IIITN NLP at SMM4H 2021 Tasks: Transformer Models for Classification of Health-Related Tweets

Varad Pimpalkhute, Prajwal Nakhate and **Tausif Diwan**
{pimpalkhutevarad, prajwalnakhate}@gmail.com and tdiwan@iiitn.ac.in
Indian Institute of Information Technology, Nagpur

Abstract

Non-availability of well annotated and balanced datasets is considered as one of the major hurdles in analysing and extracting meaningful information from health-related tweets. Herein, we present transformer based deep learning binary classifiers for distinguishing the health related tweets for the three shared tasks 1a, 4 and 8 of the 6^{th} edition of SMM4H Workshop. We evaluate the different transformer based models viz. RoBERTa (for Task 1a & 4) and BioBERT (for Task 8), along with various dataset balancing techniques. We implement augmentation and sampling techniques so as to improve performance on the imbalanced datasets.

1 Introduction

Twitter has gained a huge popularity among all the social media platforms, especially to share and discuss information related to various aspects of life, including health-related problems. Analysing these health related Tweets and extracting the meaningful information from them is an important task for offering better health related services. With the advancements in sequential deep models, Natural Language Processing (NLP) and underlying processes got benefited from it and effective automation is introduced for the various NLP processes to a great extent. Healthcare research community has developed a keen interest in processing these health related information efficiently using advancements of deep learning. The Sixth Social Media Mining for Health Applications (SMM4H) shared tasks focus on addressing such classic health related problems applied to Twitter micro-corpus (tweets) (Magge et al., 2021).

Our team participated in three different shared binary classification tasks viz. Task 1a, Task 4, and Task 8. Task 1a focuses on distinguishing tweets mentioning adverse drug effects (ADE) from other tweets (NoADE). (O'Connor et al., 2014) focused on the identification of tweets mentioning drugs having potential signals for ADEs. Task 4 focuses on distinguishing tweets mentioning adverse potential outcomes (APO) from other tweets (NoAPO). Task 8 focuses on segregating the tweets containing self-reports (S) of breast cancer from other tweets (NR). The datasets provided for the shared tasks 1a and 8 are highly imbalanced. However, dataset for the shared Task 4 is comparative balanced. Table 1 illustrates the underlying datasets characteristics for the three shred tasks.

Due to the scarcity of users tweeting on health topics, most of the datasets on these topics are highly imbalanced in nature. (Mujtaba et al., 2019) gives a broad overview on the various balancing techniques applied on various medical datasets. (Ebenuwa et al., 2019) demonstrates the effect of strategies such as oversampling and cost-sensitivity on various health-related datasets. (Amin-Nejad et al., 2020; Tayyar Madabushi et al., 2019) presents extension of this work on cost-sensitivity to allow models such as BioBERT and BERT to generalize well on imbalanced datasets. (Liu et al., 2019; Akkaradamrongrat et al., 2019; Padurariu and Breaban, 2019) also present strategies such as text generation techniques, embedded feature extraction methods to generalize the classifier on an imbalanced dataset.

We propose transformer based classification models for the binary classification for all the aforementioned tasks. We especially address the class imbalance in the datasets, for Task 1a and Task 8. We experiment with techniques such as undersampling, oversampling, and data augmentation to address the datasets imbalance for these tasks. The rest of the paper is organized as follows. Section 2 covers the underlying datasets for the three shared tasks, their characteristics, preprocessing details, and sampling techniques to address the inherent imbalance in the dataset. Section 3 presents the classification models for the shared tasks. Re-

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 118–122
June 10, 2021. ©2021 Association for Computational Linguistics

Table 1: Dataset characteristics for the shared tasks.

Task	Label	#	Sample Instance
Task1a	ADE	1300	ooh me too! rt @xyle50ul: #schizophrenia #seroquel did not suit me at all. had severe tremors and weight gain..
	NoADE	17000	I need Temazepam and alprazolam.... Is there any doctor can prescribe for me?? :/
Task 4	APO	2922	The LAST thing you wanna do is call my son "slow" or say he's "different than everyone else" because he's a preemie.. Fuck off.
	NoAPO	3565	I don't usually use the term "rainbow baby" myself but I think it's incredibly brave when people share these... https://t.co/jjktHOewDz
Task 8	S	975	@arizonadelight i'm a breast cancer survivor myself so i understand the scare.
	NR	2840	All done, we done for raising awareness, I have a good friend battling this at the moment #breastcancer.

sults and discussions are sketched in the Section 4. Section 5 conclude the paper and presents future research directions.

2 Dataset: Sampling Techniques and Preprocessing

The datasets for the shared tasks were collected in the form of English tweets. The datasets were well annotated for each of the shared tasks. We majorly employ three dataset balancing techniques viz. undersampling, oversampling, and augmentation.

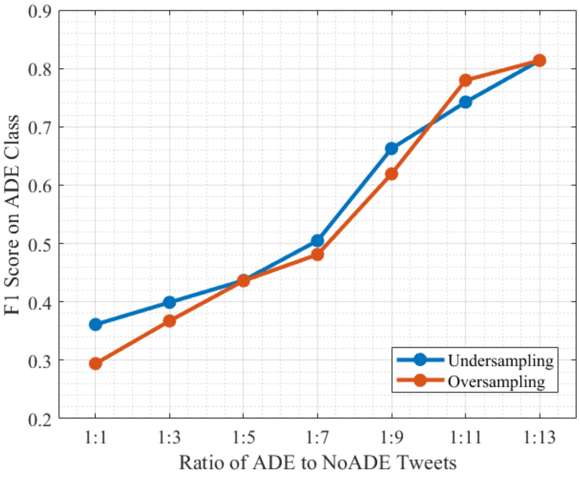
2.1 Sampling Techniques

Under-sampling is performed to balance the data by reducing the instances of the excessive class nearly equal to the rare class. Over-sampling is the approach to duplicate the rare class instances, thus increasing the number of samples of rare class to that of the excess class in the dataset. We achieved this either by addition of tweets of rare class with repetition or using Synthetic Minority Over-Sampling technique (SMOTE) (Bowyer et al., 2011). Performance of these sampling techniques for different ratios of rare to excess class for the dataset of Task 1a on applying RoBERTa model are presented in Figure 1. For our experiments, rare class is ADE / APO / S and excess class is NoADE / NoAPO / NR for three datasets corresponding to three shared tasks.

2.2 Data Augmentation

Data-Augmentation using the nlpaug library (Ma, 2019) is undertaken to balance the datasets. Synthetic data of the rare class is added by generating tweets with different spellings, synonyms, word-embedding, contextual word-embedding of words in-order to have artificial tweets look as natural as real tweets. Data Augmentation is different from Oversampling in the sense that data augmentation adds variations in input text whereas oversampling is not able to change the features of the text.

Figure 1: Sampling performance using RoBERTa model for Task 1a.



2.3 Pre-processing

Before feeding the dataset to a text classification model, we cleaned and preprocessed the tweets in each of the datasets. For each tweet in the dataset, we normalized usernames and keywords into reserved keywords[1]. We also de-emojized the tweets using the emoji package[2] to replace the emojis with relevant tags. Lastly, we expanded contractions[3]

[1] https://github.com/avian2/unidecode

[2] https://github.com/carpedm20/emoji

[3] https://github.com/kootenpv/contractions

Table 2: Datasets Characteristics for each of the three tasks.

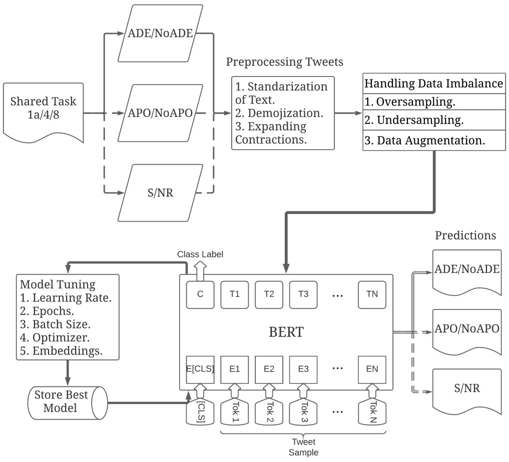
	Task 1a			Task 4			Task 8		
Corpus	**ADE**	**NoADE**	**#**	**NoAPO**	**APO**	**#**	**NR**	**S**	**#**
Train Set	1235	16150	17385	3030	2484	5514	2615	898	3513
Valid Set	65	850	915	535	438	973	225	77	302
Test Set	NA	NA	10000	NA	NA	10000	NA	NA	1204

and lower-cased the text to present the data in a much cleaner format.

3 System Description And Model

We employ transformer based models and their architectural variants for all the shared tasks, along with dataset balancing techniques described in the previous section. For all the tasks, the experiments have been performed using the scikit-learn, Tensorflow[4], PyTorch [5] and Flair (Akbik et al., 2019) frameworks. Table 2 describes the three datasets and their distribution in train, test, and validation sets for training and evaluation of transformer based sequence models.

Figure 2: Proposed model architecture.



3.1 Classification Model

We mainly experimented with various tranformer languages models such as BERT (Devlin et al., 2018), DistilBert (Sanh et al., 2019), XLNET (Yang et al., 2019), and RoBERTa (Liu et al., 2019). In addition to these routine transformer models, we also experimented on health related architectural variants such as BioBERT (Lee et al., 2019), BERT-Epi (Müller et al., 2020) and BERTweet (Nguyen et al., 2020). Table 3 presents the sample results of all these models for shared task 4. In the subsequent section, we demonstrated the results for the best preforming transformer models for each of the shared tasks. Furthermore, we penalized the loss of the rare class with a loss weight two times the original loss weight. We kept the loss weight for the excess class as it is. We experimented each of the models on four different versions of the underlying dataset: Original, Undersampled, Oversampled and Augmented. The architecture of our proposed system is illustrated in Figure 2.

Table 3: Comparative results of various transformer based models for the shared task 4.

Architecture	xLR ($\times 10^{-6}$)	F1	Prec	Recall
BERT	10	0.872	0.843	0.902
BERTweet	10	0.899	0.896	0.906
DistilBERT	50	0.835	0.839	0.831
RoBERTa	6	**0.924**	0.897	**0.952**
XLNET	5	0.903	**0.922**	0.886
BioBERT	5	0.874	0.859	0.890

3.2 Hyperparamter Tuning

All the experiments have been performed on Flair Framework. We tried various ensemble of models – where, there were three models in each ensemble – but, this didn't draw good results on the validation set, thus, we choose the final model as a single transformer language model. Ensembling didn't work well as majority of the incorrectly predicted samples were predicted incorrectly by most of the models in the ensemble. For Task 1 and Task 4, we choose the final transformer model as RoBERTa, and for Task 8 we made use of a health related model trained on COVID19 related tweets – BioBERT. We experimented with various hyperparamter settings such as learning rate, learning rate decay, early stopping, varying batch size, and number of epochs. Based on the various experiments, we settled that the learning rate in the range of 0.000006 - 0.00001, batch size of 8, patience of 2

[4] https://www.tensorflow.org/

[5] https://pytorch.org/

and 3 epochs of training gave the best performance on the models. The performance was measured across standard metrics such as precision and recall, with the final determining metric being the harmonic mean of precision and recall (F1-score) for the rare classes.

4 Results & Discussions

All the experiments were performed on an Intel core i5 CPU @2.50GHz, 8GB RAM machine having 4 logical cores. The task wise results can be presented as follows:

4.1 Task 1a: Adverse Drug Effect Mentions.

Table 4: Task 1a using RoBERTa (Learning Rate = 1 × $10^{-}5$, Epochs = 3).

Validation set			
Dataset	**F1**	**Precision**	**Recall**
Undersampled	0.5048	0.5561	0.4623
Oversampled	0.4361	0.4186	0.4553
Original	0.8136	**0.9057**	0.7385
Augmented	**0.8433**	0.8209	**0.8572**
Test set			
Dataset	**F1**	**Precision**	**Recall**
Original	0.3	0.473	0.217
Augmented	0.4	0.405	0.401
Median	**0.44**	**0.505**	**0.409**

As we know, Task 1 is a highly imbalanced dataset with the ratio of ADE to NoADE tweets being about 1:13. Table 4 presents the metrics on the validation as well as test data for Task 1a. As it can be observed, RoBERTa shows the best performance on Augmented Dataset. Undersampling results in underfitting the training model whereas oversampling results in model overfitting. The probable reason behind this is the sparse ADE samples present in the dataset for the shared Task 1a. In contrast, data augmentation results in increasing variations in the training dataset, thus, we are able to generalize well as compared to the original dataset.

4.2 Task 4: Self-reporting Adverse Pregnancy Outcome.

Similar to Section 4.1, RoBERTa model shows the best performance on the validation set for the shared Task 4 also, represented using Table 5. As Task 4 dataset was comparatively balanced, there was little motivation for using sampling techniques on the dataset. Surprisingly, augmenting the data couldn't draw better F1 score.

4.3 Task 8: Breast Cancer Self-reports.

Task 8 is also an imbalanced dataset with the ratio of Self-Reports to Non-Relevant Tweets being about 1:3. Thus, similar to Section 4.1, we experiment with all the four variations of the dataset. The metrics on the validation and test data are presented in Table 6. It can be seen that the model with the best performance is on the augmented dataset. As the imbalance in Task 8 was significantly lower than that in Task 1, we observe better results for this task.

Table 5: Task 4 using RoBERTa (Learning Rate = 6 × $10^{-}6$, Epochs = 5).

Validation set			
Dataset	**F1**	**Precision**	**Recall**
Original	0.9437	0.9251	0.9631
Augmented	0.9279	0.9028	0.9543
Test set			
Dataset	**F1**	**Precision**	**Recall**
Original	**0.93**	0.9149	**0.9412**
Augmented	0.92	0.8919	0.948
Median	0.925	**0.9183**	0.9234

Table 6: Task 8 using BioBERT (Learning Rate = 5 × $10^{-}6$, Epochs = 10).

Validation set			
Dataset	**F1**	**Precision**	**Recall**
Undersampled	0.8182	0.7273	0.9351
Oversampled	0.828	0.8125	0.8442
Original	0.8707	**0.9143**	0.8313
Augmented	**0.8947**	0.9067	**0.8831**
Test set			
Dataset	**F1**	**Precision**	**Recall**
Original	0.83	0.8441	0.8216
Augmented	0.84	**0.8706**	0.8084
Median	**0.85**	0.8701	**0.8377**

5 Conclusions

We proposed a text classification pipeline while also making an attempt to handle dataset imbalance corresponding to three different shared tasks in SMM4H'21 (Magge et al., 2021). We conclude that data augmentation gives best performance on highly imbalanced datasets. Moreover, augmentation provides better results in case of comparatively balanced datasets. As part of future work, additional experiments are planned to further analyze strategies to improve the performance of the model on the dataset.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

S. Akkaradamrongrat, P. Kachamas, and S. Sinthupinyo. 2019. Text generation for imbalanced text classification. In *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 181–186.

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708, Marseille, France. European Language Resources Association.

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat. 2019. Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE Access*, 7:24649–24666.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

H. Liu, M. Zhou, and Q. Liu. 2019. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, 6(3):703–715.

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. *CoRR*, abs/1907.11692.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

Ghulam Mujtaba, Liyana Shuib, Norisma Idris, Wai Lam Hoo, Ram Gopal Raj, Kamran Khowaja, Khairunisa Shaikh, and Henry Friday Nweke. 2019. Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications*, 116:494–520.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

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*.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, , Karen Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. pages 924–33.

Cristian Padurariu and Mihaela Elena Breaban. 2019. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.