Joseph Prieto

CS 410 Text Information Systems

October 27, 2023

Project Proposal

Title: Medical Device Reviews Data Set Creation for Sentiment Analysis

Team Members:

Joseph Prieto, NETid: jprieto3, Team Captain

Project Type:

Data Set Creation

Project Overview:

The purpose of this project is to create a data set of medical device reviews to be used by sentiment analysis algorithms. Medical devices are such a niche topic that there aren't any data sets readily available to be used to test sentiment analysis algorithms on. This is important because algorithms can be developed to analyze sentiment and then recommend the best medical devices and save lives by being able to test their algorithms on this data set. There are data sets for amazon product reviews, and this may be one of the data sets that is closest to the one I will create but I believe medical device language differs from general products and may yield different results when testing sentiment analysis algorithms.

Data Set Description:

The data set I intend to create is a data set that includes medical devices review for various products. It will contain the text data and any type of ranking such as a star rating or a rating out of 10. The data will come from reviews of the devices on the internet. The data will be formatted in .txt files so that it is easy to read and work with. The size of the data set will be 100,000 lines. This data set will enable the testing of sentiment analysis algorithms in the niche topic of medical devices.

Data Collection:

To collect my data, I plan on using a web scrapper to scrape data from the reviews online of medical devices.

Data Cleaning and Annotation:

After retrieving the raw data from the web, I will use python tools to clean and analyze the data. I will convert the data into a format that is easy to work with using python by cleaning it of any unwanted characters. Also using analytics in python, I will make sure there is a good balance of positive and negative sentiment data.

Expected Outcome:

I plan on creating a robust data set of medical device reviews so that it is readily available for anyone testing sentiment analysis algorithms in this specific field.

Final Deliverables:

I will provide the data set along with some statistics on the data set.