Bayes at FigLang 2022 Euphemism Detection shared task: Cost-Sensitive Bayesian Fine-tuning and Venn-Abers Predictors for Robust Training under Class Skewed Distributions

Paul Trust
University College Cork
Cork, Ireland

Kadusabe Provia
Worldquant University
Louisiana, USA

Kizito Omala
Makerere University
Kampala, Uganda

Abstract

Transformers have achieved a state of the art performance across most natural language processing tasks. However, the performance of these models often degrades when being trained on data that exhibits skewed class distributions (class imbalance) common social media data. This is because training tends to be biased towards head classes that have majority of the data points . Most of the classical methods that have been proposed to handle this problem like re-sampling and re-weighting often suffer from unstable performance, poor applicability and poor calibration. In this paper, we propose to use Bayesian methods and Venn-Abers predictors for well calibrated and robust training against class imbalance. Our proposed approach improves $f1-$score over the baseline RoBERTa (A Robustly Optimized Bidirectional Embedding from Transformers Pretraining Approach) model by about 6 points (79.0% against 72.6%) when training with class imbalanced data.

1 Introduction

The phenomena of skewed class distribution also known as class imbalance is ambiguous and common in most real-world datasets and natural language processing (NLP) tasks (Tayyar Madabushi et al., 2019). Instead of preserving an ideal uniform distribution over each category of labels, most large-scale datasets exhibit skewed class distributions with a long tail having some target distributions with significantly more observations than others (Yang and Xu, 2020).

Although transformer-based models (Vaswani et al., 2017) have achieved a state of the art performance across several tasks in NLP, their performance tends to degrade when trained on long-tailed data. The main challenge lies in the sparsity of tail classes leading to estimation of the decision boundaries severely biased towards head classes (classes with more observations) (Pan et al., 2021a).

Class imbalance problem can be tackled at either model training or model inference phases. Approaches to handle class imbalance at training phase can be classified into re-weighting or re-sampling and those at model inference phase are mostly calibration techniques (Menon et al., 2020; Tian et al., 2020) which adjusts a classifier's confidence scores without changing the internal weights or architectures (Pan et al., 2021b) of the underlying models.

Post-processing calibration techniques have been found to be efficient since they requires no further training of the model and are effective on multiple class imbalanced classification benchmarks in computer vision (Kang et al., 2020; Pan et al., 2021b). Inspired by the success of post-processing calibration techniques, we experiment with techniques that are theoretically known to produce well calibrated predictions; Bayesian inference for neural networks (Blundell et al., 2015; Wen et al., 2018; Gal and Ghahramani, 2016) and Venn-Abers predictors (Vovk and Petej, 2014, 2012).

We test these methods by participating in the shared task at the third Workshop on Figurative Language Processing 2022 at EMNLP 2022 (Conference on Empirical Methods in Natural Language Processing). The training dataset exhibited a long tail distribution with 70% of the training texts containing euphemism (Gavidia et al., 2022; Lee et al., 2022).

Euphemisms are mild or indirect expressions that are used in place of more unpleasant or offensive ones common in social media data. They are used to show politeness when discussing sensitive topics or as a way to make unpleasant things sound better for example saying "laid to rest" instead of "buried" or "armed conflict" instead of "war" (Lee et al., 2022). With the need to curb inappropriate material on social media, people use these euphemisms to bypass media censoring software and thus automatically identifying texts containing

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 94 - 99

these statements is a timely task. Several computational techniques have been proposed for the euphemism task (Gavidia et al., 2022; Lee et al., 2022; Zhu and Bhat, 2021). To the best of our knowledge, this is the first attempt to combine Bayesian transformers and Venn-Abers predictors for this task. The contributions of this work are:

- We show that fine-tuning transformers with Bayesian methods boosts performance over naive training in imbalanced class setting.
- We propose an an approach to combine Bayesian transformers and Venn Predictors for long tail distribution learning.
- We propose a euphemism detection method with considering of the class imbalance.

2 Background and Related Work

2.1 Euphemism Detection

Machine learning approaches have been proposed for euphemism detection (Kapron-King and Xu, 2021; Magu and Luo, 2018; Gavidia et al., 2022; Lee et al., 2022). Sentiment analysis methods have been utilized to recognize and classify euphemistic language in text (Felt and Riloff, 2020; Lee et al., 2022). Magu and Luo, 2018 used word embeddings and network analysis to identify euphemisms in the context of hate speech (Magu and Luo, 2018). Self supervised methods (Zhu and Bhat, 2021; Zhu et al., 2021) have also been employed. Our methodology is different from methods in literature in that we consider the long tailed distribution nature of the task and we also present apply novel techniques from Bayesian inference and Venn predictors which have not been used before in this task.

2.2 Learning under skewed class distributions

The dominant solutions to learning data with long-tailed distributions can be classified into re-sampling, re-weighting, confidence calibration and regularization. Re-sampling strategies flatten the data distribution, popular techniques are over-sampling (Buda et al., 2018; Byrd and Lipton, 2019; Shen et al., 2016) and under-sampling (He and Garcia, 2009; Haixiang et al., 2017). However, under-sampling may discard most of the data points and over-sampling results into over-fitting on the minority classes.

Cost sensitive learning (loss re-weighting) is another widely used method which works by assigning weights for different training samples. class-balanced loss assigns weights to classes proportional to the inverse of their frequency in the dataset (Huang et al., 2016, 2019). But optimizing deep learning models with this method under extreme class class imbalance may deteriorate performance (Zhong et al., 2021). Focal loss (FL) is a weighted version of cross-entropy loss with sample-specific weight. Label distribution-aware margin loss (LDAM) derives a generalization error bound for imbalanced training and proposes a margin-aware weighted cross-entropy loss (Cao et al., 2019) by minimizing margin-based generalization bound achieving significant performance boost over unweighted cross-entropy loss.

Post-processing methods of handling class imbalances re-calibrate the posterior distribution from the predicted confidence scores at test time. Examples of the methods are are logit adjustment (Menon et al., 2020) and posterior calibration (PC) (Tian et al., 2020).

2.3 Bayesian modeling with transformers

Deep learning models especially those based on the transformer architecture (Vaswani et al., 2017) have achieved a state-of-the-art performance across several tasks. BERT (Devlin et al., 2019) (Bidirectional Embedding from Transformers) and RoBERTa (Liu et al., 2019) (Robustly Optimized BERT Pretraining Approach) are among the most influential transformer variants in NLP. Despite their impressive performance, deep learning models tend to be produce over-confidence scores that are not calibrated which may deteriorate performance in imbalanced learning settings (Blundell et al., 2015).

Unlike the traditional neural networks trained with Maximum Likelihood Estimation (MLE) that fit a point estimate for the neural network's weights, Bayesian inference puts a prior distribution $p(w)$ over the weights and approximates the posterior distribution $p(w|D) \propto p(w)p(D|w)$. The predictive distribution of an unknown label $\tilde{y}$ of a test data item $\tilde{x}$ is given by $p(\tilde{y}|\tilde{x}) = E_{p(w|D)}[p(\tilde{y}|\tilde{x}, w)]$, we observe that taking an expectation over the posterior distribution of the weights is equivalent to using an ensemble of unaccountably infinite number of neural networks which would results into a boost in performance over a single neural network

Model	Precision	Recall	f1-score
BERT-base	0.712	0.714	0.713
RoBERTa-base (Baseline)	0.745	0.719	0.726
RoBERTa-Platt Scaling	0.702	0.710	0.706
RoBERTa-Venn-Abers	0.736	0.728	0.731
RoBERTa-bayesian	0.732	0.761	0.743
RoBERTa-LDAM	0.769	0.779	0.774
RoBERTa-bayesian-LDAM	0.769	0.819	0.787
RoBERTa-Bayesian-LDAM-Venn-Abers (Ours)	**0.794**	**0.786**	**0.790**

Table 1: Accuracy, precision and $f1-$score in percentages on the test data set for baseline model (RoBERTa-base) and our proposed approach (RoBERTa-Bayesian-LDAM-Venn-Abers), LDAM stands for label-distribution-aware margin loss

(Blundell et al., 2015).

However computing the posterior distribution over the weights often involve high dimensional integrals that are intractable and cannot be obtained in closed form. Popular approaches that have been proposed to produce approximates of these distribution are based on monte-carlo estimates and variational inference. Popular methods that utilise Bayesian principles for approximating the posterior distribution over neural networks are Bayes by Backprop (Blundell et al., 2015) and Flipout (Wen et al., 2018) and monte-carlo dropout (Gal and Ghahramani, 2016).

Flipout (Wen et al., 2018) is an efficient method for decorrelating the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each example. Bayes by Backprop (Blundell et al., 2015) learns a probability distribution on the weights of the neural networks by minimizing the expected lower bound on the marginal likelihood. Monte Carlo dropout (Gal and Ghahramani, 2016) casts dropout training during training of neural networks as approximate Bayesian inference in deep Gaussian processes.

2.4 Venn-Abers Prediction

Venn-Abers predictors (Vovk and Petej, 2012) are a special case of Venn predictors (Vovk and Petej, 2014) which are distribution-free probabilistic predictors that have a guarantee of being valid under a sole assumption of the training examples being exchangeable. They work by transforming the output of a scoring classifier which in our case is a machine learning model into a multi-probabilistic prediction that has calibration guarantees.

More formally, assume we are given training samples $D = \{(x, y)\}_{i=1}^{n}$ consisting of two components; a data point $x \in X$ and its label $y \in Y$. Given a test data point x_{n+1}, the Venn predictor outputs a multi probabilistic prediction in the form of a probability distribution over possible values of the label.

A venn taxonomy B is a measurable function B that assigns to each $n \in \{1, 2, ...\}$ and each sequence $(d_1, ...d_n) \in D^n$ an equivalence relation $\sim$ on $\{1, ..., n\}$. The relation has to be equivariant in the sense that for each n and each permutation ϕ of $\{1, ..., n\}$,

$$(i \sim j|d_1, ...d_n) \Rightarrow (\phi(i) \sim \phi(j)|d_{\phi(1)}, ..., d_{\phi(n)}) \tag{1}$$

where $(i \sim j|d_1, ...d_n)$ means that i is equivalent to j under the relation assigned by B to $(d_1, ...d_n)$. A venn predictor with a Venn taxonomy B outputs a pair (p_0, p_1) where

$$p_y = \frac{|\{i \in B(n+1|d_1, ..., d_n, (x_{n+1}, y))|y_i = 1\}|}{|B(n+1|d_1, ...d_n, (x_{n+1}, y))|} \tag{2}$$

where $B(j|d_1, .., d_n)$ the class of the equivalence of j is defined as follows:

$$B(j|d_1, .., d_n) = \{i \in \{1, ..., n\}|(i \sim j|d_1, ...d_n)\} \tag{3}$$

p_0 and p_1 express the predicted probabilities of the test object x_{n+1} belonging to a certain class.

3 Methodology

The dataset $D = \{(x, y)\}_{i=1}^{n}$ is divided into 3 splits; D_{train} for training the model, $D_{validation}$ for selecting the best models and calibration step, D_{test} for testing our approaches. We fine-tune RoBERTa (Liu et al., 2019) with standard cross entropy loss and with label-distribution-aware margin loss (LDAM) function (Cao et al., 2019). We first experiment with training our models in non-

Bayesian way using the standard maximum likelihood estimation and also in a Bayesian way by applying Bayesian layers in our neural network. The Bayesian layers used for our experimentation are Monte carlo dropout (Gal and Ghahramani, 2016).

To calibrate our predictions, we perform inference on the validation dataset $D_{validation}$ of size k with our trained model and obtained uncalibrated confidence scores denoted as $\{z_1, ..., z_k\}$ for each test data point x. Venn-Abers predictors proceeds by fitting an isotonic regression on the set $(z_1, y_1), ...(z_k, y_k), (z, 0)$ and the computing the score $s(x_i)$ for each calibration data points (x_i, y_i). Let g be an increasing function on the set $s(x_1), ...s(x_k)$ that maximizes the likelihood $\prod_{i=1}^{k} p_i$ where:

$$p_i = \begin{cases} g(s(x_i)) & \text{if } y_i = 1 \\ 1 - g(s(x_i)) & \text{if } y_i = 0 \end{cases} \quad (4)$$

Thus the multi-probabilistic prediction for x is the pair

$$(p_0, p_1) = (g_0(s_0(x)), g_1(s_1(x))) \quad (5)$$

The estimated label for a text data point x is the probability that minimizes the regret of the loss function calculated as in Equation 6.

$$p = \frac{p_1}{1 - p_0 + p_1} \quad (6)$$

4 Results and Discussion

4.1 Datasets

The dataset used for experiments is an Euphemism detection (ED) dataset (Gavidia et al., 2022; Lee et al., 2022) released by Third Workshop on Figurative Language Processing 2022 at EMNLP 2022 shared task on Euphesim Detection. This was a binary classification problem for identifying text expression that was euphemistic. The training data consisted of 1572 training points and test data consisted of 393 texts. Of the 1572 training texts, only 466 (30%) were did not contain euphemism.

4.2 Experimental Setup

We conduct experiments with pretrained transformer language models; RoBERTa (Liu et al., 2019), Bayesian methods and Venn-Abers predictors . Experiments are done for 50 epochs, max length of 512, batch size of 50 and the learning rate was set at 0.0005. The final submission were evaluated using $f1$-score. Transformers are implemented using hugging-face transformer library (Wolf et al., 2019), bayesian layers are implemented using Bayesian torch and baal (Krishnan and Tickoo, 2020; Atighehchian et al., 2022) and conformal predictors were implemented using reliabots (Shafer and Vovk, 2008).

4.3 Discussion

To assess the impact of Bayesian fine-tuning and Venn predictors, we perform experiments on the euphemisms detection dataset (Lee et al., 2022) described in section 4.1. Table 1 shows a combination of different models and their results on the test set. F1-score, recall and accuracy measures were used to evaluate the performance of different models as shown in Table 1. RoBERTa achieves a a slightly better performance compared to BERT (72.6% versus 71.3%). The observation is re-enforced by the impact of the architecture design of the pre-trained model on downstream tasks.

Experiments results on the test as shown in Figure 1 reveal that calibrating confidence scores of RoBERTa using Venn Abers predictors improves performance of the model by 1.2%. This is consistent with other results that report improved performance with post-hoc posterior calibration but naive calibration using platt scaling degrades performance of the model (Tian et al., 2020). Fine-tuning RoBERTa with a Bayesian layer boosts performance (about 2%) compared to the traditional fine-tuning, This is because Bayesian layers in a neural networks can be seen an ensemble of many networks at test time.

The biggest performance boost comes from training our models with a label distribution aware margin loss function (LDAM) and differed weighting, and this demonstrated the importance of cost sensitive learning when the data distribution is skewed. Finally our best system which we submitted for competition to the euphemism shared tasks was a combination of RoBERTa, Bayesian learning, cost sensitive learning and Venn Abers Predictors (*RoBERTa-bayesian-LDAM-Venn-Abers*) with an $f1-$score of 79% as shown in Table 1.

5 Conclusion

In this work, we have presented an approach for improving classification performance of transformer model when the data exhibits skewed class distributions. Data exhibits skewed class distribution when

majority of the data points belong to some classes while other classes have very few data points. The situation makes naive training of neural networks hard since they tend to biased towards head classes. Our approach is based on cost sensitive Bayesian learning with Venn predictors for robust training against the class imbalance. Experiments the Euphemisms detection dataset which had class imbalance show that this method improves over traditional fine tuning by about 6% in terms of $f-$ score (79.0% versus 72.6%). As future work, we would like to investigate how these finding extend beyond the euphemisms detection dataset.

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

Parmida Atighehchian, Frederic Branchaud-Charron, Jan Freyberg, Rafael Pardinas, Lorne Schell, and George Pearse. 2022. Baal, a bayesian active learning library. https://github.com/baal-org/baal/.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

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.

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.

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022. Cats are fuzzy pets: A corpus and analysis of potentially euphemistic terms. *arXiv preprint arXiv:2205.02728*.

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.

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.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*.

Anna Kapron-King and Yang Xu. 2021. A diachronic evaluation of gender asymmetry in euphemism. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 28–38, Online. Association for Computational Linguistics.

Ranganath Krishnan and Omesh Tickoo. 2020. Improving model calibration with accuracy versus uncertainty optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 18237–18248.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022. Searching for PETs: Using distributional and sentiment-based methods to find potentially euphemistic terms. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Seattle, USA. Association for Computational Linguistics.

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*.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics.

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*.

Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. 2021a. On model calibration for long-tailed object detection and instance segmentation. *Advances in Neural Information Processing Systems*, 34:2529–2542.

Tai-Yu Pan, Cheng Zhang, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. 2021b. On model calibration for long-tailed object detection and instance segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 2529–2542. Curran Associates, Inc.

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

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.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive BERT for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China. Association for Computational Linguistics.

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.

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.

Vladimir Vovk and Ivan Petej. 2012. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*.

Vladimir Vovk and Ivan Petej. 2014. Venn-abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, page 829–838, Arlington, Virginia, USA. AUAI Press.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. 2018. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*.

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*.

Yuzhe Yang and Zhi Xu. 2020. Rethinking the value of labels for improving class-imbalanced learning. In *Conference on Neural Information Processing Systems (NeurIPS)*.

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.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 163–168, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 229–246. IEEE.