Analisi del dataset

Prima cosa ML che va bene:

* Rimozione stop-words
* Tokenizzazione
* TF-IDF
* Naive-Bayes per la classificazione visto i pochi dati

Classificazione con LLM:

* Classificare usando le soglie, il primo che supera la soglia dà la classificazione
* Calcolare tutte le probabilità e usare quella più alta
* Calcolare tutte le probabilità e prendere la media più alta tra agree e disagree

Il primo metodo di classificazione con LLM è inutile. Tanto vale provare gli altri due. In quanto devo eseguire su tutti i rumor in quanto userò il loro programma per il calcolo delle metriche.

LLM

Per la parte di LLM non c’è bisogno di andare a fare il fine tuning. Per come è posta la challenge, ovvero con un dev set e un numero massimo di volte che si può sottomettere il proprio modello al test, che il dev set deve essere usato per fare il fine tuning. Ma mi è stato detto di fare prompt engineering.

MODELLI

I modelli di LLM che intendo utilizzare sono un modello piccolo, uno medio e uno grande:

Il più grande è facebook/bart-large-mnli, che è quello con più parametri (137B), più layer (96). Richiede grande potenza di calcolo ma molto accurato e adatto a problemi di NLP.

Il modello medio è joeddav/xlm-roberta-large-xnli, con 1.5B parametri e 24 layer. Più veloce ma meno accurato per compiti complessi.

Infine quello piccolino è typeform/distilbert-base-uncased-mnli, 66M parametri e solo 6 layer. Velocissimo, ma meno accurato.

RISULTATI

Per essere validati i risultati devo usare il loro programma e devo formattare il file come lo vogliono loro:

* Id
* Predicted\_label
* Predicted\_evidence (autorità, id, testo, score)

PERFORMANCE LLM

Performance col il primo metodo di classificazione ovvero di usare come label la label più probabili e prendere i 5 tweet più probabili con quella label.

I tempi di esecuzione per 32 rumors sono all’incirca di 11 minuti per Bart, 7,5 minuti per Roberta e 3 minuti per distilBert.

Il modello Bart ha ottenuto 0,428 di Macro\_F1 e 0,280 della Strict Macro\_F1.

R5 = 0,449 AP=0,373

Il modello Roberta ha ottenuto 0,317 di Macro\_F1 e 0,128 della Strict Macro\_F1.

R5 = 0,228 AP=0,158

Il modello distilBert ha ottenuto 0,3 di Macro\_F1 e 0,155 della Strict Macro\_F1.

R5 = 0,326 AP=0,165

Bart è andato benino, mentre gli altri malino.

Labels utilizzate “SUPPORTS”, “REFUTES”, “NOT ENOUGH INFO”.

TEST BART NUOVE LABEL

Uso le nuove label che tendono alla conferma. Le label sono “CONFIRM”, “OPPOSE”, “NOT ENOUGH INFO”.

Le performance sono scarsissime.

Macro\_F1=0,108 Strict Macro\_F1=0,076 R5=0,186 AP=0,098

TEST ROBERTA NUOVE LABEL

Le nuove labels di Roberta dovrebbero essere più precise. Le nuove labels sono “VALIDATES”, “DISAGREE”, NOT ENOUGH INFO”.

Anche qui le performance sono scarsissime.

Macro\_F1=0,108 Strict Macro\_F1=0,040 R5=0,149 AP=0,075

TEST BART CON LE LABEL PENSATE ALL’INIZO

Test giusto per sfizio. Le label sono “AGREE”, “DISAGREE”, “OFF TOPIC”.

Macro\_F1=0,335 Strict Macro\_F1=0,165 R5=0,265 AP=0,223

STESSO TEST CON ROBERTA

Macro\_F1=0,158 Strict Macro\_F1=0,089 R5=0,230 AP=0,119

STESSA COSA PER DISTILBERT

Macro\_F1=0,395 Strict Macro\_F1=0,282 R5=0,405 AP=0,283

MACHINE LEARNING

Ora le cose del Machine learning.

Fatte ma le performance sono mediocri, molto mediocri. In ogni caso faccio anche la evidence retrieval per poter eseguire i loro test e avere le metriche. Risultati:

Macro\_F1=0,257 Strict Macro\_F1=0,222 R5=0,298 AP=0,196