CS 410 Project Progress Report, Fall 2022

Team: Tetra

Members: Dillon Harding (dth3), Kowshika Sarker (ksarker2), Nikhil Garg (nikhilg4)

Captain: Vaibhav Karanam (karanam5)

Progress Made

Data Source: We investigated several potential web sources for crawling disease-specific diet recommendations. We looked into clinical websites like Mayo Clinic and some other sources, but due to the complexity of URL structures and lack of dedicated sections on diet recommendations, these resources did not seem feasible within the project's time constraint.

Finally, we have identified Wikipedia as the source of disease-specific diet recommendations. We have identified some lifestyle diseases, because onset and prevention of such diseases are more likely to be associated with diet habits than other kinds of diseases such as genetic disorders. For the preliminary phase of our work, we have shortlisted 11 lifestyle diseases to work with, because Wikipedia pages of these diseases have a dedicated diet section or subsection - which will facilitate the crawling.

Sentiment Analysis:

Pretrained Model: We have formulated the sentiment analysis of disease-specific diet recommendations as a 'entailment analysis' problem. We plan to use a pretrained entailment analysis model named RoBERTa. RoBERTa takes two texts as input - *premise* and *hypothesis*. RoBERTa checks whether the statement of the *hypothesis* is a true fact or not with respect to the knowledge contained in the *premise*. This is done as a three class classification task, the classes being - neutral, entailment, contradiction. Entailment refers to the case when the facts expressed in *hypothesis* are true based on the information of *premise*, contradiction means *hypothesis* is false with respect to *premise*, and neutral denotes that *premise* does not provide enough information to crosscheck *hypothesis*. RoBERTa returns three probabilities - one for each of the classes. Some examples of the RoBERTa model are mentioned here.

Diet Sentiment Analysis: For our task, say we have the nutrient and ingredient list of a food item available on Walmart website. We will iterate over each nutrient/ingredient and form a dummy positive/negative sentence. For example, with an ingredient 'sugar', some positive dummy sentences may be 'sugar is beneficial', 'sugar is recommended', 'high sugar is recommended' and some negative dummy sentences can be 'sugar is harmful', 'sugar is risky', 'high sugar is harmful' etc. Using the diet recommendation as the premise text, we will obtain the probabilities of neutral/entailment/contradiction sentiments. We will use the sentiment with the maximum probability to classify the food association with the disease. Based on some preliminary experiments with diet recommendation of Type 2 Diabetes, we have found RoBERTa to perform significantly well at detecting such dietary sentiments. The code of the preliminary experiments are posted on our project GitHub repository.

Customization: In most cases, diet recommendation of a disease is a multi-line paragraph where each nutrient or food item is usually mentioned in only one or two sentences. We checked that if individual sentences are used as premise, RoBERTa performs significantly well in extracting sentiments, but when the entire paragraph is used as a premise, the verdicts are usually neutral, even if the premise contains information regarding that ingredient. The reason is, for each ingredient most sentences in the paragraph mentions nothing and so the information content about that nutrient contained only in one or two sentences gets diluted. For this reason, we have decided to use each individual sentence as premise for each ingredient and aggregate the obtained sentiment probabilities through maximization to get the final output.

Website: In order to express the data gathered via web-scraping, we created a web page to allow users to interface with the data in a user-friendly way. A simple HTML web page was created as a placeholder until the data is ready to interface with the website.

Web Scraper: Beyond the creation of the HTML web page, progress has been made towards the web-scraper tool we will be using to retrieve text-data from Wikipedia. The beginning stages of the web-scraper have been initiated, with preliminary web-scraping tests ready to be run soon. Research into some of the necessary URLs to feed the web-scraper has begun, and will need to be finalized before completion of the scraper.

Remaining Tasks

In the past week, our team has established the base level of the project by finalizing the sources that web crawler will use to gather data and create a repository for diseases and food. We have also created a project roadmap which addresses the tasks that would be done each week and the final cutoff for the project to be finished. The three major tasks that constitute the project are - building of the web crawler, sentiment analysis of the data, and deploying of the webpage for users to use the product.

Web Crawler: After finalizing the web sources that will be used. We will start building the crawler in python. Right now we are in the process of deciding the libraries and the pattern of the information. We expect the crawler to be finished in the next coming week.

Sentiment Analysis: We have standardized the input and output of the crawler and are using that output to run it through the sentiment analysis code. The team has been experimenting with a couple of libraries and from scratch sentiment analysis to see which one gives the most accurate results. This process will be completed in the next couple of weeks and ready to be merged with the crawler.

Webpage: After the crawler and sentiment analysis module is finished, the team will shift its addition to deploying webpage on github to host the product and create user interaction with the product. This would be the last stage of the product and will be completed in the final week of development along with any modifications throughout the process to improve integration.

The team feels prepared to tackle each of the tasks remaining. While the beginning weeks were spent to solidify our approach, we felt it was necessary to make sure we iron out details before jumping into development. Now that we have a concrete idea of what the project will look like we can start the development process from now to the end of the deadline without having to worry about any resistance from lack of details.

Challenges

***Web Crawler*:** Our web crawler has had several challenges such as finding the right source to run the crawler on to collect information. Several sites were considered, such as Mayo Clinic, WebMD, and Wikipedia. The crawler needed a site which was easily traversable and listed the food items that were good for a particular condition in a recognizable pattern. After further consideration, Wikipedia was chosen as the site to use the crawler on, since the information about diet could often be found after looking for certain keywords.

***Sentiment Analysis*:** One important challenge for sentiment analysis is how to categorize the output from the crawler. Since the crawler would return all information about diet for a particular condition, sentiment analysis would need to sort out the foods into foods that are good for a condition, or foods that are detrimental. The important challenge of sentiment analysis has been finding the right keywords to determine if a sentence containing a food has a positive or negative connotation to it.

***Web Page*:** Since the web page will show the results of the sentiment analysis to the user, an important aspect of creating the web page is to decide how the results should be displayed. While the team has decided what to use to create the web page, we still need to consider how exactly the page will appear to the user and how the user will interact with the page.