Title: Detecting Political Party of Reddit Users Based On Reddit Posts and Comments

Marek Hulva, Jeremy Glebe

10/9/2020

On my honor as a UNT student, I have neither given nor received unauthorized assistance on this work.

Note regarding unauthorized assistance: I would prefer that you learn as much about this subject as possible through your project. Therefore, I encourage you to talk to everyone and anyone who is willing to discuss (or even just hear about) your project. Feel free to involve collaborators somewhat deeply, but in order to learn, ***you will of course need to do a significant amount of work yourself****. I might expect certain written assignments to be your own work and reflect* ***your*** *understanding of that work. That said, your paper should reflect feedback from your class peer group as well as from me. Please include proper citations for any writing or other intellectual property that came from someone else. Missing citations have to be reported to the office of academic integrity and will have a substantial negative impact on your grade.*

You can either answer the following questions in essentially a QA bullet point format, or if you prefer, you can answer them implicitly in writing similar to the writing in a research paper.

Please replace this example abstract with your own.

Abstract

Content	*Example*
One sentence about the state of the world or prior research.	*Prior research was successful in detecting political party from Twitter posts with around 70% accuracy.*
One sentence about the problems with or deficiencies of that state or research.	*This study was conducted based on Twitter posts and not Reddit posts. Twitter posts have a limitation of 280 characters per tweet, which can be harder to determine the correct political party, because the user is limited with word usage and sometimes has to leave out important keywords.*
Possibly one sentence about solution enablers.	*Using data from Reddit's post and comments will enable us to analyze more keywords in the post and therefore more accurately detect the political party of the users*
One (or *maybe* two) sentence(s) about your proposed or implemented solution.	*We propose a system which, trained on Reddit posts, will analyze and detect the political party of the author of a given text post using key words and phrases.*
One or two sentences about your (anticipated) findings, if at all possible, in the context of a meaningful alternative.	*I anticipate that over 70% accuracy in determining the political party of users from the sample being analyzed*
One sentence about the actual or anticipated impact on the world or field.	*Detection of political affiliation may also serve as a filter for political bias in texts related to current events. It could be used to alert readers of any potential biases the author may be influenced by.*

Introduction

What is the problem you are addressing?

In this project, we are aiming to detect support of political parties of Reddit users based on collecting data from political threads from Reddit.

What very specific claim do you hope to make when your work is complete?

We will show that our analysis can successfully identify the political affiliation of the author of a text post, within some percentage threshold. (80% is a somewhat arbitrary goal going forward)

Who will care about your system?

Political parties and their campaign coordinators.

Why do they care?

This project will benefit for political campaigns of both parties because they will know how many supporters of the opposite parties are on Reddit and therefore they will be able to do specific campaign ad targeting, as well as creating their own content on Reddit to try to convince voters to change their mind.

What do you know about previous attempts to solve this problem? List key publications of which you are aware.*

Publication: Predicting users' political support from their Reddit comment history

Authors: Aaron Acosta, Silviana Ciurea-Ilcus, Michal Wegrzynski

http://cs229.stanford.edu/proj2016/report/AcostaIlcusWegrzynski-Predicting%20user%27s%20political%20support%20from%20their%20Reddit%20comment%20history-report.pdf

Indicate what papers you plan to review as part of your project and why you plan to review them / how they are relevant.*

What have previous results been?*

Data

Describe your data,

Reddit posts and comments.

Have you already collected your data or will it be trivial to collect?

Data collection will be trivial thanks to a data scraping tool called PRAW.

Where will you get the data and is it licensed for public distribution?

As mentioned earlier, we will be using a Python package called PRAW - Python Reddit API Wrapper. PRAW allows for simple access to Reddit"s API and is easy to use and designed to follow all of Reddit's API rules and policies. Therefore we will not have to further worry about any licensing.

How much data do you think you will need\use?

PRAW only allows for scraping the most recent 1000 posts, however that will be enough for the purpose of our project.

Is the data already annotated with class/category?

The dataset obtained from Reddit through PRAW will be structured into "submissions" and "comments." Each row in "submission" will represent a single post, with columns containing "postID", "postTitle" and "postBody" - which will contain the text of the post. Each row in "comments" will represent a single comment in a post, with columns of "commentDate", "user", "body" - which will be again the text of the comment.

What is the class distribution?

This will be determined after our project is finished as it is hard to estimate the class distribution of republicans on reddit and twitter prior to the research.

What attributes/features do you plan to use initially?

Counts of specific words that would represent each political party, like for example for Republicans (Trump, Pence) connected with positive verbs like or phrases like (helps, improves, etc..) to detect promoting of the Republican political party and the same thing done for Democratic political party. These keywords will be extracted from the text of the post.

Have these features already been extracted; if not, what is required to do so?

These features have not been extracted yet. We need to manually make our own research and come up with words that are most likely to be used for promoting each political party and look for the same words in the data that we will collect.

Are there any other key details that I should be aware of regarding the data?

For training, there will be collected from reddit and twitter posts and after that, there will be another set of data collected for testing.

General Approach to Solving the Problem

What is your planned high-level approach to solving the problem?

Supervised ML will be crucial in determining key connections between keywords that would be representing one of two political parties.

What will the specific output of your system be?

The specific output of the system will be to determine whether a user posting comment or thread on twitter is Republican or Democrat. We will get a percentage of Republicans and Democrats from our sample of population.

What learning algorithms do you plan to utilize, enhance, evaluate, etc.

We will be performing supervised learning in which we manually correct the system and help to increase its accuracy during classification. This will ensure that there is a human element which can, at least initially, better identify what posts belong to which political groups.

MapReduce will serve for modeling the data and will help us generate datasets of keywords.

Map function will map the data and show how many times a certain word appeared in the post. The reduce function will then take the results from the map stage and then combines them together.

An additional map function that tokenizes words and then reductions based on patterns of expressions could also be applied given enough time and training.

Experimental Design and Evaluation

What experiments do you *plan* to run – what is the control condition or baseline comparison method?

Experiments will consist primarily of training with known data, then examining results on unknown data, We will cycle between these two phases until we can bring accuracy to an acceptable level. (Our current, somewhat arbitrary, objective is an 80% success rate)

We will divide data into training, tuning and test sets. Posts of users whose political representation will be determined manually will serve as a training data set and the rest of the data will be categorized as test data. In some cases, even if data is not pre-classified we may be able to assume that most posts within a category fit one classification. For instance, posts made in r/democrats are likely written by democrats and likewise for republicans. We will also create our own posts, one set of posts promoting Republican party and one set promoting Democratic party, and test how the machine detects those.

How specifically will you use your training, development and test dataset in your experiments?

The machine will make guesses against the inputs from the training set and those guesses will be corrected by a human supervisor. Training will be typically labeled as Republican/Democrat, in regards to categorizing the political party it aims to promote and support.

How do you intend to ensure that your results are sound - that the claims you make are accurate and that applying your solution to new data of the type your claims are based on will produce similar results?

The application of “gold standard” data items, with obvious/intuitive classifications, will serve to ensure that our results are consistent with the current American political climate. Essentially, we should be able to gather posts from subreddits such as r/democrats for training, but we can also use it to ensure that the system is performing as expected. (When tested with new data from r/democrats, it should identify it as democrat) If the system continues to guess correctly on data which we know the classification of, but which it has not been previously trained with, then we will know that the classifier produces accurate results.

Work Plan

What is your schedule of work? What tasks will you perform; how much time do you think they will take; and when do you plan to complete them?

Currently we are working on acquiring the Reddit data. (Both the retrieval of existing Reddit data and creating a system to collect real time data) Once we have our data collected and stored, we will immediately start working on training cycles and then designing the machine. Data collection should be accomplished by mid-November. It may be difficult, but ideally a prototype system to detect biases will be completed by the end of this term in December.

Summary

Key Points of our project will be to collect relevant data, construct software that is capable to categorize the data into different groups and determine the political preference of users based on their Reddit’s comments and posts.