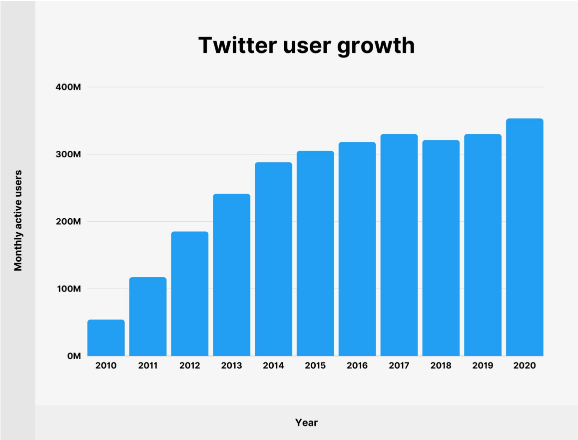
Transformers Models for Classification of Health-related Tweets

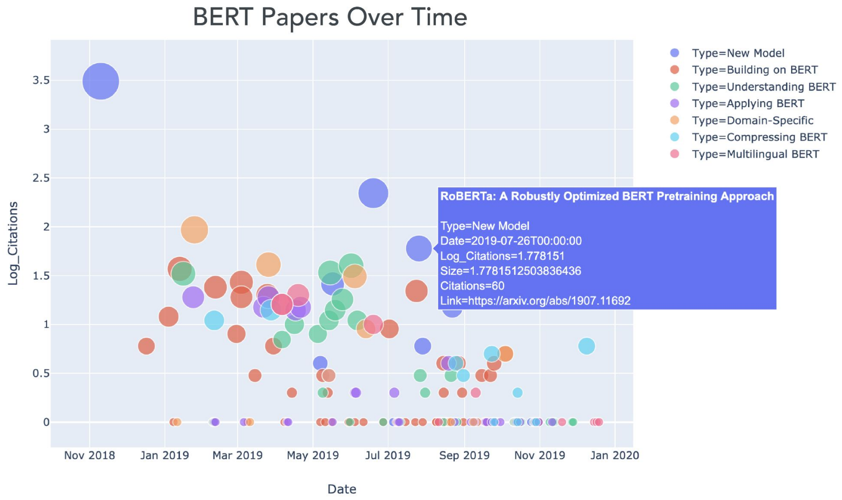
HUGGING FACE

- Varad Pimpalkhute BT17ECE071
- Prajwal Nakhate BT17ECE070
- Project Guide : Dr.Tausif Diwan
- Indian Institute of Information Technology, Nagpur

flair

Motivation behind mining tweets in Health Applications





- Increase in use of twitter over the years.
- Advancements in Natural Language Processing (NLP) and sequential deep models.

Problem Statement

- We participated in the Sixth Social Media Mining for Health Applications (SMM4H) shared tasks focus on addressing such classic health related problems applied to Twitter micro-corpus.

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

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

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

- The goal was to improve the performance of our binary classification model on each of the three tasks such that they perform well on the test set.

Goal

1. Developing binary classifier models for the shared tasks.
2. Dealing with imbalanced annotated datasets.
3. Fine tuning and optimizing performance of proposed models.

Studying the dataset of each task

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.

Studying the dataset of each task

	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

- Shortcomings of the dataset and plausible disadvantages while training on a text classifier model.
- Proposing techniques to counter data imbalance.

Countering Data Imbalance

A) Undersampling Dataset

- Reduce samples of majority class in the dataset.
- A reduction of ratio from 1:16 to 1:4 or 1:6 results in better performance than ratio of 1:1

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.
- Ratio is adjusted according to bias present.

C) Data augmentation

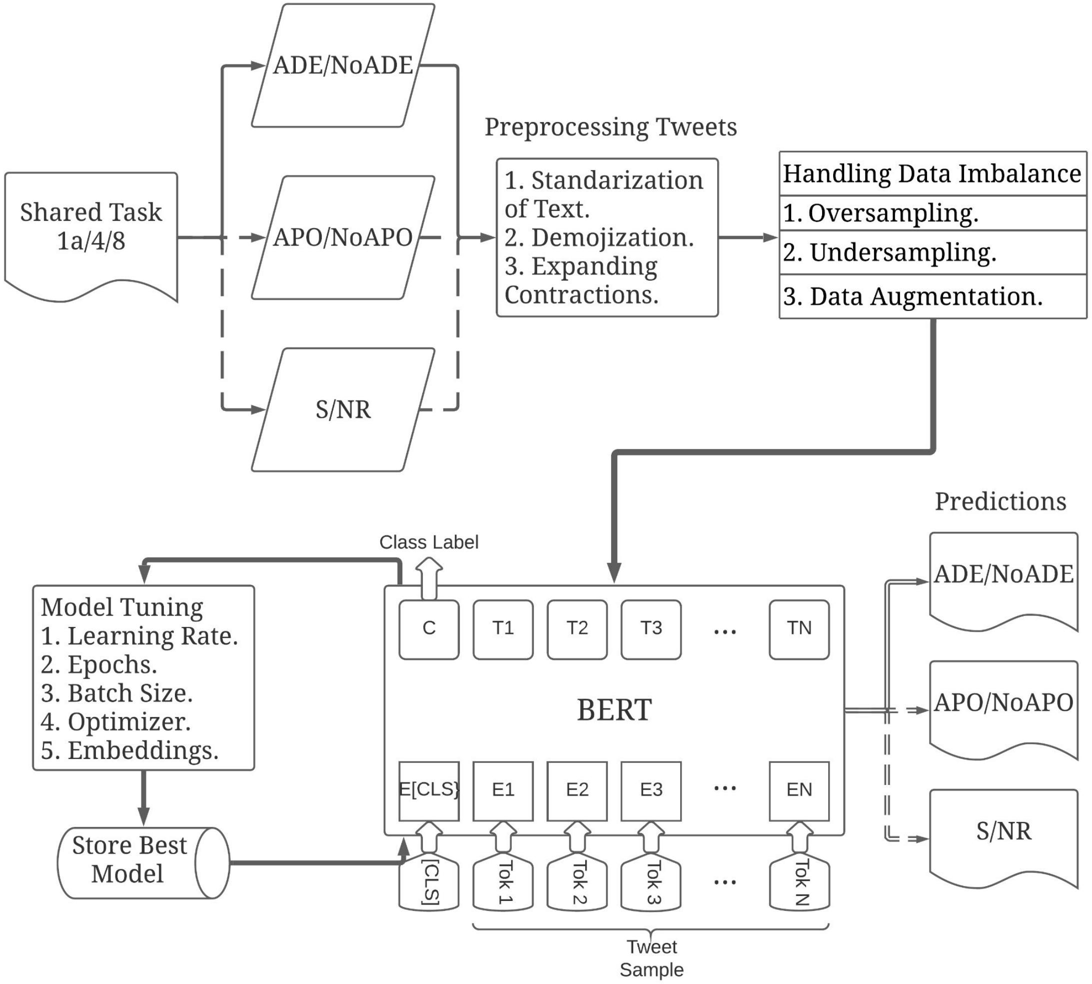
- Make use of *nlpaug* library for augmenting dataset.
- Synthetic data is generated by making use of spelling variations, word-embedding, synonyms, etc.

Preprocessing datasets.

For each tweet in our dataset:

- Normalized usernames and keywords into reserved keywords.
  - Library used: unidecode
  - Example: S’il vous plaît :: Sil vous plait
- De-emojized the tweets to the emojis with relevant tags.
  - Library used: emoji
  - Example: 😂 :: :face_with_tears_of_joy:
- Expanded contractions for faster processing.
  - Library used: contractions
  - Example: I don’t want to go to Paris :: I do not want to go to Paris.
- Lower-cased all tweets.

Proposed general model architecture for SMM4H shared tasks 1a, 4 & 8.



System Description

- Machine Description:
  - Intel core i5
  - CPU @2.50GHz
  - 8GB RAM
  - 4 logical cores.

- Framework & Libraries used:
  - NLPAUG
  - Tensorflow
  - PyTorch
  - Flair
  - KTrain

System Description

Classification Model

- We experimented on various types of embeddings.
- RoBERTa and BioBERT worked the best in our case.

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

Hyperparameter Tuning

1. Penalize loss weights of rare class two times.
2. Experiment of all 3 variations of dataset.
3. Learning Rate: 6e-6 - 1e-5
4. Batch size: 8
5. Patience: 2
6. Max Epochs: 3

Results & Observation: Task 1a

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**

Discussion:

- The ratio of rare to excess class is 1:13.
- Embedding name: RoBERTa.
- Learning Rate: 1x10^(-5)
- Max epochs: 3
- Undersampling and oversampling don’t tend to work well.
- Augmentation seems to give good results.

Results & Observation: Task 4

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

Discussion:

- The ratio of rare to excess class is close to 1:1
- Embedding name: RoBERTa
- Learning Rate: 6x10^(-6)
- Max epochs: 5
- Less motivation to make use of sampling techniques.
- Surprisingly, augmenting data doesn't produce good results.

Results & Observation: Task 8

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**

Discussion:

- The ratio of rare to excess class is 1:3
- Embedding name: BioBERT
- Learning Rate: 5x10^(-6)
- Max epochs: 10
- Undersampling and oversampling don’t tend to work well.
- Augmentation seems to give best results.

Conclusion & Future Work

- We proposed a text classification pipeline while also making an attempt to handle dataset imbalance corresponding to three different shared tasks in SMM4H'21.
- 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

[1] Health Language Processing Lab @ Penn IBI

[2] FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP| Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, Roland Vollgraf

[3] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding | Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

[4] Edward Ma. 2019. Nlp augmentation

[5] Dealing with Data Imbalance in Text Classification l Cristian Padurariu, Mihaela Elena Breaban

[6] RoBERTa: A Robustly Optimized BERT Pretraining Approach| Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov

[7] BioBERT: a pre-trained biomedical language representation model for biomedical text mining| Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang

Questions