# Social Media Mining for Health Application (#SMM4H)

Sougata Saha, Souvik Das,Prashi Khurana
University at Buffalo – SUNY
sougatas@buffalo.edu, souvikda@buffalo.edu , prashikh@buffal.edu,

## Abstract

This paper presents techniques and models developed for the two tasks at SMM4H 2020. By fine-tuning BERT for task2 we were able to distinguish tweets that have an adverse reaction to the drug from those that do not. By using BERT NER tagger and BERT Q/A we were able to implement task3 which helped us to find the exact adverse drug reaction and then map it to the corresponding MedDRA Preferred Term (PT).

## 1 Introduction

In recent years, with the boom of social networks, the communication between companies and the user has improved. Especially in the context of health mining, social networks like Twitter can prove useful. Social networks can be used to raise awareness about existing drugs, new drugs, and health concerns. Social network analysis of the posts on Twitter can help in public monitoring of general health. In this project, we will analyze this data to focus on health monitoring. Since we will be dealing with data written by normal users, the data would not be clean. Hence pre-processing the data would be an important step.

We will be looking into Task2 and Task3 of the Social Media Mining for Health Applications (#SMM4H) Shared Task 2020.

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

This binary classification task involves distinguishing tweets that report an adverse effect (AE) of a medication (annotated as “1”) from those that do not (annotated as “0”), taking into account subtle linguistic variations between AEs and indications (i.e., the reason for using the medication). This classification task has been organized for every past #SMM4H Shared Task, but only for tweets posted in English. This year, this task also includes distinct sets of tweets posted in French and Russian.

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

This task is an end-to-end task that involves extracting the span of a text containing an adverse effect (AE) of medication from tweets that report an AE, and then mapping the extracted AE to a standard concept ID in the MedDRA vocabulary (preferred terms).

Since task2 and task3 are closely related, the paper presents an end-to-end system, which given a tweet, will identify the adverse effect, followed by getting the adverse drug reaction phrase, and then mapping the adverse drug reaction to the correct MedDRA Preferred Term (PT).

As per the Overview of the Fourth Social Media Mining for Health (#SMM4H), Shared Task at ACL 2019, the state of the art for task2 F1 score was 0.645, while for task3 was 0.432.

## 2 Related work

Since this is an emerging field of study there have been many previous breakthroughs that have been achieved in this field.

In Towards text processing pipelines to identify adverse drug events-related tweets: the University of Michigan @ SMM4H 2019 Task 1 trained two variations of neural network models — a bidirectional LSTM model, and a bidirectional LSTM model with a convolutional neural network (CNN) layer. They also compared the performance of both models using pre-trained GloVe word embedding and using pre-trained Word2vec Twitter word embedding. To generate the tweet representation for deep learning they have taken POS tag embedding, first character embedding, and created a one-hot representation of tweets using the medical dictionary SIDER. The best F1 score obtained from this method is 0.537.[1]

In Adverse Drug Reaction Classification With Deep Neural Networks using different neural network (NN) architectures for ADR classification. In particular, they proposed two new neural network models, Convolutional Recurrent Neural Network (CRNN) by concatenating convolutional neural networks with recurrent neural networks, and Convolutional Neural Network with Attention (CNNA) by adding attention weights into convolutional neural networks. On the Twitter dataset, all the NN architectures perform similarly. But on the ADE dataset, CNN performs better than other more complex CNN variants.[2]

In Entity-Level Classification of Adverse Drug Reaction: A Comparative Analysis of Neural Network Models used the concept of Aspect-level sentiment analysis, which aims to determine the sentimental class of a specific aspect conveyed in user opinions, has been actively studied for more than 10 years. They proposed a number of Neural Network architectures like (1) Target Dependent LSTM, (2) Interactive Attention Network which consists of two LSTM layers for the representation of the sentence and the target entity and of layers with cross attention the combined outputs of which are passed to a layer with the softmax function for making the classification decision, this method showed a F1 score of around 0.81 (different corpus) on the Twitter dataset. (3) The deep memory network (MemNet) , which repeatedly applies the attention mechanism to the input layer of word embeddings, and the output of the last attention layer is passed to the layer with the softmax function for making the classification decision. (4) The network with recurrent attention memory (RAM) extends the model MemNet by additional LSTM layers and multiple applications of the attention mechanism to the outputs of these layers. This method showed an F1 score of around 0.83 on the Twitter dataset (different corpus).[3]

In MSA: Jointly Detecting Drug Name and Adverse Drug Reaction Mentioning Tweets with Multi-Head Self-Attention proposed a neural approach with hierarchical tweet representation and multi-head self-attention mechanism to jointly detect tweets mentioning drug names and adverse drug reactions. In order to alleviate the influence of massive misspellings and user-created abbreviations in tweets, they proposed to use a hierarchical tweet representation model to first learn word representations from characters and then learn tweet representations from words. In addition, they proposed to use the multi-head self-attention mechanism to capture the interactions between words to fully model the contexts of tweets. The effectiveness of the attention mechanism is evident from the graph illustrated below.[4]

In Adverse drug reaction detection via a multihop self-attention mechanism they proposed a multihop self-attention mechanism (MSAM) model that aims to learn the multi-aspect semantic information for the ADR detection task. First, the contextual information of the sentence is captured by using the bidirectional long short-term memory (Bi-LSTM) model. Then, via applying the multiple steps of an attention mechanism, multiple semantic representations of a sentence are generated. Each attention step obtains a different attention distribution focusing on the different segments of the sentence. Meanwhile, their model locates and enhances various keywords from the multiple representations of a sentence. Their model achieved an F-measure of 0.853, 0.799, and 0.851 for ADR detection for TwiMed-PubMed, TwiMed-Twitter, and ADE, respectively.[5]

Ristad and Yianilos (1998) in Learning string-edit distance incorporated edit-distance when mapping similar texts. [6]

The most popular knowledge-based system for mapping texts to UMLS identifiers is MetaMap (Aronson, 2001). This linguistic-based system uses lexical lookup and variants by associating a score with phrases in a sentence. But this does not deal with numbers and also does not handle WSD. [7]

In 2016, Nut Limsopatham and Nigel Collier in Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation used semantic representation to get the vectors of the word and feed the vector to an RNN to get the best medical term. [8]

In 2017, Carson Tao, Kahyun Lee, Michele Filannino, Kevin Buchan, Kathy Lee, Tilak Arora, Joey Liu, Oladimeji Farri, Özlem Uzuner in Extracting and Normalizing Adverse Drug Reactions from Drug Labels, used a combination of MetaMap and Sub-Term Mapping Tools (STMT). The STMT is a generic toolset designed to find all sub-terms in a specific corpus and map synonymic variations of the corpus to UMLS concepts[14]. [9]

In 2018, Sarker et al. and other authors, in (Data and systems for medication-related text Twitter: classification and concept normalization from insights from the Social Media Mining for Health (SMM4H)-2017 shared task). They used a multinomial logistic regression model, 3 variants of recurrent neural networks (RNNs), and an ensemble of the 2 types of models. But they were not able to beat the state of the art. [10]

In 2019, Mert Tiftikci, Arzucan Özgür, Yongqun He, and Junguk Hur Corresponding (Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels) implemented an integrated deep learning and dictionary/rule-based approach for entity detection and normalization. The machine learning technique achieved 82.6% micro-averaged F1 score while the rule-based system achieved and 77.6% F1 scores.[11]

In 2019, Zulfat Miftahutdinov and Elena Tutubalina in Deep Neural Models for Medical Concept Normalization in User-Generated Texts studied the use of deep neural models, i.e., contextualized word representation model BERT (Devlin et al., 2018) and Gated Recurrent Units

(GRU) (Cho et al., 2014) with an attention mechanism, paired with word2vec word embeddings and contextualized Elmo embeddings (Peters et al., 2018). With the SMM4H dataset using BERT w/ TF-IDF (MAX) they were able to achieve an accuracy of 89.64. [12]

## 3 Data Analysis and Augmentation

Below is the data distribution for the tasks:

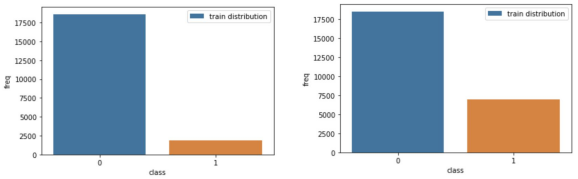
| Task2 | Total | Class 1 | Class 0 |
|---|---|---|---|
| **English** | | | |
| Training | 20,544 | 1,903 | 18,641 |
| Validation | 5,134 | 474 | 4660 |
| **Russian** | | | |
| Training | 6,090 | 533 | 5,557 |
| Validation | 1,522 | 133 | 1,389 |
| **French** | | | |
| Training | 1,941 | 31 | 1,910 |
| Validation | 485 | 8 | 477 |

| Task3 | Total | ADR | No-ADR |
|---|---|---|---|
| Training | 2,246 | 1,464 | 782 |
| Validation | 560 | 365 | 195 |

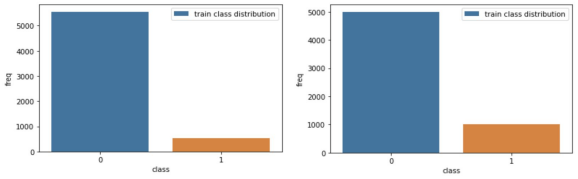
But the datasets given for task2 and task3 are highly imbalanced. Thus we need to augment the dataset for both task2 and task3.

**Task2**: Automatic classification of multilingual tweets that report the adverse effect

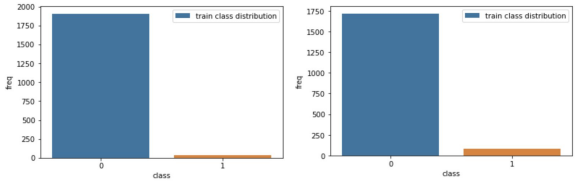
For Fine Tuning BERT we used this year's tweets(20,000 training and 5,000 dev set). Besides that, we also combined tweets from task 3(2,200 training and 5,60 dev). We combined all the tweets and used the 22,760 data points as training data and 5,000 as dev data.
The class distribution of ADR and non-ADR tweets in the training data was 10% and 90% respectively. We up-sampled the minority class of the training dataset to 25% from 10% by randomly duplicating tweets.




**Class distribution of training data before and after upsampling(English)**




**Class distribution of training data before and after upsampling(Russian)**

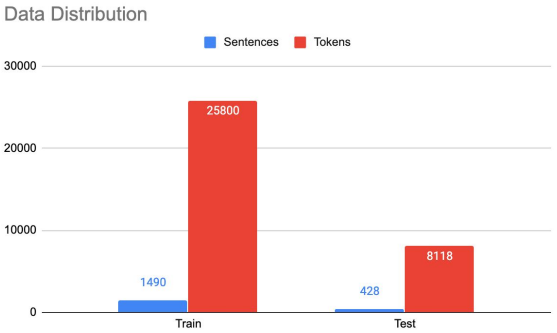



**Class distribution of training data before and after upsampling(French)**

**Task3**:
There are two parts to the task:

<u>Part1: Using the BERT NER tagger and BERT Q/A to get the ADR from the tweets.</u>

For fine-tuning the transformer NER models we have tokenized all the tweets present in the training and evaluation dataset based on a BIO encoding scheme, the distribution of the tokens is visualized as follows:




**Token Distribution of for task 3 part 1**

Although there were 2,200, 560 sentences in the training and evaluation dataset respectively, we

have reduced it to 1490 and 560 sentences respectively. This is because many tweets have multiple ADR mentions. We basically created a unique set of tweets so that it is easier for the NER tagger to produce results.

| | tokens | labels |
|---|---|---|
| 0 | [my, nigga, dante, addicted, to, that, nicotine] | [O, O, O, B, O, O, O] |
| 1 | [i, feel, soo, much, better, today,, cymbalta,... | [O, O, O, O, O, O, O, B, I, O, O, O, O] |
| 2 | [@theotherrift, it, sort, of, can., :(, you, t... | [O, O, O, O, O, O, O, O, O, O, O, O, O, O, ... |
| 3 | [@sumiyyahiqbal, @shahbaigg, difference, is, i... | [O, O, O, O, O, O, O, O, O, O, O, O, O, O, O] |
| 4 | [rt, @fightforfood:, what, i, lack, in, money,... | [O, O, O, O, O, O, O, O, O, O, O, O, O, O, ... |

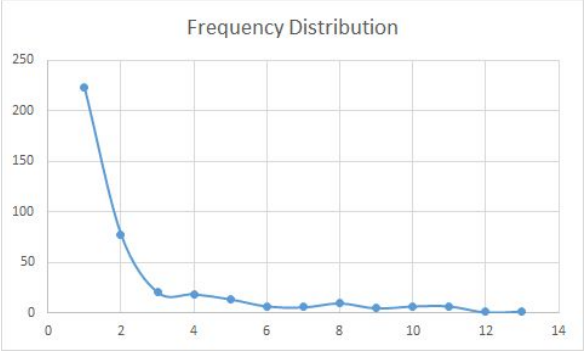
**BIO Tagging**

Part2: Mapping the ADR extract to the correct MedDRA term.

*MedDRA Codes in the Dataset*:
The training dataset for this year contained 417 classes but not all classes were in the validation dataset. So to expand the MedDRA codes, we collected the classes from all previous year datasets and this year's dataset giving us a list of 475 unique MedDRA Id's.

Initially, the extracts in the training set per class in the training dataset are as below.

While some meddra_id classes have more than 20 examples, others have just one.




**Frequency Distribution of examples and MedDRA Id**

More than 50% of the MedDRA id's have 1 extract example per class only.

We need to augment the dataset. Examples for the different MedDRA terms were extracted from the *UMLS API, CADEC*, and the user-generated text from *last year's dataset*. (SMM4H 2019 task3 dataset).

For example, MedDRA Id: 10047522. Let us see the extracts and tweets for this meddra id.

Extracts in the training set:

Extract: Lost Vision

Tweet: #cymbalta withdrawal has reached a peak, lost vision, and almost crashed my car from a brain zap. thanks a zillion #elililly #bigpharma

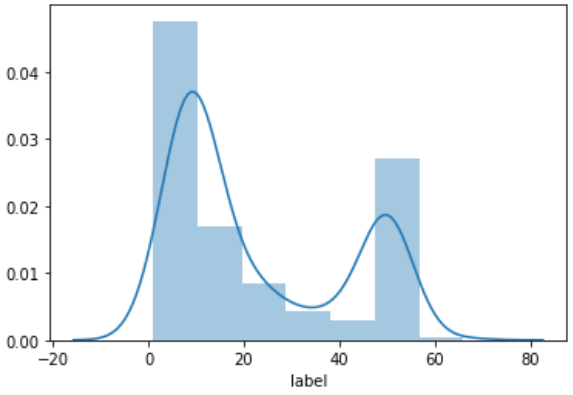
UMLS Terms:

loss vision things occur', 'episodic loss vision', 'progressive loss peripheral vision onset first decade', etc

CADEC Terms:

Extract: Sudden loss of vision
Tweet: Very blurred vision and lower back pain. I had a sudden loss of vision and attributed it to getting older. After reading literature on another medication a light bulb went off and I realized the blurry vision was probably due to the drug Lipitor and am stopping the drop today. My vision is so blurry it feels like I need glasses.

Frequency distribution of the dataset is after augmentation:

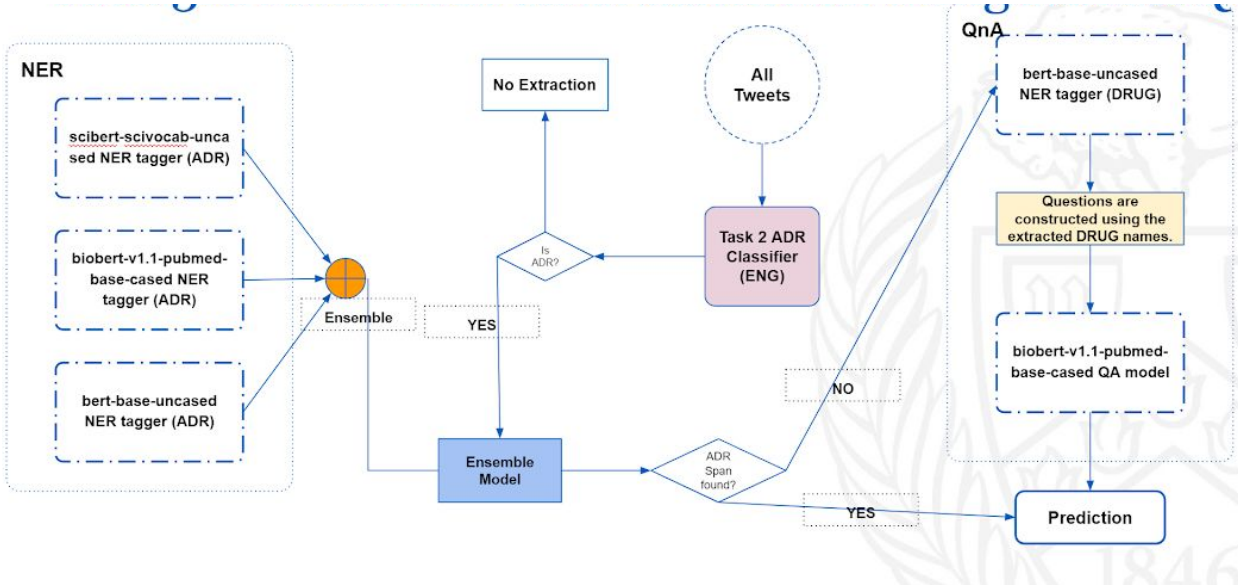



**Frequency Distribution of examples and Meddra Id**

## 4 Process

**Pre-Processing**:
Social media data is always noisy, hence we had to cleanse the data before we could ingest it. We cleansed the tweets using Ekaphrasis, regex, and NLTK. The cleansing involved the following:

1. Lowercase
2. Normalize 'email', 'money', 'phone', 'time', 'date'.
3. Normalize elongated and repeated characters
4. Normalize hashtags
5. Remove special tokens like 'rt'
6. Unpack contractions
7. Remove URL, mentions, smileys & emojis
8. Convert data to ASCII
9. Remove duplicates
10. Remove tweets from the training set that are in the validation set




**Flow of Task3 Part 1**

## 5 Models

**Task2**:

English:

We have experimented with many transformer models. For our final pipeline, we have implemented an ensemble of roBERTa large, bert base uncased, sciBERT with scivocab, and bioBERT base v1.1. For roBERTa, we have pooled the last 6 hidden layers and then used a linear layer to perform the final classification. We achieve an ensemble F1 of 0.65 and a F1 Of 0.66 with roBERTa large alone. We intend to submit our predictions using the ensemble as well as roBERTa large to the final SMM4H 2020 leaderboard and keep the one that performs best on the test data.

Russian and French:

For the Russian and French tweets, we have used ruBERT and camemBERT respectively. For bothe of them, we have applied a similar pooling strategy as roBERTa for English tweet classification. Below are our results for the same.

| Language | Model | Method | F1 |
|---|---|---|---|
| Russian | ruBERT | Model trained on Russian Text | 0.42 |
| French | camemBERT | Model trained on French Text | 0.22 |

**Flow of Task3 Part 1**

Since our English tweet classifier performs better than the Russian and French classifiers, in the future we intend to perform language translation on the multilingual tweets and then perform classification using the English language classifier.

**Task3**:

Part1: Using Transformer Based Ensemble NER tagger and bio best Q/A to get the ADR from the tweets.

This part can be divided into 3 modules:

I. ADR classification.
II. Transformer based Ensemble NER tagger.
III. BioBERT based question answering system.

I. ADR Classification: As the task 2's ADR classifier is able to classify the ADR tweets with decent accuracy, we passed all the evaluation tweets through the task 2 ADR classifier and ran NER of the ADR classified tweets. This reduced the chances of getting false positives. However, in our final run before submitting to SMM4H we might need to re-think this strategy based on the data.

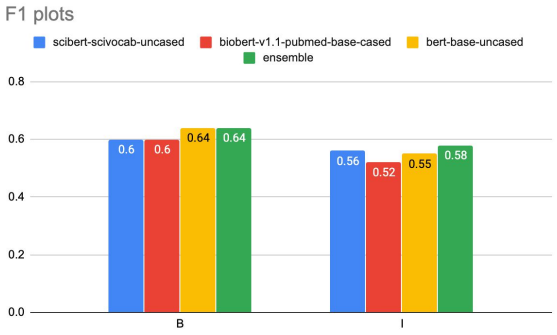
II. Transformer based Ensemble NER tagger:Here, we ensembled 3 transformer-based models:
   A. Scibert-scivocab-uncased.
   B. Biobert-v1.1-pubmed-base-cased.
   C. Bert-base-uncased.

We fine-tuned the above models with the tokenized training data and created an ensemble mechanism. The ADR tweets

from the step I is passed through this ensemble system to get the ADR extraction terms. The tweets in which this NER system failed is passed through the question answering system.

**Model Params**

max_seq_length = 128
epochs = 4
validation_fraction = 0.1
learning_rate=3e-5,
train_batch_size=16,
eval_batch_size=16




**Individual NER model results**

III. Bio-BERT based question answering system: In the final filtering step, we tried to find a relationship between the drug and ADR mentions. For that, we leveraged the bio-BERT based question answering system. The following steps are followed:

A. A drug NER tagger is built using the previous strategy but here we used the bert-base-uncased model. An F1 score of 0.92 was observed.
B. For a given tweet we extracted out the drug mention and constructed a question like this: **`Which is the adverse effect of <drug_name>?`**
C. The generated question along with the tweet as a context is sent to the QA model to get the prediction of the ADRs and the spans.

*For training*: Using the training data we fine-tuned the bio-BERT QA model to better our predictions.

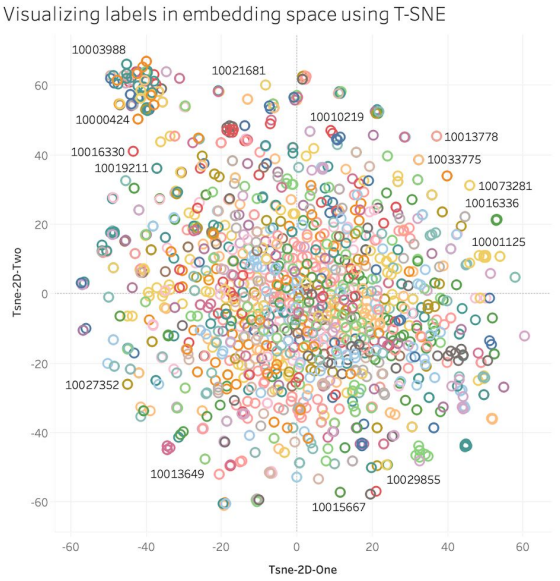
**Training Process**

- We used the standard Bio-BERT cli to fine tune the model using our training data
- `! python biobert/run_qa.py --do_train=True --do_predict=True --vocab_file=$BIOBERT_DIR/vocab.txt --bert_config_file=$BIOBERT_DIR/bert_config.json --init_checkpoint=$BIOBERT_DIR/biobert_model.ckpt --max_seq_length=384 --train_batch_size=6 --learning_rate=5e-6 --doc_stride=128 --num_train_epochs=1.0 --do_lower_case=False --train_file=$QA_DIR/train_data_smm4h.json --predict_file=$QA_DIR/test_data_smm4h.json --output_dir=$OUTPUT_DIR`

**QnA model training params**

Part 2: Mapping the ADR extract to the correct MedDRA term.

Considering the growth in the field of NLP, we have some widely and commonly used embeddings available. The embeddings we experiment are BERT(trained on the tweets in the dataset, pre-trained), BIO-BERT, FAST-TEXT(trained on the tweets in the dataset, pre-trained).



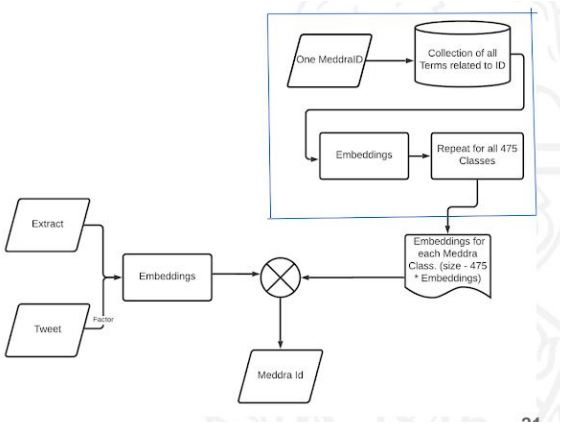

**Embeddings of extracts in the BERT Embedding space.**

We are using an ensemble of cosine similarity, hierarchical GRU, SVM, and logistic regression.

| Model | Dataset | Accuracy | Macro F1 |
|---|---|---|---|
| Ensemble | Validation | 44.93 | 38.93 |
| Ensemble | Training | 57.85 | 52.8 |

**Result of Ensemble**

Model 1: Cosine Similarity




**Flow of cosine similarity algorithm**

| Model | Accuracy | Macro F1 |
|---|---|---|
| Char2vec | 30.05 | 19.52 |

| | | |
|---|---|---|
| BioBert | 26.02 | 21.55 |
| **FastText** | **47.75** | **38.93** |

**Cosine Similarity with different embeddings**

Model 2: SVC and Logistic Regression on FastText Embeddings

| **Model** | **Accuracy** | **Macro F1** |
|---|---|---|
| SVC | 45.75 | 32.30 |
| Logistic Regression | 46.30 | 34.63 |

**Results of SVC and Logistic Regression on FastText Embeddings**

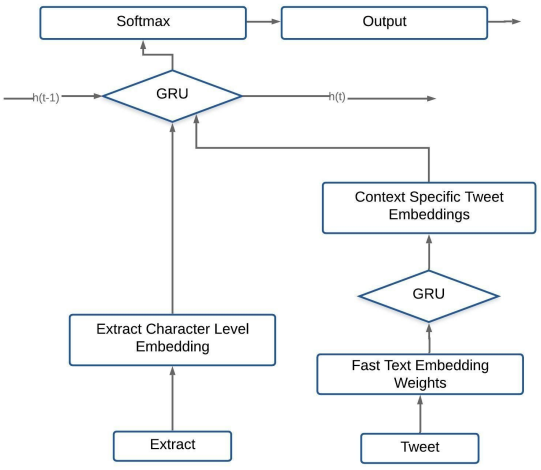
But as we can see there are overlaps in the embedding space, we need to consider the context of the tweet also. Consider the example given below. We see the extracts are almost the same, but their usage in the tweets is different and hence they map to the same meddra Id.

| **Extract** | **MeddraId** | **Tweet** |
|---|---|---|
| Addiction | 10001125 | also, yay addiction. it's drugs i need for living, but it's still dependance. every time i get a "paxil headache" i realize this. oh well. |
| Addictive | 10012336 | rt @silkius: @ouch_uk didnt know lamotrigine was addictive stopped as didnt think were helping @clusterheads 3 days of hell before realized? |

**Example of similar extract in different meddra Id.**

Hence we can come up with a hierarchical GRU which takes the context of the tweet in consideration.

Model 3: Hierarchical gru




**The flow of Hierarchical GRU.**

| **Model** | **Accuracy** | **Macro F1** |
|---|---|---|
| Self Trained Fast Text | 45.47 | 31.63 |

**Results of Hierarchical GRU on self-trained fastText Embeddings**

## 6 Results

The individual results from the tasks are as follows:

| **Task2** | **F1 Score** |
|---|---|
| English | 0.66 |
| Russian | 0.42 |
| French | 0.22 |
| **Task3** | **Macro F1** |
| Part1: Using BERT NER tagger and BERT Q/A to get the ADR from the tweets. | 0.72 |
| Part 2: Mapping the ADR extract to the correct MEDDRA term. | 0.39 |

**Results**

| | Relaxed | | | Strict | | |
|---|---|---|---|---|---|---|
| **Team** | **F1** | **P** | **R** | **F1** | **P** | **R** |
| KFU NLP | **0.432** | 0.362 | **0.535** | **0.344** | 0.288 | **0.427** |
| myTomorrows-TUDelft | 0.345 | 0.336 | 0.355 | 0.244 | 0.237 | 0.252 |
| TMRLeiden | 0.312 | **0.37** | 0.27 | 0.25 | **0.296** | 0.216 |
| GMU | 0.208 | 0.221 | 0.196 | 0.109 | 0.116 | 0.102 |

Last Year's result from SMM4H 2019 tasks

| Model | Macro F1 Score |
|---|---|
| End to End | 0.262 |

**Results of an End To End Workflow**

## 7 Future Work

In the future, we intend to experiment with the following things before submitting to the SMM4H 2020 leaderboard.

1. For Russian and French tweet classification, we want to experiment with an ensemble of models. Specifically, we want to develop a translation model, which will enable us to translate the tweet from one language to English, and then make a prediction on the English translated tweet.

2. For 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.

3. For better predictions from the Meddra Mapper, we will train a model on the CADEC dataset. Then fine-tune it with our tweet training dataset.

## 8 Contributions

| Team Member | Task |
|---|---|
| Sougata | Task2 (both parts) |
| Souvik | Task3 (part 1) |
| Prashi | Task3 (part 2) |

## 9 References

[1]Vydiswaran, V.G.Vinod & Ganzel, Grace & Romas, Bryan & Yu, Deahan & Austin, Amy & Bhomia, Neha & Chan, Socheatha & Hall, Stephanie & Le, Van & Miller, Aaron & Oduyebo, Olawunmi & Song, Aulia & Sondhi, Radhika & Teng, Danny & Tseng, Hao & Vuong, Kim & Zimmerman, Stephanie. (2019). Towards Text Processing Pipelines to Identify Adverse Drug Events-related Tweets: University of Michigan @ SMM4H 2019 Task 1. 107-109. 10.18653/v1/W19-3217.

[2]Huynh T, He Y, Willis A, Uger S. Adverse Drug Reaction Classification With Deep Neural Networks. In: COLING. Osaka: 2016: 877-87.

[3]Alimova, I. & Tutubalina, Elena. (2019). Entity-Level Classification of Adverse Drug Reaction: A Comparative Analysis of Neural Network Models. Programming and Computer Software. 45. 439-447. 10.1134/S0361768819080024.

[4]Wu, Chuhan. (2019). MSA: Jointly Detecting Drug Name and Adverse Drug Reaction Mentioning Tweets with Multi-Head Self-Attention. 10.1145/3289600.3290980.

[5]Zhang, T., Lin, H., Ren, Y. et al. Adverse drug reaction detection via a multihop self-attention mechanism. BMC Bioinformatics 20, 479 (2019).

[6] Ristad, E.S. and Yianilos, P.N. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20: 522--532, 1998.

[7] Aronson,A.R. ( 2001 ) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc. AMIA Symp. , 17 –21.

[8] N. Limsopatham and N. Collier. Normalizing medical concepts in social media texts by learning semantic representation. In Proceedings of the 54th

[9] Extracting and Normalizing ADRs from Drug Labels C Tao, K Lee, M Filannino, K Buchan, K Lee, T Arora, J Liu, O Farri, …2017 Text Analysis Conference, Gaithersburg, MD

[10] Sarker A, Belousov M, Friedrichs J, et al. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. J Am Med

Inform Assoc. 2018;25(10):1274-1283. doi:10.1093/jamia/ocy114

[11] Tiftikci M, Özgür A, He Y, Hur J: Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels. BMC Bioinformatics (VDOS 2018).

[12] Zulfat Miftahutdinov and Elena Tutubalina. Deep neural models for medical concept normalization in user-generated texts. arXiv preprint arXiv:1907.07972, 2019.