Natural Language Processing – CS 4420

Assignment 1

Shrey Chhabra

File Guide –

I've added different files for every step of the assignment, i.e., there are code and dataframe CSVs for scraping, pos tagging, ner, er, and coref. I've also saved all these together in a single dataframe – ***results.csv*** *(this does not include the coref text and coref chains, which are stored separately in* ***ozempic_neural_coref_with_chains.csv****). All manual annotations done with labelstudio.*

1.b. My scraper uses selenium, beautifulsoup4, and pandas. I opted to scrape Google, Bing, and MSN news articles to get a total of 150 articles (even though the requirement is 100, a lot of articles are just *stubs* which are more difficult to clean, which is I scraped several extra articles as well). The reason to use these three news aggregators is that since selenium allows me to use a headless webdriver (basically running chrome without the browser window actually opening), scraping the news articles is akin to browsing these on google news. This ensures that the articles that are scraped successively are the most recent possible articles, and I would not have to spend time setting up logic for the program to filter out older articles. After this, I only needed to specify the search queries (eg. "Ozempic", "Ozempic FDA approval") to direct the webdriver to scrape the articles with these matches. Finally, I used bs4 to parse the HTMLs and pandas to store the articles in a CSV. Some stub articles still remain in the csv.

2.d. For entity resolution, I opted to use fuzzy matching. It works best for this specific use case since different articles refer to the same entities in different ways. There are several examples that support this, for instance; "Novo Nordisk" (a Danish pharma company) is also often referred to in different ways, i.e., "Novo-Nordisk Ltd.", "Novo Pharma", etc. A lot of the scraped articles are written by entirely different people, from different countries and different writing styles. Fuzzy matching thus becomes the best method for entity resolution (well, specifically the Levenshtein Distance method of fuzzy matching), measuring string similarity, and is flexible, albeit slow for large datasets. The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. If the similarity score is above a certain threshold (90% in my case), then the entities are considered to be a match, and are stored in the dataframe.

3.a. Here are my results for the evaluation metrics of the spacy NER model:

Precision: 0.2050

Recall: 0.0251

F1-score: 0.0438

These scores are expectedly bad, and the reasons for this are quite clear. First, the precision is about 20%, which is better after some changes in eval logic (more on this later), which means that compared to my manual annotations, the model is only able to match about 20% of the

NER entries. The recall, and the subsequent f1 score are also bad, implying that the model missed about 97.5% of the entities that were manually labelled, and is this not great at recognizing said entities. What could be the cause for this? First, the pre trained spacy model is likely only trained on general data (eg. news, wikipedia), and not medical/chemical terms. The most common mistake the model made was related to this, it almost always labelled "ozempic" as an organisation, similar to other drug names. Given that of the scraped articles these drug names were the most likely recognized entities, the model got nearly every entity NER tag wrong.

3.b. Coreference resolution is a famously difficult NLP problem, and this is evident from the results of the coref model (I've used Stanza CoreNLP's neural algorithm for the best possible accuracy). As it happens, as the size of sentences and articles increases, it becomes more difficult for the coref algorithm to determine exactly which nouns and pronouns need to be replaced in the text. While the neural coref model is better than the earlier rule based approach, it is still quite inaccurate, and struggles with long range dependencies – where an entity and its reference are far apart in the text. Consequently, ambiguous pronouns are also poorly handled, for example, consider the sentence - Ozempic is manufactured by Novo Nordisk. It is a leading pharmaceutical company." Stanza may mistakenly resolve "it" to "Ozempic" instead of "Novo Nordisk.". It is also likely that errors in sentence splits cause coref chains to break, leading to incorrect resolutions. Given that most of the my articles are quite long, multiple references spread across multiple paragraphs also become easy to miss. Evidences of this can be seen in the file *Ozempic_neural_coref_with_chains.csv,* where quite a few coref chains seem to be extremely incorrect.

3.c.

The NER model extracted various entity types, including:

Organizations (ORG): Novo Nordisk, FDA, Eli Lilly, CNN, WHO, MedPage Today

Persons (PERSON): Dave Moore, Kåre Schultz, Donald Trump, Frederik Duch Bromer

Locations (GPE): United States, Denmark, Canada, China, Texas

Drugs & Chemical Products (PRODUCT): Ozempic, Wegovy, Mounjaro, Zepbound, Semaglutide, Tirzepatide

Dates & Time (TIME): 2022, 2023, 2024, February 21, August, April 22

Monetary Values (MONEY): $6.5 billion, $1,000, $1,400, $100 per vial

The biggest insights from the identified entities were that Ozempic and wegovy were the most frequently mentioned drugs, alongwith the chemicals semaglutide and tizepatide in articles discussing weight loss treatments. Additionally, it also provided information about the economies of these drug markets, with several companies (Eli Lilly and Novo suggesting corporate competition in this market). Regulatory and legal focuses were also prevalent (seen in the frequent use of the ORG tags), with the FDA frequently appearing in the news about drug shortages and supply issues, compounding pharmacy regulations, and concerns of off-

brand drugs as well as competition between manufacturers. To support these, several instances of monetary values (MONEY tags) were also very prevalent in the articles, indicating that most of the news surrounding weight loss drugs heavily skews towards their manufacture, sales, economics, and market trends. There were fewer medical research articles collected, which is why there is only a small collection of articles also mentioning side-effects of the drugs, such as osteoporosis and muscle loss. The second most used tag was “GPE – geopolitical entity”, for instance: Denmark & United States: Most frequent mentions due to Novo Nordisk’s manufacturing and U.S. regulatory policies and China, Canada & Texas: Referenced in relation to pharmaceutical supply chains, lawsuits, and research studies, etc.

Additional Insights (non-NER)-

The following are some insights I’d like to develop that aren’t possible to glean from NER:

1. Sentiment Analysis on entity mentions, for example, it recognizes "Novo Nordisk," but doesn’t indicate whether the article is praising or criticizing the company’s pricing strategies.
2. Entity Relationships, NER finds "FDA" and "Ozempic" in the same text, but it doesn’t establish how the FDA regulates Ozempic. Knowledge graph construction would likely help with this.
3. Financial Analysis tools – as mentioned before, mentions of the PRODUCT are inevitably tied with mentions of the MONEY tag. While these are detected, no insight is provided into the meaning of these values, such as stock price movements or how drug supply changes impact competition and market sentiment. Financial analysis tooling would be quite helpful.