Capstone Project 2 - Final Report

Problem Statement

In the lead up to the 2016 US Election, and for a short time afterwards, social media users in he US were targeted by a disinformation campaign by a Russian "troll factory," designed to "sow disinformation and discord into American politics via social media."

> Earlier this year, as part of special counsel Robert Mueller's investigation, the Justice Department charged 13 Russian nationals with interfering in American electoral and political processes. The defendants worked for a well-funded "troll factory" called the Internet Research Agency, which had 400 employees, according to one Russian news report. From a bland office building in St. Petersburg, the agency ran a sophisticated and coordinated campaign to sow disinformation and discord into American politics via social media. This often involved Trump's favorite medium: Twitter. (via FiveThirtyEight.com, Why We're Sharing 3 Million Russian Troll Tweets, July 31, 2018)

I will use the Tweets to explore questions about the nature of the disinformation campaign, such as:

- Did the tweets increase in frequency or volume around the time of major events?
- Did other trolls retweet and amplify troll tweets?
- Can clusters be made of Twitter handles/'users' grouped with similar features?
- Can common topics or themes be identified?
- What were the most-used hashtags?
- Did the tweets predominantly support one candidate or political party, or seek to undermine the other?

By exploring the patterns, topics and methods of the disinformation campaign, I will seek to create insight into these efforts and understand how to recognize, identify and potentially avoid future disinformation attacks.

Client

My client for this project is the American voter, and I intend to provide analysis to aid them in discerning manufactured disinformation from "real" opinion and information.

Data

The data I will be using for this project is data that has been made available to the public by Five Thirty Eight, on their GitHub at https://github.com/fivethirtyeight/russian-troll-tweets/ .

The data was originally gathered by two professors at Clemson University; Darren Linvill and Patrick Warren, and shared with FiveThirtyEight.

> Using advanced social media tracking software, they pulled the tweets from thousands of accounts that Twitter has acknowledged as being associated with the IRA. The professors shared their data with FiveThirtyEight in the hope that other researchers, and the broader public, will explore it and share what they find. (via FiveThirtyEight.com, Why We're Sharing 3 Million Russian Troll Tweets, July 31, 2018)

Solution Approach

I plan to use Python data analysis and manipulation techniques to manipulate, aggregate and clean the data. I will then use Natural Language Processing and text mining techniques to extract topics and sentiments, and data visualization libraries and techniques to create visualizations to support my findings.

Deliverables

A GitHub repository will be created, containing this project proposal, as well as the following project deliverables:

- Python Code
- Russian Troll Tweets data set
- Project Presentation paper and slide deck

Github Repo: https://github.com/chudzikr/Capstone2

==========

Milestone Draft

Problem Statement

In the lead up to the 2016 US Election, and for a short time afterwards, social media users in he US were targeted by a disinformation campaign by a Russian "troll factory," designed to "sow disinformation and discord into American politics via social media."

> Earlier this year, as part of special counsel Robert Mueller's investigation, the Justice Department charged 13 Russian nationals with interfering in American electoral and political processes. The defendants worked for a well-funded "troll factory" called the Internet Research Agency, which had 400 employees,

> according to one Russian news report. From a bland office building in St. Petersburg, the agency ran a sophisticated and coordinated campaign to sow disinformation and discord into American politics via social media. This often involved Trump's favorite medium: Twitter. (via FiveThirtyEight.com, Why We're Sharing 3 Million Russian Troll Tweets, July 31, 2018)

I will use the Tweets to explore questions about the nature of the disinformation campaign, such as:

- Did the tweets increase in frequency or volume around the time of major events?
- Did other trolls retweet and amplify troll tweets?
- Can clusters be made of Twitter handles/'users' grouped with similar features?
- Can common topics or themes be identified?
- What were the most-used hashtags?
- Did the tweets predominantly support one candidate or political party, or seek to undermine the other?

By exploring the patterns, topics and methods of the disinformation campaign, I will seek to create insight into these efforts and understand how to recognize, identify and potentially avoid future disinformation attacks.

Client

My client for this project is the American voter, and I intend to provide analysis to aid them in discerning manufactured disinformation from "real" opinion and information.

Data

The data I will be using for this project is data that has been made available to the public by Five Thirty Eight, on their GitHub at https://github.com/fivethirtyeight/russian-troll-tweets/ .

The data was originally gathered by two professors at Clemson University; Darren Linvill and Patrick Warren, and shared with FiveThirtyEight.

> Using advanced social media tracking software, they pulled the tweets from thousands of accounts that Twitter has acknowledged as being associated with the IRA. The professors shared their data with FiveThirtyEight in the hope that other researchers, and the broader public, will explore it and share what they find. (via FiveThirtyEight.com, Why We're Sharing 3 Million Russian Troll Tweets, July 31, 2018)

Data Wrangling

List other potential data sets you could use

Explain your initial findings