KUL@SMM4H2024: Optimizing Text Classification with Quality-Assured Augmentation Strategies

Sumam Francis
KU Leuven
sumam.francis@kuleuven.be

Marie-Francine Moens
KU Leuven
sien.moens@kuleuven.be

Abstract

This paper presents our models for the Social Media Mining for Health 2024 shared task, specifically Task 5, which involves classifying tweets reporting a child with childhood disorders (annotated as "1") versus those merely mentioning a disorder (annotated as "0"). We utilized a classification model enhanced with diverse textual and language model-based augmentations. To ensure quality, we used semantic similarity, perplexity, and lexical diversity as evaluation metrics. Combining supervised contrastive learning and cross-entropy-based learning, our best model, incorporating R-drop and various LM generation-based augmentations, achieved an impressive F1 score of 0.9230 on the test set, surpassing the task mean and median scores.

1 Introduction

The Social Media Mining for Health (SMM4H-24) (Xu et al., 2024) shared task 5 aims to explore data sources for assessing the link between pregnancy exposures and childhood disorders in real-time and at scale. Many children face lifelong disorders, with 17% in the U.S. diagnosed with developmental disabilities and 8% with asthma. This task involves a binary classification to automatically distinguish tweets from users who reported their pregnancy and have a child with childhood developmental disorders (annotated as "1"), from tweets that merely mention a disorder without evidence of a diagnosis (annotated as "0").

SMM4H Task 5 includes datasets of English tweets with annotations on the presence or absence of childhood disorders. Recent studies (Wu et al., 2021; Francis and Moens, 2023; Liu et al., 2022) have shown that data augmentation enhances training data diversity and model robustness. It's crucial to ensure the quality of these augmentations to maintain original meanings and diversify training data. Augmented data using model-based augmentations together with regularised dropout (R-drop) (Wu et al., 2021) serve as regularization methods, mitigating overfitting in machine learning models. Model-based augmentations involve generating new data samples using pre-trained language models (LMs) that have rich semantic and syntactic knowledge stored in their parameters. This can create variations of the original data while preserving its semantic content.

2 Methodology

This section outlines our methodology consisting of 3 steps: 1. enhancing textual data with LM-based augmentation techniques, 2. assessing augmentation quality, and, 3. finetuning a transformer-based (Vaswani et al., 2017) model Bertweet (Nguyen et al., 2020) integrating supervised contrastive loss, cross-entropy loss, and regularised dropout (R-Drop) (Wu et al., 2021) in the training process.

2.1 LM-based Data Augmentation

The textual augmentations leveraging LMs used to enhance training data are **masked language modeling (MLM)**: where we mask certain tokens or spans in the input text and predict them using context from surrounding tokens (Devlin et al., 2019). We used the BERT-large model for MLM, which predicts the masked tokens based on the surrounding context. Tokens to be masked are selected based on their importance and identified using POS tags[1]. VERB and ADJECTIVE tags are masked, while NOUN phrases are left intact to preserve vital classification information. The next method is **text replacement using LMs:** This technique replaces specific words or phrases in the input text with semantically similar alternatives predicted by an LM, generating diverse and meaningful variations of the input text. For text replacement, the GPT-

[1] https://www.nltk.org/

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 142–145
August 15, 2024 ©2024 Association for Computational Linguistics

2 (Radford et al., 2019) model was employed to predict semantically similar alternatives, differing from MLM as it directly replaces words rather than predicting masked tokens. The third approach is **back translation:** where the English-French Marian MT[2] translation model is used to translate input sentences to French and then back to English, creating nuanced paraphrasing (Sennrich et al., 2016).

2.2 Evaluation Metrics for Augmentation Quality

Ensuring the quality of generated augmentations is essential to ensure they retain the original meaning and enhance the training data. We use several methods to evaluate augmentation quality. **Semantic similarity:** measures the degree of similarity between original and augmented text based on meaning, using Sentence BERT (Reimers and Gurevych, 2019) to calculate embedding-based cosine similarity. **Perplexity:** measures how well an LM predicts text samples, with lower perplexity indicating better performance and more meaningful text. We use a pre-trained GPT-2 (Radford et al., 2019) model to calculate the perplexity. **Lexical diversity**: measures the variety of vocabulary used in the text, with higher lexical diversity indicating a richer expression and ensuring useful augmentations. We use token overlap as a measure to calculate the diversity (McCarthy and Jarvis, 2010). Further, these data are used to train the classification model incorporating R-drop.

Table 1: Precision (P), Recall (R) and F1 scores (F1) on the validation set of the SMM4H2024 Task 5 with BERTweet model.

Augmentation	F1	P	R
-	0.9230	0.9078	0.9301
+ LM-aug	0.9309	0.9143	0.9471
+ LM-aug+aug-ql	0.9358	0.9225	**0.9497**
+ LM-aug+aug-ql+R-drop	**0.9398**	**0.9403**	0.9333

Table 2: Precision (P), Recall (R) and F1 scores (F1) on the test set of the SMM4H2024 Task 5 with BERTweet model.

Augmentation	F1	P	R
+ LM-aug+aug-ql+R-drop	0.923	0.906	0.940
Posteval			
+ LM-aug(pos) +aug-ql+R-drop	0.938	0.927	0.949
Task mean results	0.822	0.818	0.838
Task median results	0.901	0.885	0.917

3 Experiments and Results

The dataset comprises a training set ($7,398$ tweets), a validation set (389 tweets), and a test set ($10,000$ tweets). For pre-processing, we removed URLs, retweets, mentions, extra spaces, non-ASCII words, and characters. We also lower-cased the text, trimmed white spaces, and inserted spaces between punctuation marks.

For classification, each model was fine-tuned over 10 epochs with a learning rate of $5e-5$ using the Adam optimizer. The batch size is set to 32, and the maximum sequence length to 128. We use PyTorch and HuggingFace [3] library for training the BERTweet large model (Nguyen et al., 2020), applying both cross-entropy loss and supervised contrastive loss (Khosla et al., 2020). Model checkpoints are saved every 200 step based on the validation set's F1-score. The BERTweet loss function was adapted to include KL divergence for R-drop regularization. Training data was enriched with LM augmentations from section 3 (LM-aug). Quality checks are integrated into the augmentation pipeline to filter out poor quality augmentations (aug-ql).

Threshold values for the augmentation quality checks were empirically determined based on the performance of the validation set. This involved experimenting with different thresholds and observing their impact on the model's performance metrics (Precision, Recall, F1 scores) on the validation data. We set the following threshold values for augmentation quality check: semantic similarity > 0.7, perplexity < 100, and $0.4 <$ lexical diversity < 0.75. The semantic similarity threshold (> 0.7) ensures that the augmented text retains a high degree of similarity in meaning to the original text. The perplexity threshold (< 100) ensures that the augmented text is coherent and grammatically correct. We calculated the perplexity using a pre-trained GPT-2 model. Lower perplexity indicates better language model prediction quality. When experimenting with higher thresholds the generated text was less meaningful. A threshold of 100 was chosen because it balanced coherence with augmentation diversity. The lexical diversity range ($0.4 - 0.75$) ensures a balance between the diversity and relevance of the vocabulary used in the augmented text. A range of 0.4 to 0.75 was set to avoid too much similarity (which would defeat the purpose of augmentation) and too much dif-

[2] Helsinki-NLP/opus-mt-en-fr

[3] https://huggingface.co/models

ference (which might change the context or meaning). These thresholds were optimized iteratively, with each adjustment followed by re-evaluating the model's performance on the validation set to ensure the thresholds contributed positively to the model's overall effectiveness.

Incorporating LM augmentations and R-drop into BERTweet yielded improved performance for detecting childhood disorder diagnoses in tweets (see Tables 1 and 2) compared to baseline model setup. Augmentations diversified the data, enhancing the model's robustness and generalization. The metrics for evaluating augmentation quality further improved model performance by ensuring high-quality training data. The diversity from augmented data reduced the risk of overfitting and the model's reliance on specific patterns. Contextually generated LM-augmented examples facilitated better language understanding. The integration of R-drop with supervised contrastive loss and cross-entropy loss further promoted the learning of more generalized features by capturing various aspects of the data, enhancing the precision of the model. In the post-evaluation, the LM augmentations were performed only on the positive class (pos) which improved results. The results surpassed the shared task's mean and median scores, demonstrating the effectiveness of this approach.

4 Conclusion

In this work, we developed a classification model enhanced with R-drop and LM-based augmentations to mitigate label imbalance and avoid overfitting, thereby improving performance. We incorporated evaluation metrics like semantic similarity, perplexity, and lexical diversity to ensure the quality of the augmentations, adding only those that meet set thresholds. Our approach of integrating data augmentation and rigorous filtering strategies showed superior performance compared to the baselines we set up. This can be attributed to the enriched dataset, which reduces overfitting and enhances generalization.

Similar performance levels were achieved by the BERTweet model in (Klein et al., 2024), highlighting that the choice of model architecture and hyperparameter tuning is crucial. However, our LM-based augmentations provide a significant edge in data diversity and model robustness. Furthermore, by augmenting only the positive class, our model demonstrated even more significant improvements in the overall F1 scores. The use of LM-based augmentations introduced meaningful variations in the training data, helping the model learn to generalize better from a more diverse set of examples. The rigorous filtering strategies ensured that only high-quality augmented data were used, preserving the original meaning and enhancing the model's ability to handle diverse inputs. Our approach significantly enhances model generalizability, achieving an impressive F1 score of 0.923 on the test set.

References

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, NAACL-HLT*. Association for Computational Linguistics.

Sumam Francis and Marie-Francine Moens. 2023. Text augmentations with r-drop for classification of tweets self reporting covid-19. *CoRR*, abs/2311.03420.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *CoRR*, abs/2004.11362.

Ari Z Klein, José Agustín Gutiérrez Gómez, Lisa D Levine, and Graciela Gonzalez-Hernandez. 2024. Using longitudinal twitter data for digital epidemiology of childhood health outcomes: An annotated data set and deep neural network classifiers. *J Med Internet Res*, 26.

Zhiwei Liu, Yongjun Chen, Jia Li, Man Luo, Philip S. Yu, and Caiming Xiong. 2022. Improving contrastive learning with model augmentation. *CoRR*, abs/2203.15508.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.

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 NLP*. ACL.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3980–3990. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Sai Tharuni Samineni, and Graciela Gonzalez-Hernandez. 2024. Overview of the 9th social media mining for health applications (#SMM4H) shared tasks at ACL 2024. In *Proceedings of the 9th Social Media Mining for Health Applications Workshop and Shared Task*.