Capstone Project 2 - Milestone Report

A Few Months on the Troll Farm

An Analysis of Russian Troll Tweets in the 2016 US Election

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 necessarily broad: the American voter. 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 (NLP) 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 has been created, containing this Project Milestone Report, 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

Data Wrangling

A handy data dictionary was provided in the FiveThirtyEight GitHub repository, which greatly aided in the understanding of the data and structure for Exploratory Data Analysis (EDA):

Header	Definition
external_author_id	An author account ID from Twitter
author	The handle sending the tweet
content	The text of the tweet
region	A region classification, as [determined by Social Studio](https://help.salesforce.com/articleView? id=000199367&type=1)
language	The language of the tweet
publish_date	The date and time the tweet was sent
harvested_date	The date and time the tweet was collected by Social Studio
following	The number of accounts the handle was following at the time of the tweet
followers	The number of followers the handle had at the time of the tweet
updates	The number of “update actions” on the account that authored the tweet, including tweets, retweets and likes
post_type	Indicates if the tweet was a retweet or a quote-tweet
account_type	Specific account theme, as coded by Linvill and Warren
retweet	A binary indicator of whether or not the tweet is a retweet
account_category	General account theme, as coded by Linvill and Warren
new_june_2018	A binary indicator of whether the handle was newly listed in June 2018

Data Cleaning & Data Wrangling

Joining Data
The data is provided in 13 separate CSV files, to deal with GitHub's file upload limit. The first step in the Data Wrangling process is to join these files are joined together into one dataframe for further analysis. This step is only required one time at the start of the analysis, and the full dataset is used from this point forward.

Missing data
An evaluation was conducted of missing data. There were several variables that had missing data. However, after further scrutinizing these variables, only one instance of missing data – an instance of an NA in the 'content' column – was removed. The other missing values in other columns were acceptable, representing an absence of a value, rather than a missing value.

For example, according to the data dictionary, the 'post_type' variable "indicates if the tweet was a retweet or a quote-tweet." This definition indicates that the abasence of a value represents an original tweet, rather than a retweet or a quote-tweet. So missing values in this case are perfectly valid.

Feature Engineering
Engineered a user table aggregating and summarizing each users' total tweets, followers and following counts…

Additional Data Cleaning
In a typical analysis project, I would perform the data cleaning and processing in the EDA phase of the project. However, because data pre-processing for NLP is so tightly coupled with the NLP libraries and the NLP pipeline, the data pre-processing will be conducted in the upcoming NLP phase of the project.

Data Shape
After cleaning the data, the remaining data set contained 20 features and 2.94 million rows.

List other potential data sets you could use

The Russian Troll Twitter data is a historical look at past activity by 'users' later identified as Russian Trolls. Twitter provides an API for Twitter data that could be mined for more current tweets or user activity. For example, an analysis could be done of users that the trolls interacted with (Retweet, Quote, etc.), who are still active.

Additionally, an excellent Github repo has been created by a user 'bet4a' titled russian-troll-tweets-by-author, which has a primary focus separating the tweet author data

from tweet data. It also contains user data from the US Congressional Nov 2018 House Intelligence Committee dataset and the June 2017 Intelligence Committee data set. This represents the user and tweet data in a format that is more easily analyzed, and the Congressional data could enrich the FiveThirtyEight data set.

Initial Findings

Authors
A count of the author/user accounts and total tweets found that the 3 million (2.94 million) troll tweets were created by only 2,843 authors, and that the top 25 most active trolls accounted for 22.5% of all the troll tweets.

Timeline of Tweet Activity
The primary assumption of this data and its analysis is that this troll activity is event-based. By the very nature of the data and it's analysis (i.e., "An Analysis of Russian Troll Tweets in the 2016 US Election") there is an assumption that these tweets represent activity around an event – the 2016 US presidential election.

Evaluating the dates for the tweet data, we see that the data set contains data from 2012-02-02 - 2018-05-30. While this is useful to identify how early the Russian trolls established their accounts in the lead-up to the election cycle, and how long after the election they remained in place, it is helpful to narrow the time window to focus on the election cycle. For the purpose of visualizing the election cycle timeline and the trolls interactions with the election cycle, the time window was narrowed to June 2015 to January 2018. These dates represent the first and last major periods of activity by the trolls: June 2015 represents when then candidate Trump announced his candidacy, and January 2018 represents when the troll accounts were shut down by Twitter.

A timeline was then created for this time period, and then annotated with labels of major events in the election cycle, and how they coincided with increased activity by the trolls. This view is particularly revealing of the trolls tactics to sow confusion and discord in the national discussion of the Presidential campaign.

The GitHub Repository for this project can be found at
https://github.com/chudzikr/Capstone2 .