Transformer Models for Classification on Health-Related Imbalanced Twitter Datasets

Varad Pimpalkhute[1], Prajwal Nakhate[1], Tausif Diwan[1]

[1]Indian Institute of Information Technology, Nagpur

Motivation

1. Increased interest of healthcare community in processing health related information efficiently using NLP/DL.
2. The 6th SMM4H Workshop (Magge et. al, 2021) shared tasks focused on addressing such classic health related problems on Twitter corpus.

Goal

We participated in three shared tasks, where the goal was to build a system for binary classification of the given dataset. Our goal was to:

1. Developing binary classifier models for the shared tasks.
2. Address the imbalance present in the datasets and rectify the imbalance.
3. Fine tuning and optimizing performance of proposed models.

Shared Tasks

A) Task 1a: Classification of Tweets mentioning Adverse Drug Effects.

Labels: ADE, NoADE

Corpus	ADE	NoADE	#Total
Train set	1235	16150	17385
Valid set	65	850	915
Test set	NA	NA	10000

B) Task 4: Classification of self-reported adverse pregnancy outcomes.

Labels: APO, NoAPO

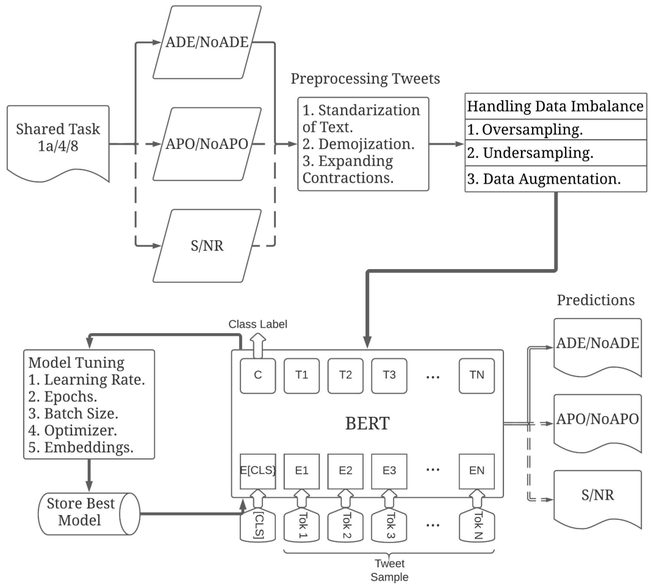
Corpus	APO	NoAPO	#Total
Train set	2484	3030	5514
Valid set	535	438	973
Test set	NA	NA	10000

C) Task 8: Classification of self-reported breast cancer tweets.

Labels: S, NR

Corpus	S	NR	#Total
Train set	898	2615	3513
Valid set	77	225	302
Test set	NA	NA	1204

Proposed Model Architecture for Tasks 1a, 4 & 8.



Addressing Imbalance in datasets

A) Undersampling Dataset

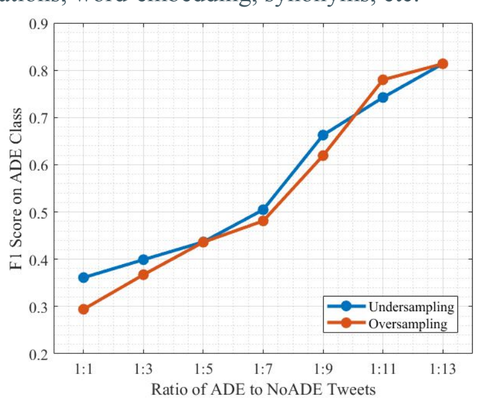
- Reduce samples of majority class in the dataset.

B) Oversampling Dataset

- SMOTE doesn't work well on text data due to its nature of high dimensionality.
- Thus, increasing samples of minority class (duplication) is executed.

C) Data augmenting Dataset

- Make use of *nlpaug* library for augmenting dataset.
- Synthetic data is generated by making use of spelling variations, word-embedding, synonyms, etc.



Evaluation and Results

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 and 3 epochs of training gave the best performance on the models.

Task 1a using RoBERTa (Learning Rate = 1 × 10−5, Epochs = 3).

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**

Task 4 using RoBERTa (Learning Rate = 6 × 10−6, Epochs = 5).

Dataset	F1	Precision	Recall
Original	**0.933**	0.9149	**0.9412**
Augmented	0.92	0.8919	0.948
Median	0.925	**0.9183**	0.9234

Task 8 using BioBERT (Learning Rate = 5 × 10−6, Epochs = 10).

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**

Conclusion

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

References

[Magge et al., 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.