ZUM_NLP: PROJECT GUIDELINES

The aim of the project is to create a model of sentiment analysis based on tweets about current events in the world.

The full project consists of 4 stages, but depending on the expected final grade, it is enough to follow the stages indicated below:

3 – stages 1B, 2 and 3

4 – stages 1B, 2-4

5 – stages 1A, 2-4

STAGE 1: DATA COLLECTION

1A – data in Polish/English/Spanish/Swedish or Portuguese (so that I can understand it ☺)

Data acquisition concerns the collection of tweets. Each person scraps tweets (about 20k) to create a dataset for further processing. Tweets should be about current events, such as the war, NATO etc.

1. Adding class labels: Collected data is not tagged as positive/negative/neutral.
    a. Select the number of target classes (2 or 3 if we include neutral).
    b. Clean data and remove stopwords
    c. Create word embeddings for vectorized representation of words simillar in meaning // OR we use pretrained model for language of choice
    d. Use K-MEANS to create clusters and use k=2 or k=3 depending on the numer of target classes
    e. Based on clusters tag data and manually fix clusters if necessary

It is a good idea to limit the number of words as much as possible and possibly manually tag some of them too.

2. Data cleaning: normalisation, special characters removal, punctuation, URL, emails, duplicates, lowercase text and choose type of tokenizer. NOTICE: this stage is necessary BEFORE the creation of word embeddings.

1B – ready data

Use ready dataset (from Kaggle etc.)

ETAP 2: CLASSIC ML

Choose 3 models to fit data and present the results with confusion matric and roc curve. Just as in class.

ETAP 3: NEURAL MODEL

Choose type of neural network to train, and through validation decide on the best set of parameters. It is not enough to just build a model and get results. Fine-tuning is necessary too.

In a loop we save the best model according to cost of validation.

ETAP 4: LANGUAGE MODEL

The last stage is to use selected language model, e.g. BERT, to create a sentiment analysis classifier.

DEADLINE: check assigments in Teams

SUBMISSION**: GitHub repository** – if its private, make sure to share with dwnuk@pjwstk.edu.pl. Then add url to repo in Teams assigment.

NOTICE! If you aim for 4 or 5, the project can be done in groups of up to 3 ppl, but with the requirement of using more advanced models (e.g., usage of two neural networks for comparison) and more than 1 language model.

BEFORE SUBMISSION MAKE SURE THAT THE REPOSITORY CONTAINS:

- scripts/python files .py,

- README.md file with z project description and instructions how to use it,

- saved models (as long as possible due to storage limits)

- data used in the project in csv format or compressed

- you can briefly describe achieved results in README file along with project overview, without going into too much detail.