Protopopescu Theodor

Company Classifier

Dificultatea rezolvarii acestui task vine din faptul ca nu exista label-uri de referinta pentru companii, astfel incat nu se poate utiliza un algoritm de invatare supervizata din prima. Pentru a ajunge la solutia finala de implementare, m-am gandit la mai multe abordari si la neajunsurilor acestora, pe care le voi detalia mai jos:

1. Utilizarea unui model de clustering (K-means)

Desi intuitiva, in principiu, aceasta metoda presupune clasificarea fiecarui exemplu drept apartinand unei singure clase, care poate sa nu fie suficient de cuprinzatoare pentru respectiva companie, mai ales in cazul companiilor cu multiple arii de expertiza.

1. Utilizarea unui LLM 0-shot

Desi performanta LLM-urilor evolueaza rapid, iar rezultatele acestora in scenarii de 0-shot si few-shot generation devin tot mai bune, in acest caz este dificila validarea clasificarii realizate, intrucat nu exista valori de referinta.

Prin urmare, solutia la care am apelat este una hibrida, in doua etape:

1. O clasificare initiala folosind tehnici NLP

Pentru aceasta etapa, se combina toate campurile existente despre companii: description, business\_tags, sector, category si niche intr-un singur string final, numit “combined\_text”. Acest camp este procesat, eliminand stop words, si realizand lematizare.

Acest text procesat, salvat in coloana “cleaned\_text”, impreuna cu label-urile din taxonomie sunt encodate folosind SentenceTransformer. Pentru stabilirea label-urilor pentru un exemplu, se calculeaza cosine similarity intre encoding-uri. Daca similaritatea este peste threshold-ul de 0.5, label-ul este asignat. In urma mai multor experimente in care am variat valoarea threshold-ului, am constatat ca aceasta valoare, desi limiteaza numarul final al exemplelor care sunt adnotate, ofera o clasificare performanta. Scaderea thresholdului asigura clasificarea mai multor exemple, dar mai putin precisa. In urma acestei etape, dintre cele **9494** de exemple, sunt adnotate **2824** de intrari utilizand **193 / 220** label-uri din taxonomie. Aceste date sunt salvate in fisierul ml\_insurance\_challenge\_annotated\_cosine.csv.

1. Antrenarea unui model BERT

Pornind de la exemplele adnotate in primul pas, vom antrena un model BERT pentru predictia claselor pentru celelalte exemple din setul de date. Setul de date din fisierul ml\_insurance\_challenge\_annotated\_cosine.csv este impartit in doua dataframe-uri, pentru exemplele labeled si cele unlabeled. Cele 2824 de exemple deja clasificate sunt folosite pentru antrenare si validare, intr-un raport 80:20. Este folosit modelul BERT base pentru clasificare multi label, utilizand “combined\_text” ca date de input. Graficul evolutiei loss-ului pentru validare se poate observa in figura urmatoare:

A graph with a line going up

AI-generated content may be incorrect.Utilizand acest model, se prezic clasele pentru celelalte exemple neadnotate.

Desi este departe de a fi o solutie performanta, aceasta abordare hibrida are cateva avantaje:

1. Este eficienta, intrucat utilizarea similaritatii cosine prin procedee clasice NLP este ieftina din punct de vedere computational, si totodata rapida. In acelasi timp, aceasta metoda garanteaza o clasificare controlata, verificand similaritatea semantica a denumirii clasei cu termenii prezenti in toate campurile exemplelor, acestea fiind concatenate.
2. Antrenarea unui model BERT pe aceste date este, de asemenea, eficienta, intrucat modelul are dimensiuni mici, performante ridicate pentru task-uri de clasificare, si este preferabila, in acest caz, utilizarii unui LLM. De asemenea, utiizand supervised learning, putem garanta ca vor fi folosite clasele din taxonomie si nu vor exista halucinatii.
3. Utilizand adnotarile generate folosind modelul BERT, ambele sisteme de adnotare se pot imbunatati prin finetuning cu aceste date.

Cateva din neajunsurile acestei solutii sunt:

1. Doar 194/220 clase din taxonomie sunt utilizate in adnotarea initiala, restul nefiind prezente in datele de antrenare ale modelului BERT. Prin urmare, celelalte 26 de clase nu vor fi atribuite niciunui exemplu.
2. Clasificarea initiala este foarte susceptibila schimbarilor prin varierea thresholdului, astfel incat este greu de spus care este configuratia initiala optima pentru adnotarea datelor ce vor fi folosite pentru BERT.

Din pacate, rezultatele obtinute in urma adnotarii folosind BERT nu sunt performante (modelul clasifica cu aceleasi 5 clase toate exemplele). Acest fapt se datoreaza fie numarului mic de exemple adnotate cu fiecare clasa, fie unei erori in partea de procesare a datelor si antrenare a modelului, pe care nu am reusit inca sa o identific. Toate exemplele din fisierul final\_insurance\_labels.csv contin clasele prezise de BERT (inclusiv cele 2824 de exemplele adnotate initial bine folosind cosine similarity). Adnotarea initiala a acestor 2824 de intrari poate fi vazuta in fisierul ml\_insurance\_challenge\_annotated\_cosine.csv.