Fake News Challenge
M2-IASD – NLP course

Amina Tadjer | Mohamed Ali Benrekia | Aymen Djelid | Abdelkader Benamara

February 13, 2022

1 Problem Statement

Fake news can shortly be defined by the fact of making a story that has the intention to deceive and usually this can be used for negative purposes. The goal of our project is to make part from this challenge called shortly (**FNC-1**).
Now let us suppose that we are given a statement and our task is to check what ever the source that gave this statement can be trusted or not, in other terms how can we detect liars about incoming news that supports this initial claim ? so in facts this **FNC-1** challenge can be seen as an attitude detector where we are given two different statements and our goal is to try to classify a relationship between them.

2 Data Description

Our data is already split into two different sets (training set with over **49k** rows and **28k** for testing) ,both Train and Test Set are Pairs of headline and body text with the appropriate class label for each.

The data provided is (headline, body, stance) instances, where **stance** is one of unrelated, discuss, agree, disagree. The Dataset is provided as two **CSVs**:
train_bodies.csv :
This file contains the body text of articles (the articleBody column) with corresponding IDs (**Body ID**).
train_stances.csv :
This file contains the labeled stances (the Stance column) for pairs of article headlines (Headline) and article bodies (**Body ID**, referring to entries in **train_bodies.csv**).
In the following figure we have a sample of our dataset for more visualisation purposes :

	text_a	text_b	label
20163	Michael Phelps' girlfriend was born male but d...	After being officially dead for 48 minutes and...	3
2765	Angry mob hacks off alleged rapist's genitals ...	When Apple unveiled its Apple Watch, the compa...	3
14177	AUSTRALIA: 600-POUND WOMAN GIVES BIRTH TO 40-P...	Christian Bale will not be playing Steve Jobs ...	3
59250	Dual citizenship should be allowed	I hold dual citizenship as do two of my three ...	0
16774	Brian Williams: No, Our Meteorologist Was Not ...	Along with unveiling the Apple Watch earlier t...	3

Figure 1
A snapshot of our training data

2.1 Exploratory data analysis

Distribution of the data : In order to have a more precise idea of what we are going to work with let us try to visualise our data . The distribution of Stance classes in **train_stances.csv** is as follows:

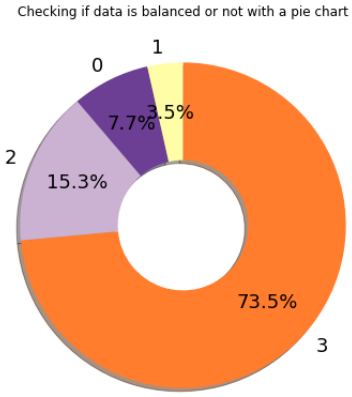
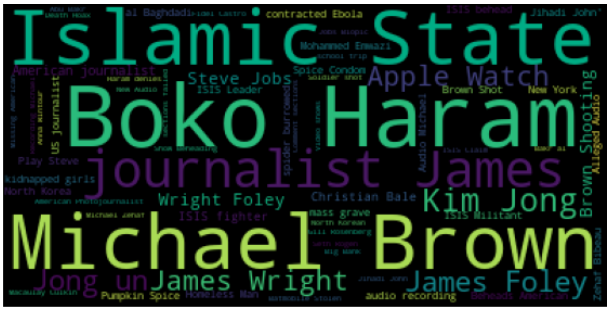


Figure 2
Classes (labels) distribution pie chart

Note. The classes are already encoded and here are the decoded version.

```
data_labels = {
                0 : "agree",
                1 : "disagree",
                2 : "discuss",
                3 : "unrelated"
              }
```

And in order to visualize the sequences and statements that appear the most in our headlines and bodies we can do it using the following **WordCloud** :



(a) Word Cloud For Article A

(b) Word Cloud For Article B

Figure 3
WordCloud showing the most common sequences of words appearing in the texts **A** and **B**

3 Tasks

The main tasks that we are planning to do during this project are listed above :

1. Data Exploratory Analysis (Visualisation).
2. Data Prepossessing (Split,Normalize,Tokenize,...).s
3. Use of a pretrained word embedding (not decided which one yet)
4. Build and Train some ML models from ($\subsetneq$) the following :
   (a) Transfer Learning (BERT,XLNET,RoBERTa,...)
   (b) Baseline ML/DL models (LogisticRegression,NaiveBayes,MLP,...)
   (c) Seq2Seq models recurrent models with attention
   (d) LSTM based models
5. Expiremntal process and results (Hyper-parameters tuning, models evaluation and results discussion (using errors analysis)).