**Tema de predicție**

Vîrlan Adrian

1. **Metoda folosită**
   1. **Setul de date**

Setul de date „*train.csv”* este format din 2 coloane. Prima coloana conține știri in limba franceza, despre schimbări climatice. Cea de-a doua coloana conține o eticheta, cu tipul știrii: „fake”, „biased” și „true”.

Exista un total de 1948 de înregistrări(linii) valide, dintre care:

* fake: 514
* biased: 249
* true: 1185

Din aceasta analiza putem deduce ca setul de date este dezechilibrat, existând mult mai multe înregistrări de tip „true”, decât „fake” sau „biased”.

* 1. **Modelul folosit**

Pentru a realiza tema de clasificare am folosit modelul **T5-base**. Acesta este un model Encoder-Decoder pentru generare de text, dar poate realiza și sarcini de clasificare. Modelul a fost antrenat pe C4, set de date in limba engleza. Din acest motiv, voi face fine tuning folosind setul de date oferit.

* 1. **Antrenare**

Pentru a putea evalua acuratețea pe parcursul antrenării, am extras 10% din datele de antrenare. Selecția acestor date a fost aleatorie, păstrând-se rația originală dintre cele trei tipuri de știri.

A urmat o prelucrare a datelor: trecerea textului in litera mica și eliminarea „stop-words”(cuvintele uzuale care nu conțin informație). Am încercat „Stemming” și eliminarea semnelor de punctuație, dar rezultatele finale au fost mai slabe.

1. **Rezultate**