UB Health Miners@SMM4H'22: Exploring Pre-processing Techniques To Classify Tweets Using Transformer Based Pipelines.

Roshan Vivek Khatri, Sougata Saha, Souvik Das, Rohini K. Srihari
Department of Computer Science and Engineering
University at Buffalo, Amherst, NY 14260
{roshanvi,sougatas,souvikda,rohini}@buffalo.edu

Abstract

Here we discuss our implementation of two tasks in the Social Media Mining for Health Applications (SMM4H) 2022 shared tasks – classification, detection, and normalization of Adverse Events (AE) mentioned in English tweets (Task 1) and classification of English tweets self-reporting exact age (Task 4). We have explored different methods and models for binary classification, multi-class classification, and named entity recognition (NER) for these tasks. We have also processed the provided dataset for noise, imbalance, and creative language expression from data. Using diverse NLP methods, we classified tweets for mentions of adverse drug effects (ADEs) and self-reporting the exact age in the tweets. Further, extracted reactions from the tweets and normalized these adverse effects to a standard concept ID in the MedDRA vocabulary.

1 Introduction

Since 2005 in the US, the percentage of American adults using social media has risen from 53% in 2012 to 72% in 2021 according to the research and tracking by Pew Research Center. Tremendous amounts of data are being generated on social media, where people express different aspects of their lives to their social circle. Many people also express their thoughts on various health-related topics. All this data presents significant opportunities for studying it to monitor people's health and the factors affecting their health. Our work is inspired by the current research using transformer architectures, and the results that Autobots Ensemble Team (Saha et al., 2020) had achieved at SMM4H 2020 and the KFU NLP Team (Miftahutdinov et al., 2019) had achieved at SMM4H 2019 using BERT (Kenton and Toutanova, 2019) as well as by the benchmark work and system configuration in ReportAGE (Klein et al., 2022).

Task 1 consists of 3 subtasks, Classification, Extraction, and Normalization. For a binary classification task, distinguishing tweets that report an adverse effect (AE) of medication are annotated as "1", from those that do not are annotated as "0", subtle causal language variation in the tweets needs to be addressed between AEs and indications. For extractions, the beginning and end offsets of these spans and the actually extracted spans are provided in the dataset for training the model and removing the spans for the unseen data. For the Normalization subtask, the extracted spans, and the normalized standard concept IDs of MedDRA[1] vocabulary are provided.

Task 4 is a binary classification task that automatically distinguishes tweets that self-report the user's exact age, annotated as "1", from those that do not, annotated as "0". Automatically identifying the actual age, rather than their age groups, would enable the large-scale use of social media data for applications that do not align with the predefined age groupings of extant models, including health applications such as identifying specific age-related risk factors for observational studies, or selecting age-based study populations.

2 Models

2.1 Task 1a: Classification of tweets that report adverse effects

The dataset (Magge et al., 2021) provided for this task has 17,120 annotated tweets for classifying tweets containing Adverse Drug Events (ADEs) related to drugs from the ones that do not contain ADEs. For classification, we used a deep neural network classifier based RoBERTa-based (Liu et al., 2019), pre-trained transformer model from Hugging Face[2] (Wolf et al., 2019).

For the RoBERTa-based classifier, we pre-processed the tweets by normalizing the hashtags to words, emoticons to expressions, and removing

[1] https://www.meddra.org/
[2] https://huggingface.co/

Proceedings of the 29th International Conference on Computational Linguistic, pages 114–117
October 12–17, 2022.

Model Type	Precision	Recall	F1-score
RoBERTa	0.74	0.43	0.54
RoBERTa + Downsampling	0.38	0.88	0.53
RoBERTa + Downsampling + Preprocessing	0.4	0.83	0.54
RoBERTa + Downsampling + Preprocessing + Data Augmentation	0.54	0.71	0.62
RoBERTa + Downsampling + Preprocessing + Data Augmentation + Paraphrasing	0.38	0.86	0.52
RoBERTa + Preprocessing + Data Augmentation	0.53	0.73	0.61

Table 1: Task 1a results for different system configurations on the validation set

the duplicates, and URLs from the dataset. After getting the embeddings of the pre-processed tweet, we took the mean of the sequence output to be the pooled output and passed it to the hidden layer, dropout layer (drop rate of 0.5), and a linear layer that predicts the class of each tweet. For training, we used Adam Optimizer (Kingma and Ba, 2014), 10 epochs, warmup of 0.2, the learning rate of 1e-5, and weight decay of 0.001.

As the data set was quite imbalanced, with approximately 10% data of "ADEs" class, and the remaining 90% data for the "non-ADEs" class, we have tried different techniques to balance the dataset by downsampling the major class, data augmentation of the positive class by back-translation (Aji et al., 2021) method to german as well as Russian, paraphrasing using the parrot paraphraser, and assigning weights to the positive class. Table 1 contains our result for task 1a and the best F1 score of 0.62 was obtained by downsampling, pre-processing, and data augmentation by back-translation methods from both German and Russian.

2.2 Task 1b: Extraction of spans in the tweets

For this task, only 1,708 tweets in the dataset contain an extraction of spans in the tweets containing ADEs. We trained a Named Entity Recognition model using this dataset. For this, we converted the dataset in a data frame to the Spacy required format containing the text and the entities containing the offset of the entities. We trained our model for the new entity over the "en_core_web_sm" model of Spacy and specifically the "ner", "trf_wordpiecer", "trf_tok2vec" pipelines of the model. Table 2 shows the achieved Relaxed F1 score of 0.28 for the model we trained.

Model Type	Precision	Recall	F1-score
Spacy NER	0.19	0.48	0.28

Table 2: Task 1b results on validation set.

Training Data	Precision	Recall	F1-score
2022 Dataset	0.03	0.09	0.03
2022 + 2020 Dataset	0.22	0.27	0.23
2022 + 2020 + CADEC Dataset	0.45	0.47	0.45

Table 3: Task 1c results for different datasets on the validation set

2.3 Task 1c: Mapping the spans of adverse effects to standard concept IDs in the MedDRA vocabulary

The dataset contains 1,711 tweet spans and mapped MedDRA ids for training. For this mapping, we developed a multi-class classifier where the number of classes is equal to the number of unique MedDRA terms available in the dataset. This multi-class classifier is also a deep neural network classifier based on the RoBERTa-based pre-trained transformer model from Hugging Face. We have padded/truncated (Gattepaille, 2020) the embeddings to the length of 16 to get the best results. After getting these padded/truncated embeddings of the preprocessed spans, we took the mean of the sequence output to be the pooled output and passed it to the hidden layer, dropout layer (drop rate of 0.5), and a linear layer that predicts the class of each tweet. For training, we used Adam Optimizer, 16 epochs, warmup of 0.2, the learning rate of 2e-5, and weight decay of 0.001.

As the dataset was quite small, we also used the 2020 dataset along with the CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al., 2015) MedDRA dataset, which in total 8,815 for training. Table 3 shows the results for this task and the best F1 score achieved was 0.45.

2.4 Task 4: Classification of tweets self-reporting exact age

The provided dataset comprises 11,000 annotated tweets for the experiment. Amongst them are 8,800

Model Type	Precision	Recall	F1-score
RoBERTa	0.85	0.92	0.88
RoBERTa + Preprocessing	0.86	0.90	0.88
RoBERTa + Downsampling	0.75	0.97	0.85
RoBERTa + Downsampling + Preprocessing	0.77	0.94	0.85
RoBERTa + Preprocessing + Data Augmentation(german)	0.82	0.92	0.87
RoBERTa + Downsampling + Preprocessing + Data Augmentation + Paraphrasing	0.79	0.92	0.85

Table 4: Task 4 results for different system configurations on the validation set

tweets for training and 2,200 as validation dataset for the classification of "age" and "no age" tweets. For classification, we used a deep neural network classifier based on the RoBERTa pre-trained transformer model from Hugging Face.

For the RoBERTa-based classifier, we preprocessed the tweets by normalizing the hashtags to words, emoticons to expressions, and usernames to the "@USER_" special token, along with normalizing the URLs and removing the duplicates from the dataset. After getting the embeddings of the preprocessed tweet, the mean of the sequence output is considered as the pooled output and passed to one hidden layer, dropout layer (drop rate of 0.5), and a linear layer that predicts the class of each tweet. For training, we used Adam Optimizer, 10 epochs, warmup of 0.2, the learning rate of 1e-5, and weight decay of 0.001.

As the dataset was quite imbalanced amongst the classes, with 32.20% data for the "age" class and the remaining 67.80% data for the "no age" class, we have tried different techniques to balance the data set by downsampling, and data augmentation by back-translation method to german as well as Russian and paraphrasing using the parrot paraphraser. All these techniques did not help in increasing the F1 score of the model as seen in Table 4 and therefore, only by preprocessing the data did the best F1 score of 0.88 with the best precision of 0.86 obtained.

3 Error Analysis

We observe that the model performs poorly to classify the tweets containing drug names and other medical terms. For the Named Entity Recognition, the model using Spacy fails in the cases where multiple spans must be extracted from the tweet. For the Normalization, for improving the F1 score, more data is required as we have only 8,815 data points including the dataset of SMM4H 2022, SMM4H 2020, and CADEC corpus trained to classify a total of 1,023 unique MedDRA IDs.

4 Result and Analysis

The performance evaluation metrics for the tasks are precision (P), recall (R), and F1-score (F1) computed on the positive class which is the minority class. The above result tables report the various techniques and the scores of those respective models. Different hyperparameters were tested to get the optimum F1 score of each of the models and different padding/truncating length lines 40, 32, and 16 of the embedding for the Normalization task were explored to see if that affected the results for the Tasks and padding/truncating to embedding length of 16 gave the best results.

5 Conclusion

All the datasets that were provided were highly imbalanced amongst the classes and many techniques were explored to overcome these imbalances, namely Random downsampling of the majority class, increasing the number of minority classes by paraphrasing the data of those classes, data augmentation by a round trip translation from English to German and back to English and this was also tried with Russian to create more data points. From the whole project, it was learned that the balance of the data between classes does not guarantee the performance of the model. It can also be seen that the preprocessing step is also very important to preserve the context of the tokens with respect to each other. For future scope, a medical domain-based pre-trained language model, with the present ones in an ensemble architecture might work to improve the score of the system.

References

Alham Fikri Aji, Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Radityo Eko Prasojo, and Tirana Fatyanosa. 2021. Bert goes brrr: A venture towards the lesser error in classifying medical self-reporters on twitter. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 58–64.

Lucie Gattepaille. 2020. How far can we go with just out-of-the-box bert models? In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 95–100.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, pages 4171–4186.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ari Z Klein, Arjun Magge, and Graciela Gonzalez-Hernandez. 2022. Reportage: Automatically extracting the exact age of twitter users based on self-reports in tweets. *PloS one*, 17(1):e0262087.

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*.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Deepademiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2019. Kfu nlp team at smm4h 2019 tasks: Want to extract adverse drugs reactions from tweets? bert to the rescue. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 52–57.

Sougata Saha, Souvik Das, Prashi Khurana, and Rohini K Srihari. 2020. Autobots ensemble: Identifying and extracting adverse drug reaction from tweets using transformer based pipelines. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 104–109.

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.

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*.