CSE 635

Team Naam-pAI
Analyzing social media posts for drug adverse reaction mentions

University at Buffalo
School of Engineering and Applied Sciences

Introduction

- The Social Media Mining for Health Applications (#SMM4H) Shared Task involves natural language processing challenges of using social media data for health research, including informal, colloquial expressions and misspellings of clinical concepts, noise, data sparsity, ambiguity, and multilingual posts.

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

SMM4H 2020 details

- The Social Media Mining for Health Applications (#SMM4H) workshop serves as a venue for bringing together researchers interested in automatic methods for the collection, extraction, representation, analysis, and validation of social media data (e.g., Twitter, Facebook) for health informatics.

- The 5th #SMM4H Workshop, co-located at COLING 2020

- Submission(extended) deadline: 1 July 2020

- Workshop dates: 12-13 December 2020

- Location: Barcelona, Spain

Task 2

Automatic classification of multilingual tweet that report adverse effects

Task 2: Multilingual Tweet Classification

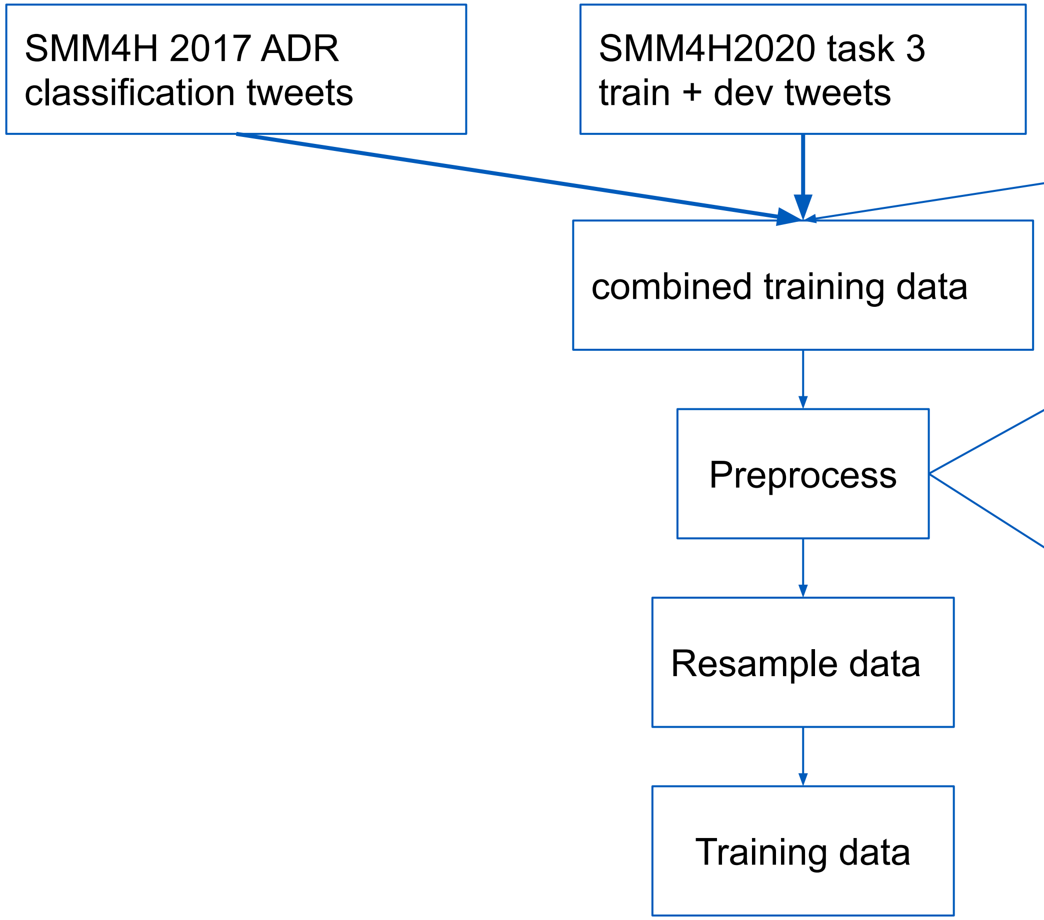
- The task includes binary classification of tweets that report adverse drug reactions of medication.
- Example dataset:

tweet_id	user_id	class	tweet
349220537903489025	323112996	0	@jessicama20045 right, but cipro can make things much worse...and why give bayer more of your money? they already screwed you once w/ essure
491775200610893825	2484689840	1	5. so what caused the #estrogen surges in #nuvaring ? did any one wonder why? do you know that a surge in #estrogen can cause a #bloodclot?

tweet_id	user_id	class	tweet
1163750273015263239	263170463	0	@policedutweet ça rappelle les innombrables mentions de sponsors le soir quand t'attends ton émission à la télé... "et parce qu'une diarrhée n'arrive jamais au bon moment, le Conseil des Ministres ce soir vous est présenté par Imodium..." 🤨
1128724741236432899	1321638810	1	@armance64 @DocNebulleuseP @doctoctocbot Pourquoi? Perso je le préfère au tramadol qui pour moi est un truc vraiment merdique qui en plus fait criser. Ou j'utilise la codeine.

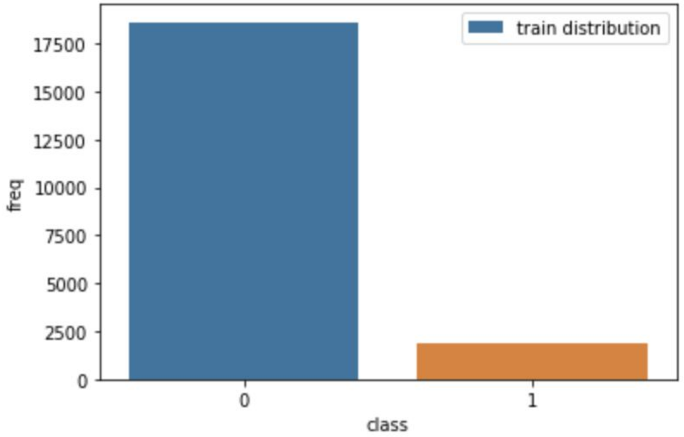
	tweet_id	tweet	class
0	760402871867367424	Настало время для ингаляторов. Дружок, Сальбутамол, где ты?	0
1	1035908416869462016	15) На прошлой зимней олимпиаде большинство лыжников приехало со справкой о том что у них якобы астма. Сделано это было для того, чтобы легально принимать сальбутамол (то же что и я принимаю в ингаляторах) который расширяет объём легких. По сути допинг для здорового человека.	1

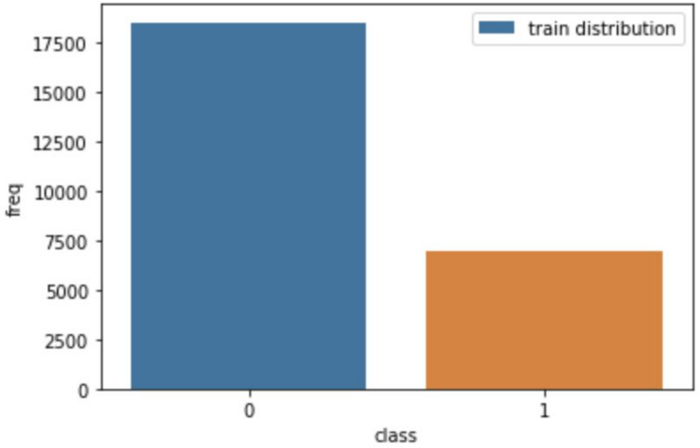
Task 2A(EN): Data Preprocessing pipeline

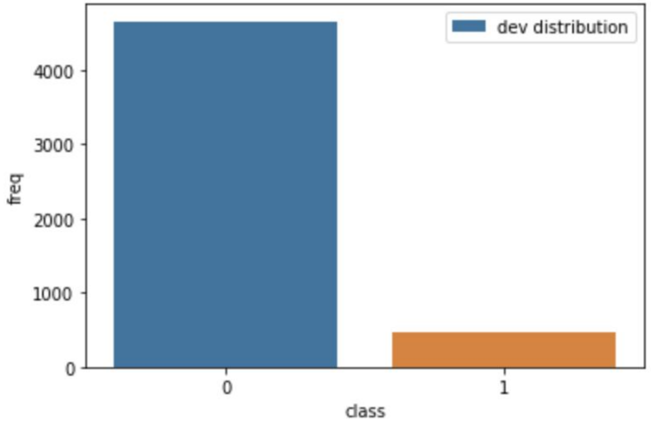


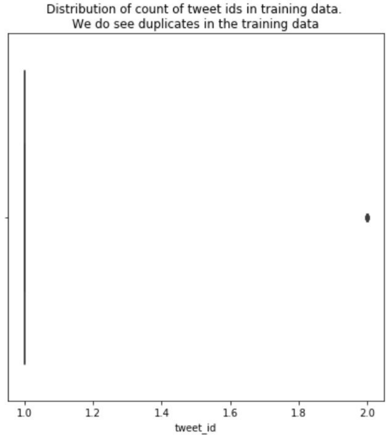
- Lowercase
- Normalize 'email', 'money', 'phone', 'time', 'date'.
- Normalize elongated and repeated characters
- Normalize hashtags
- Remove special tokens like ‘rt’
- Unpack contractions
- Remove url, mentions, smileys & emojis
- Convert data to ascii
- Remove duplicates
- Remove tweets from training set that are in validation set

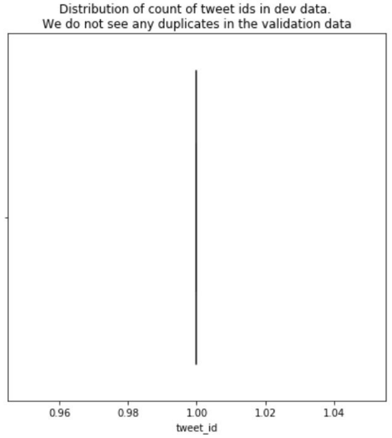
Task 2A(EN): Data distributions

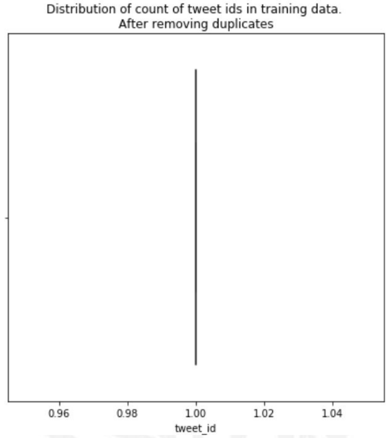












Task 2A(EN): Tweet Classification - Modelling

- We have implemented an ensemble approach for classifying a tweet as containing ADR mentions.
- Our ensemble contains of the following models:
  - roBERTa large
    - We pool the last 6 hidden layers and pass it through a linear layer to make the final class prediction. We also apply dropout before passing through the linear layer.
    - Optimizer: AdamW with weight decay of 0.01 on weights
    - Max gradient norm = 1
  - BERT base uncased
    - Standard Classification head. No pooling
  - sciBERT base with sci-vocab
    - Standard Classification head. No pooling
  - bioBERT v1.1 base
    - Standard Classification head. No pooling
- Universal parameters:
  - Learning rate: 2e-5
  - epochs: 4

Task 2A(EN): Tweet Classification - Results

Model	**F1**
Roberta Large	**0.66**
Bert Base	0.60
SciBert	0.58
BioBert	0.57
Ensemble	**0.65**

Team	F1	P	R
ICRC	**0.6457**	0.6079	0.6885
UZH	0.6048	0.6478	0.5671
MIDAS@IIITD	0.5988	0.6647	0.5447
KFU NLP	0.5738	**0.6914**	0.4904
CLaC	0.5738	0.5427	0.6086
THU_NGN	0.5718	0.4667	**0.738**
BigODM	0.5514	0.4762	0.655
UMich-NLP4Health	0.5369	0.5654	0.5112
TMRLeiden	0.5327	0.6419	0.4553
CIC-NLP	0.5209	0.6203	0.4489
UChicagoCompLx	0.4993	0.4574	0.5495
SINAI	0.4969	0.5517	0.4521
nlp-uned	0.4723	0.5244	0.4297
ASU BioNLP	0.4317	0.3223	0.6534
Klick Health	0.4099	0.5824	0.3163
GMU	0.3587	0.4526	0.2971

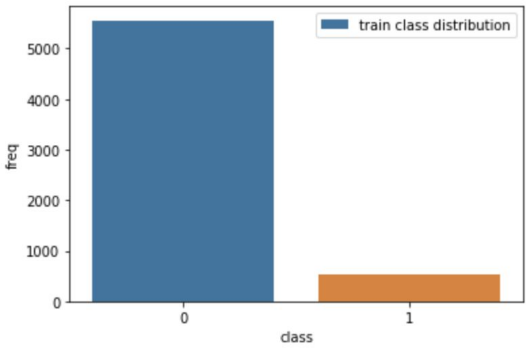
SMM4H 2019 leaderboard

Task 2B: Tweet Classification - Russian & French

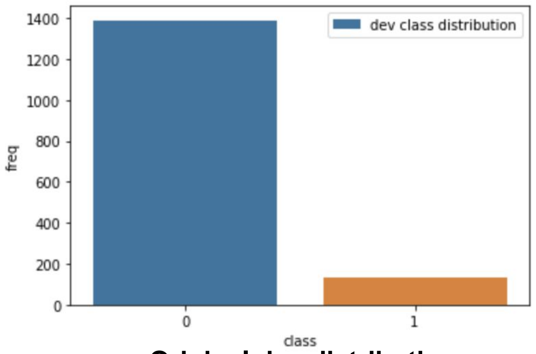
For the Russian and French tweet classification, we have performed the following preprocessing

- Normalize 'email', 'money', 'phone', 'time', 'date'.
- Remove special tokens like ‘rt’
- Remove url, mentions, smileys & emojis
- Remove duplicates

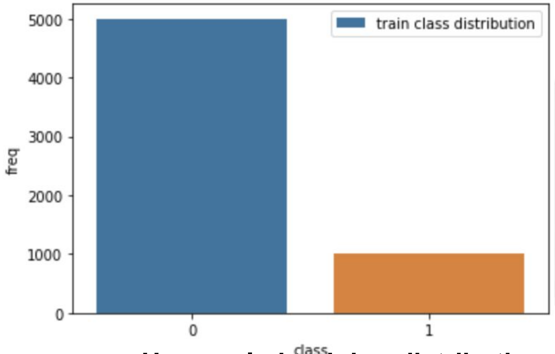
Russian tweets



Original training distibution

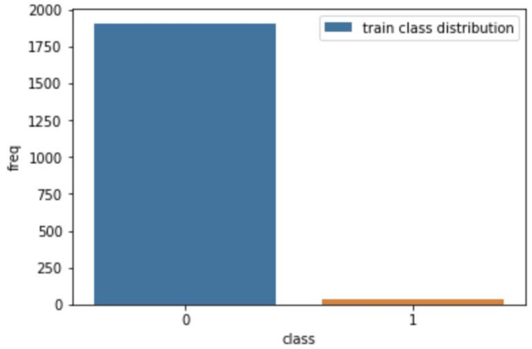


Original dev distribution

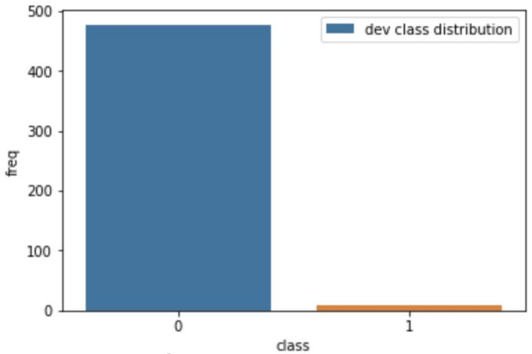


Upsampled training distribution

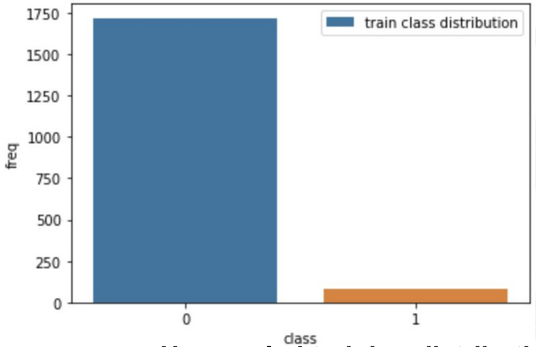
French tweets



Original training distibution



Original dev distribution



Upsampled training distribution

Task 2B: Tweet Classification - Modelling

Russian tweets:

- For classifying Russian tweets we use ruBERT base: A BERT model pre trained on russian text by deepPavlov(http://docs.deeppavlov.ai/en/master/features/models/bert.html)
- We pool all the hidden layers and concatenate the final layer [CLS] hidden representation
- We pass the hidden representation through a linear layer which performs the final classification
- We achieve a **test F1** score of **0.42**

French tweets:

- The data for French tweet is very low and highly imbalanced.
- Out of the 1746 training data, only 28 belong to ADR class, and only 3 belong to ADR class out of the 195 dev data.
- We fine tuned camemBERT base and achieved a **test F1** score of **0.22**

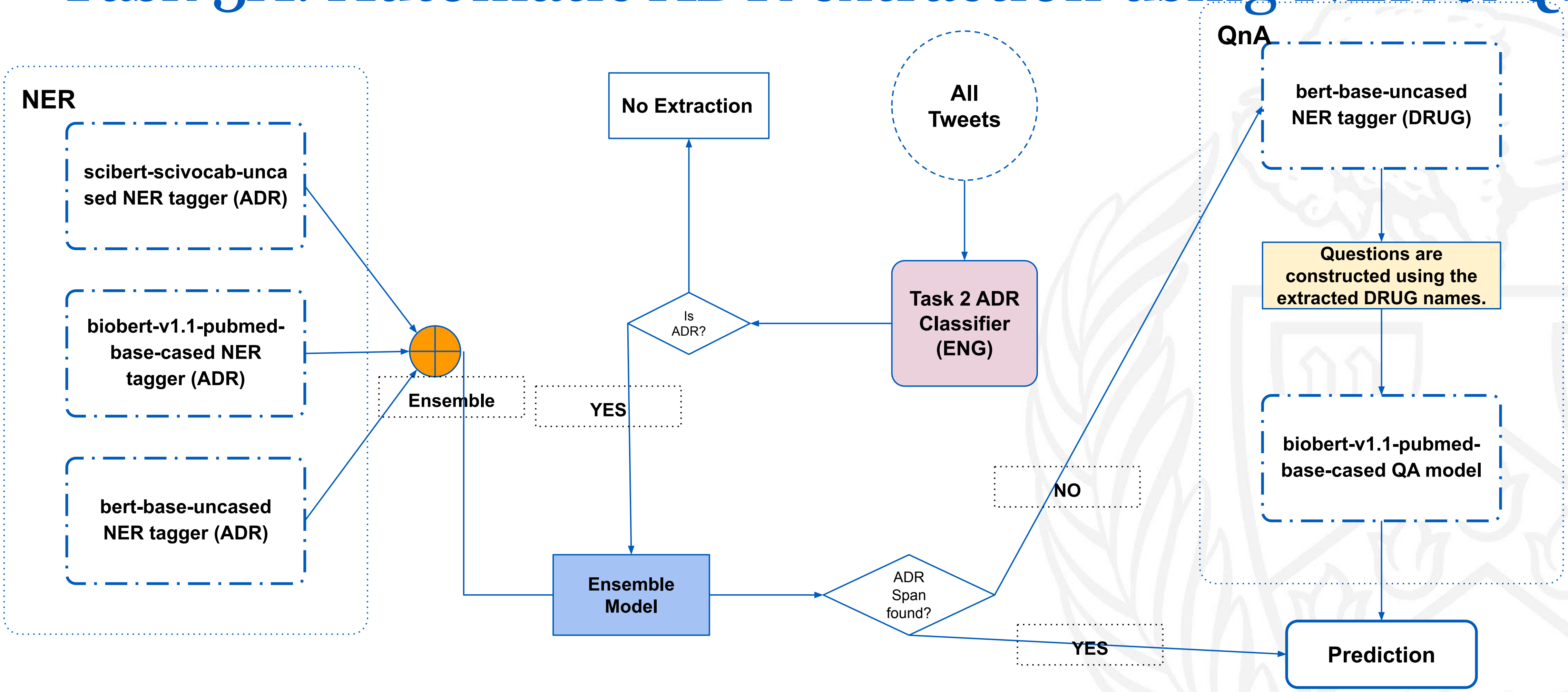
Note:

- Since these multilingual tasks are introduced this year in SMM4H, we are unable to provide a comparison study.
- We propose betterment to the existing classification techniques in the future work section

Task 3

Automatic extraction and normalization of adverse effects in English tweets

Task 3A: Automatic ADR extraction using NER & QnA



Task 3a: Training the NER Taggers

Data Preparation:

- Tweets are cleaned using the data preprocessor used in task 2.
- Then the tweet ADR are labeled using a BIO encoding scheme.
- Then the data is used to fine tune **bert-base-uncased, scibert-scivocab-uncased** and **biobert-v1.1-pubmed-base-cased** models.
- Fine tuning is performed using a library called BERT-sklearn.

Model Params

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

	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, ...

Task 3a: QnA Training process

Drug NER

- A similar strategy is followed like ADR NER tagger training, here drug names are only BIO encoded.
- We are using **bert-base-uncased** model here with the same model params like ADR NER tagger.

Question Construction and Dataset generation

- Using the drug name extracted from the NER tagger we constructed sentences in the following manner:

Which is the adverse effect of <drug_name>?

- Then a QnA dataset in constructed where the context of the question is the tweet text


```
{
    "data": [{
        "paragraphs": [{
            "context": "do you have any medication allergies? \"asthma!!!\" me: \"........\" pt: \"no wait. avel
            "qas": [{
                "id": 1,
                "question": "What is the adverse effect of avelox?"
            }]
        }, {
```


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
  ```
  ```

Task 3a: Results - Classification

Tweets classification ADR/ non-ADR ensemble

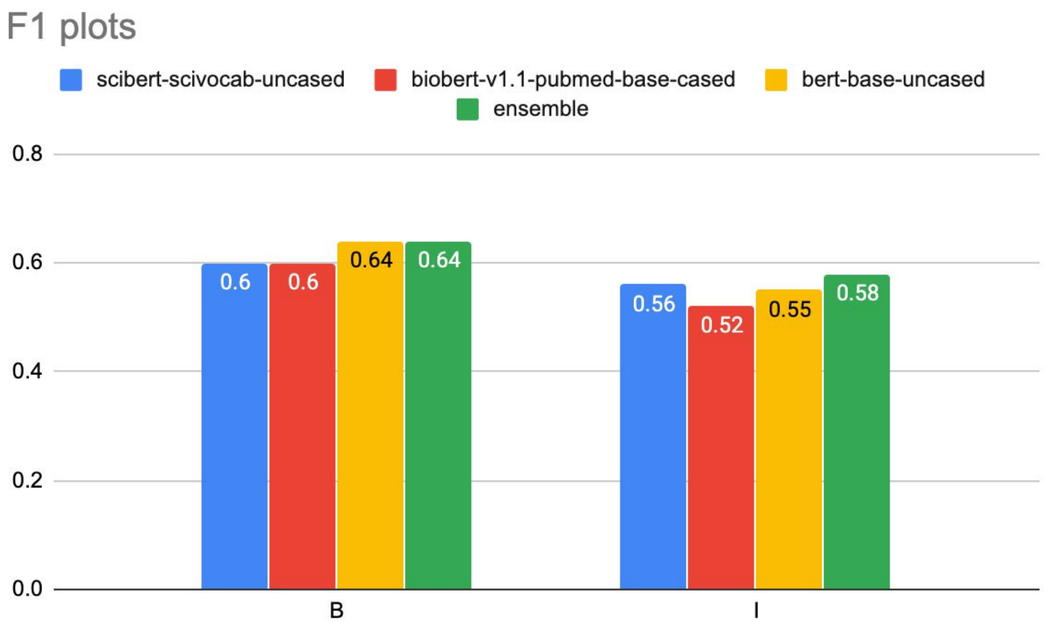
F1	0.846335
Accuracy	0.84813
Precision	0.9421
Recall	0.76824

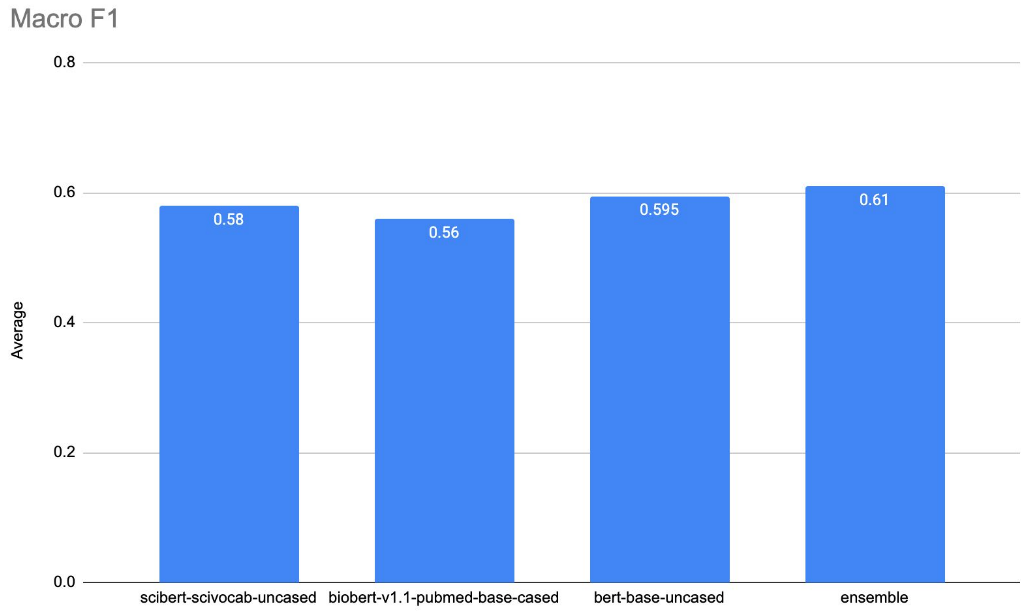
Total validation tweets = 428

ADR classified tweets = 190

No ADR classified tweets = 238

Task 3a: Result - Individual Models





Task 3b: Result - Overall(Combining QnA)

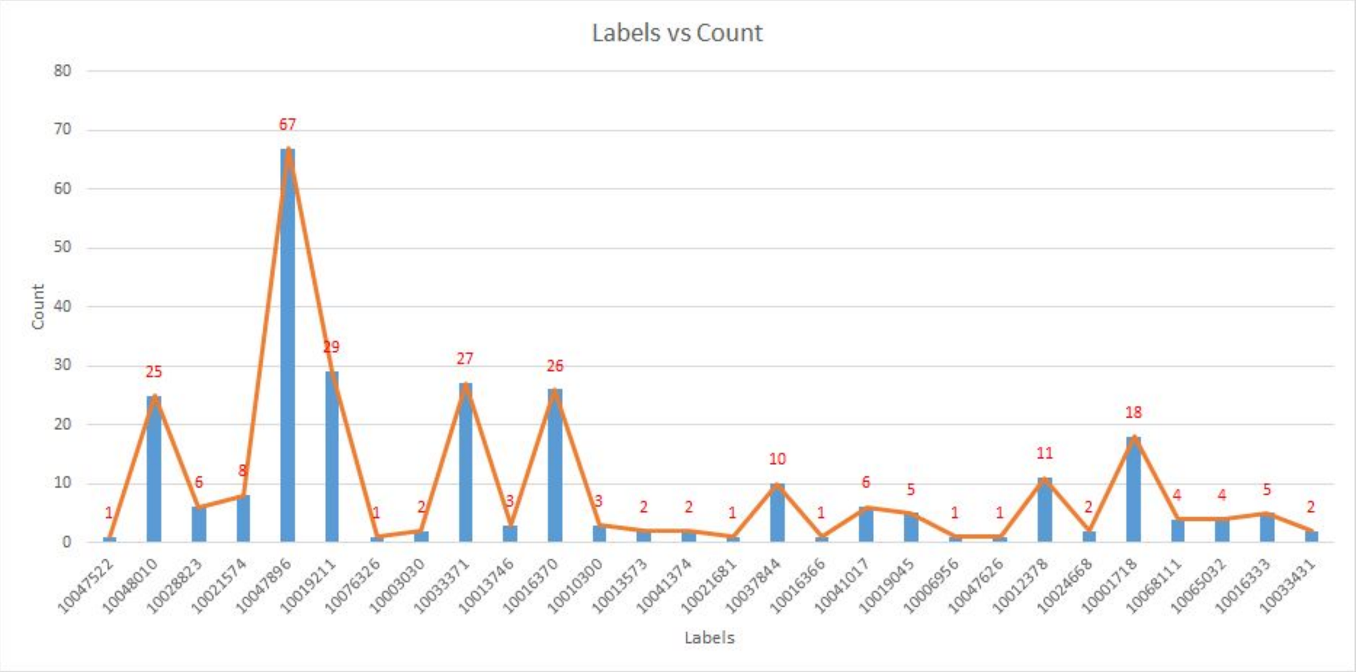
Estimated precision, recall and F1(Strict)

	Final(Strict)	Baseline(BiLSTM-CRF)
F1	0.7184	0.334

Previous Year's Leaderboard

	Relaxed			Strict		
Team	**F1**	**P**	**R**	**F1**	**P**	**R**
KFU NLP	**0.658**	0.554	**0.81**	**0.464**	0.389	**0.576**
THU_NGN	0.653	0.614	0.697	0.356	0.328	0.388
MIDAS@IIITD	0.641	0.537	0.793	0.328	0.274	0.409
TMRLeiden	0.625	0.555	0.715	0.431	0.381	0.495
ICRC	0.614	0.538	0.716	0.407	0.357	0.474
GMU	0.597	0.596	0.599	0.407	0.406	0.407
HealthNLP	0.574	**0.632**	0.527	0.336	0.37	0.307
SINAI	0.542	0.612	0.486	0.36	**0.408**	0.322
ASU BioNLP	0.535	0.415	0.753	0.269	0.206	0.39
Klick Health	0.396	0.416	0.378	0.194	0.206	0.184

Task 3B: Normalization

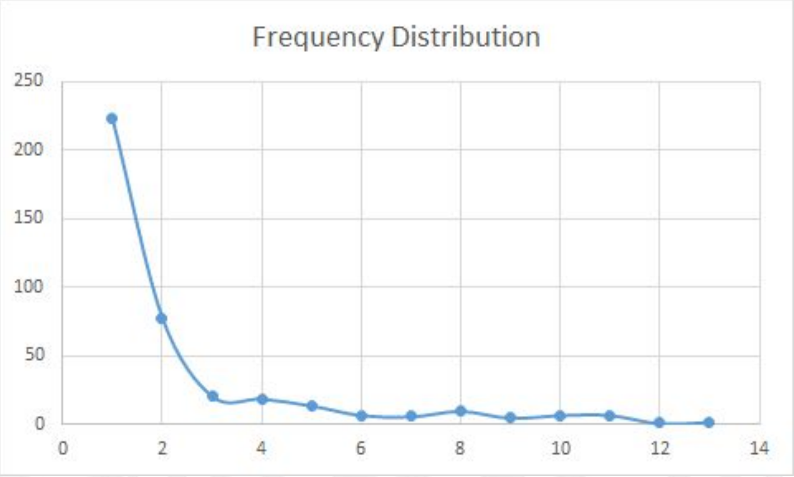


Data Analysis:

1. The training data is same as task 3a
2. But data augmentation was needed. Data was augmented from UMLS, and CADEC.
3. After augmentation, the distribution of data per class was in the range of 5-55 examples per class.

We are using ensemble of

- cosine similarity,
- hierarchical gru,
- svm
- logistic regression.



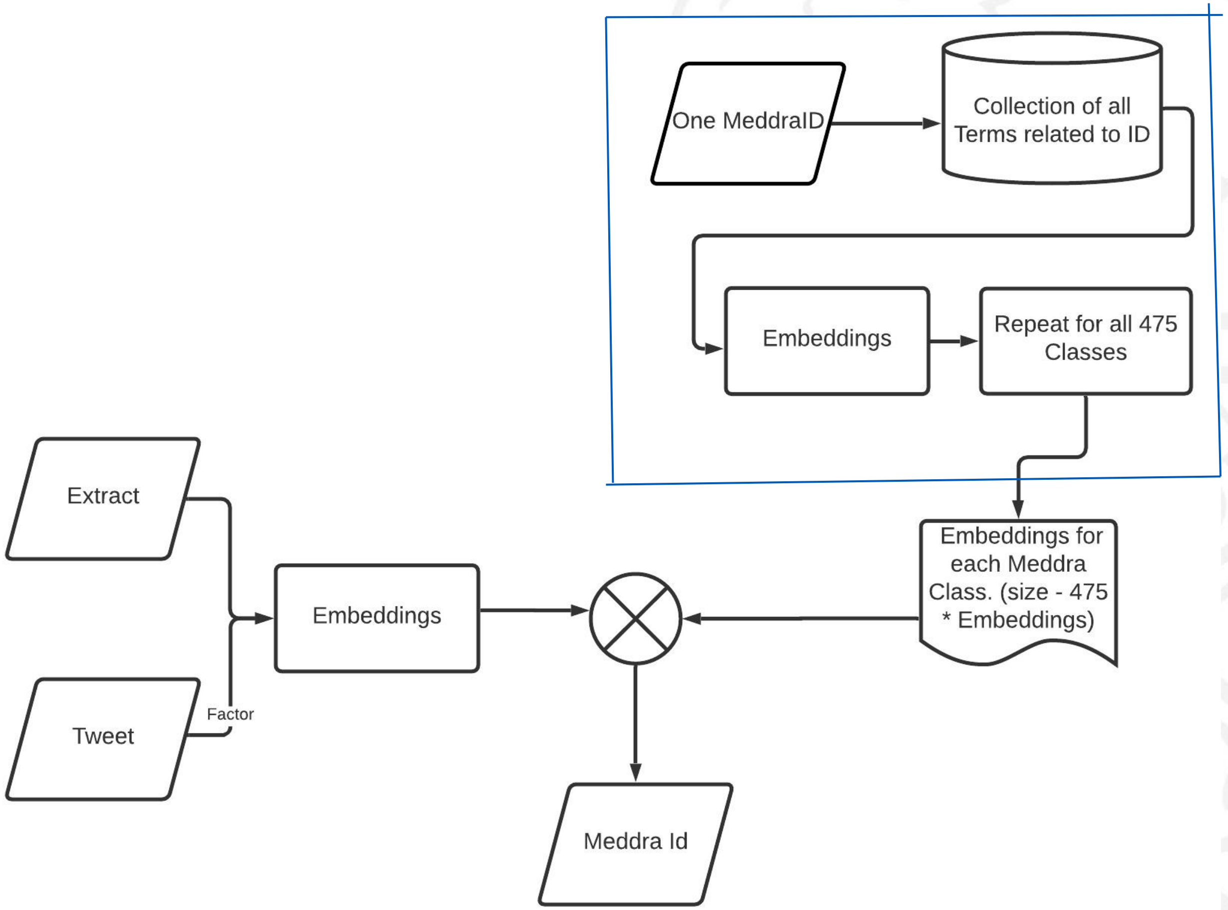
Model	Dataset	**Accuracy**	**Macro F1 Score**
Ensemble	Validation	44.93	38.93

Task 3 - Mapping with MEDRA id - Method 1

Model	**Accuracy**	**Macro F1**
Svc	45.75	32.30
Random	37.8	23.84
Gaussian NB	39.72	29.01
Logistic Regression	**46.30**	**34.63**
Decision Tree	33.15	18.7
KNN	41.09	30.74

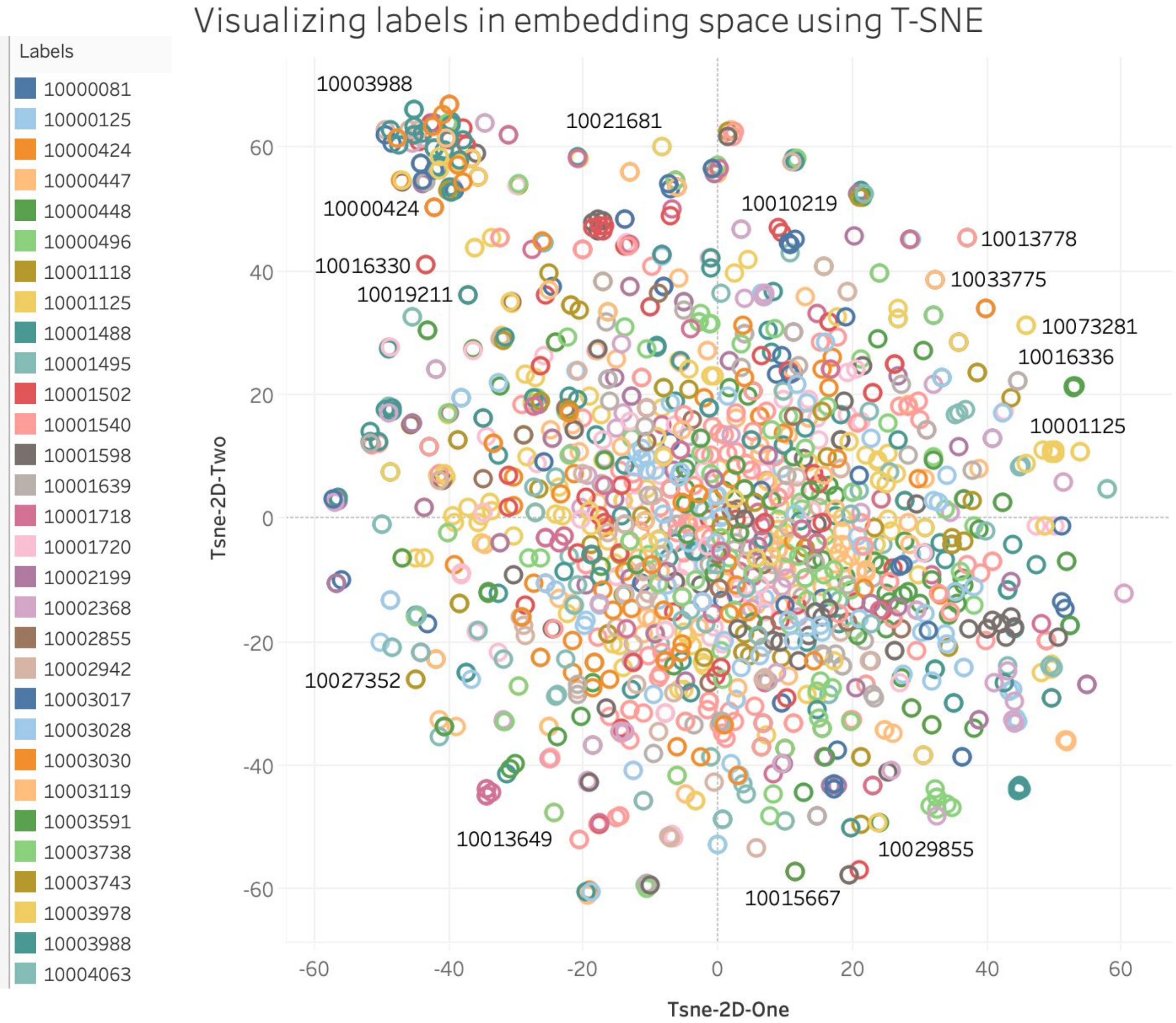
Task 3 - Mapping with MEDRA id - Method 2

Model	Accuracy	Macro F1
Char2vec	30.05	19.52
BioBert	26.02	21.55
FastText	**45.75**	**38.93**



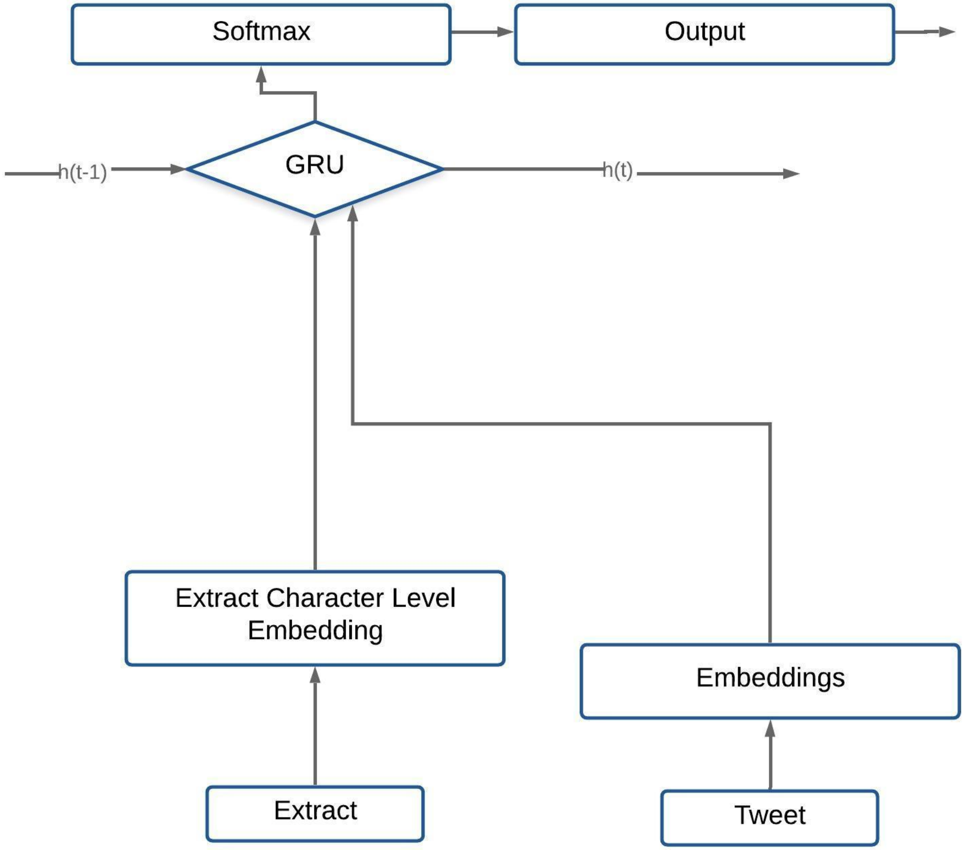
Visualization:

Extract	Meddra Id	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?

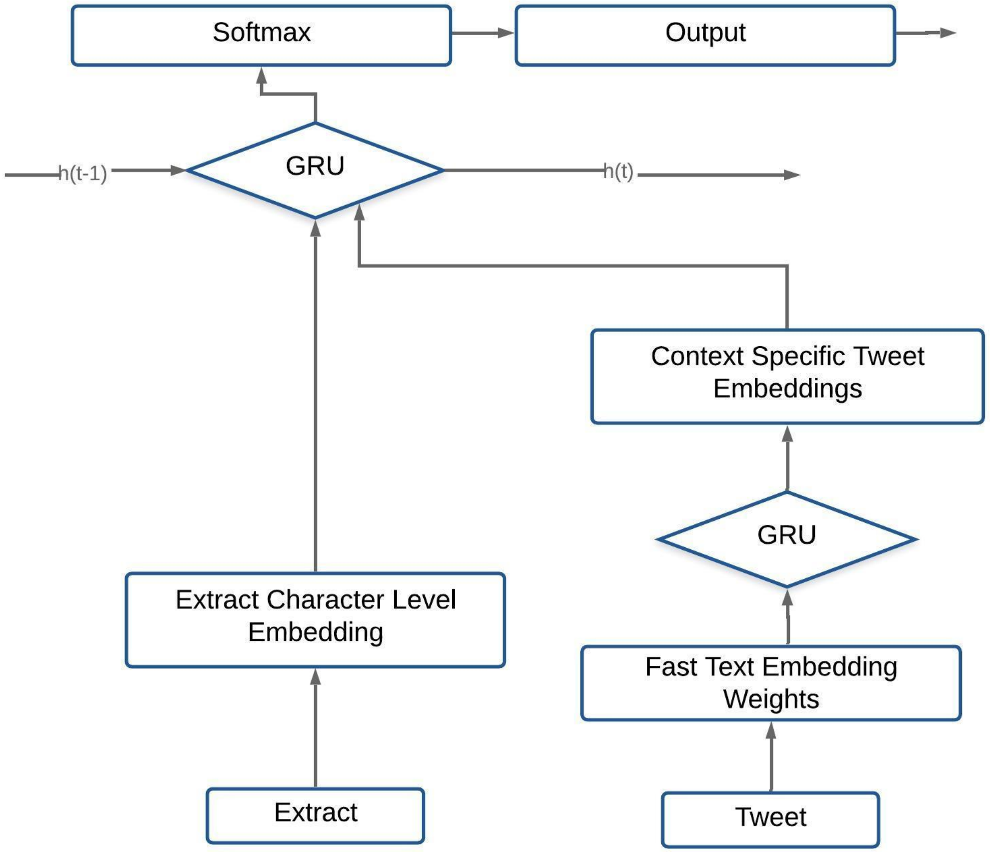


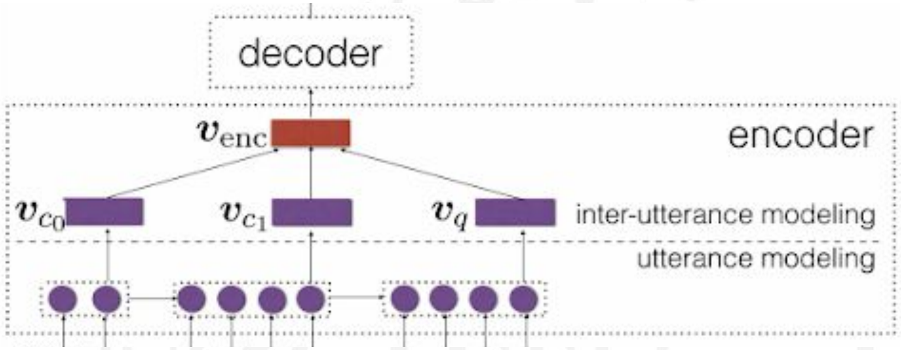
Task 3 - Mapping with MEDRA id - Method 3

Model	**Accuracy**	**Macro F1**
Self Trained Bio Bert	43.30	25.53
Self trained Bert	41.91	25.25
FastText	36.16	19.9
FastText self trained	41.91	26.67
Bio Bert	**45.47**	28.22
Bert based uncased	43.83	**29.27**
Bert large uncased	42.46	28.61



Task 3 - Mapping with MEDRA id - Method 4





Model	Accuracy	F1 Score
Self Trained FastText	45.47	31.63

Results

Model	Dataset	Macro F1 Score
Ensemble	Test (On NER_QA_prediction)	0.265

	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

Future Work

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

- 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.

- 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 feature and information sharing across tasks. For example, the NER would make less false positive classification for non ADR tweets.

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


MAY THE 4TH
BE WITH YOU