Autobots Ensemble: Identifying and Extracting Adverse Drug Reaction from Tweets using Transformer Based Pipelines

Sougata Saha[∗], **Souvik Das**[∗], **Prashi Khurana**[∗], **Rohini K. Srihari**
Department of Computer Science and Engineering
University at Buffalo, Amherst, NY 14260
{sougatas, souvikda, prashikh, rohini}@buffalo.edu

Abstract

This paper details a system designed for Social Media Mining for Health Applications (SMM4H) Shared Task 2020. We specifically describe the systems designed to solve task 2: Automatic classification of multilingual tweets that report adverse effects, and task 3: Automatic extraction and normalization of adverse effects in English tweets. Fine tuning RoBERTa large for classifying English tweets enables us to achieve a F1 score of 56%, which is an increase of +10% compared to the average F1 score for all the submissions. Using BERT based NER and question answering, we are able to achieve a F1 score of 57.6% for extracting adverse reaction mentions from tweets, which is an increase of +1.2% compared to the average F1 score for all the submissions.

1 Introduction

With the world adapting to the new normal, social media is proving to be a key resource for humans. With more people sharing their life experiences in social media platforms, pharmaceutical firms can benefit by leveraging the power of deep learning and natural language processing for digital pharmacovigilance. In this paper, we showcase our systems for task 2 & 3 of the Social Media Mining for Health Applications Shared Task 2020 (Klein et al., 2020). Inspired by the current research using transformer architectures, and the results that KFU NLP Team (Miftahutdinov et al., 2019) had achieved at SMM4H 2019 using BERT (Devlin et al., 2018), we experimented with a suite of different transformer architectures. Transformers (Vaswani et al., 2017) are solely based on attention mechanisms, which dispense recurrence and convolutions entirely, enabling parallel processing and state of the art models. Liu et al. (2019) in their research uncovered that BERT was significantly under trained, and proposed RoBERTa: A Robustly Optimized BERT Pretraining Approach. We fine tuned RoBERTa on the English training tweets to classify a tweet as containing adverse reaction mention or not. For the Russian and French tweets classification tasks, we fine tuned RuBERT (Kuratov and Arkhipov, 2019) and CamemBERT (Martin et al., 2019) respectively.

For extracting the adverse reaction mentions from a tweet, we devised an end to end pipeline by posing the task as a named entity recognition (NER) task as well as a question answering task. We fine tuned BERT, SciBERT (Beltagy et al., 2019) and BioBERT (Lee et al., 2019) and created an ensemble NER, and fine tuned BioBERT QA for the question answering module. Post adverse mention extraction, we normalised the mention to the MedDRA code using pre-trained fastText (Bojanowski et al., 2016) embeddings and cosine similarity. The paper is organized as follows. In section 2 we describe the problem statements. Section 3 describes the architectures and methods that were implemented for each of the tasks, and showcase our results in section 4. We discuss some of the challenging aspects of the problems in section 5, and finally conclude in section 6.

[∗]Equal contribution

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 104–109
Barcelona, Spain (Online), December 12, 2020.

Task	Training Set			Test Set
	% Positive	% Negative	Total examples	
Task 2-English	9.25%	90.75%	25,672	5,000
Task 2-Russian	8.75%	91.25%	7,612	1,903
Task 2-French	1.61%	98.39%	2,426	607
Task 3-Resolution(NER + Norm)	51.01%	48.61%	2,376	1,000

Table 1: Data distribution for each task.

2 Task and Data Description

2.1 Task 2: Automatic classification of multilingual tweets that report adverse effects

This task involved developing a system which is capable of distinguishing tweets that report an adverse reaction to medication from tweets that do not. This task was subdivided into 3 tasks by language of tweet: English, Russian and French. Table 1 shows the distribution of training and testing data samples, and the split of positive and negative examples in the training data set.

2.2 Task 3: Automatic extraction and normalization of adverse effects in English tweets

This task consisted of two parts. The first being extraction of the specific adverse reaction of a drug from English tweets. The second being mapping the extracted adverse reaction to a standard concept ID in the MedDRA vocabulary. Table 1 shows the distribution of training and testing data samples, and the split of positive and negative examples in the training data set.

3 Methods

3.1 Task 2-Automatic classification of multilingual tweets that report adverse effects: English

Twitter data is almost always noisy, hence we cleansed and pre-processed the tweets before training the classifier. Using Ekphrasis (Baziotis et al., 2017), regex and NLTK, we converted tweets to lowercase, normalized elongated characters, repeated characters and hashtags, unpacked contractions, removed URL, mentions, smileys and emojis, and removed special tweet tokens like ‘rt’.

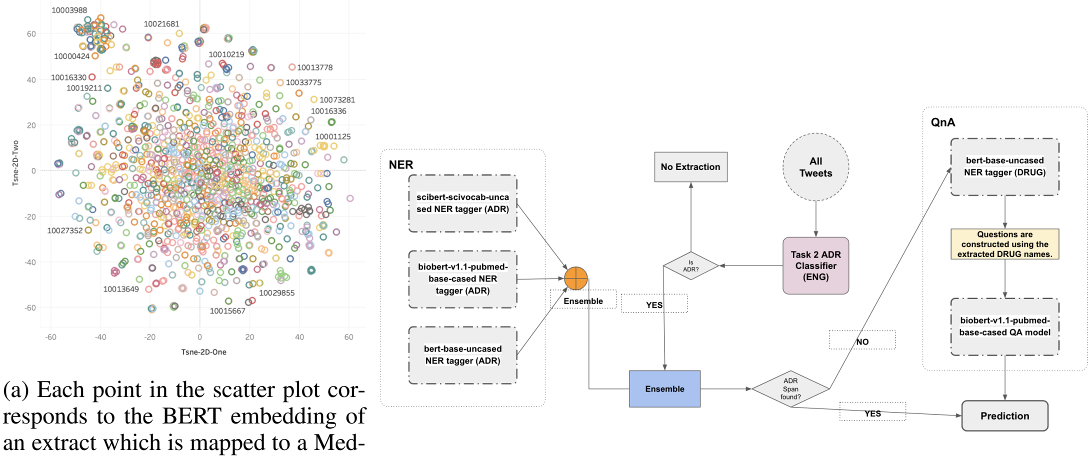
We experimented with different transformer models like BERT base uncased, SciBERT with scivocab, BioBERT base v1.1 and RoBERTa large, and achieved best validation results with RoBERTa large. We fine tuned the RoBERTa large model using the pre-processed English tweets. We sum pooled the last 6 layers of RoBERTa, and performed classification by passing the pooled representation through a linear layer. We trained the model for 6 epochs with a learning rate of 2e-5. Table 2 demonstrates the model performance on the validation set, and Table 3 demonstrates the model performance on the test set.

3.2 Task 2-Automatic classification of multilingual tweets that report adverse effects: Russian and French

For the Russian and French classification, we pre-processed the tweets by removing special tweet tokens like ‘rt’, URL, mentions, smileys and emojis. We experimented with different transformer models like multilingual BERT, RuBERT, CamemBERT and FlauBERT (Le et al., 2019), and got best validation results using RuBERT for Russian tweets, and CamemBERT for French tweets. For both the models we had sum pooled the last 4 layers of the transformer, and trained for 6 epochs with a learning rate of 2e-5. For the French tweets we achieved a validation F1 of 0.22, but unfortunately could not classify any tweets correctly in the test data set. Table 2 demonstrates the performance of the Russian tweet classifier on the validation set, and Table 3 demonstrates the model’s performance on the test set.

3.3 Task 3-Automatic extraction and normalization of adverse effects in English tweets: Extraction

We devised a three step extraction pipeline for this task, which is illustrated in Figure 1b. We detail the three steps below:


10003988
10021681
10000424
10010219
10013778
10016330
10033775
10019211
10073281
10016336
10001125
10027352
10013649
10029855
10015667
60
40
20
0
-20
-40
-60
Tsne-2D-Two
-60
-40
-20
0
20
40
60
Tsne-2D-One
NER
scibert-scivocab-uncased NER tagger (ADR)
biobert-v1.1-pubmed-base-cased NER tagger (ADR)
bert-base-uncased NER tagger (ADR)
Ensemble
No Extraction
YES
Is ADR?
All Tweets
Task 2 ADR Classifier (ENG)
Ensemble
ADR Span found?
NO
YES
QnA
bert-base-uncased NER tagger (DRUG)
Questions are constructed using the extracted DRUG names.
biobert-v1.1-pubmed-base-cased QA model
Prediction

(a) Each point in the scatter plot corresponds to the BERT embedding of an extract which is mapped to a MedDRA code.

(b) ADR extraction process flow.

Figure 1: Pipeline for extracting adverse mentions from tweets and visualizing the adverse mentions extracts in 2-D using T-SNE

- **Classifying tweets as containing ADR mentions:** We cleanse the tweets using the same preprocessing pipeline as mentioned in section 3.1, and use the RoBERTa classifier trained from section 3.1 to classify tweets as containing adverse reaction mentions.

- **Transformer based ensemble named entity recognition (NER) tagger:** We fine tuned SciBERT, BioBERT and BERT base to create an ensemble of NER taggers for extracting the ADR mention extract. Each tweet was tagged using the 'BIO' scheme, where 'B' denoted the start token of an extract, 'O' denoted tokens outside the extract, and 'I' denoted tokens inside the extract. We fine tuned each of the models in the ensemble for 5 epochs with a learning rate of 3e-5. Only the tweets that are classified as containing adverse drug reaction mentions from the previous stage are passed through the ensemble NER tagger to extract the adverse mentions.

- **Transformer based question answering system:** The tweets that were classified as containing adverse mentions, but did not yield any extracts from the previous ADR mention extraction NER stage were passed through the question answering stage for adverse reaction extraction. This stage is sub divided into the following two steps:
  - **NER tagger for drug detection**: We trained a BERT NER tagger for detecting drug names in a tweet. We tagged each tweet using the standard 'BIO' scheme to distinguish tokens containing drug names from other tokens, and trained the classifier for 5 epochs with a learning rate of 3e-5. We passed tweets through this tagger to extract the drug name and passed the drug name and tweet to the below step.
  - **BioBERT question answering**: Given the tweet and the drug name as context, we fine tuned BioBERT question answering on our training dataset to extract the adverse reaction to the drug. For example, after identifying the drug name (for example Tylenol) through the drug NER tagger, we constructed the question "What is the adverse effect of Tylenol?". Given the constructed question and the tweet as context, we fine tuned the BioBERT question answering model for 3 epochs with a learning rate of 5e-6, to extract the adverse reaction mention from the tweet.

3.4 Task 3-Automatic extraction and normalization of adverse effects in English tweets: Normalization

The essence of this task was to assign the most probable MedDRA code to the extracted adverse reaction mention from a tweet. The distribution of the number of training examples per MedDRA code followed a long tail distribution, which led to majority of the MedDRA codes having insufficient training examples.

We overcame this problem by enriching the training data set with additional CADEC (Karimi et al., 2015) and UMLS (Bodenreider, 2004) adverse reaction to MedDRA code mapping pairs. We ensured the number of examples are approximately 50 for each MedDRA code. We leveraged pre-trained fastText word embeddings to map the extracted adverse reactions to the most probable MedDRA code. We denoted each of the 475 MedDRA codes by a 300 dimensional vector, which was computed by mean pooling the fastText word embeddings of all the 50 adverse reaction extracts associated with the MedDRA code. For each adverse reaction mention extract, we mean pooled the 300 dimensional fastText embedding of the extract and the tweet in a heuristically determined proportion of 10:1 and represented it as a fixed 300 dimensional vector. Finally cosine similarity between the 300 dimensional extract vector and all the 300 dimensional vectors for MedDRA codes was performed to determine the closest MedDRA code for the extract. The extract normalization process can be formulated by the following formulas.
During training:

$$x_{extracts} = (x_{extract\_1}, x_{extract\_2}, ..., x_{extract\_n})$$

$$x_{extract\_i} = [x_1, x_2, ..., x_{300}]^T$$

$$x_{MedDRA\_code\_i} = 1/n * \sum_{x_{extract\_i} \, \epsilon \, x_{extracts}} x_{extract\_i} = [\hat{x_1}, \hat{x_2}, ..., \hat{x_{300}}]^T$$

$$\mathbf{X}_{MedDRA\_code} = [x_{MedDRA\_code\_1}, ..., x_{MedDRA\_code\_475}]$$

During validation/testing:

$$x_{extract\_embedding} = [p_1, p_2, ..., p_{300}]^T$$

$$x_{tweet\_embedding} = [q_1, q_2, ..., q_{300}]^T$$

$$x_{extract\_embedding\_contextual} = [(10p_1 + q_1)/2, ..., (10p_{300} + q_{300})/2]^T$$

$$closest\_MedDRA\_code = argmax \frac{x^T_{extract\_embedding\_contextual} \bullet \mathbf{X}_{MedDRA\_code}}{norm(x_{extract\_embedding\_contextual}) * norm(\mathbf{X}_{MedDRA\_code}, dim = 0)}$$

4 Results

Strict and relaxed F1 scores are used for evaluating the models. Under strict mode of evaluation, ADR spans are considered correct only if both start and end indices matches with the indices in the gold standard annotations. Under relaxed mode of evaluation, ADR spans are considered correct only if spans in predicted annotations overlapped with the gold standard annotations. In our system, this leads to significant differences between the two F1 scores. Our RoBERTa based English tweet classifier for task 2 outperforms most systems, and achieves a test F1 score of 0.56. With the multi staged adverse reaction extraction pipeline, we are able to achieve a relaxed F1 score of 0.576, which is above the average F1 of all the other submitted systems. Unfortunately, due to highly imbalanced training data, our French tweet classifier is not optimally trained, and fails to correctly identify any French tweets containing adverse reaction mentions.

Below we summarize the performance of our systems on all the different tasks. Table 2 summarizes our system's performance on the validation set. In table 3 we summarize our system's performance on the test set, and also show a comparison between our system's performance, and the average performance of all the systems in each of the tasks.

Task	**Strict**			**Relaxed**		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Task 2-English	0.648	0.630	0.668	-	-	-
Task 2-Russian	0.423	0.411	0.436	-	-	-
Task 3-Resolution(NER + Norm)	0.138	0.194	0.107	0.265	0.373	0.205

Table 2: Results on validation set.

Task	Strict			Relaxed		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Task 2-English	**0.56**	**0.50**	**0.63**	-	-	-
Task 2-English(Avg others)	0.46	0.42	0.59	-	-	-
Task 2-Russian	0.36	0.3350	0.3976	-	-	-
Task 2-Russian(Avg others)	0.427	0.362	0.583	-	-	-
Task 3-Detection(NER)	0.291	0.317	0.269	**0.576**	**0.614**	0.543
Task 3-Detection(NER)(Avg others)	-	-	-	0.564	0.607	0.557
Task 3-Resolution(NER + Norm)	0.157	0.171	0.145	0.216	0.235	0.200
Task 3-Resolution(NER + Norm)(Avg others)	-	-	-	0.292	0.312	0.29

Table 3: Results on test set, and comparison against arithmetic mean of best submissions made by other teams.

5 Discussion

As represented in Table 1, although most of the classification tasks had imbalanced training data the transformer models performed well, as they are more resilient to imbalanced classes compared to traditional machine learning models. For the French tweet classifier, the models learning capabilities were seriously hampered by the highly imbalanced data set, and re-sampling techniques did not help.

Adverse drug reaction extract normalization was a particularly challenging task. We experimented with several hierarchical recurrent neural network based architectures, transformer architectures and fixed embedding based similarity architectures. We finalised on using embedding based similarity architecture as the other architectures did not boost the score much. In order to understand more about the problem, we looked into the data closely and uncovered the following patterns.

5.1 Extracts with similar meaning mapped to distinct MedDRA codes

There are extracts which are very similar in meaning, yet mapped to different MedDRA codes. For example, the extracts ‘addiction’ and ‘addictive’ are very similar in meaning, but are mapped to MedDRA codes 10001125 and 10012336 respectively. To make our normalization algorithm resilient to such differences, we included the embedding of the tweet as context, along with the embedding of the extract while mapping the extract to the MedDRA code. As discussed in section 3.4, we heuristically determined a weight of 10:1 for pooling the extract and tweet embedding to generate more contextual vector representation of the extract. Figure 1a illustrates the problem of overlapping extract embeddings in 2-D using T-SNE. We can see the extracts forming clusters, which makes the MedDRA code mapping problem hard.

5.2 Potential training data set issues

Consider the tweets ‘addicted to nicotine badly’ and ‘... dante addicted to that nicotine’. In both the tweets ‘nicotine’ is tagged as the drug, and ‘addicted’ is the adverse reaction extract. Intuitively, both the tweets should map to the same MedDRA code. However in the training data, the first tweet maps to MedDRA code 10012336, which stands for ‘dependence addictive’, and the second tweet maps to MedDRA code 10001125, which stands for ‘addiction’. These examples hamper the learning capabilities of the models.

6 Conclusion

In this work, we have experimented with different transformer models for classifying ADR tweets as well as extracting ADR terms. We have leveraged several transformer based pre-trained models like RoBERTa, BioBERT etc. Also, we have devised a multi step pipeline for extracting the ADR terms from a tweet.

We can immediately think of two future improvements, firstly for the non English tasks in Task 2 we can develop a translation model to translate the tweets into English before classifying using the Task 2 English tweet classifier. Secondly, bettering the NER and MedDRA mapping, we want to incorporate a model that will be jointly trained to perform multiple tasks. For example, given a text, the model should be able to extract the ADR extracts as well as classify the tweet as ADR or non-ADR, as well as map it to the correct MedDRA code. Also, we will add a relationship extraction task, where we will identify the relation between the drug and ADR. We hypothesize that such a model should outperform a standard model as it will incorporate features and information sharing across tasks. For example, the NER would make a less false positive classification for non-ADR tweets.

References

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–D270, Jan. 14681409[pmid].

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (smm4h) shared tasks at coling 2020. *Proceedings of the Fifth Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task.*

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french.

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. *Bioinformatics*, Sep.

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.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model.

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, Florence, Italy, August. Association for Computational 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.