Taylor Imhof

Bellevue University | DSC 680

Project 2 Milestone One (Proposal and Data Selection)

11/02/2022

TOPIC

For this project, I will be conducting some sentiment analysis on open text collected from the popular website reddit.com (specifically the subreddit r/politics). Sentiment analysis is a branch of the machine learning domain natural language processing (NLP). The core technology used in this project is the Python programming language, as there are rich NLP libraries from which I can leverage to perform this type of analysis.

BUSINESS PROBLEM

A company has asked me to create an NLP model that is able to gauge whether the comments on their website are positive or negative in order to understand how their customers are reacting to the products and services they release.

DATASET

In order to obtain the reddit data, I will utilize the Python Reddit API Wrapper (PRAW) package. PRAW is used to collect all kinds of different data from reddit's API that they expose for use by third-party developers. The type of data I will be collecting is the headlines of the posts on the r/politics subreddit.

METHODS

As outlined in the step above, the dataset will first be obtained from reddit's API. The data will have to be read into a Pandas dataframe so that it can be more easily processed using Python. After the data is read into a dataframe, I will perform some exploratory data analysis in order to obtain a better understanding of the underlying data trends. For this, I plan on leveraging data visualization libraries such as Matplotlib and Seaborn. After visualizing my data, I plan on using the Valence Aware Dictionary and sEntiment Reasoner (VADER) model to ascertain whether the headlines are positive, negative, or neutral. I will train the VADER model on the processed headlines and examine the performance of the model on the headlines themselves.

ETHICAL CONSIDERATIONS

Since reddit is open for pretty much anyone to contribute, I will have to be cognizant of the content of the headlines. Some of the posts might be derogatory or contain hate speech. There are moderators of this subreddit, but since I am collecting the data real-time, there is a potential that they have not caught these types of posts in time. Moreover, there still do not exist many robust rules or regulations surrounding the use of people's online data. I will be using the text written by people online and perhaps they would not have consented to me using their posts. As such, their usernames and other data associated with them specifically will not be used during this project.

CHALLENGES/ISSUES

One of the main challenges that I see with this type of project is ensuring that the model is trained properly. Human language is so nuanced that it could be hard for me to pick up on the headline's sentiments, let alone the machine learning model. As such, even after examining and tweaking the model, there is the potential that I have overlooked some glaring mistakes that will cause the model to predict the sentiment of the customer's posts incorrectly.

Another issue is that the PRAW package only allows me to make a call for approximately 1,000 posts on the subreddit. While certainly enough to begin training the model, a dataset with only 1000 observations would be considered relatively small. Perhaps I could run the PRAW to obtain comments on different days and then store them in separate <.csv>’s and later join them into a larger dataset

References

Pandas. (2022). pandas documentation. Retrieved from https://pandas.pydata.org/docs/

Praw. (n.d.). PRAW: The Python Reddit API Wrapper. Retrieved from https://praw.readthedocs.io/en/stable/

Reddit Politics. (n.d.). r/politics [Sub-reddit]. Retrieved from https://www.reddit.com/r/politics/

vaderSentiment. (n.d.). VADER-Sentiment-Analysis [Github Repository]. Retrieved from https://github.com/cjhutto/vaderSentiment