. The subtitles in the literature review are written with Times New Roman font 11.5, and the contents are Times New Roman font 11 (1.15 spaces). The results are not raw data, but data that have been processed and interpreted in the form of statistical data, either in the form of tables, graphs, charts, sketches, and photographs combined with relevant theories. While the discussion is the result of data analysis based on relevant theory. The content of the results and the discussion should address the research issues and find the appropriate analysis for the solution/positive impact on the development of science and technology in society.

5. CONCLUSION

Resampling is a simple way to handle imbalanced data, either by oversampling or undersampling. Resampling allows us to create a balanced dataset to simplify the classification process. However, resampling has some flaws. Oversampling can lead to model overfitting, since it will duplicate instances from minority class, while undersampling can end up leaving out important instances that provide important differences in the majority class. Ultimately, there is no one-size-fits-all method for the imbalanced problems, we just have to try out each method and see their effect on specific use cases and metrics. From this experiment, it is known resampling techniques on imbalanced data text generally improve the classifier performance. The best oversampling method is SMOTE, the best undersampling method is TomekLinks, and the best combined resampling method is SMOTETomek. Interesting things here is that the oversampling method techniques give negative impact on kNN, especially with SMOTE, SVM SMOTE, Borderline SMOTE, KMeans SMOTE. It's happens because SMOTE generate their synthetic samples by interpolating between existing data points of the minority class that are closer to each other. It is very possible that the data generated is so close to other classes, so that it is difficult for KNN to classify new data from resampling results.

.

REFERENCES

Al-Azani, S. & El-Alfy, E. 2017. Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text. *Procedia Computer Science*. 359-366. doi: 109.10.1016/j.procs.2017.05.365.

Batista, G., et al. 2004. A Study of The Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations*, *6*(1), 20-29. doi: 10.1145/1007730.1007735.

Blagus, R. & Lusa, L. 2013. SMOTE for High-Dimensional Class-Imbalanced Data. *BMC Bioinformatics, 14,* 106. doi: 10.1186/1471-2105-14-106.

Chawla, et al. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16,* 321-357. doi: 10.1613/jair.953.

Fernández, A., et al. 2017. An Insight into Imbalanced Big Data Classification: Outcomes and Challenges. *Complex & Intelligent Systems*. doi: 10.1007/s40747-017-0037-9.

Han, H., et al. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Advances in Intelligent Computing,* 878-887. doi: 10.1007/11538059_91.

He, H, et al. 2008. Adasyn: Adaptive Synthetic Samplingapproach For Imbalanced Learning. *International Joint Conference on Neural Networks*, June. 10.1109/IJCNN.2008.4633969.

Krawczyk, B. 2016. Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence, 5,* 221–232. doi: 10.1007/s13748-016-0094-0.

Last, F., et al. 2017. Oversampling for Imbalanced Learning Based on K-Means and SMOTE.

Li, Y., et al. 2010. Data Imbalance Problem in Text Classification. *Third International Symposium on Information Processing,* 301-305. doi: 10.1109/ISIP.2010.47.

Loyola-González, O. 2016. Study of the Impact of Resampling Methods for Contrast Pattern based Classifiers in Imbalanced Databases. *Neurocomputing, 175,* 935-947. doi: 10.1016/j.neucom.2015.04.120.

Padurariu, Cristian & Breaban, Mihaela. 2019. Dealing with Data Imbalance in Text Classification. *Procedia Computer Science, 159,* 736-745. doi: 10.1016/j.procs.2019.09.229.

Suh, Y, et al. 2017. A Comparison of Oversampling Methods on Imbalanced Topic Classification of Korean News Articles. *Journal of Cognitive Science, 18.* 391-437. doi: 10.17791/jcs.2017.18.4.391.

Tomek, I. 1976. Two Modifications of CNN. *iIEEE Transactions on Systems, Man, and Cybernetics, SMC-6*(11), 769-772. doi: 10.1109/TSMC.1976.4309452.

Wilson, D.L. 1972. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-2*(3), 408-421, doi: 10.1109/TSMC.1972.4309137.

Xie, J., et al. 2020. Fused Variable Screening for Massive Imbalanced Data. *Computational Statistics & Data Analysis. 141*. doi: 10.1016/j.csda.2019.06.013.

Yanminsun, Y. 2011. Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*. doi: 10.1142/S0218001409007326.

Zhang, C., et al. 2018. A Cost-Sensitive Deep Belief Network for Imbalanced Classification.