What are the Trolls up to?

Jordi Serra[a], Sebastian Wolf[a], Felix Adam[a]

[a] *Barcelona Graduate School of Economics, Barcelona, Spain*

Abstract

Abstract to be written here.

Keywords: Textmining, Russian Interference, Polarization

1. Introduction

We analyze russian inteligence agency efforts on social media / twitter

Which strategies are used, which goals are being followed

Focus on polarization

Discuss the goals of social media campaigns from a geopolitical and strategic point of view, with an eye on polarization

We then analyse polarization from a social sciences point of view

One of the hard questions is how to counteract against these measures, so first goal should be to identify polarizing tweets

if one can identify these, could help for targeting them from a technical point of view, but also for further analysis, like building polarization index

we make use of a large scale tweet database, whose users are thought to be related to Russian Intelligence agencies

use textmining techniques to uncover targeted efforts to polarize the discussion

Email addresses: jordi.morera@barcelonagse.eu (Jordi Serra), sebastian.wolf@barcelonagse.eu (Sebastian Wolf), felix.adam@barcelonagse.eu (Felix Adam)

our findings show that...

the rest of this report is structured as follows

2. The Strategy behind Russian Social Media Efforts

In order to identify efforts to polarize US politics through social media, we need to understand the strategies behind such efforts. In particular, it is crucial to know why Russia would use such strategies, and how they could potentially influence domestic and foreign politics in the US.

Unlike the predicted "End of History", (**?**) conflicts between liberal democracies in the west and Russia reemerged in the beginning of the 21st century. Following the fall of the Soviet Union, instead of integrating into the West, Russia tried to reestablish it's position as a superpower. Issues like the NATO integration of eastern European countries or the establishment of a missile defense shield in easter Europe lead to rising tensions and mistrust of Russian leaders towards the West (sources). These tensions became more pronounced in the beginning of the second decade of the 21st century, with hot conflicts in the middle East, most notably Syria, and the Maidan Revolution in Ukraine. At the heart of these conflicts were and still are diverging interests of the United States and Russia. Be it the integration of the Ukraine into the European Union (EU) and NATO or the Russian support for the Syrian leader Assad. These points of conflict can be seen as a motivator for the Russian interference in the 2016 US presidential elections.

In the aftermath of the 2016 elections, the US intelligence community published a number of assessments of the Russian strategy. Most notably, the Central Intelligence Agency (CIA), the Federal Bureau of Investigation (FBI) and the National Security Agency (NSA) published an assessment of Russian activities, stating that the goals of Russia were to weaken public faith in the US democratic process and denigrate presidential candidate Hillary Clinton (**?**). The idea behind these efforts was to undermine the US-led liberal democratic order, posing a threat to Putin's regime. In particular, Putin saw a potential presidency of Hillary Clinton as a threat to his ambitions in Ukraine and Syria, due to Clinton's foreign policy positions as Secretary of State. Trump on the other hand was seen more friendly towards Russia and thus favoured over Clinton. As part of these

efforts, Russia used social media networks like Facebook and Twitter, to promote radical discontent with US politics, polarize the discussion and denigrate Hillary Clinton. The report further states, that the Russian strategy changed over time. In the beginning, the goal was to undermine public institutions. However, with Clinton leading over Trump, the Russian efforts shifted towards the defamation of Clinton, trying to harm her electability and potential presidency. Interestingly, the report states that after Trump had won, the efforts to undermine public institutions stopped. The three intelligence agencies pick out the so called Internet Research Agency (IRA), as one of the main sources of these social media accounts. The IRA is described as a private agency with ties to the Russian government, engaging in targeted social influence efforts. The findings of the so-called Mueller investigation (**?**) support this assessment. According to the Mueller Report, the IRA carried out social media campaigns to amplify social discord in the US. In order to do so, Twitter accounts linked to the IRA tweeted on divisive US political and social issues, such as illegal immigration and racial injustice. According to **?**, the IRA used social media accounts in two manners. Some accounts were designed around fictitious US personas, posting original content on divisive topics, promoting radical ideas and denigrating or promoting political candidates. Other accounts weren't used for original content, but rather to promote and amplify the impact of the "original" content. Some of the accounts posting such content had a large follower share, such as TEN_GOP, an account pretending to be related to the Tennessee Republican Party. TEN_GOP was clearly used to promote polarization by pushing extreme content and promoting Donald Trump. Some tweets included:

- "White girl burned alive by Black gang members They should pay! Why media remains silent?"
- "Wake up America before it's too late! Europe has already lost its chance! #Ban-Islam #StopIslam #filibuster"
- "Donald Trump: "I will be the greatest jobs-producing president that God ever created"

The strategy clearly worked, since tweets of the IRA we're picked up by major news outlets in the US and thus shaped at least daily discussions (**?**). We can thus summarize

the Russian strategy as follows: Polarize the political discussion in the US, undermine Clintons authority and electability and promote Trump.

Now the question remains how these strategies, if successful, would affect US foreign policy, in particular towards Russia. In the subsequent analysis, we'll focus on the idea of polarization. **?** define polarization as the simultaneous presence of opposing or conflicting principles, tendencies or points of view. In a quantitative manner, polarization can be seen as an increase of variance of ideas and attitudes towards poltical questions. **?** argue, that this reduces the probability of group formation at the center of the political spectrum and increases the formation of groups with irreconcilable preferences. This can have peculiar effects on domestic and foreign policy. **?** argues that polarization leads to a weakened international position of the US. First, international endeavours such as the promotion of trade agreements, UN resolutions or even military campaigns crucially depend on domestic support. Without domestic support, international allies of the US discount promises, while the US appears weak towards enemies (**?**). **?** supports these findings. He argues that domestic political polarization leads to three issues. First, it is more difficult to get bipartisan support for risky undertakings. Second, it gets harder to agree on lessons from failures, complicating efforts to learn. And lastly, the risk of dramatic policy swings complicates the ability to make long term commitments. The overall effect of polarization can thus be summarised as follows: A polarized society is caught up with fighting against itself. It can't reconcile large differences to find a common foreign policy strategy.

Consequently, Russia's global standing would greatly increase from a polarization of US politics. Actually, we can already see these effects in motion, especially in Syria. The US failed to gather international support for UN resolutions and seems to have no clear strategy, while Russia continues pushing its ally Assad (**?**). Interestingly, the other strategies described by Martin and Shapiro, such as defamation and persuasion can also be seen as tools of polarization.

Having discussed the effects of polarization, we can now try to identify Russian polarization efforts through social media.

3. Methodology

Informally, polarization in society can be defined as a situation where we have high within-group similarity, and low across-group similarity. More formal measures have been proposed, for example by **?**, that relate specific characteristics of density distributions to polarization, and provide a more thorough treatment of the concept. But for our purposes the informal definition shall suffice. More importantly, we note the close relationship between polarization and clustering techniques in the field of unsupervised learning, or community detection in the field of network science. Clustering techniques in the field of unsupervised learning aim to detect data points that naturally belong together because they are close by some measure in some higher dimensional space, and that are far from other data points by that same measure. Similarly, in network science we try to detect communities of nodes that display high similarity in terms of their links. Nodes that form a community share many common friends, and they share very few common friends with nodes outside of their community. Essentially, polarization is the social science term to describe clustering of data points. This means that to detect polarization we can employ unsupervised learning methods to detect clusters, or network science tools to detect communities.

In this report we aim to make use of this insight by attempting to identify clusters in the body of tweets which could provide evidence of Russian efforts to polarize the American society ahead of the 2016 presidential election. A priori, the most natural clusters we would expect to emerge are two: one cluster pulling the debate towards the extreme left, and another cluster pulling the debate towards the extreme right. However, this would not be the only type of clustering that would yield evidence for polarization. Russian polarization efforts could also have been organized along topic lines, creating two or more clusters within a chosen set of topics. Further, any effort to polarize the debate would have to be hidden among a cloud of noise, in order to conceal orchestrated nature of the effort, and to appear like a body of real tweets to the American twitter users. Therefore, we do not expect the clusters to yield perfect class separation, or yield very clear separation of opinions on topics. In support of our hypothesis we would expect that compared to a non-orchestrated discussion on Twitter, the IRA tweet corpus shows more pronounced clustering. Similarly, the level of clustering should be comparable to what are known

to be very polarized debates, for example the discussion among fans ahead of a football match of two rival teams.

To measure the level of clustering, we propose to use the average cosine distance between tweets in the two clusters. Other measures such as the distance between the centroids, or the distance between certain quartiles could also be considered, but we restrict ourselves to the average distance as this captures the across-group separation very well, giving the same weight to all tweets. Since our scope in this report is limited, we attempt to argue for polarization without comparing the average cosine distance we find in the IRA tweet corpus to other corpora, even though that would have been our preferred approach had we had more time. Instead, we check how close the average cosine distance is to the maximum theoretically possible distance (the cosine distance ranges from 1 to -1, with 0 being the largest distance).

The space that we measure the cosine distance in can be defined by any projection of the corpus of tweets onto a vector space, such as a simple term frequency encoding. For this report we use the Word2Vec package to embed our tweets based on the combination of words used in the tweets. Word2Vec uses a simple two layer neural network that is trained to predict the context a certain word, or a combination of words is most likely to appear in. We train Word2Vec using the full corpus, and then use the vector representations of tweets predicted by the model to measure cosine similarity.

4. Data

Discuss the data here and how it can help us to identify polarization efforts.

To identify efforts by Russia to polarize public opinion on social media ahead of the 2016 US presidential election, we make use of a dataset of X tweets linked to the Internet Research Agency and made publicly available by Twitter covering the time period XX-XX.

5. Results

6. Discussion