project-016: I523: Proposal
Data Analysis of Interaction Dynamics for Persuasion

F16-DG-4064, Carlos Sathler
Graduate Student, MS in Data Science
Indiana University at Bloomington
cssathler@gmail.com

ABSTRACT
This is a proposal for the term project for the Fall 2016 Section of the class Big Data Applications and Analytics. We propose to explore the work of a group of researchers from Cornell University published on the paper Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faithe Online Discussions [7]. Our exploration will entail reproducing and commenting on many of their findings, particularly regarding interaction dynamics. Project deliverables will include working code, exploratory data analysis and predictive analytics.

1. INTRODUCTION
In February and March of 2016 a group of scientists from Cornell University made the news [1][8] when they reported the results of a study about persuasion in which they used online data collected from a Reddit community called "Change My View" [4]. The researchers were interested in understanding patterns of interaction and language that ultimately led to one user changing another user's opinion about a given topic. The study behind the news is documented on a paper [7] presented at the 25th International Conference on World Wide Web.

The project we are proposing will use the same dataset these researches used in their study [6]. Our primary goal is to reproduce the original research results around interaction dynamics among online participants in the community, and **possibly** explore some new hypotheses. Since this is a class project that needs to be completed in two months we may not be able to delve deep into analysis of text and language, but we will do so **if time permits**.

When analysing language used in the posts, we are particularly interested in exploring if it will be possible to identify empathic responses as defined by Marshall Rosenberg in his work on Non-Violent Communication [5]. If we are successful in identifying empathic responses through analysis of text, we would like to further determine if these responses are clearly more persuasive. If we are not successful we will document why and offer suggestions for future work.

2. DATASET

2.1 Source
The dataset for this work was released in January of 2016 and distributed with the original research paper that inspired this project [7]. It can be downloaded from one of the author's (Mr. Chenhao Tan) web page [6] at the following url: https://chenhaot.com/pages/changemyview.html

2.2 Data Structure
The dataset used on the project is a collection of online discussions. Each discussion consists of a tree structure where the root node is the initial post containing an opinion from a community user along with a invitation abbreviated by the initials "CMV", as in "Change My View".

First users who respond to the post initiating each discussion will have the option to start a new thread under the original post or offer their opinion under an existing thread created by a prior user. This logic does not change as the tree grows; every user has a chance to expand the tree by either adding a sibling node or a child node, each node being a post.

The physical data on these discussions is captured in the project dataset as a collection of JSON records.

2.3 Definitions
We will borrow some of the definitions from the original research that inspired our project [7, p 616, 617]:

- Discussion Tree - Each discussion originated by a user with the invitation "CMV" (as in "Change My View").
- Original Post (OP) - The post which initiated each discussion tree.
- Root Reply - A direct reply to the Original Post.
- Path - All replies from OP to a given leaf.
- Root Path - A full path from a root reply to a given leaf.
- Challenger - Any users who accepts the invitation in the Original Post and adds a post to the discussion tree.
- Root Challenger - User who posted a Root Reply.

Other terms that are important for the project:

- Comment - Any post which is not the Original Post.
- Reply - Same as Comment.

- Delta - An acknowledgement offered by one user to indicate that another user changed their view.

Note that Deltas can be granted by the users responsible for the Original Post to express a Challenger was successful, but they can also be granted by other users to each other. Additional information on how the Reddit community Change My View [4] works can be found in the community Wiki https://www.reddit.com/r/changemyview/wiki/index.

2.4 Dataset Statistics

The following dataset statistics are reported in [7, p 615]:

- Number of Discussion Trees: 20,626
- Number of nodes: 146,847.533
- Number of unique participants in the discussions: 86,888

The research dataset is actually partitioned into a training dataset and a testing dataset ("heldout"). The above numbers are grand totals. We plan to confirm these counts and generate a few others, especially relative to number of words in the discussions. We will preserve partitions.

Additionally, we will reproduce 4 monthly activity graphs for the training dataset, which are present in the original research [7, p 616]:

1. Number of posts per month
2. Average numboer of replies per post
3. Average number of challengers per post
4. Average delta percentage

The size of the training partition of the dataset is 300.5 MB; the size of the test ("heldout") partition of the dataset is 2.23 GB.

3. INTERACTION DYNAMICS

3.1 Exploratory Data Analysis

We will explore several simple statistics around interaction dynamics in the training dataset, such as trees per user, deltas per tree, deltas per user (given and received), deltas per user per post (given and received) and we will explore the relationship between discussion tree attributes such as size and depth and number of deltas awarded. Finally we will reproduce the following important results of the original research [7, p 616]:

1. Delta % vs. number of unique challengers
2. Delta % based on tree structure

3.2 Predictive Analytics

We will elaborate on the feasibility of performing predictive analytics based on interaction dynamics features, even though that was not attempted in [7, p 615].

One of the important findings from the original research is that "situations where there is more than one reply in a rooted path-unit correspond to a higher chance that the OP will be persuaded" [7, p 617]. We will explore if it makes sense to attempt predictive analytics based on number of root replies.

4. LANGUAGE CONTENT

4.1 Exploratory Data Analysis

We will at a minimum explore statistics involving word counts, since that won't require specific knowledge of natural language processing (NLP).

However, if time permits, we will use the project to learn NLP techniques in Python using resources on the web, such as the tutorial "Working With Text Data" available from the Scikit-Learn website [9] and two Kaggle tutorials on NLP techniques [2][3].

4.2 Predictive Analytics

Time permitting, we will run predictive analytics according to the following prioritization:

1. Word count on discussion thread.
2. Original post content.
3. Content of comments/responses leading up to a Delta.

Number 2 in the list above is one of the subjects explored in the original paper [7]. In their research the authors were able to analyze OP language and predict what posts indicate the author is more susceptible to persuasion [p 620].

5. PROJECT DELIVERABLES

These are the project deliverables:

1. Python code
2. Analysis (Python code output, reproducible)
3. Predictive analytics (Python code output, if time permits - also reproducible)
4. High-level design document
5. Project report (explaining analysis and presenting conclusions)

6. LEARNING OBJECTIVES

These are the project learning objectives:

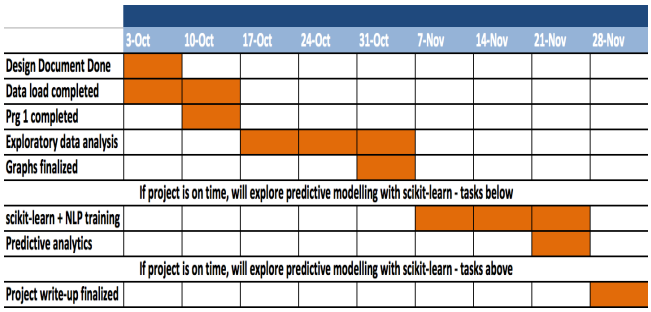
1. Learn enough python to complete the project
2. Learn how to use regex with python

3. Develop a better understanding of how to do NLP with python

4. Learn at least one package/tool to do predictive modelling in Python; perform at least one predictive modelling exercise.

7. PROJECT SCHEDULE

A high level project schedule is presented below.



8. REFERENCES

[1] Ana Swanson, The Washington Post. How to change someone's mind, according to science. Web Page, February 2016.

[2] Bag of Words Meets Bags of Popcorn | Kaggle. Web Page.

[3] Sentiment Analysis on Movie Reviews | Kaggle. Web Page.

[4] Change My View (CMV). Web Page.

[5] M. B. Rosenberg. *Nonviolent Communication: A Language of Life: Life-Changing Tools for Healthy Relationships*. PuddleDancer Press, 2015.

[6] C. Tan. Chenhao Tan's Homepage - changemyview. Web Page.

[7] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[8] Tanya Basu, Science of Us. A subreddit sparked a scientific inquiry into how to change someone's mind. Web Page, February 2016.

[9] Working With Text Data - scikit-learn 0.18 documentation. Web Page.