Prediction contest

- descrierea solutiei -

Solutia aleasa realizeaza clasificarea textului folosind un model BiLSTM. Pentru implementarea solutiei am urmat 3 pasi:

1. Preprocesarea datelor

Textul este incarcat din fisierele CSV (train.csv si test.csv). Este utilizata biblioteca spaCy pentru eliminarea de stop-words si a caracterelor nerelevante.

1. Antrenarea modelului

A fost ales un model de tipul BiLSTM, deoarece acesta poate analiza textul bidirectional (si forward si backward), captand mai bine contextual si relatiile dintre cuvinte.

In acest pas, modelul BiLSTM este construit si antrenat utilizand datele procesate. Labelurile sunt encodate numeric. Se realizeaza tokenizarea textelor.

Arhitectura modelului:

* Un strat embedding pentru vectorizarea textelor
* Un strat LSTM bidirectional pentru captarea relatiilor dintre cuvinte
* Straturi Dense si Dropout pentru generalizarea modelului si reducerea overfittingului

Se utilizeaza Cosine Decay pentru reducerea learning rate-ului, iar functia de pierdere este CrossEntropy. Setul de date este impartit in 5 fold-uri pentru antrenare si validare.

Modelul si tokenizer-ul sunt salvati.

1. Evaluarea modelului

Modelul si tokenizer-ul sunt incarcati pentru realizarea predictiilor. Textele de test sunt tokenizate, apoi sunt realizate predictiile. Acestea sunt salvate in fisierul test.csv

Link git: [mateidumitru1/prediction-contest](https://github.com/mateidumitru1/prediction-contest)