VA

by Ruth Wetters

Submission date: 10-Jan-2021 11:55PM (UTC+0000)
Submission ID: 141479348
File name: 189635_Ruth_Wetters_VA_1673889_935696997.docx (480.67K)
Word count: 2897
Character count: 15493

Measuring the Impact of Russian Troll Tweets on the 2016 US Election

Ruth Wetters

Abstract—In this paper I use data provided by FiveThirtyEight to investigate Russian troll activity prior to the 2016 US election. I will identify some key characteristics, focusing on polarising activity and identifying key topics. I will use sentiment analysis to look for correlation between Russian troll activity and approval ratings of Hillary Clinton and Donald Trump.

1 Problem Statement

Disinformation is one of the biggest threats to governance in recent history. Social media has contributed significantly to this change, with 97% of people using social media, and 1 in 4 using social media as their main news source. Much has been written about the formation of intellectually homogeneous 'echo chambers' and their contribution to political polarisation and the growth of populism worldwide.

Although disinformation can come from anywhere, there is a significant subset which is produced and disseminated by state actors. The most high profile of these is Russian 'troll farms' which are widely credited with influencing the outcome of the 2016 US election.

In 2018, Twitter released a dataset of over 9 million tweets from 3,800 accounts that it said were linked with the Russian Internet Research Agency (IRA). Of these, 3 million were re-released by the election forecasting company FiveThirtyEight in collaboration with researchers Linvill and Warren from Clemson University.

I will attempt to investigate whether the main aim of Russian troll activity was to disrupt the 2016 US election, and whether it had a measurable impact on the election. Firstly, I will analyse the corpus of troll tweets to identify trends, especially in the USA. I will use topic modelling and sentiment analysis to further elucidate these trends. Then, I will try to isolate tweets relating to the US election, and compare them to aggregate candidate approval polls in 2016 to investigate correlation.

2 State of the Art

The paper that most informed my approach to analysis was 'Social Media Visual Analytics' by Chen, Lin and Yuan [1], which gave a state of the art for social media analytics, improving on the original social media state-of-the-art descriptions by Schrenk et al (2013) and Wu et al (2010). It divided social media analytics into three main sections: network visualisation, spatio-temporal data and text visualisation. Of these, each is further divided into three subsections. The first is divided into follower networks, diffusion (messaging) networks and reposting networks, which will be of limited use in this study, since the data does not include details of networks. The second category is divided into geographic, spatial temporal event analysis, and movement analysis. Of these, spatial temporal event detection is most useful, since I am trying to identify anomalies in troll behaviour. Finally, the third category, text classification, covers keywords, topic, and sentiment. This is the most pertinent to this data, as the most important variable in the dataset is tweet content. Chen, Lin and Yuan argue that words and topics are distinct and both have value in understanding meaning. Topics can be considered in terms of hierarchy or interaction, and this paper is especially useful for understanding the interplay between topic cooperation and competition, or 'coopetition'. They point out that word clouds are flawed in topic detection, and must be considered alongside other methods.

Chen, Lin and Yuan refer to Dou et al.'s definition of an event for detection purposes: user, topic, time, and location. Combined with the aforementioned specific topic visualisation, I have used this definition in event detection.

Secondly, I read Kerren, Kucher and Paradis' 'State of the Art in Sentiment Visualisation' [2], which describes 132 peer-reviewed techniques. I used this paper mostly to evaluate the drawbacks of my models, due to time limitations.

Finally, this dataset was originally used in the paper 'Troll Factories' by Linvill and Warren [4]. They categorised the accounts by their purpose into the categories of: LeftTroll, RightTroll, NewsFeed, HashtagGamer and Fearmonger. They explain how they have derived these categories and give some characteristics of each type, which was useful for an initial overview. They went into a small amount of detail on the events they detected, but these were mostly in service of account categorisation. This analysis is useful for future researchers but of itself is not sufficient to merit congratulations.

3 Properties of the Data

The dataset consists of three million tweets identified by Twitter as being associated with the Russian IRA, from 2,848 accounts, active between 2012 and 2018. The majority of these tweets are from 2015-2017, when the accounts became much more active. Twitter deleted the tweets and suspended the accounts permanently after they were identified as state-

sponsored accounts, but the researchers at Clemson University were able to trace back to the original tweets so have provided a field linking to the original URL (which now shows 'item unavailable'). There are a small number of null values for content, and more null values in the 'tco' columns.

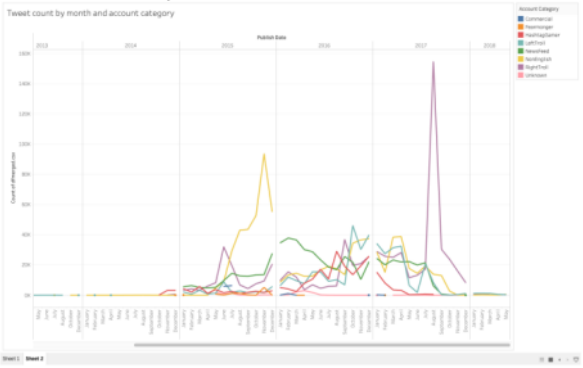


Fig. 1 tweets by month and account category

In terms of the characteristics of the data, the important fields are author, publish date, content, region, and account category. Publish date has a precise timestamp, but the region is limited to country-level data, and the second-biggest category is 'Not Defined'. Linvill and Warren also collected 'local' tweets by scraping for specific locations. The dataset is large, although there are some imbalances: the most active account tweeted 160,000 times, compared to a median of 4,000.

As Linvill and Warren explained, they categorised accounts into several categories, of which the largest are 'RightTroll','NonEnglish','LeftTroll','HashtagGamer', and 'NewsFeed', in that order. These accounts behave differently and tweet on different subjects in different styles from different locations.

Many previous papers have used node-link visualisations, but this dataset does not include interactions between accounts. It would be interesting to run a second experiment visualising the networks of people associated with these accounts: Twitter announced that they had contacted 1.4 million people who had been in contact with these accounts, so it is reasonably safe to assume that although there may have been interplay between accounts, the vast majority of interactions would have been with legitimate Twitter users.

As for the opinion poll data, after having researched many different pollsters, I decided to use FiveThirtyEight's aggregate opinion polls for 2016, which give a score to each pollster and an adjusted score based on their weighting. FiveThirtyEight has been criticised in recent years for its inaccurate predictions, which are based on weight-adjusted estimates as they evaluate other pollsters, but in 2016 it gave a much higher likelihood of Trump winning than any other outlet. The use of polls is also flawed as a metric, as they were widely criticised following the 2016 election for failing to predict the actual outcome. However, even though the actual values are inaccurate, they can still show trends in public opinion.

Fig 2: wordcloud of original dataset content

4 Analysis

4.1 Approach

I started with some preprocessing and exploratory data analysis (EDA) using Tableau. The EDA has a high level of human interaction, since I need to focus on my question and the most suitable way to use the data to answer it. I supplemented this with Python, particularly focusing on text attributes. I created two smaller datasets on the topics of Clinton and Trump from the main data, in order to conduct comparative analysis.

I also inspected the FiveThirtyEight polling data, which provides a raw and adjusted poll result; I decided to use the adjusted result, since the adjustments were slight and likely to be more reflective of public opinion.

After isolating the useful data, I used topic modelling to discover the most commonly discussed topics in relation to the election. I chose to use Latent Dirichlet Allocation (LDA) as it is effective for large datasets and easily interpretable. I also responded to changes, conducting this on the 'before' dataset and then adapting it as appropriate.

Once I had identified topics, it gave me some idea of the methods used: on the right, heavy identification with pro-Trump right-wing media, and on the left, topics including police brutality and the Black Lives Matter movement.

I then used the VADER (Valence Aware Dictionary for Sentiment Reasoning) tool in Python to generate sentiment scores for each tweet in my two corpuses, using visual inspection to check the results. I chose VADER because of its broadness and sensitivity to polarity.

Finally, I used Pearson's correlation to investigate the correlation between sentiment and approval, with the hypothesis that negative sentiment was likely to be tied to a decrease in approval rating. I used Pearson's correlation

assuming a linear relationship.

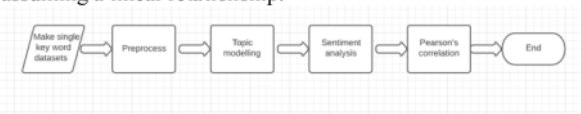


Fig. 3. Workflow diagram

4.2 Process

First I dropped around half of the columns from the original dataset, most of which were administrative: tco, URLs, date harvested etc. I identified the most relevant data, removed columns, removed null values, preprocessed word attributes etc. I decided not to do feature engineering as the provided features proved sufficient. I started with EDA, viewing the tweet number by date, account category, and author. I initially wondered if it would be feasible to look at UK tweets, but on inspection found that the majority of these tweets were in the NewsFeed category, so dismissed this idea.

From the main dataset, I pulled all tweets containing 'Hillary' or 'Clinton' into one dataset, and all tweets containing 'Donald' or 'Trump' into a second. Interestingly, Hillary was known more often by her first name, and Trump by his surname. I then conducted all further analysis in parallel using the two datasets.

I preprocessed tweets content by normalizing, tokenizing, removing stopwords etc in preparation for text processing, and built a wordcloud to visualise. After viewing the wordcloud, it was clear that some of the most common words were 'tco','https', and 'http' so I removed these.

EDA revealed that although there were peaks in troll activity in June 2016, February 2016 and October 2016, these continued after the election, with later peaks in December 2016 followed by the biggest spike, in August 2017. I then used Python and my own knowledge to investigate these peaks: the August 2017 one was the day of the far-right Charlottesville rally, during which one woman was killed, and the spike was particularly represented by RightTrolls.

I then used LDA to visualise topics, to see where Russian trolls are focusing their energies. I adapted the LDA to remove the words 'tco','https' and 'http' as these were all in the 30 most common terms. I also found that many of the most used terms were in Russian, so I dropped all tweets in the account category 'NonEnglish' which removed 1 million tweets. LDA also shows correlation between topics.

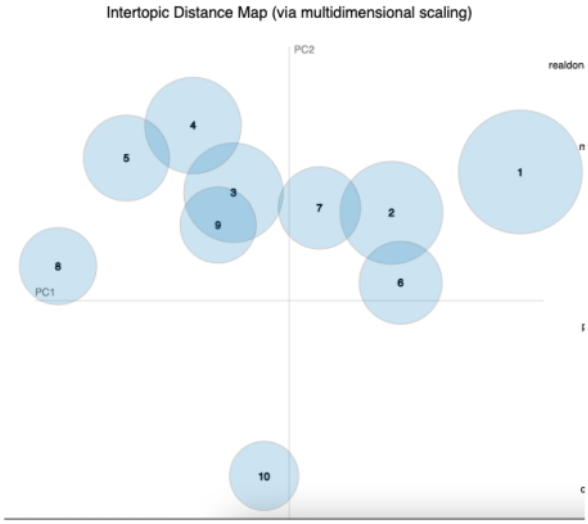


Fig. 4: Intertopic Distance Map for Trump.

In fact, LDA showed that the vast majority of topics were closely interrelated, more so for Clinton than Trump. Trump's LDA showed clear topics such as 'McDonald's' and 'Korea', whereas 8/10 of Clinton's showed high frequency of 'emails' 'Benghazi' and 'Obama'. This could be purely because Trump has always been an outrage candidate, whereas Clinton was an establishment candidate and so remained more consistent for the majority of her campaign.

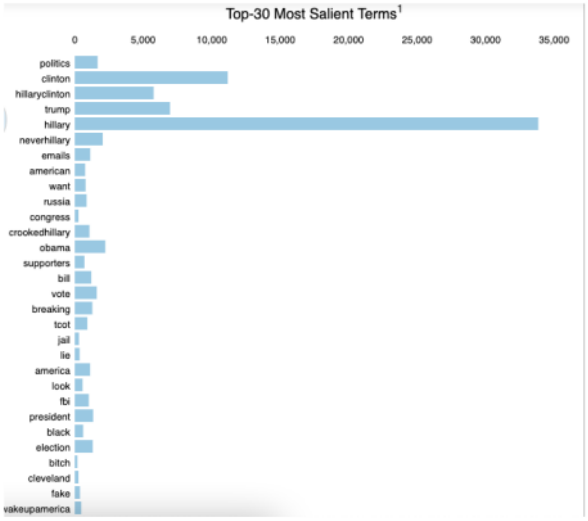


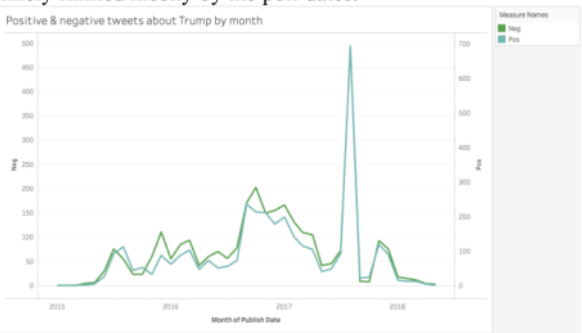
Fig. 5 Most salient topics for Clinton

I then generated sentiment analysis scores for each tweet using VADER. Russian trolls are often credited with polarising discourse, so I wanted to see whether most of them were negative or not. VADER generates four metrics: positive, negative, neutral, and compound, the latter of which is a measure of polarity. All these are generated on a scale of 0 to 1, which was extremely useful for my analysis. I did find

that tweets were more likely to be positive or negative than others. Interestingly,

I then wanted to know whether tweets were correlated with changes in election polling, so I did a Pearson's correlation on tweets and election polling, merged on date. This is imperfect as some polls are taken over long periods, so a visual analysis is also necessary.

When making the Pearson correlation, I needed an equally sized matrix, so converted the publish_date from the tweets and the enddate of the polls to datetime in Python and merged. I got around 700 results for Clinton, and 690 for Trump, which was likely limited mostly by the poll dates.



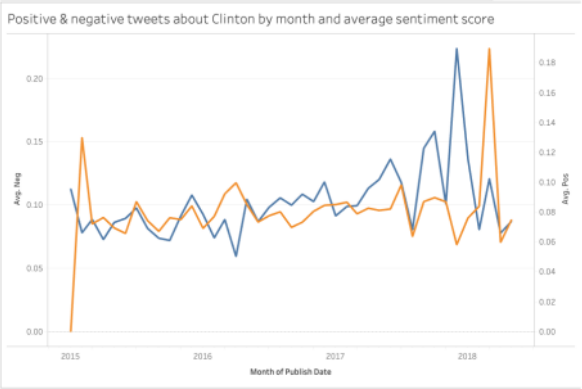


Fig. 6 negative tweets about Trump by month (sum)

Fig. 7 negative tweets about Clinton by month and average sentiment score.

Most interestingly, both negative and positive tweets peaked on the same days for Clinton and Trump, suggesting a general polarisation of discourse rather than significant support for either candidate.

4.3 Results

From my final result, it can be seen that the Russian trolls did tweet about Clinton and Trump in the runup to the election, but the findings are not robust enough to be able to say that Russian trolls influenced the election. More broadly, the trolls tweeted in ways that could be considered divisive political rhetoric, but again this is not statistically significant enough to be relevant. Furthermore, the most activity for both candidates was actually found after the 2016 election, which suggests that it was far from the main aim of the IRA.

The Pearson correlation for Clinton tweets was (0.037714943 08177879, 0.29466670188706423); for Trump it was (-0.084 94731929086179, 0.025656483906285554).

A p-value of less than 0.05 for the Trump tweets indicates a statistically significant negative correlation between the sentiment analysis of Trump and his opinion polling. Meanwhile, the Clinton data shows a statistically insignificant positive correlation, so is not useful here.

In conclusion, there is not enough evidence to support the hypothesis that Russian disinformation campaigns influenced the 2016 election. At least concerning Twitter activity, it is likely that the IRA did not influence the final result.

5 Critical Reflection

There are many aspects which were not ideal in optimising my workflow. Firstly, I realise that calling only tweets with the keywords 'Hillary', 'Clinton' and 'Donald','Trump' will probably have left out a proportion of tweets discussing the election. For example, those with a negative view of Trump have in the past called him 'Drumpf', which could potentially have affected results, resulting in a more balanced sentiment analysis. By widening these criteria, it is possible that my results would have been affected. Secondly, I could have incorporated the tweet authors, as some give an insight into their content, for example 'COVFEFENATIONUS' is likely to be tweeting about Trump.

Thirdly, although LDA is useful in visualising topics, it has some drawbacks. It is static, so cannot change over time, which would have been useful in this analysis. The number of topics is fixed, which reduces adaptability. It is also hard to evaluate compared to other algorithms. Its interpretability and probabilistic nature made it useful here, but my results make it clear there is a lot of overlap between topics, and in fact most of the tweets are expressing more or less the same sentiment on the same topic.

Additionally, I used Pearson's correlation to measure correlation between sentiment and approval, assuming a linear relationship, but this could have been inaccurate.

Fourthly, my training set for sentiment analysis could be improved on in future. I used the VADER corpus, but there are some terms that could be specified as negative or positive for this dataset, for example 'emails' and 'FBI' are likely to have negative connotations in this specific context, although they do not actually carry a negative meaning. I could not find a specific training corpus for US politics, but if this existed, it would improve results.

Finally, there are often flaws in datasets. In this case, it is not possible to visualise the networks of trolls, so it is impossible to know whether their accounts are linked. We also do not have information about the 'updates': these could have been retweeted by bots and had very little influence, especially since Twitter estimates that 500 million tweets are sent every day. This also applies to the polling data: I chose to use a mean average, since it seemed that the data points were close together, but it is possible that this created skewed results. Most frustratingly, I had to use a smaller dataset due to hardware limitations, which will undoubtedly have skewed my results. I would like to repeat this investigation in future with a larger corpus.

Table of word counts

Problem statement	237
State of the art	395
Properties of the data	420
Analysis: Approach	288
Analysis: Process	622
Analysis: Results	186
Critical reflection	427

REFERENCES

[1] S. Chen, L. Lin, and X. Yuan, 'Social Media Visual Analytics', *Computer Graphics Forum*, vol. 36, no. 3, pp. 563–587, 2017, doi: https://doi.org/10.1111/cgf.13211.

[2] K. Kucher, C. Paradis, and A. Kerren, 'The State of the Art in Sentiment Visualization', *Computer Graphics Forum*, vol. 37, no. 1, pp. 71–96, 2018, doi: https://doi.org/10.1111/cgf.13217.

[3] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski, 'The State-of-the-Art in Predictive Visual Analytics', *Computer Graphics Forum*, vol. 36, no. 3, pp. 539–562, 2017, doi: https://doi.org/10.1111/cgf.13210.

[4] D. L. Linvill, 'Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building'.

[5] R. Python, 'Invalid Syntax in Python: Common Reasons for SyntaxError – Real Python'. https://realpython.com/invalid-syntax-python/ (accessed Jan. 10, 2021).

[6] 'Visual Text Analysis in Digital Humanities - Jänicke - 2017 - Computer Graphics Forum - Wiley Online Library'. https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.12873 (accessed Jan. 10, 2021).

[7] N. Willems, H. van de Wetering, J.J. van Wijk. Visualization of vessel movements. *Computer Graphics Forum*, 28(3): 959-966, Jun. 2009.

VA

ORIGINALITY REPORT

8%	6%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to City University Student Paper	2%
2	Janis Peksa. "Autonomous Open Data Prediction Framework", 2019 IEEE 7th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), 2019 Publication	1%
3	scholarspace.manoa.hawaii.edu Internet Source	1%
4	onlinelibrary.wiley.com Internet Source	1%
5	openaccess.city.ac.uk Internet Source	1%
6	smappnyu.org Internet Source	1%
7	Submitted to De Montfort University Student Paper	1%

8	Siming Chen, Lijing Lin, Xiaoru Yuan. "Social Media Visual Analytics", Computer Graphics Forum, 2017 Publication	<1 %
9	vis.pku.edu.cn Internet Source	<1 %
10	Natalia Andrienko, Gennady Andrienko, Georg Fuchs, Aidan Slingsby, Cagatay Turkay, Stefan Wrobel. "Visual Analytics for Data Scientists", Springer Science and Business Media LLC, 2020 Publication	<1 %
11	www.intechopen.com Internet Source	<1 %
12	Kostiantyn Kucher, Rafael M. Martins, Carita Paradis, Andreas Kerren. "StanceVis Prime: visual analysis of sentiment and stance in social media texts", Journal of Visualization, 2020 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		