**Tema de predicție**

Vîrlan Adrian

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

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

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

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

Din aceasta analiză am dedus faptul 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. Pentru a adapta modelul la cerința temei, am făcut „*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 în literă mică 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.

Am încercat să reduc problema dezechilibrului setului de date prin a face „*oversampling*” pe clasele minoritare(„*fake*” și „*biased*”). Am duplicat înregistrări pana când toate clasele au avut un număr similar de elemente. După analiza rezultatelor, am remarcat scoruri puțin mai bune a acurateței, față de cazul când foloseam setul de date original, dezechilibrat.

Am antrenat modelul în 7 epoci, folosind acuratețea ca metrică de evaluare. La finalul antrenării, s-a ales automat modelul care a avut cea mai bună performanță.

1. **Rezultate**

Modelul „*fined-tuned”* cel mai bun a obținut o acuratețe de 79.48%, în 7 epoci.

Folosind acest model, am realizat o etichetare a setului de date de test.

1. **Surse**

Codul folosit pentru realizarea temei este disponibil aici:

<https://github.com/adrianvirlan200/French_News_Classification>