5880 Ellsworth Avenue, Suite 2
Pittsburgh, PA 15232

NLP Analysis of Consumer posts on Diabetes Dataset

Author: Anupam Singh, anupams@113industries.com

Date: Sep 20, 2024

Background

Diabetes is a pervasive disease for millions. Over 34 million Americans, or 10.5 percent of the U.S. population, have diabetes. 1 in 5 don't know they have it. 122 million Americans, or 37 percent of the U.S. population, live with diabetes or prediabetes. Including prediabetes, it is the most common underlying chronic condition in the U.S.

What is worse is that Diabetes and other related chronic conditions are nearly twice as common among communities of color than among white Americans. Diabetes and related conditions have been researched thoroughly from a clinical perspective and we have volumes of data on the socio-economic factors related to the condition. For example, we know that Diabetes prevalence is inversely related to household income level. Living in poverty in the two years prior to diagnosis increases the risk of developing Type 2 diabetes by nearly 25 percent. Low-income communities in America have fewer grocery stores, more convenience stores, and less transportation access to reach supermarkets than higher income areas. Low-income census tracts have half as many supermarkets as wealthy tracts.

This project is an introduction to how we can better understand the disease and its impact on those suffering from diabetes from a different perspective… a "consumer" or patient/caregiver lens. To keep the project focused and allow for practical applications of the NLP and AI concepts, we are proposing a project with a slimmed down scope focused on using tools to better understand what patients, caregivers and loved ones are saying about diabetes. Extensions of this project to enable the broader understanding of the disease and its link to socio-economic factors may include augmenting the dataset with additional geo-tagging, integrating the dataset with other datasets such as clinical, lab and consumer purchase panel data etc.

5880 Ellsworth Avenue, Suite 2
Pittsburgh, PA 15232

Dataset

This dataset is natural language text data of consumer posts, comments and replies on Diabetes on online channels including social media (X/Twitter, Instagram, Facebook etc.), forums, community groups and blogs.

The data file captured data from social media and online channels such as Forums, Blogs, Twitter, Instagram and Facebook (under Column "**Source**"). The period of data capture is between Jan 1, 2023 to Dec 31, 2023. The export is from an overall set of 3,664,800 Posts (5,575,900 Mentions). All posts are in English even if they are from non-US countries.

You will find two sets of data files with each file containing approximately 10K posts. There are two sets because of a nuance in the nature of the posts vis-à-vis the geography/location from whence the content was posted. In order to facilitate some geography based analysis, we have tried to include the geography when available. Unfortunately, reddit DOES NOT Include a geography tag. Since reddit is a significant source of overall volume AND also of relevant data with rich insights, we definitely want you to analyze data from reddit. Therefore, we have had to do two different forms of exports: one for reddit (with no Geography) and the other for ALL other sources, some of whom do provide the Geography tag.

Files with the name

Diabetes Geo US No Reddit 2023 50K Rows.xlsx (all sources except Reddit)

And

Diabetes Reddit Only Q1 2023.xslx through Diabetes Reddit Only Q4 2023.xslx

(Reddit exports only - one for each Quarter in 2023)

NOTES:

1. There could be some duplicate posts in the files due to the nature of how the data has to be captured and exported. You should eliminate them before processing.
2. The details, Sources on the dates etc. are at the bottom of each Excel file. Remove them before ingesting them into your programs.

The natural language data to be processed and analyzed is under Column "**Sound Bite**" and potentially the Column "**Title**". The latter is only applicable for forums and sometimes Blogs as well. In addition, you are free to use other structured data elements for a more advanced analysis (e.g. **Source, Author Gender, Sentiment, Published Date** etc.).

The goal of the dataset is to focus on data from regular consumers, NOT from news sources and professional experts. Filters have been applied to discard some of the non-relevant data (e.g. discard news sources and professional reviewers). Further, the underlying tool used to capture the data has to make inferences on how to exactly match before capturing or discarding data, the relevance to the topic varies and will sometimes pull in non-relevant data.

Due to the nature of how consumers talk and communicate, the language is not precise.

Some notes on pre-processing data before analysis:

1. You can eliminate more of the non-relevant data by eliminating posts from authors that have a large number of followers (Column: **No. of Followers/Daily Unique Visitors)**: they will likely be professionals or celebrities.
2. There are many columns with no significant data. You can eliminate those.
3. TBD: Some of the captured data may be marketing or "promotional" rather than from patients or loved ones or family members. You could identify and remove the "promotional" content.

Tasks

Your task is to process and analyze the natural language data under "Sound Bite" and "Title". In particular, here are some comparisons and questions that you could tackle:

1. Pre-process the data as defined above to the best of your ability. Most of this can be done in Excel.
2. Quantify the type of diabetes mentioned explicitly: Type 1, Type 2 and gestational.
3. What other diseases/comorbidities were mentioned? For example: obesity/weight management, hearth health, high blood pressure, kidney issues, obesity etc. How many of each?
4. Do some spot checks or Excel based analysis of the Reddit and non-Reddit datasets separately for some of the tasks above. What are your thoughts on the ease or accuracy of processing and analysis for these two sets? Would one set of sources be easier or more accurate than the other? Which one?
5. How many people are mentioning needing or taking insulin? What do they say about the availability and cost of insulin? Are there mentions of the insulin cap for medicare? How many?
6. How many conversations relate to these three topics in reference to diabetes: a) cost b) price/expense c) pharma companies d) government/public health agencies? Describe and document the approach you took to analyze this. NOTE: Think about applying Topic Modeling to extract the various topics being discussed about Diabetes instead of the three topics above.
7. This task is to introduce LLMs for analysis. Use ChatGPT, Claude or any LLM Chat interface of your choice. Your task is to get the LLM to extract the Sentiment towards the Continuous Glucose Monitor (CGM) they are using AND the comorbidities/diseases being mentioned from a few of the "SoundBite" rows (3 to 10). The output should be of the form:

CGM: Values can be: Dexcom, Libre, FreeStyle Libre, etc.
Sentiment: Values can be: Positive, Negative, Neutral, Unknown
Comorbidity: Example values: “High Blood pressure”, Obesity, CardioVascular,... If there are multiple comorbidities mentioned, they should be separated by commas. Multiple words should be enclosed in double quotes.

8. Extra Credit: Write a prompt to use an LLM to input a file (say, a subset of the files in this dataset) and create an output containing the Sentiment towards the Continuous Glucose Monitor (CGM) they are using AND the comorbidities/diseases being mentioned in the format described above and output it in a json or csv file. Essentially, this is the same as the task above except the input data and output results are formatted and in a file.

9. Extract information for the Sentiment Diabetes patients express and feel. You will notice a column named **Sentiment** with the values positive, negative and neutral. These are pre-computed sentiment values by a platform. For those that have programming background, you should compute these yourself using **Vader** and **Stanford** CoreNLP for Sentiment Analysis and compare the results from those mentioned in the data. What percent agree? What is your analysis of the cause of the differences?

10. Extra Credit: Note the columns **Positive Objects? And Negative Objects?**
This attributes the identified sentiment to specific objects/things or brands or companies in the post. This is also known as aspect based sentiment analysis (ABSA). You see these all the time on eCommerce platforms such as Amazon or when you go to any travel booking site. How would you do this yourself? You are not expected to implement this but provide an overview of why it is important, what the challenges are and some potential ways of implementing it.

11. Extra Credit: Advanced Topic Modeling: In the Topic Modeling steps above, you may have come across many different topics/aspects. How would you group or align them into more readily accepted terms in Public Health? Do some secondary research of publications and reports from industry groups (such as the American Diabetes Association, ADA) or government agencies such as NIH or CDC and extract 5 to 10 topics of importance. An example may be the phrase “Disease Management”. Patients are unlikely to use this exact phrase. What terms and phrases do you see in the dataset that might be related to this phrase? Now, use Topic Modeling and other NLP techniques such as Named Entity Recognition (NER) to map and quantify all the topics you extracted using NLP to those from the secondary research. Summarize your approach and findings.

References:

https://towardsdatascience.com/aspect-based-sentiment-analysis-using-spacy-textblob-4c8de3e0d2b9

https://textblob.readthedocs.io/en/dev/