Enhancing Disease Symptom Analysis in Thai Text: Methods for Text Oversampling in Imbalanced Data for Disease Detection

Ekkarat Boonchieng (Senior Member, IEEE)
Department of Computer Science
Chiang Mai University
Chiang Mai 50200, Thailand
ekkarat.boonchieng@cmu.ac.th

Wanchaloem Nadda
Department of Computer Science
Chiang Mai University
Chiang Mai 50200, Thailand
wanchaloem.nadda@gmail.com

Wongthawat Liawrungrueang
Department of Orthopaedics
School of Medicine, University of Phayao
Phayao 56000, Thailand
mint11871@hotmail.com

Waraporn Boonchieng*
Faculty of Public Health
Chiang Mai University
Chiang Mai 50200, Thailand
waraporn.b@cmu.ac.th
(Corresponding Author)

***Abstract*— This study employs machine learning and natural language processing (NLP) in the field of disease detection and management while focusing on Thai language texts. We have introduced two innovative text oversampling methods to address the challenges associated with imbalanced datasets in medical diagnostics: Text Oversampling Using Keyword Selection and Text Oversampling Using Synonym Words. The Keyword Selection method identifies and selectively removes specific misspelled words and stop words, while the Synonym Words method replaces certain words with their synonyms, thereby enhancing data quality and the effectiveness of model training.**

Our approach utilizes unstructured data obtained from electronic health records. It addresses the complex issue of symptomatic overlap for certain diseases such as dengue hemorrhagic fever, common migraine, common cold, influenza, and tonsillitis. The concept of symptomatic overlap presents significant challenges in achieving accurate disease diagnoses. This study primarily focused on the Saraphi Hospital dataset, which was characterized by a high imbalance ratio. Notably, the Keyword Selection method demonstrated superior performance over other techniques in terms of the f1-score, underscoring its efficacy.

By reducing data complexity and preventing overfitting, the Keyword Selection method enhanced the accuracy of disease classification. Our findings suggest that these novel oversampling methods can significantly improve the process of disease identification and management, marking an important advancement in the integration of machine learning in disease detection.

Keywords*— *Machine Learning, Text Classification, Oversampling, Digital Disease Detection, Imbalanced Data Problem

I. Introduction

In an era where digital innovation intersects with healthcare, data analysis and artificial intelligence (AI) play pivotal roles in disease detection and management. The synergy between AI and the vast repositories of electronic health records has been essential in enabling health care professionals to enhance disease detection and diagnosis with unprecedented precision [1], [2]. The COVID-19 pandemic has underscored the critical role of digital tools, particularly in interpreting radiological reports, which have become a cornerstone of the diagnostic process [1]. Beyond imaging, textual analysis has also become crucial. This has resulted in an increase in the development of algorithms that can capture language subtleties, which can then be employed to identify specific health issues [3].

Social media platforms have been transformed into vital instruments for public health surveillance, reflecting the health-related concerns of populations in real-time [4]. The ongoing dialogue on these platforms about common illnesses, such as the flu and dengue, has provided researchers with an invaluable dataset for health monitoring. Nonetheless, distilling actionable insights from the vast, unstructured content of social media requires the use of advanced text classification techniques capable of navigating this complexity [5]. Our approach is founded on advancements in computational models that have significantly improved disease detection and prediction [6]. Machine learning, in particular, has spurred the evolution of epidemiological data analysis, leading to more robust models for forecasting the spread of diseases [7]. The integration of AI into diagnostic workflows has not only sped up the diagnostic process but has also increased its precision, resulting in better patient outcomes through early and tailored interventions [8].

In healthcare, Natural Language Processing (NLP) has emerged as a transformative force, facilitating the extraction of vital information from medical texts and streamlining the classification of diseases from clinical narratives [9]. Its role extends to the realm of enhancing communication between patients and healthcare providers by translating complex medical information into understandable language [10].

The convergence of data science and healthcare has given rise to sophisticated health informatics tools. Meanwhile, text analytics has become indispensable in capturing the health trends identified from social media and other digital sources.

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00068

The integration of machine learning with linguistic data opens new avenues for automated health advice systems, and AI's entrance into predictive modeling signals a shift toward a more proactive and preventative healthcare framework.

Distinguishing between diseases, such as the common cold, flu, dengue, and tonsillitis, is further complicated by their symptomatic overlap, necessitating more refined diagnostic methods [11]. This intricate task is further compounded by the varied symptom descriptions that exist across different populations, presenting a significant diagnostic challenge [12]. Therefore, it is crucial to develop classification tools that can discern between these conditions with a high degree of accuracy, ensuring the appropriate treatment [13].

Imbalanced datasets, where instances of certain diseases are rare, present a significant challenge in disease classification. To address this, the research community has turned to oversampling techniques like SMOTE, which has revolutionized the field by creating synthetic samples to improve model performance [14], [15]. However, applying such techniques to textual data is not straightforward due to the unique properties of language [16]. This has led to the development of innovative methods that have modifies existing records [17].

Our study will introduce a novel oversampling approach tailored for textual data. By strategically omitting keywords from input records, we can generate new and varied data instances. This method meticulously preserves the semantic integrity of the original data while fostering variations that strengthen the classifier's ability to distinguish between diseases, a critical factor when dealing with the nuanced language of symptom descriptions.

II. Related Works

In this section, we will explain the methods used for text classification in this research study.

A. Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) [8] is a technique in natural language processing (NLP). Its main purpose is to convert text in a document or a record into a vector of real numbers, making it more comprehensible for machine learning models. Below is a detailed explanation of how TF-IDF functions:

- Term Frequency (TF): This metric measures how often a term appears within a document. To calculate TF, divide the number of times a term occurs in the document by the total number of terms in that document. This normalization adjusts for differing document lengths, ensuring a fair representation of term frequency.

- Inverse Document Frequency (IDF): This represents the significance of a term across a set of documents. It is computed by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. This step lessens the influence of terms that appear frequently in many documents (such as "the", "is", "at"), as they typically offer less unique information.

TF-IDF score is derived by multiplying the TF and IDF values. A high TF-IDF score indicates that a term is not only prevalent in a specific document but also relatively rare across various documents, suggesting its greater importance and distinctiveness in characterizing that document [18], [19].

B. Imbalanced data

In machine learning and data classification, the issue of imbalanced data is a significant challenge. This occurs when the number of records in one class exceeds the number of records in another within a dataset, leading to models that may unintentionally favor the more populous class [16].

The problem of imbalance is especially noticeable in areas like disease detection, where accurately predicting the less represented class is vital. This difficulty arises because many machine learning algorithms are initially designed to work with an equal distribution of instances among various classes [15].

To address this imbalance, various algorithms have been developed. These include methods to increase the number of instances in the minority class (known as oversampling), decrease instances in the majority class (undersampling), or implement advanced techniques like the Synthetic Minority Over-sampling Technique (SMOTE) [14]. The main objective of these approaches is to create a dataset that is more balanced across classes, facilitating the development of models that are fair and unbiased in their predictions.

The paper proposes a new ensemble oversampling method combining logistic regression and three-way decisions (3WD) for automatic keyword extraction from policy texts, addressing the issue of unbalanced data sets and demonstrating superior performance over traditional methods in various experiments [20].

The proposed T5W paraphrasing approach for oversampling significantly improves the performance of text classification algorithms by effectively balancing imbalanced textual datasets through automated paraphrasing and integration with a Robotic Process Automation tool [21].

III. Our Text Oversampling Methods

Our methods concentrate on analyzing the Thai language text that is used to describe the symptoms of patients believed to have one of five diseases: dengue hemorrhagic fever, common migraine, common cold, influenza, and tonsillitis. Each patient record contains symptoms (approximately 10 words), along with numerical data on body temperature, the month of service, blood pressure, gender, and the age of the patient.

A. Text Oversampling Using Keyword Selection

For this method, we utilize a list of Thai words from the PyThaiNLP library, a Python library containing Thai language words. Additionally, we have used a list of Thai stop words from the PyThaiNLP library. These stop words, similar to commonly

used words like 'the' and 'is' in English, are often filtered out in text processing as they provide little meaningful information for specific tasks such as those involving search and analysis. Removing these words can guide researchers to focus on the more significant words present in the text. We will demonstrate the 'Text Oversampling Using Keyword Selection' method for the text in the minority classes and provide an example focusing on symptoms, presented in English, as follows:

1) Receive the text detailing a patient's symptoms. For example, consider data from a tonsillitis patient with symptoms “Fever, cough with phlegm, sore throat, and has mucus,” with a body temperature of 37 degrees Celsius, and gender noted as “male”.

2) Tokenize the text into a list of words, for example, [“Fever”, “cough”, “phlegm”, “sore throat”, “and”, “has”, “mucus”].

3) Identify words that are not in the word list (meaning they are misspelled) and the words that are stop words, for example, [“Fever”, “cough”, “phlegm”, “sore throat”, “and”, “has”, “mucus”].

4) Subsequently, we identify 7 different scenarios for the potential removal of each word from this text, as has been illustrated in Table I.

5) Then, apply the TF-IDF (Term Frequency-Inverse Document Frequency) transformation to convert each symptom into a vector and incorporate the numerical features into each vector.

B. Text Oversampling Using Synonym Words

Headings, We will demonstrate the 'Text Oversampling Using Synonym Words' method for the text in the minority classes and provide an example focusing on symptoms, presented in English, as follows:

1) Receive the text detailing a patient's symptom example, consider data from a tonsillitis patient with symptoms “Fever, cough with phlegm, sore throat, and has mucus,” with a body temperature of 37 degrees Celsius, and gender noted as “male”.

2) Tokenize the text into a list of words, for example, [“Fever”, “cough”, “phlegm”, “sore throat”, “and”, “has”, “mucus”].

3) Identify the synonyms for each word using the PyThaiNLP library, and then compile a list of these synonyms for each word. For example,

[[“Fever”], [“cough”],

[“phlegm”, “sputum”],

[“sore throat”],

[“and”], [“has”, “have”],

[“mucus”, “mucous secretion”]].

4) Subsequently, we identified 7 different scenarios for replacing each word in this text with a synonym, as has been illustrated in Table II.

5) Then, apply the TF-IDF (Term Frequency-Inverse Document Frequency) transformation to convert each symptom into a vector and incorporate the numerical features into each vector

TABLE I. EXAMPLE DATA FOR TEXT OVERSAMPLING USING KEYWORD SELECTION

No.	List of Words are Removed	New Text of Symptoms	Gender	Body Temperate (degrees Celsius)	Class
1	[“Fever”]	[“cough”, “phlegm”, “sore throat”, “and”, “has”, “mucus”]	Male	37	Tonsillitis
2	[“and”]	[“Fever”, “cough”, “phlegm”, “sore throat”, “has”, “mucus”]	Male	37	Tonsillitis
3	[“has”]	[“Fever”, “cough”, “phlegm”, “sore throat”, “and”, “mucus”]	Male	37	Tonsillitis
4	[“Fever”, “and”]	[“cough”, “phlegm”, “sore throat”, “has”, “mucus”]	Male	37	Tonsillitis
5	[“Fever”, “has”]	[“cough”, “phlegm”, “sore throat”, “and”, “mucus”]	Male	37	Tonsillitis
6	[“and”, “has”]	[“Fever”, “cough”, “phlegm”, “sore throat”, “mucus”]	Male	37	Tonsillitis
7	[“Fever”, “and”, “has”]	[“cough”, “phlegm”, “sore throat”, “mucus”]	Male	37	Tonsillitis

TABLE II. EXAMPLE DATA FOR TEXT OVERSAMPLING USING KEYWORD SELECTION

No.	New Text of Symptoms	Gender	Body Temperate (degrees Celsius)	Class
1	["Fever", "cough", "phlegm", "sore throat", "and", "has", "mucus secretion"]	Male	37	Tonsillitis
2	["Fever", "cough", "phlegm", "sore throat", "and", "have", "mucus"]	Male	37	Tonsillitis
3	["Fever", "cough", "phlegm", "sore throat", "and", "have", "mucus secretion"]	Male	37	Tonsillitis
4	["Fever", "cough", "sputum", "sore throat", "and", "has", "mucus"]	Male	37	Tonsillitis
5	["Fever", "cough", "sputum", "sore throat", "and", "has", "mucus secretion"]	Male	37	Tonsillitis
6	["Fever", "cough", "sputum", "sore throat", "and", "have", "mucus"]	Male	37	Tonsillitis
7	["Fever", "cough", "sputum", "sore throat", "and", "have", "mucus secretion"]	Male	37	Tonsillitis

IV. EXPERIMENTS

Our model's training and testing data comprises patient medical records obtained from Saraphi Hospital in Chiangmai Province, Thailand, spanning from January 2010 to May 2021. The features integrated into our models include the patient's gender, age at the time of service, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), the month of service, and the symptoms reported during their visit. Table III and Table IV present essential statistical information about this dataset. The classification models' outputs are categorized into five classes: dengue hemorrhagic fever, common migraine, common cold, influenza, and tonsillitis.

TABLE III. MEAN AND STANDARD DEVIATION OF NUMERICAL FEATURES OF INPUT DATA

Features	Average	Standard Deviation
Body temperature	36.760	5.760
Age	41.160	24.040
SBP	83.932	59.469
DBP	49.472	35.260
Length of symptom of each record	10.929	8.048

TABLE IV. FREQUENCY OF EACH VALUE OF CATEGORICAL FEATURES OF DATASET

Columns	Frequency of each value
Gender	{male: 6256, female: 7006}
Month of Service	{01: 1586, 02: 1311, 03: 1228, 04: 685, 05: 586, 06: 710, 07: 817, 08: 870, 09: 1111, 10: 1340, 11: 1525, 12: 1493}
Diseases	{dengue hemorrhagic fever: 65, common migraine: 277, common cold: 10740, influenza: 61, tonsillitis: 2119}

To train the classification models, we divided the data into two sets: a training set comprising 80% of the data and a test set consisting of the remaining 20%. We only applied oversampling to the training set, before training each classification model using 5-fold cross validation. Our method, which included 'Text Oversampling Using Keyword Selection' and 'Text Oversampling Using Synonym Words', will be compared with SMOTE [14] and a basic oversampling method that involved duplicating records in the minority class. Subsequently, we will then calculate the F1-score to measure the performance of each model.

V. RESULTS AND DISCUSSION

In this section, we present the F1-scores for four classification models: Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Decision Tree, and Multi-Layer Perceptron (MLP). These models were tested with four different oversampling methods, including our 'Text Oversampling Using Keyword Selection' and 'Text Oversampling Using Synonym Words,' as well as SMOTE, and a basic oversampling method

that involves duplicating records in the minority class. The results are displayed in Table V.

TABLE V. F1-SCORE OF EACH CLASSIFICATION AND EACH OVERSAMPLING METHOD

Classification Models \ Oversampling methods	copy	SMOTE	Text Oversampling Using Keyword Selection	Text Oversampling Using Synonym Words
SVM	0.434	0.405	**0.441**	0.426
k-NN	0.342	0.329	**0.351**	0.327
MLP	0.366	0.405	0.410	**0.434**
Decision Tree	0.345	0.358	**0.366**	0.330

The results for the Saraphi dataset, which have been characterized by a high degree of imbalance (Imbalance Ratio = 10740/61 = 176.065), indicate that the Text Oversampling Using Keyword Selection method achieved a higher f1-score when compared to the other methods. This improvement may be attributed to the method's ability to eliminate stop words and correct misspellings, thereby reducing data complexity and mitigating overfitting in the classification models.

VI. CLINICAL APPLICATION AND DISCUSSION

A current review conducted by a qualified medical team is very important in medical diagnosis. ICD-9 (International Classification of Diseases, 9th Revision) and ICD-10 (International Classification of Diseases, 10th Revision) are systems used for medical classification worldwide. They categorize diseases, injuries, medical diagnoses, and health-related issues for various purposes like billing, statistical tracking, and epidemiological studies. ICD9 and ICD10 are keywords that healthcare workers must understand. However, the key difference across national languages, such as Thai, exists in the variations that are present in the medical terminology spelling of these terms, while maintaining the same meanings. As a result, the code may include an error, which can be a challenging point. The usage of expert ICD coders in Thai languages has become a significant challenge for management teams due to the transition from ICD9 to ICD10. Several hospitals have attempted to use AI text-based models to perceive visual content such as photographs or handwritten text directly. There is a wide array of AI-driven tools and software that can proficiently identify and comprehend handwritten writing. To use Optical Character Recognition (OCR) technology, consider using tools such as Adobe Acrobat, Microsoft OneNote, Google Keep, or specialized OCR software like Tesseract or Abbyy FineReader. These programs have the capability of analyzing photos that include handwritten text and transform them into text that can be edited. However, the use of Thai language has been challenging and has proven to be susceptible to several errors. Therefore, this study makes use of electronic or computerized health records.

Authors' unorganized data have been extracted from electronic health records. This step tackles the intricate problem of overlapping symptoms in certain illnesses including dengue hemorrhagic fever, common migraine, common cold, influenza, and tonsillitis. This study offers numerous advantages for medical diagnosis, revolutionizing the healthcare industry in various ways including Improved Accuracy, Early Detection and Diagnosis, Efficiency and Speed, Personalized Medicine, Assistance in Decision-Making, Enhanced Imaging Analysis, Reduction in Diagnostic Errors, and Research and Development. The presence of similar symptoms has produced considerable difficulties in accurately diagnosing diseases. All authors have presented a unique oversampling technique specifically designed for textual data in our research. We have produced novel and diverse data instances by deliberately excluding certain terms from the input records. This strategy carefully maintains the meaning and accuracy of the original data while promoting modifications that can enhance the classifier's capacity to differentiate between illnesses. This is particularly important when working with the intricate language that is typically used in symptom descriptions. To our knowledge, this is the first medical diagnostic investigation of its kind produced in Thailand. Nevertheless, other diseases may need to be included in the medical diagnostic system in order to broaden its application in future research work.

VII. CONCLUSION

This study showcases an innovative approach that involves analyzing Thai language text to identify symptoms of patients with certain diseases such as dengue hemorrhagic fever, common migraine, common cold, influenza, and tonsillitis. The research has proposed two methods for oversampling text data: 'Text Oversampling Using Keyword Selection' and 'Text Oversampling Using Synonym Words'. These methods were specifically applied to texts in minority classes to tackle the challenges posed by the highly imbalanced Saraphi dataset (Imbalance Ratio = 10740/61 = 176.065). The results were particularly noteworthy, with the Keyword Selection method outperforming others in terms of f1-score. This method proved effective in reducing data complexity and preventing overfitting in classification models, which was achieved by eliminating stop words and correcting misspellings.

REFERENCES

[1] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, and M. T. Martín-Valdivia, "COVID-19 detection in radiological text reports integrating entity recognition," Comput Biol Med, vol. 127, p. 104066, 2020, doi: https://doi.org/10.1016/j.compbiomed.2020.104066.

[2] A. Borjali, M. Magnéli, D. Shin, H. Malchau, O. K. Muratoglu, and K. M. Varadarajan, "Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation," Comput Biol Med, vol. 129, p. 104140, 2021, doi: https://doi.org/10.1016/j.compbiomed.2020.104140.

[3] H. Wang et al., "Diagnosis of dairy cow diseases by knowledge-driven deep learning based on the text reports of illness state," Comput Electron Agric, vol. 205, p. 107564, 2023, doi: https://doi.org/10.1016/j.compag.2022.107564.

[4] S. Muñoz and C. A. Iglesias, "A text classification approach to detect psychological stress combining a lexicon-based feature framework with

distributional representations," Inf Process Manag, vol. 59, no. 5, p. 103011, 2022, doi: https://doi.org/10.1016/j.ipm.2022.103011.

[5] Y. Li, H. Guo, Q. Zhang, M. Gu, and J. Yang, "Imbalanced text sentiment classification using universal and domain-specific knowledge," Knowl Based Syst, vol. 160, pp. 1–15, 2018, doi: https://doi.org/10.1016/j.knosys.2018.06.019.

[6] N. N. Qomariyah, A. S. Araminta, R. Reynaldi, M. Senjaya, S. D. A. Asri, and D. Kazakov, "NLP Text Classification for COVID-19 Automatic Detection from Radiology Report in Indonesian Language," in 2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2022, pp. 565–569. doi: 10.1109/ISRITI56927.2022.10053077.

[7] A. Alessa, M. Faezipour, and Z. Alhassan, "Text Classification of Flu-Related Tweets Using FastText with Sentiment and Keyword Features," in 2018 IEEE International Conference on Healthcare Informatics (ICHI), 2018, pp. 366–367. doi: 10.1109/ICHI.2018.00058.

[8] S. Amin et al., "Recurrent Neural Networks With TF-IDF Embedding Technique for Detection and Classification in Tweets of Dengue Disease," IEEE Access, vol. 8, pp. 131522–131533, 2020, doi: 10.1109/ACCESS.2020.3009058.

[9] X. Zhang, J. Wang, N. Cheng, and J. Xiao, Improving Imbalanced Text Classification with Dynamic Curriculum Learning. 2022. doi: 10.48550/arXiv.2210.14724.

[10] L. Da Quach, A. Quynh, K. Nguyen, and A. Nguyen, "Using the Term Frequency-Inverse Document Frequency for the Problem of Identifying Shrimp Diseases with State Description Text," International Journal of Advanced Computer Science and Applications, vol. 14, p. 2023, May 2023, doi: 10.14569/IJACSA.2023.0140577.

[11] M. Torii et al., "Risk factor detection for heart disease by applying text analytics in electronic medical records," J Biomed Inform, vol. 58, pp. S164–S170, 2015, doi: https://doi.org/10.1016/j.jbi.2015.08.011.

[12] A. López Pineda, Y. Ye, S. Visweswaran, G. F. Cooper, M. M. Wagner, and F. (Rich) Tsui, "Comparison of machine learning classifiers for influenza detection from emergency department free-text reports," J Biomed Inform, vol. 58, pp. 60–69, 2015, doi: https://doi.org/10.1016/j.jbi.2015.08.019.

[13] W. Nadda, W. Boonchieng, and E. Boonchieng, "Influenza, dengue and common cold detection using LSTM with fully connected neural network and keywords selection," BioData Min, vol. 15, no. 1, 2022, doi: 10.1186/s13040-022-00288-9.

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

[15] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," in 2021 Sixth International Conference on Informatics and Computing (ICIC), 2021, pp. 1–8. doi: 10.1109/ICIC54025.2021.9632912.

[16] S. Akkaradamrongrat, P. Kachamas, and S. Sinthupinyo, "Text Generation for Imbalanced Text Classification," in 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2019, pp. 181–186. doi: 10.1109/JCSSE.2019.8864181.

[17] G. Xu, Z. Niu, X. Gao, and H. Liu, "Imbalanced text classification on host pathogen protein-protein interaction documents," in 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE), 2010, pp. 418–422. doi: 10.1109/ICCAE.2010.5451921.

[18] G. Sun, Y. Cheng, Z. Zhang, X. Tong, and T. Chai, "Text classification with improved word embedding and adaptive segmentation," Expert Syst Appl, vol. 238, p. 121852, 2024, doi: https://doi.org/10.1016/j.eswa.2023.121852.

[19] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf Process Manag, vol. 24, no. 5, pp. 513–523, 1988, doi: https://doi.org/10.1016/0306-4573(88)90021-0.

[20] Liang, D., Yi, B., Cao, W., & Zheng, Q. (2022). Exploring ensemble oversampling method for imbalanced keyword extraction learning in policy text based on three-way decisions and SMOTE. Expert Systems with Applications, 188, 116051. doi:10.1016/j.eswa.2021.116051.

[21] Patil, A. P., Jere, S., Ram, R., & Srinarasi, S. (2022). T5W: A Paraphrasing Approach to Oversampling for Imbalanced Text Classification. 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 1–6. doi:10.1109/ CONECCT55679.2022.9865812.