A Structural Approach for Classifying Russian Social Media Narratives in the Baltic Region

Timur Chabuk[1] and Adam Jonas[2] and Sue Kase[3]

[1] Perceptronics Solutions, Falls Church VA, USA
[2] Network Engagement Team, Training and Doctrine Command (TRADOC), Ft Eustis, VA
[3] Computational and Information Sciences Directorate, CCDC Army Research Laboratory
timc@percsolutions.com

Abstract. The social media revolution of the past decade has made information operations (IO) more accessible than any time in history for both state and non-state actors. In order to effectively and efficiently identify and counter information operations, capabilities to detect and understand unfolding political narratives are essential. We compare two methods for classifying discovered social media narratives as applied to Russian-language Twitter data pertaining to the Baltic region. In the first method, narratives are manually classified based solely on content. Challenges to performing this classification are discussed with real-world narratives. In the second method, social network analysis and community detection are used to discover distinct groups within the social media data. These groups are then manually coded as being pro- or anti-Kremlin, and then the narratives are automatically labeled based on which groups are primarily contributing content to those narratives. We compare the results of the manual classification of narratives to the automatic classification of narratives based on group participation in a political influence campaign. The automated classification successfully confirmed all classifications that were performed manually; classified narratives that manual coding found ambiguous; and highlighted additional potentially relevant narratives. Collectively, these findings suggest that results derived from structural analysis and community detection of political social media content may represent a more efficient and accurate way to automate rather narrative discovery than using content alone.

Keywords: Social Media, Narrative Classification methods, Social Network Analysis.

1 Introduction

State and non-state actors around the world are increasingly carrying out online information operations (IO) to shape public opinion and influence world events. Full spectrum IO involves coordinated action across a diverse set of different mediums ranging from cyber-attacks, to press releases, to simple sponsored advertisements in magazines. Social media is a particularly important medium in which IO is conducted

as modern social media platforms offer unprecedented opportunities for IO to reach large numbers of people with messaging that can be micro-targeted to individual users and groups. Recent events including Russian interference in the United States Presidential election (Bessi, 2016), Russian IO conducted in support of military operations in Ukraine (Jaitner and Geers, 2015), ISIS recruitment and radicalization through social media, and Iranian influence operations targeting the American public, have drawn public attention to this issue. Recent studies suggest that numerous countries around the world actively participate in manipulating social media space for political gains (Bradshaw, 2017).

The social media information environment is challenging to make sense of because it is complex, noisy, and constantly changing and evolving. Social media is shaped by a wide range of actors often with diverse objectives, strategies, and tactics. Narrative is an account of a sequence of events in the order in which they occurred so as to make a point (Labov and Waletsky 1967; Polletta & Chen 2012). Narratives and counter-narratives are created, amplified, and countered continuously and quickly in response to unfolding events. These stories can be told and capitalized by states and social movement actors to persuade or change people's opinion challenging the status quo (Polletta & Chen 2012). Various groups and communities respond to narratives in different ways further shaping the information environment through their likes, shares, and comments. Effectively monitoring the social media space to detect and understand ongoing IO requires being able to identify these different actors, groups, and communities, and their narratives as they change and evolve in real-time. It is essential the United States and its allies develop methodologies for detecting, analyzing, and monitoring social media IO and the associated background narratives (Chabuk and Jonas, 2018).

The next section of the paper briefly overviews the Baltic region previously established as a population targeted by Russian IO (Radin, 2017; Standish, 2017). Then, collection of a Russian-language social media dataset focused on the Baltic region is described. Section 3 discusses two different approaches for classifying the Baltic region dataset narratives as either pro- or anti-Kremlin. The concluding section reviews key findings derived from the pro- versus anti-Kremlin analysis while identifying important directions for future research and technology development.

2 Baltic Region Social Media Data Collection

As an example of how national and political leaders can use social media to create and counter-narratives to influence the views of a target population, a Russian-language Twitter dataset was collected for examining indications of Russian IO in the Baltic social media space.

2.1 Baltic Region Background

The Baltic region is a hotbed of tension between Russia and the West (Chivvis, 2015; Noack, 2017). Historically, the countries of Latvia, Estonia, and Lithuania were all part of the Soviet Union. After the fall of the Soviet Union in the early 1990s, these countries all gained independence and ultimately became members of the North Atlantic Treaty Organization (NATO). Russia has attempted to reassert their influence in these countries and undermine their membership in NATO. In the Baltic region, it is widely believed that Russia advances political influence by constructing narratives in social media space that are designed to reinforce Russian-speaking people's identification with Russia and undermine Baltic nation citizen's confidence in their own country (Jaitner and Mattsson, 2015). These narratives seek to drive a wedge between Russian-speaking people and their host nations, promote distrust of the West alleging that the West betrayed the former Soviet nations in the 1990's, and erode trust in the host nation governments through allegations of corruption and incompetence (Helmus, 2018).

2.2 Data Collection and Overview

The social media data collection for the Baltic region was performed using the web-based OssaLabs analysis platform (Chabuk, 2019). OssaLabs was developed specifically to help analyst users identify, analyze, and monitor social media IO. The platform's development has been funded by Small Business Innovative Research funding from the Defense Advanced Research Projects Agency, Office of Naval Research, Army Research Laboratory, and other Department of Defense agencies. Currently, OssaLabs developers are collaborating with the US Army Training and Doctrine Command Intelligence Directorate (TRADOC G2) to transition the analysis platform to the Army.

Russian-language social media data relevant to the Baltic region was collected from Twitter using OssaLabs by targeting high-level search terms related to Lithuania such as the Russian equivalents of English words "Lithuania," "Lithuanian," "Klaipeda," and "Vilnius." Similar data collection was performed for Estonia and Latvia as well, though this paper focuses on Lithuania. The analysis reported here examined a two-month span ranging from March 22, 2018 through May 22, 2018. All collected tweets were automatically translated from Russian to English using the OssaLabs platform.

A total of 13,680 tweets were collected that matched the search criteria described above. The top 15 words comprised just under half (48%) of the instances of the top 100 words. The top 15 words included: Lithuania, Russian, Lithuanian, Vilnius, diplomats, Ukraine, Russia, Latvia, Estonia, NATO, send, Poland, member, and expulsion. Several of the words were then used to drive addition data collections. The presence of words such as "diplomat," "expulsion," and "member" in the top 15 is discussed in a later section.

3 Classifying IO Narratives

A total of 15 narratives potentially advancing political influence were identified in the Baltic region social media data collected by OssaLabs. Two different classification approaches were used to classify the narratives as either pro-Kremlin or anti-Kremlin. In the first approach, manual classification was used to label the narratives as either pro-Kremlin or anti-Kremlin. The manual classification resulted in several challenges and unlabeled narratives. In the second approach, community detection algorithms were first applied to the social network underlying the social media data. An examination of the detected communities resulted in a more comprehensive labeling of the narratives.

3.1 Manual Classification of Narratives

OssaLabs uses emergent topic detection algorithms to group similar posts together into topic narratives, which are presented to the analyst user for review. The topic detection capability used by OssaLabs is a proprietary Perceptronics Solutions Inc., algorithm that is an extension to the well-known KeyGraph method (Sayyadi, 2009; Sayyadi, 2013). The algorithm clusters similar tweets together into topic narratives based on statistical analysis of keyword co-occurrence in the corpus of tweets. Additionally, a representative tweet that best exemplifies the topic narrative is identified and presented to the analyst for review.

The fifteen largest automatically detected narratives identified were manually coded to determine whether each narrative fit either pro-Kremlin or anti-Kremlin agendas. Some narratives were easily recognizable as pro- or anti-Kremlin. Other narratives could be classified but only after conducting additional background research. And, still other narratives could not be classified even with additional research.

Of the fifteen largest narratives that were identified, three were clearly pro-Kremlin and four were clearly anti-Kremlin. The rest were not immediately classifiable into these two groups. Examples of some of the largest pro-Kremlin and anti-Kremlin narratives are illustrated in Figures 1 and 2.

The two largest unambiguously anti-Kremlin narratives are displayed in Figure 1, with the tweet that best represents the overall topic displayed. The narrative on the left was also the largest narrative overall, while the one on the right was the sixth largest narrative overall. For both narratives, the representative tweet for each topic was authored by the @Vitauskas_A account. The “Celebrating NATO Membership” narrative (Figure 1, left side) claims that had Lithuania not joined NATO, Russian troops would have invaded Lithuania under the pretense of protecting Russian-language speaking peoples. This presentation of Russia as aggressor with territorial ambitions to the Baltics is clearly an anti-Kremlin narrative. The “Forest Brothers” narrative (Figure 1, right side) appears to be designed as a reminder of Russia’s historical aggression toward and conflict with Lithuania. It shows a collage of war-torn Lithuania during World War II. The pictures show: 1) damage done by the Soviet Union when they invaded Lithuania to simultaneously expel the Germans and occupy the country; 2) dead members of the militia group known as the “Forest Brothers,”

who were young men of the Baltics who fought against the Soviet occupation. The forest brothers have emerged as a great source of national pride for Lithuanians; and 3) mass forced deportations of Lithuania people to Siberia. Collectively, this narrative seems designed to convey the message that Russia is not a historical friend to Lithuania.

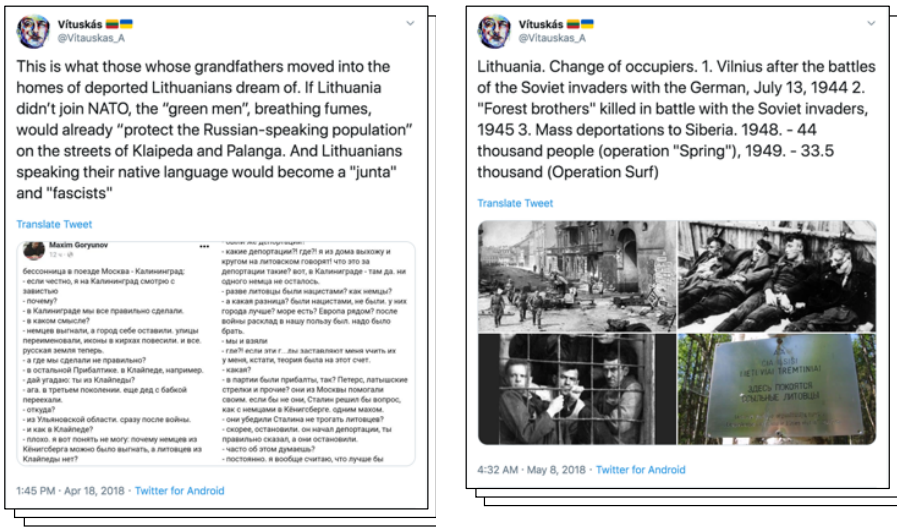

Fig. 1. Examples of largest Anti-Kremlin narratives discovered.

Fig. 2. Examples of largest Pro-Kremlin narratives discovered.

In contrast, the two largest unambiguously pro-Kremlin narratives are illustrated in Figure 2. These narratives were the fifth and seventh largest narratives detected, overall. Both of these narratives appear to be in response to several countries' decisions to expel Russian diplomats following the apparent poisoning of former Russian spy Sergei Skripal in England on March 4th, 2018, and Russia's subsequent decision to expel western diplomats. In the "Call to Expel Lithuanian Diplomats" narrative (Figure 2, left side), the sentiment that all Lithuanian diplomats should be expelled from Russian, and in fact should already have been expelled prior to the Skripal incident, is shared along with derogatory language being directed at Lithuania. The "Insults to

Countries that Expelled Russian Diplomats" narrative (Figure 2, right side) essentially consists of name-calling directed at countries that expelled Russian diplomats and other ad hominem attacks directed at France's PM Macron that seek to question the motives of his decision to expel Russian diplomats.

The remaining eight narratives could not immediately be classified as pro-Kremlin or anti-Kremlin. However, three of them (the second, third, and fourth largest narratives discovered) could be classified after a limited investigation was conducted. At first glance, the "Lithuanian Special Forces" narrative (Figure 3, top left) appears to be very supportive of the Lithuanian special forces. However, the included video shows soldiers (presumably Lithuanian) trying and failing to knock down a door until one of the soldiers casually walks up and opens the door. Far from praising the Lithuanian special forces, this narrative is actually mocking them. This narrative contains sarcasm and highlights one of the major challenges in understanding underlying sentiment.

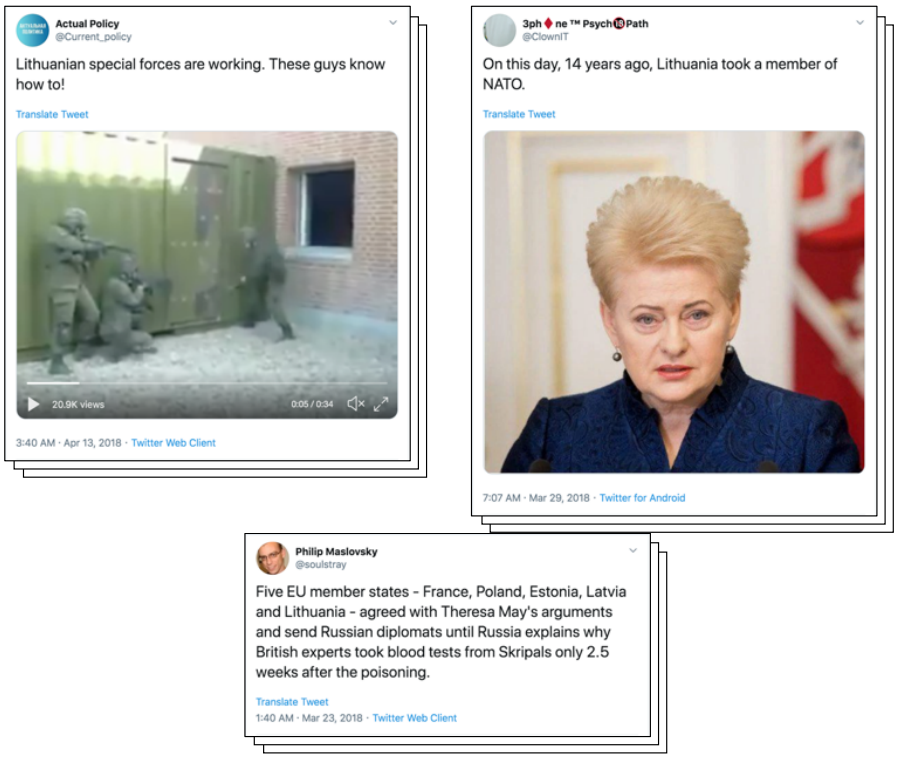

Fig. 3. Narratives that were initially challenging to understand.

The "Member of NATO" narrative (Figure 3, top right) appears to be celebrating the inclusion of Lithuania as a member of NATO. However, deeper examination reveals that the Russian word that translates to "member" is often used as slang for male sexual anatomy. With this in mind, the narrative appears to be mocking the Lithuani-

an politician who played a large role in joining NATO. This narrative highlights another challenge in understanding narratives that include slang and unfamiliar jargon, particularly when the original tweet is in foreign language and the true meaning of the narrative is lost in translation. The "Skripal" narrative (Figure 3, bottom) seeks to discredit the British claim that Sergei Skripal was poisoned by Russian operatives. In doing this, the narrative highlights the claim that British experts did not collect blood samples from Skripal for nearly 2.5 weeks after he was poisoned. The insinuation is that something is suspicious about this delay, and perhaps the British fabricated lab results or else even poisoned Skripal after he was hospitalized. Understanding this narrative required rather additional research into the Skripal case and the conspiracy theories surrounding it. This narrative highlights the challenge of oftentimes requiring deep familiarity with the nuances being debated in the narrative in order to understand its meaning.

Five out of the largest 15 narratives could not be classified as pro- or anti-Kremlin, even with additional research. In many instances the ambiguity stems from not being able to discern the underlying motivations for which a narrative was shared. It is not clear if the information contained in these narratives was being shared because people were happy about them, outraged by them, or laughing at them.

When monitoring social media-based information operations, it is important for human analysts to be able to quickly identify emerging narratives and talking points, and to understand which side of political influence these narratives are supporting. The above examples enumerate several of the challenges involved with manually classifying narratives: 1) the use of sarcasm and slang can obfuscate the intended meaning of narratives; 2) narrative battles are often waged over very specific points that require a nuanced understanding of the domain to appreciate; and 3) imperfect translation of foreign languages exacerbates these problems. In sum, manual classification of narratives is time and labor intensive, and sometimes unsuccessful.

3.2 Structural Approach to Classifying Narratives

To mitigate the challenges associated with manual classification, an alternative classification approach was investigated. The approach uses a structural analysis of the social network that underlies the social media data to identify distinct communities used for the narrative classification. In this way, social network analysis forms the foundation of the classification.

Social network analysis (SNA) can quantitatively answer questions such as: who has direct influence online; who are the individuals influencing the influencers; what sub-groups exist in a network representing social/ideological divides; and what members of the network were disproportionately targeted or elevated by bot activity (Jonas 2017). In developing this approach, the OssaLabs platform was used to construct a network characterized by: each node in the network is a social media user that appeared in the Baltic region dataset; 2) users are connected by an edge if they retweeted the same tweet; and 3) weight of that edge reflects how many times such co-re-tweeting occurred. This network was exported from OssaLabs and visualized using Gephi, as shown in Figure 4.

Fig. 4. Network analysis of the collected data reveals two distinct groups among several others, one is pro-Kremlin and the other is anti-Kremlin.

The Louvain community detection algorithm (Blondel 2008) was applied to identify distinct parts of the network that are more tightly connected to each other than to other parts. Each of these distinct communities was randomly assigned a color. By manually reviewing the biographies of the most retweeted users and the most retweeted tweets within each community, we were easily able to label two important communities according to their interests. The red-colored community is pro-Kremlin and the blue-colored community is anti-Kremlin. The most retweeted users in those communities were @RussianHeart, @sashakots, @Current_policy in the pro-Kremlin group, and @natohawks, @Vitauskas_A, and @Skrepo_history in the anti-Kremlin group. Unsurprisingly, several of these user accounts are recognizable from our previous discussion of manually classified narratives. Many of the other communities either had no clear defining interest or else seemed to be interested in general Lithuania related news.

Having identified these two distinct communities and their influencers, we were then able to use this information to filter the data and focus on content related to each community. Using OssaLabs, we retrieved a list of Twitter users who follow each of the above six influencers. Then, a new dashboard was constructed only for narratives with at least 50% of their constituent tweets attributed to those pro-Kremlin followers. A similar dashboard was constructed for anti-Kremlin content. In each dashboard, the narratives were ranked according to the volume of tweets coming from the followers of each group's influencers. In this manner, all of the largest 15 narratives were classified as being either pro-Kremlin or anti-Kremlin. Note that while we anticipated having to construct more elaborate classification rules, the 50% threshold was sufficient to clearly and unambiguously classify all 15 of the largest narratives.

The pro-Kremlin and anti-Kremlin dashboards confirmed the findings for the 10 narratives manually classified. All four of the narratives manually classified as being anti-Kremlin in nature appear among the ten largest narratives in the anti-Kremlin

dashboard. All six of the pro-Kremlin narratives appeared among the ten largest narratives in the pro-Kremlin dashboard also. This included the narratives that previously could only be manually classified after conducting additional research.

Furthermore, the pro- and anti-Kremlin dashboards shed light on the nature of the five unclassifiable narratives. Three of the unclassifiable narratives were among the ten largest anti-Kremlin narratives, including one related to banning entry of Russian citizens, another related to dismantled powerlines to Belarus, and the last one related to discovering a Russian tractor with radioactive materials. The two other unclassifiable narratives were among the ten largest pro-Kremlin narratives, including one related to the expulsion of Russian diplomats, and the other related to Lithuania being offended by Russia's decisions to expel diplomats. Knowing who was participating in the narrative discussion, increased our understanding of the motivation behind the narrative.

In summary, there was no ambiguity in the automatic classification of the narratives based on pro- and anti-Kremlin group participation. Each of the 15 narratives were classified as belonging to one of the two groups; were in the top 10 largest narratives of the groups to which they were classified; and contained content exclusively authored by members of the group they were assigned to. The 10 largest narratives in these group-focused dashboards also included relevant narratives that were not part of the overall largest 15 narratives. For example, the 10th largest narrative in the anti-Kremlin dashboard discussed Lithuania's official decision to start using the name "Sakartvelo" to refer to the country of Georgia—a move designed to reject the "Russian" name of Georgia and instead embrace the name that Georgian's give their own country.

4 Conclusion

Social media-based IO will continue to be an increasingly important part of the modern battlespace and the constant competition occurring below the threshold of warfare. Techniques to effectively and efficiently identify and classify narratives are an important part of monitoring and countering these IO campaigns.

This paper compared two approaches for classifying narratives. The first approach involved attempted manual classification of narratives based on examination of content only. There were several challenges to manual classification including sarcasm, the loss of meaning due to translation, and the requirement of deep domain knowledge to understand the motivation behind the narrative. To mitigate these challenges a second classification approach was developed based on first understanding and classifying the communities that exist in the social network that underlies the social media data. Using the second approach all of the 15 narratives were correctly classified including the unclassifiable narratives from the manual approach. These results demonstrate the promise of using a more structural analysis to understand narratives rather than content alone.

There are several important directions for future work in this area. First, further study in different linguistic and topical contexts will be needed to clearly identify if

these methods of classification can be generalized. Second, in this study the discovered narratives clearly belonged to one group or the other. It remains unclear how common place it is for narratives to span multiple groups and how best to apply the methods discussed here to such instances. Third, in this work we compared and contrasted the two methods qualitatively, while comparative quantitative methods may offer further insight. Follow-on work should examine broader applications of these methods, while applying additional novel variations that are assessed using a mix of quantitative and qualitative methodologies.

References

1. Bessi, A., Ferrara, E. (2016) Social bots distort the 2016 US presidential election online discussion. First Monday 21.11-7.
2. Blondel, V., et al. (2008) Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiments. 2008-10.
3. Bradshaw, S., Howard, P. (2017) Troops, trolls, and troublemakers: a global inventory of organized social media manipulation. University of Oxford, Computational Propaganda Research Project.
4. Chabuk, T., Jonas, A., Kase, S. (2019) Monitoring Russian online information operations: an OssaLabs case study. No. ARL-TR-8732. CCDC Army Research Laboratory Aberdeen Proving Ground United States.
5. Chabuk, T., Jonas, A. (2018) Understanding Russian information operations, SIGNAL. September 2018, pp 37-39.
6. Chivvis, C. (2015) The Baltic balance: how to reduce the chances of war in Europe. Foreign Affairs, Snapshot.
7. Giles, K. (2017) Handbook of Russian information warfare. NATO defense college.
8. Helmus, T., et al. (2018) Russian social media influence: understanding Russian propaganda in Eastern Europe. Rand Corporation.
9. Jonas, Adam B. (2017) How the hashtag is changing warfare: Armies of social media bots for hearts and minds online. SIGNAL. July 2017, pp. 34-36.
10. Labov, W., and Waletsky, J. (1976) Narrative analysis: oral versions of personal experience. In: Helm J (eds) Essays on the verbal and visual arts, Seattle, University of Washington Press.
11. Noack, R. (2017) Lithuania fears a Russian invasion. The Washington Post.
12. Polletta, F., Ching, P., Chen, B. (2012) Narrative and social movements. In: Alexander J, Jacobs R, Smith P (eds) The Oxford handbook of cultural sociology. Oxford University Press.
13. Radin, A. (2017) Hybrid warfare in the Baltics: threats and potential responses. No. RR-1577-AF. RAND Project AIR FORCE, Santa Monica.
14. Sayyadi, H., Raschid, L. (2013) A graph analytical approach for topic detection. ACM transactions on internet technology 13(2): 1-23.
15. Sayyadi, H., Hurst, M., and Maykov, A. (2009) Event detection and tracking in social streams. In: Icwsm.
16. Standish, R. (2017) Russia's neighbors respond to Putin's 'hybrid war'. Foreign Policy.