ML for NLP final project 2022 – Yasmine Houri, Mathis Sansu

Context:

Data: French MPs tweets of the 15th legislature of the Vth Republic (2017-2022)

We create a list of the Twitter handles of MPs thanks to several data sources: database from previous academic research, scraping from website Nosdeputes.fr, and manual updates. This list is saved in a csv file.

We use the Academic Research Product Track V2 API Endpoint of Twitter (<https://developer.twitter.com/en/portal/dashboard>). We use the R-package ‘academictwitteR’ (https://cran.r-project.org/web/packages/academictwitteR/index.html) to extract all tweets of MPs of the current legislature (2017-2022), using the function ‘get\_all\_tweets’. We can provide the R code on demand. After merging the database on all MPs, we obtain a complete RDS dataframe, where each line corresponds to a single tweet. From another academic database, we also joined sociodemographic and political variables about MPs. We attached the first 100 lines of this database to this email.

Th final database is then composed of around 50 variables (tweets content, metrics on tweets, tweets and authors identifiers, sociodemographic and political variables, etc.), one row corresponding to a unique tweet. For now, with a database ranging from May 2017 to December 2021, it counts almost 2M tweets.

Key questions: Which emotions do MPs express in their tweets? Which words are more frequently linked to which emotions? Are there key periods in time when MPs instrumentalize emotions more than usual (election campaigns, …)? What other elements does CamemBERT shed light on (syntax, …)?

Tasks:

First and foremost, we will exclude retweets from our database, and work only with tweets and answers to other tweets.

This, being done, we intend to perform sentiment analysis on French textual data, using CamemBERT. We will first need to manually annotate a sample of our data in order to fine-tune CamemBERT and adapt it to our corpus. This will require us to define the emotion labels we want to predict (whether a binary prediction of positive / negative emotions, or more).

Then, we will split our data into two sets (train and test), and train and test it on each set respectively. We may want to validate our model on a new dataset, composed of other tweets from French politicians (mayors or others) if relevant.

Project Framing: You must frame your project yourself

Textual Data

Descriptive Statistics

Embedding techniques

Baseline model: FastText. Final model: CamemBERT.

Task Specific Modeling

Sequence classification: sentiment analysis (positive, negative, other) over tweets

Evaluation

Quantitative evaluation: metrics to look at