Observations on the Datasets

Difficulty of Rules-Based Classification

The initial takeaway I get from looking at the dataset is that is seems specifically tailored to be difficult for a rules-based system. Many of the non-antisemitic tweets have words typically used by antisemites to express antisemitism and the vocabulary overlaps between the antisemitic tweets and the tweets that simply mention Judaism. There are a few words that appear to only come up in the antisemitic tweets, such as mentions of George Soros, but those words appear in few tweets and even in the case of filtering out mentions of Soros, that would also affect innocent mentions of him.

The files in the counters folder can be used to examine the vocabulary differences between the antisemitic and non-antisemitic texts. The folder is split into four files, one for each type of antisemitism identified in () et al. Each word that appears in a tweet of that type is included in the CSV, next to the percentage of the time it appeared in the tweets of that type subtracted by the percentage of the time it appeared in the non-antisemitic tweets. An initial idea for a rules-based system may use the most comparatively common of these words and classify texts where they appear more than usual (with 'usual' being defined as the baseline identified with the word frequencies in the non-antisemitic tweets), possibly using some sort of hypothesis testing to see if the word frequencies appear to come from the same distribution. The problem with this is that, as exemplified by the appearance of these common words in non-antisemitic tweets in the dataset, the use of these words for classification is not very useful without the context they are used in, which is the primary advantage of a machine learning approach to text classification.