OPEN ACCESS
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 8, No. 1, June 2022. pp. 20-29
doi:10.21108/ijoict.v8i1.622

Overcoming Data Imbalance Problems in Sexual Harassment Classification with SMOTE

Irfan Dwi Wijaya[1], Aji Gautama Putrada [2*], Dita Oktaria [3]

[1,3]*School of Computing, Telkom University*
Bandung, Indonesia

[2]*Advanced and Creative Networks, Telkom University*
Bandung, Indonesia

[*] ajigps@telkomuniversity.ac.id

Abstract

Delivery of justice with the help of artificial intelligence is a current research interest. Machine learning with natural language processing (NLP) can classify the types of sexual harassment experiences into quid pro quo (QPQ) and hostile work environments (HWE). However, imbalanced data are often present in classes of sexual harassment classification on specific datasets. Data imbalance can cause a decrease in the classifier's performance because it usually tends to choose the majority class. This study proposes the implementation and performance evaluation of the synthetic minority over-sampling technique (SMOTE) to improve the QPQ and HWE harassment classifications in the sexual harassment experience dataset. The term frequency-inverse document frequency (TF-IDF) method applies document weighting in the classification process. Then, we compare naïve Bayes with K-Nearest Neighbor (KNN) in classifying sexual harassment experiences. The comparison shows that the performance of the naïve Bayes classifier is superior to the KNN classifier in classifying QPQ and HWE, with AUC values of 0.95 versus 0.92, respectively. The evaluation results show that by applying the SMOTE method to the naïve Bayes classifier, the precision of the minority class can increase from 74% to 90%.

Keywords: synthetic minority over-sampling technique, quid pro quo, hostile work environment, sexual harassment, data imbalance, text analysis, text frequency-inverse document frequency, naïve Bayes, k-nearest neighbor

I. INTRODUCTION

SEXUAL harassment is a request to perform a sexual act verbally or physically that offends someone [1]. Some examples of sexual harassment behavior include comments about someone's body, sexual comments, sexist jokes, and constantly asking someone out. This behavior may occur online or personally and directly [2]. Victims of sexual harassment experience shame, anger, and humiliation, victims who are aware of being victims of sexual harassment should report it [3]. Based on the facts, 57% of victims of sexual harassment in Indonesia felt that their cases were unresolved. In addition, 39.9% of victims of harassment said that the resolution of the instances of harassment was with money.

Received on 21 June 2022. Revised on 5 July 2022. Accepted and Published on 19 July 2022.

The United States, in 1964, released a regulation, namely Title VII of the Civil Rights Act, which prohibits discrimination in the workplace, including dividing harassment into two, namely quid pro quo (QPQ) and hostile work environment (HWE) [4]. QPQ and HWE are labels used to describe two types of sexual harassment relevant to fundamental questions at trial [5].

Delivery of justice with the help of artificial intelligence is a current research interest [6]. Machine learning with natural language processing (NLP) can classify the types of sexual harassment experiences [7]. Putri *et al.* [8] classified QPQ and HWE on tweets with the hashtag #MeToo using the naïve Bayes algorithm with an accuracy of 88.5%. Haque *et al.* [9] proved that k-nearest neighbour (KNN) bests other methods in sentiment analysis related to sexual harassment. However, imbalanced data are often present in classes of sexual harassment classification on specific datasets [10]. Imbalance data is in a dataset when the data has skewed proportions and causes the data to split into vast majority and few minority data [11]. Data imbalance can cause a decrease in the classifier's performance because it usually tends to choose the majority class [12]. Several techniques can overcome imbalance data, for example, oversampling minority data or under-sampling majority data [13].

Based on the state-of-the-art study on imbalance, synthetic minority oversampling technique (SMOTE) has the best f1-score performance compare to methods such as random oversampling [14]. SMOTE is a method that oversamples minority data by creating artificial data [15]. SMOTE works to generate synthetic data along the line between a minority data and all its nearest neighbors [16]. The advantage of SMOTE is that it does not cause overfitting caused by data duplication as in random oversampling [17].

This study proposes the implementation and performance evaluation of the SMOTE method to improve the QPQ and HWE harassment classifications in the sexual harassment experience dataset. Improved classification performance can occur by overcoming the problem of data imbalance in the dataset between the QPQ and HWE classes. In the classification process, the term frequency-inverse document frequency (TF-IDF) method is used as the document weighting method. Then, naïve Bayes and (KNN) methods are two legacy machine learning techniques that are compared to classify the harassment experiences [8][9].

In evaluating the performance of SMOTE, we compare the classification performance of SMOTE-balanced data and unbalanced data. The measurement metrics are the receiver operating characteristics (ROC) curve, the area under the curve (AUC), confusion matrix, accuracy, precision, recall, and f1-Score.

To the best of our knowledge, previous research has not dealt with the data imbalance problem in sexual harassment classification, which is hazardous if the classification becomes a tool for delivering justice. The following are the contributions of our research:

1) A classification of sexual harassment experience based on the theprofessorisin.com dataset
2) A data imbalance improvement mechanism on sexual harassment classification using SMOTE
3) A novel sexual harassment classification method that increases the classification performance of minority class.

The remainder of this paper has the following systematics: Chapter II contains related works. Chapter III explains the proposed method. Chapter III presents the test results and discusses comparisons with the state-of-the-art studies. Finally, Chapter V highlights important findings.

II. LITERATURE REVIEW

Previous studies have analysed #MeToo tweets related to sexual harassment experiences. Putri *et al.* [8] classified QPQ and HWE on tweets with the hashtag #MeToo using the naïve Bayes algorithm with an accuracy of 88.5%. They carried out classification on a dataset crawled from Twitter. Modrek *et al.* [18] classified sexual assault and abuse based on #MeToo tweets obtained from crawling on Twitter. The study used the support vector machine (SVM) classification, but there were no imbalance problems in the dataset.

Stance classification is a classification of Twitter users' reactions to a #MeToo tweet. The classification usually detects, among others, hate speech and sarcasm. Sawhney *et al.* [19] held an internal competition regarding stance classification on the #MeTooMA dataset. There were 10 participants, and each used a different classification method, such as the convolutional neural network (CNN), recurrent neural network (RNN), logistic regression, and other methods. The winner of the competition used CNN. The winner of the competition used CNN. Basu *et al.* [20] also performed stance classification on the #MeTooMa dataset. However, the study found that data imbalance occurred. The solution to this problem was by using the focal loss method.

Reyes-Menendez *et al.* [21] made a sentiment analysis based on #MeToo tweets and divided it into three categories, namely positive indicators, neutral indicators, and negative indicators. The study used SVM for classification and Krippendorff's alpha value (KAV) to verify the classification performance. Priyanshu *et al.* [22] used several machine learning methods and compared them for stance classification on the #MeTooMA dataset. The study realized that data imbalance occurred but did not apply any method to overcome it. The process with the best type was naïve Bayes for hate speech class and random forest for sarcasm.

A comparative research presentation on sexual harassment classification is available in Table I for a clear view of the contributions mentioned. The result of standardized data can be seen in Table I.

TABLE I.
SEXUAL HARASSMENT CLASSIFICATION RESEARCH COMPARISON

Paper	Classification Problem	Database	Data Imbalance	Method
[8]	QPQ and HWE	Twitter Crawling	Yes	naïve Bayes
[18]	Sexual Assault and Abuse	Twitter Database	No	SVM
[19]	Stance Classification	#MeTooMA dataset	No	CNN
[20]	Stance Classification	#MeTooMA dataset	Yes	Focal Loss with BERT
[21]	Sentiment Analysis	Twitter Crawling	No	SVM
[22]	Stance Classification	#MeTooMA dataset	Yes	naïve Bayes
Proposed System	QPQ and HWE	Theprofessorisin.com	Yes	SMOTE with naïve Bayes and KNN Comparison

III. RESEARCH METHOD

Fig. 1 shows a flowchart that describes the research methodology. The research process is as follows: downloading the dataset from theprofessorisin.com., pre-processing, implementing TF-IDF, implementing SMOTE, conducting training and testing for KNN and naïve Bayes, evaluating the best model, and evaluating SMOTE. The result of standardized data can be seen in Fig.1.

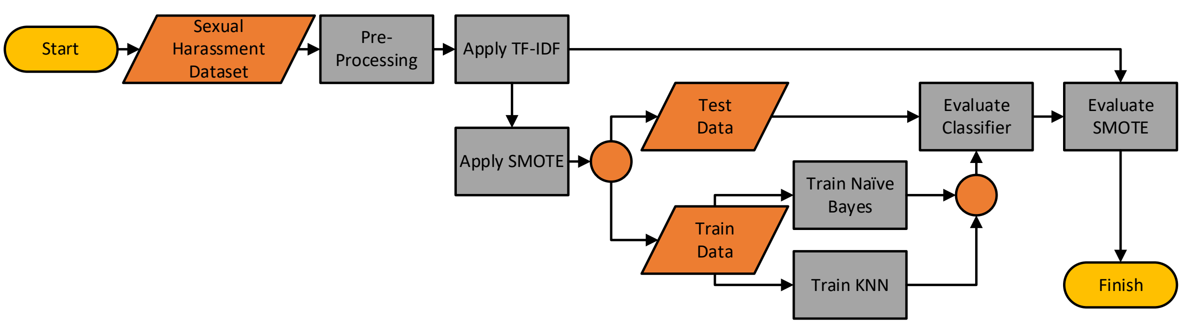

Fig. 1. Research methodology flowchart

A. Pre-Processing and TF-IDF

The dataset used in this study comes from theprofessorisin.com website and is available at the link https://github.com/amir-karami/Workplace_Sexual_Harassment [23]. There are two labels or classes in the dataset: QPQ and HWE. QPQ harassment at work is when the employer provides conditions related to sexuality to subordinates to complete a job. HWE is when an employee makes other employees uncomfortable because of illegal discrimination, such as dirty jokes, groping, and activities related to dirty photos.

There are several stages in the pre-processing step. The stages are as follows:

- Drop Null: Remove data rows that have no value
- Tokenization: Make sentences into word series
- Case Folding: Makes all letters lower case
- Data cleaning: Remove punctuation, numbers, multiple whitespaces, incomplete uniform resource locators (URL), and single chars.
- Remove Stopword: Removes words that are not main.
- Stemming: Removes word suffixes and returns a word to its root word.

Term frequency (TF) with Inverse Document Frequency (IDF) is a well-known method in text analysis for feature extraction. The TF-IDF determines a word's importance to a document in a corpus or collection [24]. TF is a method to calculate the frequency of a word's occurrence in a document, and IDF is the inverse, in which it calculates the frequency of a document containing certain words [25].

The equation of TF with the notation $tf(t,d)$ is as follows

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

where t is the resulting term, d is the document, and $f_{t,d}$ is the number of t in d. While t' is any other term besides t.

The equation of IDF with the notation $idf(t,D)$ is as follows

$$idf(t,D) = \log \frac{N}{1 + |\{d \in D, t \in d\}|} \tag{2}$$

where N states how many documents are in the corpus and the value $N = |D|$ and $|\{d \in D, t \in d\}|$ show how many documents t appears in.

B. SMOTE

SMOTE is an oversampling method to overcome imbalanced data in a dataset. Table II shows the imbalance data between QPQ and HWE in the sexual harassment experience dataset from theprofessorisin.com. The way SMOTE works is to generate synthetic samples for the minority class. This algorithm solves the overfitting problem that occurs in random oversampling. It focuses on feature space to create new data by interpolating between adjacent data. The result of standardized data can be seen in Table II.

TABLE II.
DATA IMBALANCE IN SEXUAL HARASSMENT EXPERIENCE DATASET

Class	Value Count	Percentage
HWE	1842	0.90
QPQ	211	0.10

The first step of SMOTE is calculating the k-nearest neighbour to a data x with Euclidean distance. The formula for the Euclidean distance in n dimensions is as follows

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

where $d(x, y)$ is the distance between points x and y, and y is the data in a dataset other than x.

The second step is to select the sampling rate N. N determines the number of data examples based on imbalance proportion. In a sample set A, with $x \in A$, N data will be selected from each neighbour x, while forming a set A'.

In the third step, new data x' will emerge from example x_k, where $x_k \in A'(k = 1,2,3, \ldots, N)$. The formula for generating new data x' is as follows

$$x' = x + rand(0,1) * |x - x_k| \quad (4)$$

where $rand(0,1)$ is a random number between 0 and 1.

C. *Classification and Evaluation*

The naïve Bayes classifier algorithm model has a very minimum error rate and is known for its simple, fast, and accurate calculations [26]. The use of naïve Bayes will be better if the training data is significant. Naïve Bayes builds a probabilistic model of terms [27]. The theorem combines with naivety, which assumes that the conditions between attributes are independent [28].

Here is the formula for determining a class with naïve Bayes

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (5)$$

where x is the instance to be classified, c is a specific class, $P(c|x)$ is the posterior knowledge, which is a number that classifies into one of the classes, and $P(c)$ is the prior knowledge, which results from data training.

KNN is a classification based on the shortest distance between the training data and the object to be classified [29]. Thus, this model is often called lazy learning [30]. The formula for calculating the distance between the classification target data and the train data uses the Euclidean distance formula in equation (3).

The evaluation stage uses the confusion matrix method to calculate the values of accuracy, precision, recall, and f1-score. Previous studies have used this evaluation method [31][32].

The ROC curve shows the characteristics of the true positive rate (TPR) against the false positive rate (FPR). AUC is the value that quantifies the ROC curve. The AUC formula is as follows

$$AUC = \sum_{k=1}^{N} \frac{f(x_{k-1}) + f(x_k)}{2} \Delta x_k \tag{6}$$

where x_k is the kth FPR value, $f(x_k)$ is the TPR value at x_k, and N is the number of existing TPR and FPR values. The range of AUC values is from 0 to 1. An AUC value close to 1 indicates good ROC curve performance.

IV. Results And Discussion

A. Results

TF-IDF determines how important a word is to a document in a corpus or collection. The greater the value of the IDF result, the more influential the word is to a document. Table III shows the top ten most important words in the sexual harassment experience dataset based on TF-IDF calculations. The result of standardized data can be seen in Table III.

TABLE III.
TF-IDF RESULT OF SEXUAL HARASSMENT EXPERIENCE DATASET

Word	TF-IDF
affair	0.384347
unsubstantiential	0.330024
convent	0.313159
student	0.303662
gossip	0.263081
typic	0.259116
somewhat	0.246216
rumor	0.202411
suppos	0.197836
attract	0.185225

After applying SMOTE on the dataset and the KNN and naïve Bayes training, the results are classification models. The performance metrics of the models are the ROC curve and AUC. Fig. 2 shows the resulting ROC curve. The model with the curve with the larger AUC value has a better performance. The result of standardized data can be seen in Fig.2.

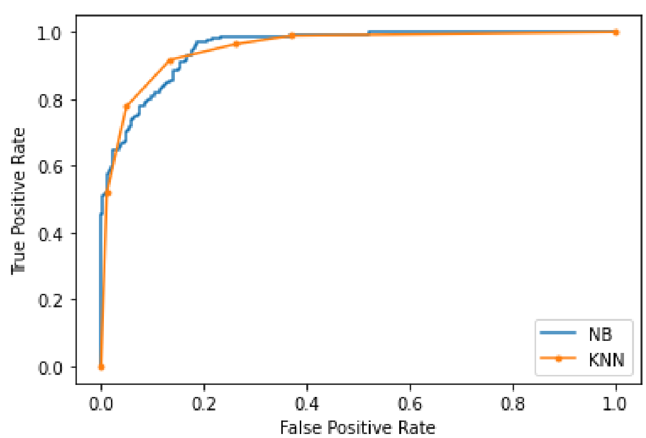

Fig. 2. ROC curve comparison of the naïve Bayes (NB) and KNN classifier.

The calculation of the AUC score based on the ROC curve uses equation (6). The AUC score of naïve Bayes is 0.954935 and the AUC score of KNN is 0.923682. From these two results, naïve Bayes has a better AUC than KNN. Table IV shows the results of the comparison. The result of standardized data can be seen in Table IV.

TABLE IV.
CLASSIFIER AUC COMPARISON

Classifier	AUC Score
naïve Bayes	0.954935
KNN	0.923682

Fig. 3 compares the confusion matrix between naïve Bayes and KNN. Based on the two confusion matrices, naïve Bayes has a higher true positive (TP) value and true negative (TN) value than KNN, which are 397 versus 351 and 411 versus 408, respectively. This indicates that naïve Bayes has a better classifier performance than KNN. The result of standardized data can be seen in Fig.3.

(a) (b)

Fig. 3. Confusion matrix comparison of naïve Bayes (a) and KNN (b) classifier

Table V compares the accuracy, precision, recall, and f1-score between naïve Bayes and KNN, where the value of naïve Bayes accuracy is 91%, and for KNN, it is 83.5%. The value of naïve Bayes precision is 91%, and for KNN, it is 85.5%. The recall value generated for naïve Bayes is 90.5%, and for KNN, it is 85%. The F-1 score generated from naïve Bayes is 91% and KNN is 85%. In these four parameters, naïve Bayes has a better performance than KNN. The result of standardized data can be seen in Table V.

TABLE V.
PERFORMANCE COMPARISON OF NAÏVE BAYES AND KNN

Classifier	Performance			
	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.91	0.91	0.91	0.91
KNN	0.84	0.86	0.85	0.85

Data imbalance will harm the minority class. In this case, the minority class is QPQ because the proportion is only 10% of the dataset. Fig. 4 shows a bar chart comparing the performance of a naïve Bayes classifier when using SMOTE and not using a balancing method. In the bar chart, SMOTE can improve the classification performance of the QPQ class, especially in terms of precision values. Comparing these values shows that, with

SMOTE oversampling, the model makes fewer mistakes in classifying QPQ as HWE. The result of standardized data can be seen in Fig.4.

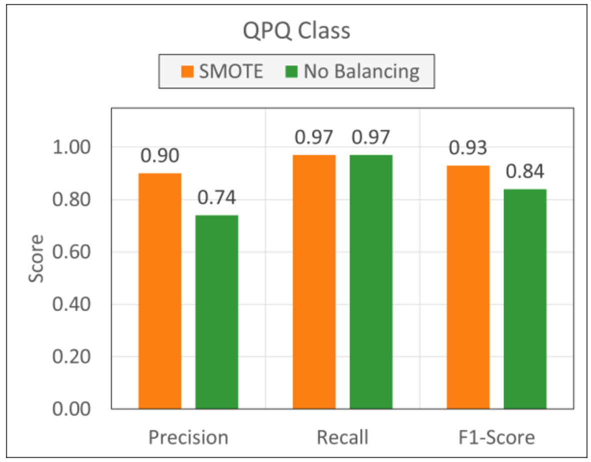

Fig. 4. Data imbalance improvement in the QPQ Class (Minority Class) with SMOTE method.

B. *Discussion*

The test results show that SMOTE on naïve Bayes is more effective than SMOTE on KNN. According to the paper [33], naïve Bayes and KNN have very different characteristics. Naïve Bayes has a high bias and low variance model. On the other hand, KNN has a low bias and high variance model. Usually, the characteristics mentioned by the naïve Bayes model are suitable for linearly separable data. The superior performance of naïve Bayes shows that the data, in this case, is linearly separable.

This research refers to Putri *et al.* [16], who had previously classified QPQ and HWE on #MeToo tweets. The study results are 89% for accuracy, 91% for precision, 97% for recall, and 94% for f1-score. Some values are higher than those obtained from the proposed study results. However, when viewed from the class composition in the dataset, the benchmark study also experienced imbalanced data. The number of datasets was 833, with 86% of the data being HWE and 14% being QPQ. The performance calculated in that study assumed that the majority class is the true positive. A setting like that would hide the weak performance of the minority class.

To make comparisons, we calculate the precision, recall, and f1-score of the minority class from the benchmark paper. Table VI shows that, by only observing the minority class, the SMOTE – Naïve Bayes method proposed in this study outperforms the performance of the minority class classification of the benchmark paper for precision, recall, and f1-Score. The result of standardized data can be seen in Table VI.

TABLE VI.
BENCHMARKING RESULTS ON MINORITY CLASS

Cite	**Method**	**Performance**		
		Precision	***Recall***	***F1-Score***
[8]	Naïve Bayes	39%	64%	48%
Proposed Method	Naïve Bayes	74%	97%	84%
	SMOTE-Naïve Bayes	90%	97%	93%

We have thoroughly checked the threats for the validity of our research results. Threats such as history, maturation, instrumentation, and testing do not apply to our case because our problem, which discusses data imbalance, does not relate to history and maturation. We also state that our research is valid in instrumentation and testing. We use the same methods and tools for both imbalanced and balanced data cases. Also, the prediction of balanced data is not affected by the results of the imbalanced data.

V. CONCLUSION

This research has successfully implemented a naïve Bayes classifier to classify QPQ and HWE on the sexual harassment experience dataset. The classifier implements SMOTE to overcome the imbalance of data contained in the dataset. The comparison shows that the performance of the naïve Bayes classifier is superior to the KNN classifier in classifying QPQ and HWE, with AUC values of 0.95 versus 0.92, respectively. The evaluation results show that by applying the SMOTE method to the naïve Bayes classifier, the precision of the minority class can increase from 74% to 90%.

REFERENCES

[1] B. Fileborn, "Justice 2.0: Street harassment victims' use of social media and online activism as sites of informal justice," *British journal of criminology*, vol. 57, no. 6, pp. 1482–1501, 2017.
[2] W. Perkins and J. Warner, "Sexual Violence Response and Prevention: Studies of Campus Policies and Practices," *Journal of School Violence*, vol. 16, no. 3, pp. 237–242, Jul. 2017, doi: 10.1080/15388220.2017.1318569.
[3] D. N. Simorangkir, M. S. Saraswati, E. Melissa, L. L. Perangin-Angin, and S. Schumacher, "RAISING AWARENESS ABOUT SEXUAL HARASSMENT IN THE MEDIA INDUSTRY," *Jurnal Sinergitas PKM dan CSR*, vol. 4, no. 3, 2020.
[4] Y. N. Pappoe, "The shortcomings of Title VII for the Black female plaintiff," *U. Pa. JL & Soc. Change*, vol. 22, p. 1, 2019.
[5] C. Girgis, "Sexual Harassment," in *Burnout in Women Physicians*, Springer, 2020, pp. 105–128.
[6] G. Chandra, R. Gupta, and N. Agarwal, "Role of artificial intelligence in transforming the justice delivery system in covid-19 pandemic," *Chandra, G., Gupta, R. and Agarwal*, no. 2020, pp. 344–350, 2020.
[7] E. Alawneh, M. Al-Fawa'reh, M. T. Jafar, and M. Al Fayoumi, "Sentiment analysis-based sexual harassment detection using machine learning techniques," in *2021 international symposium on electronics and smart devices (ISESD)*, 2021, pp. 1–6.
[8] T. A. M. Putri, U. Enri, and B. N. Sari, "Analisis Algoritma Naive Bayes Classifier untuk Klasifikasi Tweet Pelecehan Seksual dengan #MeToo," p. 10.
[9] M. Haque *et al.*, "Data Mining Techniques to Categorize Single Paragraph-Formed Self-narrated Stories," in *ICT Analysis and Applications*, Springer, 2021, pp. 701–713.
[10] J. Jang, Y. Kim, K. Choi, and S. Suh, "Sequential Targeting: an incremental learning approach for data imbalance in text classification," *arXiv preprint arXiv:2011.10216*, 2020.
[11] M. Khushi *et al.*, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021.
[12] A. Sarker *et al.*, "Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1274–1283, 2018.
[13] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.
[14] C. Zhang *et al.*, "An empirical study on the joint impact of feature selection and data resampling on imbalance classification," *Applied Intelligence*, pp. 1–13, 2022.
[15] K. Polat, "A hybrid approach to Parkinson disease classification using speech signal: The combination of SMOTE and random forests," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–3.
[16] A. S. Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1412–1422, 2019.
[17] F. Last, G. Douzas, and F. Bacao, "Oversampling for imbalanced learning based on k-means and smote," *arXiv preprint arXiv:1711.00837*, 2017.

[18] S. Modrek and B. Chakalov, “The# MeToo movement in the United States: text analysis of early twitter conversations,” *Journal of medical Internet research*, vol. 21, no. 9, p. e13837, 2019.
[19] R. Sawhney, A. K. Gautam, and R. R. Shah, “BMGC 2020 Grand Challenge: Multi-Aspect Analysis of the MeToo Movement on Twitter,” in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 481–484.
[20] P. Basu, S. Tiwari, J. Mohanty, and S. Karmakar, “Multimodal Sentiment Analysis of# MeToo Tweets using Focal Loss (Grand Challenge),” in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 461–465.
[21] A. Reyes-Menendez, J. R. Saura, and F. Filipe, “Marketing challenges in the# MeToo era: Gaining business insights using an exploratory sentiment analysis,” *Heliyon*, vol. 6, no. 3, p. e03626, 2020.
[22] A. Priyanshu *et al.*, “Stance Classification with Improved Elementary Classifiers Using Lemmatization (Grand Challenge),” in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 466–470.
[23] A. Karami, S. C. Swan, C. N. White, and K. Ford, “Hidden in plain sight for too long: Using text mining techniques to shine a light on workplace sexism and sexual harassment.,” *Psychology of Violence*, 2019.
[24] S. Qaiser and R. Ali, “Text mining: use of TF-IDF to examine the relevance of words to documents,” *International Journal of Computer Applications*, vol. 181, no. 1, pp. 25–29, 2018.
[25] D. Prabowo, M. Fhadli, M. Najib, H. Fauzi, and I. Cholissodin, “TF-IDF-Enhanced Genetic Algorithm Untuk Extractive Automatic Text Summarization,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 3, p. 208, Sep. 2016, doi: 10.25126/jtiik.201633217.
[26] A. M. Putrada, M. Abdurohman, and A. G. Putrada, “Increasing smoke classifier accuracy using naive bayes method on internet of things,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 19–26, 2019.
[27] A. P. Wijaya and H. A. Santoso, “Naive bayes classification pada klasifikasi dokumen untuk identifikasi konten e-government,” *Journal of Applied Intelligent System*, vol. 1, no. 1, pp. 48–55, 2016.
[28] K. Y. Raharja, “PERBANDINGAN KINERJA ALGORITMA GAUSSIAN NAIVE BAYES DAN K-NEAREST NEIGHBOR (KNN) UNTUK MENGKLASIFIKASI PENYAKIT HEPATITIS C VIRUS (HCV),” PhD Thesis, Universitas Muhammadiyah Jember, 2021.
[29] P. Nando, A. G. Putrada, and M. Abdurohman, “Increasing The Precision Of Noise Source Detection System using KNN Method,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 157–168, 2019.
[30] M. Ramadhani and D. H. Murti, “Klasifikasi Ikan Menggunakan Oriented Fast and Rotated Brief (Orb) Dan K-Nearest Neighbor (Knn),” *JUTI J. Ilm. Teknol. Inf*, vol. 16, no. 2, p. 115, 2018.
[31] A. G. Putrada, N. G. Ramadhan, and M. Abdurohman, “Context-aware smart door lock with activity recognition using hierarchical hidden markov model,” *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 5, no. 1, pp. 37–44, 2020.
[32] N. G. Ramadhan, A. G. Putrada, and M. Abdurohman, “Improving Smart Lighting with Activity Recognition Using Hierarchical Hidden Markov Model,” *Indonesia Journal on Computing (Indo-JC)*, vol. 4, no. 2, pp. 43–54, 2019.
[33] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, “Machine Learning Methods in Smart Lighting Toward Achieving User Comfort: A Survey,” *IEEE Access*, vol. 10, pp. 45137–45178, 2022, doi: 10.1109/ACCESS.2022.3169765.