NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001287

TMUNLP at the NTCIR-17 MedNLP-SC Task

Yong-Zhen Huang
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
Department of Nursing,
National Taiwan University Cancer Center, Taiwan
m946111005@tmu.edu.tw

Eugene Sy
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
m946111012@tmu.edu.tw

Yi-Xuan Lin
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
m946111013@tmu.edu.tw

Yu-Lun Hsieh
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
morpheus.h@gmail.com

Yung-Chun Chang*
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
changyc@tmu.edu.tw

ABSTRACT

The TMUNLP team participated in the adverse drug event (ADE) detection subtask, focusing on social media texts in English for NTCIR-17's MedNLP-SC. This paper outlines our approach to addressing the challenge. Within the ADE subtask, we primarily implemented two methods to tackle the long tail distribution issue: distribution balanced loss and data augmentation. Finally, we employed ensemble learning to enhance the performance of our model.

KEYWORDS

Social Media, Adverse Drug Event, BERT, Multi-Label, Ensemble Learning, Data Augmentation, Distribution-Balanced Loss

TEAM NAME

TMUNLP

SUBTASKS

SM-ADE-EN

1 INTRODUCTION

NTCIR-17 MedNLP-SC [20] is a shared task workshop focused on medical natural language processing, utilizing radiology reports and adverse drug event detection in social media. The objective of this task is to foster the development of practical systems that support various medical services. This year, within the MedNLP-SC track, we primarily participated in the social media (SM) Subtask: Adverse Drug Event Detection (ADE) (Languages: English).

For the detection of adverse drug events, the objective of this task is to extract ADE information from social media, classifying the information into those that contain ADE (22 symptoms) and those that do not. Furthermore, all information that includes ADE is subject to multi-label tagging for symptoms. We view this task as a multi-label classification challenge, where the goal is to identify both the drug names and the associated symptoms in tweets that reference an ADE. Through data mining, we observed that the distribution of symptoms exhibits a long-tail distribution.

Regarding the long-tail distribution issue in multi-label classification, we explored two methods to address the problem: the utilization of a distribution balance loss and data augmentation specifically for the minority classes.

2 RELATED WORK

2.1 Multi-Label Text Classification

Multi-label classification (MLTC) is a task applied to various data mining applications, such as tagging videos, images, music, and text. Unlike traditional binary or multi-class single-label classification, multi-label classification requires assigning a document to multiple categories based on its content. In MLTC, it integrates both multi-class and binary classification, as it does not impose any restrictions on the number of output labels [6, 21]. The issue of class imbalance is widespread across different classification tasks. In multi-label classification, when there is a significant imbalance between high and low instances for each label, the challenge becomes even more pronounced due to the vast number of labels and their sparse distribution [21].

Early research in machine learning MLTC algorithms mainly falls into two categories: problem transformation and algorithm adaptation. The former transforms the multi-label classification problem into a series of single-label classification problems, while the latter modifies existing single-label algorithms to cater to multi-label data. Since machine learning relies heavily on feature engineering and can be easily influenced by noise, the predictive performance still has room for improvement [11]. In recent years, with the rapid development of deep learning, MLTC algorithms based on deep neural networks have gained widespread attention. The emergence of technologies such as Convolutional Neural Networks (CNN), Recurrent Neural Networks

(RNN), and Transformers has significantly improved the performance and generalization capability of multi-label classification [2, 8, 14, 22].

2.2 Classification with Imbalanced Data

In classification tasks involving imbalanced data having characteristics of a long-tail distribution, the majority classes tend to dominate the model training process. In other words, when there is an imbalance of data between majority and minority classes,

Table 1. Distribution of symptoms in training set

ID	CUI	Symptoms	ADE	non-ADE
01	C0027497	nausea	806	7158
02	C0011991	diarrhea	547	7417
03	C0015672	fatigue	268	7696
04	C0042963	vomiting	193	7771
05	C0003123	loss of appetite	249	7715
06	C0000737	abdominal pain	354	7610
07	C0018681	headache	267	7697
08	C0015967	fever	153	7811
09	C0206062	interstitial lung disease	16	7948
10	C0023895	liver damage	28	7936
11	C0012833	dizziness	124	7840
12	C0030193	pain	237	7727
13	C0002170	alopecia	71	7893
14	C0004096	analgesic asthma syndrome	95	7869
15	C0022658	renal impairment	34	7930
16	C0020517	hypersensitivity	184	7780
17	C0917801	insomnia	99	7865
18	C0009806	constipation	71	7893
19	C0005956	bone marrow dysfunction	8	7956
20	C0010692	hemorrhagic cystitis	11	7953
21	C0015230	rash	116	7848
22	C0149745	stomatitis	57	7907

the model is inclined to favor the majority class. To address the class imbalance issue in multi-label classification, common techniques include Binary Relevance, Re-sampling, and Re-weighting.

Binary Relevance [24] is a straightforward approach that breaks down the original multi-label classification problem into multiple independent binary classification tasks (one per class label). This involves adopting popular binary imbalance learning techniques to directly address the skew between majority and minority classes.

Re-sampling [7] is a traditional solution for long-tail learning. It encompasses oversampling techniques for tail labels in training data, such as random sampling [3], nearest neighbor informed oversampling [19], and synthetic instance generation [4, 5, 16], aimed at augmenting the training data for minority classes. Another approach involves under sampling the head labels, but this method can lead to significant data loss, especially when there is limited training data.

Re-weighting technique involves adjusting the loss function of the training model [9, 13, 15, 23], allocating a higher loss to samples of tail labels in order to enhance model performance.

3 METHODS

In the following section, we present methods used in ADE.

3.1 Pretrained Language Models

In this task, we experimented with several different pre-trained language models. Ultimately, we opted for BERT-base-uncased

Table 2. Subtask 1-SM-ADE-EN training set 10-fold cross-validation binary scores on different models.

Model	Class	P	R	F	Macro F1-score
BERT-base-uncased	non-ADE	0.95	0.91	0.93	0.89
	ADE	0.82	0.89	0.85	
BERT-base-uncased + DBL	non-ADE	0.95	0.91	0.93	0.89
	ADE	0.82	0.89	0.85	
BERT-base-uncased + DA	non-ADE	0.94	0.92	0.93	0.89
	ADE	0.83	0.86	0.85	
ClinicalDistilBERT + DBL	non-ADE	0.95	0.92	0.93	0.90
	ADE	0.83	0.89	0.86	

Abbreviations: P, Precision; R, Recall; F, F1-score.

[10] and ClinicalDistilBERT [18]. BERT-base-uncased is a transformers model pre-trained on English data from Wikipedia and Book Corpus in a self-supervised manner. ClinicalDistilBERT was developed by training the BioDistilBERT-cased model in a continual learning fashion for 3 epochs using a total batch size of 192 on the MIMIC-III notes dataset.

3.2 Distribution Balance Loss

Traditionally in Natural Language Processing (NLP), Binary Cross Entropy (BCE) loss is commonly adopted for multi-label text categorization [1]. Building on this foundation, variations like the Focal loss emerged. Specifically, Focal loss, deeply anchored in BCE, prioritizes challenging-to-classify samples [15]. On a related note, the Class-balanced Focal Loss enhances the Focal loss by incorporating the notion of class balance, ensuring the model is more sensitive to less represented classes [9]. Moreover, the Distribution Balanced Loss, through the amalgamation of rebalanced weighting and Negative Tolerant Regularization (NTR), primarily diminishes redundant information inherent in label co-occurrence, essential in multi-label contexts. Following this, it downweighs the 'easy-to-classify' negative instances [13,23]. Following is the distribution-balanced loss we used for this task.

$$L_{DB}(x^k, y^k) = \frac{1}{C}\sum_{i=0}^{c} \hat{r}_i^k [y_i^k \log\left(1 + e^{z_i^k - v_i}\right) + \frac{1}{\lambda}\left(1 - y_i^k\right) log\,(1 + e^{-\lambda(z_i^k - v_i)})] \quad (1)$$

3.3 Data Augmentation

Data augmentation has demonstrated its effectiveness in numerous artificial intelligence areas. We employed "nlpaug", a toolkit

Table 3. Subtask 1-SM-ADE-EN training set 10-fold cross-validation individual per class scores on different models.

CUI / Model	BERT			BERT + DBL			BERT + DA			ClinicalDistilBERT + DBL		
	P	R	F	P	R	F	P	R	F	P	R	F
C0027497	0.86	0.93	0.89	0.86	0.86	0.86	0.87	0.90	0.88	0.87	0.93	0.90
C0011991	0.80	0.86	0.83	0.81	0.82	0.81	0.81	0.81	0.81	0.81	0.84	0.82
C0015672	0.84	0.91	0.88	0.86	0.91	0.88	0.86	0.93	0.89	0.84	0.93	0.88
C0042963	0.80	0.90	0.85	0.84	0.91	0.87	0.82	0.91	0.86	0.80	0.90	0.85
C0003123	0.80	0.85	0.82	0.81	0.84	0.82	0.84	0.84	0.84	0.80	0.88	0.84
C0000737	0.84	0.90	0.87	0.86	0.88	0.87	0.83	0.90	0.86	0.86	0.90	0.88
C0018681	0.67	0.64	0.65	0.65	0.69	0.67	0.70	0.69	0.69	0.63	0.75	0.88
C0015967	0.60	0.38	0.46	0.65	0.69	0.67	0.60	0.56	0.58	0.59	0.81	0.68
C0206062	0.69	0.39	0.50	0.56	0.64	0.60	0.61	0.50	0.55	0.55	0.61	0.68
C0023895	0.81	0.85	0.83	0.83	0.87	0.85	0.86	0.90	0.88	0.82	0.89	0.58
C0012833	0.61	0.57	0.59	0.60	0.52	0.56	0.56	0.51	0.50	0.59	0.54	0.85
C0030193	0.92	0.92	0.92	0.84	0.83	0.84	0.87	0.96	0.91	0.88	0.94	0.56
C0002170	0.93	0.96	0.94	0.92	0.95	0.93	0.92	0.98	0.95	0.92	0.98	0.91
C0004096	0.69	0.59	0.63	0.67	0.76	0.71	0.75	0.71	0.73	0.67	0.82	0.95
C0022658	0.80	0.86	0.83	0.83	0.83	0.83	0.83	0.86	0.85	0.82	0.86	0.74
C0020517	0.71	0.68	0.69	0.66	0.73	0.69	0.64	0.75	0.69	0.69	0.73	0.84
C0917801	0.79	0.79	0.79	0.77	0.93	0.84	0.76	0.77	0.77	0.77	0.82	0.71
C0009806	0.00	0.00	0.00	1.00	0.25	0.40	0.75	0.38	0.50	0.67	0.50	0.79
C0005956	0.77	0.86	0.81	0.73	0.82	0.77	0.78	0.82	0.80	0.76	0.81	0.57
C0010692	0.75	0.27	0.40	0.80	0.73	0.76	0.90	0.82	0.86	0.85	1.00	0.79
C0015230	0.78	0.87	0.82	0.77	0.91	0.83	0.76	0.84	0.80	0.79	0.91	0.92
C0149745	0.76	0.79	0.78	0.75	0.82	0.78	0.75	0.79	0.77	0.77	0.89	0.84
Micro F1-score	0.80	0.84	0.82	0.80	0.83	0.81	0.81	0.83	0.82	0.80	0.86	0.83
Macro F1-score	0.74	0.72	0.72	0.77	0.78	0.77	0.78	0.78	0.77	0.76	0.83	0.79

Abbreviations: P, Precision; R, Recall; F, F1-score.

designed for data augmentation [12], to select synonyms based on WordNet [17, 25]. For tweets in the minority categories, we utilized the Synonym Augmenter to replace words or phrases within the tweets with their corresponding synonyms.

3.4 Model Ensembles

During the model parameter tuning phase, given that we only had a training set, we employed k-fold cross-validation to thoroughly assess the model's performance and fine-tune its parameters. We selected several promising parameter combinations and trained the models using the entire training set. From these, we picked a few models with the highest performance for ensemble modeling. Since we employed two distinct methods for model training, we decided to use the hard voting approach for the model, as it was the simplest to implement.

Table 4. Subtask 1-SM-ADE-EN training set 10-fold cross-validation full per label scores on different models.

Model	Class	P	R	F1	Macro F1
BERT-base-uncased	0	1.00	1.00	1.00	0.91
	1	0.80	0.84	0.82	
BERT-base-uncased + DBL	0	1.00	1.00	1.00	0.90
	1	0.80	0.83	0.81	
BERT-base-uncased + DA	0	1.00	1.00	1.00	0.91
	1	0.81	0.83	0.82	
ClinicalDistilBERT + DBL	0	1.00	1.00	1.00	0.91
	1	0.80	0.86	0.83	

Table 5. Subtask 1-SM-ADE-EN training set 10-fold cross-validation exact accuracy on different models.

Model	Exact accuracy
BERT-base-uncased	0.8542
BERT-base-uncased + DBL	0.8439
BERT-base-uncased + DA	0.8547
ClinicalDistilBERT + DBL	0.8566

4 EXPERIMENTS

Our experiments utilized 10-fold cross-validation and examined various transformer models, focusing particularly on their performance in predicting 22 different symptoms, as outlined in Table 1. The results, tabulated from Tables 2 through 6, provide a comprehensive understanding of model effectiveness across multiple metrics. Table 2 shows that all BERT-based models achieved a Macro F1-score of 0.89, suggesting that the application of Distribution Balance Loss (DBL) and Data Augmentation (DA) techniques are effective. ClinicalDistilBERT with DBL outperformed slightly with a Macro F1-score of 0.9.

Table 3 dives deeper, highlighting the shortcomings of the baseline BERT model in predicting underrepresented symptoms like *headache*, *fever*, *interstitial lung disease*, *dizziness*, *constipation*, and *hemorrhagic cystitis*, which had poor F1-scores ranging from 0.00 to 0.65. The introduction of DBL and DA significantly improved these scores. For instance, *headache* saw an increase in F1-score from 0.65 to 0.67 and 0.69 with DBL and DA, respectively. Similarly, *fever* improved from an F1-score of 0.46 to 0.67

Table 6. Individual Per Class Scores - Micro and Macro F1-scores on the Subtask 1-SM-ADE-EN training set using the ensemble model.

Model	Micro F1	Macro F1
Hard Voting 5 model	0.8356	0.8026
Hard Voting 6 model	0.8335	0.7997
Hard Voting 10 model	0.8324	0.8042

with DBL and 0.58 with DA. Remarkably, even extremely rare symptoms like *interstitial lung disease*, which had only 16 ADE cases in the training set, improved its F1-score to 0.60 and 0.55 with DBL and DA, respectively. These improvements were not isolated but contributed to a general boost in the model's Macro F1-score, elevating it from 0.72 in the baseline to 0.77 with DBL and DA, and further to 0.79 with ClinicalDistilBERT and DBL.

Based on Full Per Label Scores, Table 4 shows consistent high performance across all models. ClinicalDistilBERT with DBL stood out slightly, achieving an F1-score of 0.83 for Class 1 and matching the highest F1-score of 0.91. In Table 5, Exact Accuracy is evaluated, and while DBL decreased BERT's performance (0.8439), DA increased it (0.8547). ClinicalDistilBERT with DBL showed the highest Exact Accuracy of 0.8566. In addition, Table 6 illustrates hard-voting ensemble models, showing that adding more models significantly improves Micro F1-score performance, while improving Macro F1-score performance slightly, indicating a more balanced performance. The results collectively affirm the efficacy of employing DBL and DA techniques, especially for improving the prediction of underrepresented symptoms, thereby enhancing the model's overall effectiveness in classification tasks.

5 SUBMISSIONS for SM-ADE-EN

We present the results of our submissions for the SM-ADE-EN test set, which is a multilingual dataset for adverse drug event (ADE) detection. We compare our submissions with three baseline models: BERT, XLM-R, and XLM-RALL. BERT and XLM-R are pre-trained language models that use English as the input language, while XLM-RALL is a multilingual model that uses four languages (Japanese, English, German, and French) as the input language [20]. Our submissions are based on ensemble techniques that combine different variants of BERT and ClinicalDistilBERT, which are fine-tuned with distribution balance loss and data augmentation.

Submission-1: This included the baseline BERT, four BERTs using distribution balance loss (with different parameters), and one BERT with data augmentation.

Submission-2: This incorporated three BERTs using distribution balance loss (with different parameters), one BERT with data augmentation, and one ClinicalDistilBERT using distribution balance loss.

Submission-3: This consisted of eight BERTs using distribution balance loss (with different parameters), one BERT with data augmentation, and one ClinicalDistilBERT using distribution balance loss.

Table 7. Results of Submissions for SM-ADE-EN

Systems	Binary Score			
	Class	Precision	Recall	F1
BERT	ADE	0.75	0.67	0.71
	non-ADE	0.87	0.91	0.89
XLM-R	ADE	0.77	0.54	0.63
	non-ADE	0.83	0.94	0.88
XLM- RALL	ADE	0.75	0.82	0.78
	non-ADE	0.92	0.89	0.91
Submission-1	ADE	0.73	0.86	0.79
	non-ADE	0.94	0.87	0.90
Submission-2	ADE	0.71	0.86	0.78
	non-ADE	0.94	0.86	0.90
Submission-3	ADE	0.72	0.86	0.79
	non-ADE	0.94	0.87	0.90
Systems	(Individual) Per Class Scores			
	Micro F1		Macro F1	
BERT	0.66		0.41	
XLM-R	0.57		0.26	
XLM- RALL	0.76		0.61	
Submission-1	0.76		0.71	
Submission-2	0.76		0.70	
Submission-3	0.76		0.69	
Systems	(Full) Per Label Scores			
	Class	Precision	Recall	F1
BERT	0	0.99	1.00	0.99
	1	0.74	0.60	0.66
XLM-R	0	0.99	1.00	0.99
	1	0.73	0.46	0.57
XLM- RALL	0	1.00	0.99	0.99
	1	0.73	0.78	0.76
Submission-1	0	1.00	0.99	0.99
	1	0.71	0.83	0.76
Submission-2	0	1.00	0.99	0.99
	1	0.70	0.84	0.76
Submission-3	0	1.00	0.99	0.99
	1	0.71	0.83	0.76
Systems	Exact Match Accuracy			
BERT	0.79			
XLM-R	0.76			
XLM- RALL	0.83			
Submission-1	0.8259			
Submission-2	0.8204			
Submission-3	0.8219			

Table 7 details the performance of our models on the SM-ADE-EN test set. The results are shown in the Table 7. We can observe that our submissions outperform the baseline models on all metrics, except for exact match accuracy, where XLM-RALL has a slightly higher score than our submissions. However, XLM-RALL uses four languages for training, while our submissions use only English. Therefore, our submissions are more efficient and practical for monolingual applications. Moreover, we can also see that our submissions have higher per ADE label F1-scores than the baseline models, especially for the ADE class. This indicates that our submissions are more effective in detecting ADEs from

sentences, which is the main goal of this task. Moreover, our submissions have higher binary F1-scores than the baseline models, which shows that our submissions can distinguish ADEs from non-ADEs with high accuracy.

Notably, our models exhibited exceptional Micro and Macro F1-scores on per Class scores, surpassing the BERT and XLM-R baselines considerably but tying with XLM-R$_{ALL}$ in Micro F1 average and exceeding in Macro F1 average. Submission-1 led with a macro average of 0.71, followed by Submission-2 at 0.70 and Submission-3 at 0.69. We attribute the success of our submissions to the ensemble technique, the distribution balance loss, and the data augmentation. The ensemble technique combines the predictions of multiple models with different architectures and parameters, which increases the diversity and robustness of the final output. The distribution balance loss penalizes the models for predicting incorrect labels with high confidence, which reduces the overfitting and improves the generalization. The data augmentation generates synthetic sentences with ADEs by replacing drugs and adverse effects with synonyms or related terms, which enlarges the training data and enhances the representation.

In summary, we present our submissions for the SM-ADE-EN test set, which achieve state-of-the-art results on various metrics. We demonstrate the effectiveness of our proposed method, which leverages ensemble techniques, distribution balance loss, and data augmentation.

6 CONCLUSIONS

This research addresses class imbalance in multi-label classification for adverse drug event prediction. Utilizing Distribution Balance Loss (DBL) and Data Augmentation (DA), we showed a significant rise in Macro F1-scores, particularly for minority labels. Among our submissions, Submission-1, which includes baseline BERT and several BERT models fine-tuned with varying DBL and DA parameters, led the pack with a Macro F1-score of 0.71. Interestingly, the performance appeared to decline slightly in Submission-2 and Submission-3, despite the incorporation of ClinicalDistilBERT and additional BERT models with varied DBL settings. This suggests that the absence of baseline BERT in these submissions might have contributed to their lower performance, indicating that even a simple model can add value to an ensemble.

While our techniques notably improved minority labels, they had a more modest impact on other metrics like Exact Accuracy, highlighting an area for future research. The observed trade-off between boosting minority labels at the expense of majority ones also deserves closer examination. Given the nuanced performance dynamics observed between the different submissions, further studies should aim to dissect the contribution of each model within the ensemble. This could lead to the development of a more balanced and performance-optimized predictive tool for multi-label classification tasks.

Acknowledgements

This study was supported by the National Science and Technology Council under grant 112-2410-H-038-007- and 112-2622-E-038-001-, as well as the National Health Research Institutes NHRI-12A1-PHCO-1823244. Yung-Chun Chang is the corresponding author.

REFERENCES

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (August 2013), 1798–1828. DOI:https://doi.org/10.1109/TPAMI.2013.50

[2] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. 3163–3171.

[3] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. 2015. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing* 163, (2015), 3–16.

[4] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. 2015. MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems* 89, (2015), 385–397.

[5] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera. 2019. Dealing with difficult minority labels in imbalanced mutilabel data sets. *Neurocomputing* 326, (2019), 39–53.

[6] Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. 2013. A first approach to deal with imbalance in multi-label datasets. In *Hybrid Artificial Intelligent Systems: 8th International Conference, HAIS 2013, Salamanca, Spain, September 11-13, 2013. Proceedings 8*, Springer, 150–160.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16, (2002), 321–357.

[8] Guibin Chen, Deheng Ye, Zhenchang Xing, Jieshan Chen, and Erik Cambria. 2017. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. IEEE, 2377–2383.

[9] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9260–9269. DOI:https://doi.org/10.1109/CVPR.2019.00949

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Lihua Duan, Qi You, Xinke Wu, and Jun Sun. 2022. Multilabel Text Classification Algorithm Based on Fusion of Two-Stream Transformer. *Electronics* 11, 14 (2022), 2138.

[12] Edward Ma. 2019. NLP augmentation. Retrieved August 10, 2023 from https://github.com/makcedward/nlpaug

[13] Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing Methods for Multi-label Text Classification with Long-Tailed Class Distribution. *arXiv.org*. Retrieved August 10, 2023 from https://arxiv.org/abs/2109.04712v2

[14] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. DOI:https://doi.org/10.1109/ICCV.2017.324

[16] Xu-Ying Liu, Sheng-Tao Wang, and Min-Ling Zhang. 2019. Transfer synthetic over-sampling for class-imbalance learning with limited minority class data. *Frontiers of Computer Science* 13, (2019), 996–1009.

[17] George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*. Retrieved August 10, 2023 from https://aclanthology.org/H94-1111

[18] Omid Rohanian, Mohammadmahdi Nouriborji, Hannah Jauncey, Samaneh Kouchaki, ISARIC Clinical Characterisation Group, Lei Clifton, Laura Merson, and David A. Clifton. 2023. Lightweight Transformers for Clinical Natural Language Processing. DOI:https://doi.org/10.48550/arXiv.2302.04725

[19] Payel Sadhukhan and Sarbani Palit. 2019. Reverse-nearest neighborhood based oversampling for imbalanced, multi-label datasets. *Pattern Recognition Letters* 125, (2019), 813–820.

[20] Shoko Wakamiya, Lis Kanashiro Pereira, Lisa Raithel, Hui-Syuan Yeh, Peitao Han, Seiji Shimizu, Tomohiro Nishiyama, Gabriel Herman Bernardim Andrade, Noriki Nishida, Hiroki Teranishi, Narumi Tokunaga, Philippe Thomas, Roland Roller, Pierre Zweigenbaum, Yuji Matsumoto, Akiko Aizawa, Sebastian Möller, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Patrick Paroubek, Shuntaro Yada, and Eiji Aramaki. 2023. NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo, Japan. *doi: 10.20736/0002001327*

[21] Adil Yaseen Taha, Sabrina Tiun, Abdul Hadi Abd Rahman, and Ali Sabah. 2021. Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *Journal of Information and Communication Technology* 20, 3 (2021), 423–456.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30, (2017).

[23] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. Springer, 162–178.

[24] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science* 12, (2018), 191–202.

[25] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level Convolutional Networks for Text Classification. DOI:https://doi.org/10.48550/arXiv.1509.01626