Practical NLP Project Proposal

Reddit user classifier

David Chu (dfc296@nyu.edu), Jasper Duan (zd793@nyu.edu), Richanshu Jha (rj1469@nyu.edu)

The Idea:

What can we tell about a Reddit user based on their posts and comments? Using NLP, we would like to be able to identify a user's subreddit affiliation (ie: /r/democrat or /r/republican) based on an NLP analysis of their post history. Our goal will be to generate a vector of values indicating how strongly associated a user's history is with each subreddit.

We would start with a few subreddits to first develop our model. Once we see positive results, we would generalize the process we used to to build a more generalized model where we can associate users with many subreddits based on the content of their posts.

At the end, we hope to try to answer hypothetical questions such as: which subreddits would Shakespeare have posted to if he had Reddit in his time? Can we know how strongly his writings would associate with each subreddit in the model? Are we able to generate meta-subreddits (ie: by aggregating comments and posts from misinformation campaigns, try to create a /r/misinformationattacker subreddit for a user to associate with)?

Problem Statement:

What a Reddit user says can give a lot of insight about the user. Therefore, given a corpus of a user's entire comment history, what information can we infer? We would like to start by identifying the strength of a user's association with individual subreddits based on their textual similarity to a subreddit.

Data source:

We will require reddit data for this project. We will attempt to retrieve reddit data using the pushshift API. An alternative is to use any of the reddit data dumps that are provided by push shift or those that are readily available on the internet. There may be additional parsing involved with this option, which is why this is our alternate option.

Application of NLP in the project:

We would begin by trying out algorithms such as Naive Bayes, SVM using NLP concepts such as bag of words, tf-idf vectors, ngram vectors, word2vec, etc. This would give us baseline accuracies for our project. We would then attempt to enhance our results by fine tuning neural models, such as LSTM or BERT for the classification, and improving the featureset by using NLTK metrics. If time permits, we would also explore using topic vectors created using topic modelling with LDA.

Evaluation of Methods:

The ground truth will be generated via the reddit data. Given that our labels are going to be the subreddits comments are posted to, we will have labelled data available to us at the extraction stage. We will use standard validation sets for the evaluation of the model.

Using push shift APIs, we can ensure that the generated dataset has data points across multiple years. This is to ensure that we do not overfit for any specific time period or event. We can validate this using our validation set as this will be a sample of our complete dataset, having data points spread across many years. We understand that the data might be time-sensitive and will have to further investigate.