NLP Analysis of Continuous Glucose Monitoring (CGM) for Diabetes

Background

Diabetes is a metabolic disease that causes high blood sugar, when the body can't produce enough insulin (Type 1 – 10%) or can't use the insulin it does make (Type 2 – 90%).In 2019, diabetes was the direct cause of 1.5 million deaths in 2019, and with the advent of the COVID-19 pandemic, is a major contributor to mortality from COVID-19. Diabetes cannot be cured, but can be managed. The goal of diabetes management is to reduce A1C levels, which measure the amount of hemoglobin (a protein in red blood cells) that has glucose attached to it. Glucose levels are usually measured with finger-stick blood glucose tests, but an emerging alternative is use of continuous glucose monitoring (CGM) which analyzes data from a sensor inserted under the skin. Since CGM is always on, glucose levels can be tracked in real time to see how glucose levels change throughout the day. CGM is usually used for type 1 diabetes.

The Data

Anupam Singh and the team at 113 Industries have kindly provided a sample of over 37K social media posts from a variety of sources related to CGMs available in Canvas as `Diabetes Continuous Glucose Monitoring – Data Export.xlsx`. Each row of the Excel file consists of a field of text called `Sound Bite Text`, along with a variety of other more structured metadata fields such as the source and when the `Sound Bite Text` field were extracted.

The Tasks

Analyze the data given to create a ten-minute presentation and a 10-20 page report that answers the following questions. Since the presentation is relatively short, you may choose to focus on a few highlights from your analysis and expand further on the details in the report.

1. General CGM analysis:
    a. What are patient expectations of CGMs? What are patient knowledge gaps with CGMs?
    b. What benefits are most important to diabetes patients?

    c. What unmet needs do patients have related to CGMs (something patients want but are not getting)?

2. CGM Product-related analysis:
    a. What are praises & complaints and features of Dexcom & Freestyle Libre?
    b. What product features are being talked about?
    c. How do consumer opinions of Dexcom and Freestyle Libre compare?
    d. What is the overall sentiment regarding the two products?
    e. Based on your analysis, which one would you recommend?
    f. What would you tell each of these brands to improve?
3. **(Extra Credit)** CGM Consumer related analysis:
    a. Can you identify different types (segments) of consumers or create segments?
    b. Are different benefits more important to different consumer types?

The intention behind the assignment is that the tools discussed in class for text preprocessing and analysis (e.g. stemming, POS tagging, topic modeling, sentiment analysis, etc.) combined with some manual review/interpretation of the outputs, should be sufficient to do this project. You are free to do some outside research to validate/supplement the findings, especially if you are unfamiliar with the domain, and to apply other more sophisticated text analysis techniques you may know of, but they should not be required.

Please submit a .pdf of your slides, as well as the final report. There are no hard guidelines for the final report, but if you are submitting more than 20 pages, you are probably including too much detail. Similarly, if your report is fewer than 10 pages, you may not be answering the questions fully. The 10-20 pages is for the text of the report, and does not include the code. The preferred submission format for the final report is a .pdf file accompanied by the code in a separate Jupyter notebook. The final report should be separate from your slides and the code

Rubric

Each team should turn in separate files for the presentation, the report and a Jupyter notebook for the code. The presentation and report will be graded based on clarity, organization and presentation/writing quality as well as content. Watch for typos and grammatical mistakes. Structure your presentation and report to tell a coherent story: with executive summary (one slide/page with key results), introduction/motivation that states why you are performing the analysis and a conclusion that reinforces key findings/recommendations.

Presentation: 10 points total, 8 points for answering the questions below, and 2 points for presentation quality. Focus the presentation on the problem you are solving and any insights/results, and leave the details to the report.

- What problem is being solved? Focus on one or two use cases your results could help with, not the broader range of problems the data could address.
- Who benefits from the problem solution?
- What does the data look like based on your exploratory data analysis?
- How as the data prepared for modeling?
- What modeling approach did you use?
- What are the results?
- How did you evaluate the results?
- What next steps would you recommend based on your results?

Report and code: 10 points total for the report/deliverable, 8 points for answering the questions above, and 2 points for report quality, including appropriate use of visualizations. Code should be well structured and easy to follow, with liberal use of comments.