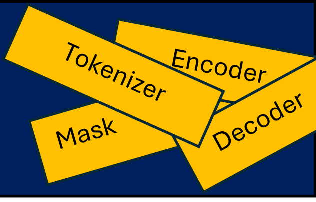
Natural Language Processing – Group Project Report

Synthetic Datasets, Entity Recognition and Classification Models in NLP



Identifying trends with Synthetic Data, Entity Recognition and NLP in the Medical Domain

This application aims to improve medical diagnosis and identify geographical disease trends. To achieve this, the app leverages the power of NLP, LLMs and Entity Recognition using synthetic data.

Data Preparation: Text Generation using LLMs

Personal medical data is a very sensitive topic. To develop this app, we need to generate a vast dataset, containing information about realistic patient descriptions alongside with other relevant information such as the hospital in which the patient is being attended, geographical information about the hospital, the disease itself and the reported date among others.

To generate this data, we leverage the power of LLMs. The first step was to use ChatGPT 4o to generate all of the general information regarding the patient's basic information when he reports to his doctor.

Secondly, and the big challenge here is to use LLMs to generate realistic patient symptom descriptions. We need the generated descriptions to be accurate, not repetitive and related to the specific disease. Our first approach was to fine-tune a GPT2 model, but the model fell short when reaching our objective. The team tried to run locally Llama 3 8B but it ran out of VRAM memory. Finally, we decided to connect to ChatGPT 3.5 API, controlling some parameters such as top_k (diversity of generated descriptions), top_p (steers the model towards generating more probable descriptions) or repetition Penalty (prevents repetitive outputs by implementing a penalty mechanism).

After this GPT 3.5T parameter tuning, we ask our LLM to generate a specific number of descriptions per disease, following the GPT4o dataset's disease distribution (where we used the spanish distribution of cases). Our final step is to merge our two LLM-generated datasets.

Unfolding Trends: Entity Recognition & Classification using Domain-specific Transformers

Named Entity Recognition (NER) is a sub-task of Information Retrieval in Natural Language Processing (NLP) that identifies and classifies named entities in a text into predefined categories such as person names, organizations and locations. The goal of NER is to extract structured information from unstructured text data and turn it into a machine-readable format. Our goal using this technology is to identify the key aspects of the patient symptom description and use them as inputs to obtain geotrends and give diagnosis suggestions.

The application developed utilizes a pre-trained medical transformer sourced from Hugging Face. This specific model is a fine-tuned version of DeBERTa, which has been trained on the PubMED dataset, a comprehensive repository of biomedical literature. This fine-tuning process prepared the model to excel in recognizing medical entities within text data.

In addition to the model, we have developed a custom function that focuses on identifying three specific entity types: DISEASE_DISORDER, SIGN_SYMPTOM and MEDICATION. The model can effectively detect these key terms within the text data of the patient descriptions.

This information will serve for two purposes: i) Geo-Trending Analysis, for which the app can generate visualizations of the most prevalent diseases across various locations, helping healthcare professionals in identifying areas with a higher concentration of specific illnesses, ii) The extracted entities, can are used as input features for a separate Machine Learning model to suggest potential diagnoses.

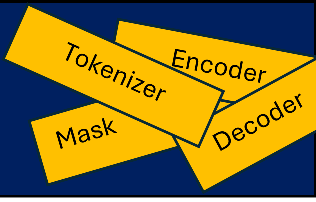
Finally, we have decided to make a classification model based on tokens. This has been possible using a TFIDF methodology with a simple logistic regression model. It has also been implemented in the app and will serve to help doctors have an initial intuition on the disease to diagnose.

Application Development

The application is based on Streamlit, a framework that allows to build easy web applications and deploy python code. The application will have the characteristics mentioned above. The potential benefits of this application are to detect outbreaks early on, as well as provide a quick disease diagnosis for patients.

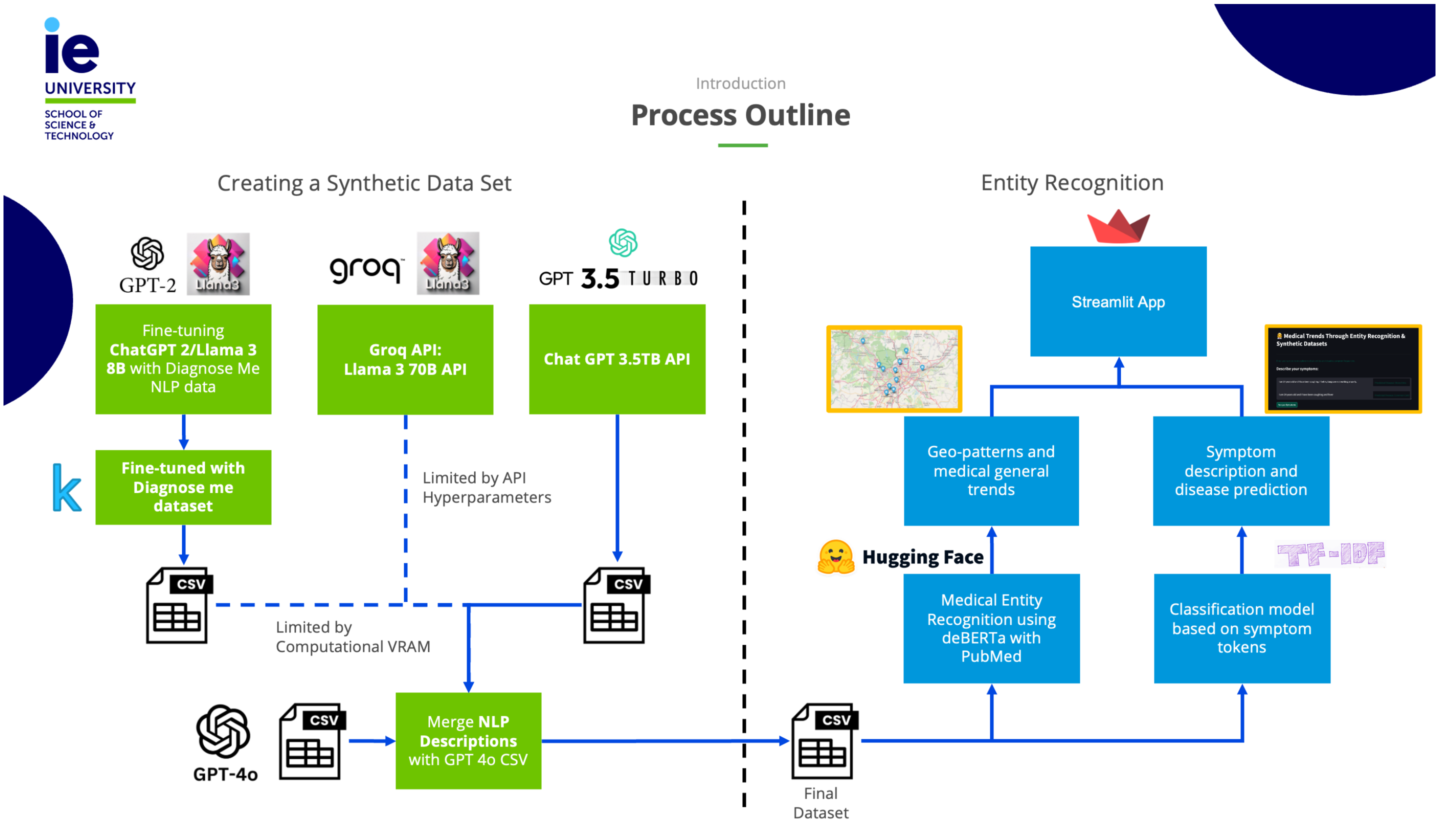
Natural Language Processing – Group Project Report

Synthetic Datasets, Entity Recognition and Classification Models in NLP



Conclusion & Potential Benefits

These use cases demonstrate the practical applications of NLP in today's world. From generating synthetic data for hard-to-obtain datasets enhancing LLMs, using domain-specific Transformers to extract geographical NLP trends, and deploying a TFIDF LOGIT classification model for diagnosing based on symptom descriptions, we've explored significant ways NLP can make a difference. These examples highlight the transformative potential of NLP in addressing real-world challenges and improving various aspects of our lives.



ie UNIVERSITY SCHOOL OF SCIENCE & TECHNOLOGY

Introduction

Process Outline

Creating a Synthetic Data Set

GPT-2 Llama3

Fine-tuning ChatGPT 2/Llama 3 8B with Diagnose Me NLP data

Fine-tuned with Diagnose me dataset

groq Llama3

Groq API: Llama 3 70B API

Limited by API Hyperparameters

GPT 3.5 TURBO

Chat GPT 3.5TB API

CSV

CSV

Limited by Computational VRAM

GPT-4o

CSV

Merge NLP Descriptions with GPT 4o CSV

CSV

Final Dataset

Entity Recognition

Streamlit App

Medical Trends Through Entity Recognition & Synthetic Datasets

Describe your symptoms:

Geo-patterns and medical general trends

Symptom description and disease prediction

Hugging Face

TF-IDF

Medical Entity Recognition using deBERTa with PubMed

Classification model based on symptom tokens

This Report was made by Ignacio Alonso López-Linares, Morten Aas-lyngby, Kangjie Yu, Harel Ben David, Andreu Artigues & Guillermo Brun.

https://medicalinsights.streamlit.app/