UCCNLP@SMM4H'22:Label distribution aware long-tailed learning with post-hoc posterior calibration applied to text classification

Paul Trust
University College Cork
Cork, Ireland

Provia Kadusabe
Worldquant University
Louisiana, USA

Rosane Minghim
University College Cork
Cork, Ireland

Ahmed Zahran
University College Cork
Cork, Ireland

Kizito Omala
Makerere University
Kampala, Uganda

Abstract

This paper describes our submissions for the Social Media Mining for Health (SMM4H) workshop 2022 shared tasks. We have participated in 2 tasks: (1) classification of adverse drug events (ADE) mentions in English tweets (Task-1a) and (2) classification of self-reported intimate partner violence (IPV) on Twitter (Task 7). We propose an approach that uses RoBERTa (Robustly Optimized BERT Pretraining Approach) fine-tuned with a label distribution-aware margin loss function and post-hoc posterior calibration for robust inference against class imbalance. We achieved a 4% and 1% increase in performance on IPV and ADE respectively, when compared with the traditional fine-tuning strategy with unweighted cross-entropy loss.

1 Introduction

Social media platforms are becoming part of our every day life with an estimation of about 4.2 billion people using some sort of social media (Hsieh-Yee, 2021). The 7^{th} Social Media Mining for Health Applications Workshop (SMM4H) 2022 organized tasks involving automatic methods for collection, extraction, representation, analysis and validation of social media data for health informatics. The tasks we have participated in utilized data from Twitter, which is one of the largest social media platforms with approximately 192 million daily active users (Conger, 2021).

Our team UCCNLP participated in 2 tasks: (1) classification of adverse drugs events (ADE) mentions in english tweets (Task-1a) and (2) classification of self-reported intimate partner violence (IPV) on twitter (Task 7). Adverse drug events (ADEs) are negative effects related to the use of drugs. ADEs mentions on social media are a useful indicator of the efficacy of medications and also can potentially reveal previously unknown side effects of drugs (Ramesh et al., 2021).

Intimate partner violence (IPV) on the other hand refers to the abuse or aggression that occurs in a romantic relationship. IPV is a serious health problem and can have lifelong impact on health and well-being of individuals. Victims of IPV sometimes share their stories on twitter making it an important tool in early identification for timely intervention and support (Ramesh et al., 2021).

Exploration of the provided training datasets of both tasks revealed that the datasets were heavily imbalanced with long tailed distributions for example the IPV data contains only 11% of the provided tweets identified as self-reported IPV.

The skewed nature of these datasets complicates efficient training even for state of the art models like BERT (Bidirectional Embeddings from Transformers) (Devlin et al., 2019) and RoBERTa (A Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019). The difficulty in model training arises from the fact that naive empirical risk minimization for imbalanced datasets tends to be biased towards learning the distribution on these head classes which deteriorates its performance on tail classes.

In this work, we proposed to solve the class imbalance in classification on both tasks using RoBERTa with label distribution-aware loss function (LDAM) combined and posterior calibration to alleviate the problem of class imbalance during training.

2 Related work

2.1 Twitter data for Intimate partner violence and Adverse drug events Classification

The task of ADE detection from English tweets featured as a shared task at SMM4H in 2021 and the top systems on the task used RoBERTa combined with under sampling and oversampling (Ramesh et al., 2021), a combination of BERT and drug embeddings obtained by chemical struc-

Proceedings of the 29th International Conference on Computational Linguistic, pages 90–94
October 12–17, 2022.

ture BERT-based encoder (Sakhovskiy et al., 2021). Other methods used BERT with naive class weights (Yaseen and Langer, 2021) and BERT combined with data augmentation (Ji et al., 2021).

The task of classification of self-reported partner violence on twitter had not appeared before as a shared task, but previous work on IPV classification fine-tuned RoBERTa (Al-Garadi et al., 2021) using a standard training procedure without consideration of class imbalance.

2.2 Long-tailed learning and class imbalance

Most of the existing algorithms for handling class imbalance can be divided in the following categories: re-sampling, re-weighting and confidence calibration and regularization.

2.2.1 Re-sampling

There are two groups of re-sampling strategies: over-sampling the the minority classes (Buda et al., 2018; Byrd and Lipton, 2019; Shen et al., 2016) and under-sampling the frequent classes (He and Garcia, 2009; Haixiang et al., 2017). The limitation of under-sampling is that it discards a large portion of the data making it infeasible when data imbalance is extreme. Over-sampling may also lead to over-fitting the minority classes.

2.2.2 Re-weighting

Cost-sensitive re-weighting assigns weights for different classes or for different samples. The traditional re-weighting scheme assigns weights to classes proportional to the inverse of their frequency (class balanced loss (CB)) (Huang et al., 2016, 2019). However, under extreme data imbalance and large-scale scenarios, re-weighting makes deep learning models difficult to optimize during training (Zhong et al., 2021). In this work, we explore efficient loss functions adapted for class imbalance mostly used in computer vision for our text classification tasks. Focal loss (FS), which a weighted version of cross-entropy loss with sample-specific weight $(1-p_i)^{\gamma}$, where p_i is the model output probability for class i (Lin et al., 2017). Label-aware distribution loss (LDAM) derives a generalization error bound for imbalanced training and proposes a margin-aware weighted cross-entropy loss (Cao et al., 2019) and Maximum Margin LDAM loss (MM-LDAM) minimizes a margin-based generalization bound by shifting decision bound (Kang et al., 2021).

2.2.3 Thresholding and calibration

These are methods that are applied at test time to calibrate confidence scores generated by the machine learning model to be robust to class imbalance. Representative methods in this category are post-hoc logit adjustment (Menon et al., 2020) and posterior calibration (PC) (Tian et al., 2020).

3 Methodology

3.1 Training phase

We fine-tuned transformer language models (RoBERTa and BERTweet) with a label distribution-aware margin loss function (LDAM). More formally, consider our classifier $f = g \odot (RoBERTa \vee BERTweet)$, which consists of two parts: a pre-trained language model that outputs hidden representations of input samples and g, which is the classification head on our imbalanced dataset, that outputs k-dimensional vector, where each dimension corresponds to the prediction confidence of a specific class.

Instead of using a standard cross entropy loss, we fine-tuned our model f with (LDAM) (Cao et al., 2019):

$$\mathcal{L}_{LDAM}((x,y);f) = -\log \frac{e^{zy-\Delta y}}{e^{z_y-\Delta_y} + \sum_{j\neq y} e^{z_j}} \quad (1)$$

where $\Delta_j = \frac{C}{n_j^{1/4}}$ for $j \in \{1,..,k\}$, Δ_y is chosen to be a label independent constant C, n_j is the number of instance per class j.

We adopted a deferred re-weighting procedure as in (Cao et al., 2019), since naive re-weighting and re-sampling were found to be inferior to the vanilla empirical risk minimization (ERM).

3.2 Prediction phase

In the prediction phase, our aim is to use the learned parameters in the training phase $\{w, \theta\}$ to make predictive inference on the test data with unknown labels. This is done by taking the most likely label on the test data defined by $h^*(x) = \text{argmax}_{y\in K}\, P_{test}(y|x)$ under the model's posterior distribution. By using parameters learned on the training data to make inference on the test data, we are assuming that test and train data are drawn from the same distribution.

This was not the case, since our training data was imbalanced and we had no information about the distribution on the test data. We therefore calibrated the posterior distribution on the test data

to be robust to class imbalance through posterior re-balancing (PC) proposed in (Tian et al., 2020).

More formally, the optimal Bayes classifier $h_{test}(x)$ on the test set is an approximation based on the training posterior distribution defined as:

$$h_{test}(x) = \underset{y \in K}{\operatorname{argmax}} \frac{p_{train}(y|x)p_{test}(y)}{p_{train}(y)} \quad (2)$$

We can view $h_{test}(x)$ as a posterior distribution from the training dataset weighted by $p_{test}(y)/p_{train}(y)$ which we denoted as a re-balanced posterior ($p_{rbal}(y|x)$). This weighting $p_{test}(y)/p_{train}(y)$ introduced at testing phase can be too big or too small depending on the label distribution on test and train datasets.

To compensate for imperfect learning due to diference in train and test distribution, we solve an approximation through Kullback-Leibler (KL) divergence by treating the re-balanced posterior and original posterior as an approximation to the true posterior proposed by (Tian et al., 2020) as follows:

$$\begin{aligned} p^*(y|x) \quad &= \operatorname{argmin}_p (1-\lambda) KL(p, p_{train}(y|x) \\ &\quad + \lambda KL(p, p_{rbal}(y|x)) \end{aligned}$$

A closed form solution to the optimization exists and is defined as follows:

$$p^*(y|x) = \frac{1}{Z(x)} (p_{train}(y|x))^{(1-\lambda)} (p_{rbal}(y|x))^{\lambda} \quad (3)$$

where $Z(x)$ is the normalization constant

We can thus balance the posterior distribution on the test data against class imbalance through a hyper parameter λ without any need to retrain the model.

4 Experiments and Results

The datasets for both tasks: (IPV and ADE) were released by SMM4H 2022 shared tasks. IPV self report classification (Task 7) consisted of 4523 training tweets, 534 validation tweets and 1291 test tweets (Al-Garadi et al., 2022). ADE detection task contained 17173 train tweets , 908 validation tweets and 10968 test tweets.

We conducted experiments with pre-trained transformer language models; RoBERTa (Liu et al., 2019) and BERTweet (Nguyen et al., 2020) with different loss functions and also with post-hoc posterior calibration (PC). Experiments were done for 10 epochs, max length of 128, batch size of 10 and the learning rate was set at 0.0005. The final submission were evaluated using a median $f1$-score over 5 runs. The system was written in PyTorch using hugging-face transformer library (Wolf et al., 2019).

Table 1 demonstrates that cost sensitive learning under class imbalance is superior to naive training with cross entropy loss with both RoBERTa and BERTweet models. Another observation is that doing posterior calibration on the test performance also improves performance. All these observations and conclusions are consistent with other works in computer vision (Tian et al., 2020; Cao et al., 2019; Lin et al., 2017) Maximum Margin LDAM (MM-LDAM-PC) with posterior calibration achieved the best performance in both datasets and with both models. RoBERTa achieves the best perfomance on the IPV dataset (69.37% versus 69.15%) while BERTweet achieves the best performance on the ADE dataset (47.81% versus 47.10%)

Model	IPV	ADE
BERTweet-CEL	62.66	46.90
BERTweet-FL	60.48	46.20
BERTweet-LDAM	67.06	47.62
BERTweet-MM-LDAM	68.74	47.77
BERTweet-MM-LDAM-PC	**69.15**	**47.81**
RoBERTa-CEL	65.34	46.29
RoBERTa-FL	66.05	46.02
RoBERTa-LDAM	67.25	47.03
RoBERTa-MM-LDAM	68.51	47.09
RoBERTa-LDAM-PC (Ours)	**69.37**	**47.10**

Table 1: The table shows results of the median f-scores on the validation sets for 5 runs on Intimate Partner violence (IPV) dataset and adverse drug effects (ADE) dataset. CEL represents cross entropy loss, FL represents focal length, LDAM represents label distribution-aware loss function, MM represents maximum margin and PC represents posterior calibration

5 Conclusion

In this work, we developed two systems based on pre-trained transformer based language models fine-tuned with label distribution aware loss functions and posterior calibration on the test set. We experimented with various loss functions as well as different transformer models. The results on the validation set revealed that incorporating class prior probabilities at both training and testing helps to boost performance under class imbalance. Further more our system achieved 62.25% $f1$-score

on IPV (Task 7) and 8% for ADE task (Task 1a) on our codalab submission.

6 Acknowledgements

We thank Science Foundation Ireland (SFI) Center for Research Training in Advanced Networks and Future communications at University College Cork for funding this research and Irish Centre for High-End Computing (ICHEC) for providing access to computing power for running some of our experiments.

References

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2021. Natural language model for automatic identification of intimate partner violence reports from twitter. *medRxiv*.

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, page 100217.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259.

Jonathon Byrd and Zachary Lipton. 2019. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Kate Conger. 2021. Twitter shakes off the cobwebs with new product plans. *The New York Times*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Ingrid Hsieh-Yee. 2021. Can we trust social media? *Internet Reference Services Quarterly*, 25(1-2):9–23.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. 2019. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2781–2794.

Zongcheng Ji, Tian Xia, and Mei Han. 2021. PAII-NLP at SMM4H 2021: Joint extraction and normalization of adverse drug effect mentions in tweets. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 126–127, Mexico City, Mexico. Association for Computational Linguistics.

Haeyong Kang, Thang Vu, and Chang D Yoo. 2021. Learning imbalanced datasets with maximum margin loss. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1269–1273. IEEE.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Sidharth Ramesh, Abhiraj Tiwari, Parthivi Choubey, Saisha Kashyap, Sahil Khose, Kumud Lakara, Nishesh Singh, and Ujjwal Verma. 2021. BERT based transformers lead the way in extraction of health information from social media. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 33–38, Mexico City, Mexico. Association for Computational Linguistics.

Andrey Sakhovskiy, Zulfat Miftahutdinov, and Elena Tutubalina. 2021. KFU NLP team at SMM4H 2021 tasks: Cross-lingual and cross-modal BERT-based models for adverse drug effects. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 39–43, Mexico City, Mexico. Association for Computational Linguistics.

Li Shen, Zhouchen Lin, and Qingming Huang. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer.

Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, and Zsolt Kira. 2020. Posterior recalibration for imbalanced datasets. *Advances in Neural Information Processing Systems*, 33:8101–8113.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Usama Yaseen and Stefan Langer. 2021. Neural text classification and stacked heterogeneous embeddings for named entity recognition in SMM4H 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 83–87, Mexico City, Mexico. Association for Computational Linguistics.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498.