REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
03-03-2022	Final Report	15-Apr-2016 - 14-Oct-2019

4. TITLE AND SUBTITLE

Final Report: Detecting civil conflict and information biases in polarized environments in social media

5a. CONTRACT NUMBER
W911NF-16-1-0174

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

University of Massachusetts - Lowell
600 Suffolk Street, Suite 226

Lowell, MA 01854 -3643

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
69164-CS.18

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Anna Rumshisky
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 978-934-3619

RPPR Final Report
as of 21-Apr-2022

Agency Code: 21XD

Proposal Number: 69164CS **Agreement Number: W911NF-16-1-0174**
INVESTIGATOR(S):

Name: Anna Rumshisky
Email: anna_rumshisky@uml.edu
Phone Number: 9789343619
Principal: Y

Organization: **University of Massachusetts - Lowell**
Address: 600 Suffolk Street, Suite 226, Lowell, MA 018543643
Country: USA
DUNS Number: 956072490 EIN: 043167352
Report Date: 14-Jan-2020 Date Received: 03-Mar-2022
Final Report for Period Beginning 15-Apr-2016 and Ending 14-Oct-2019
Title: Detecting civil conflict and information biases in polarized environments in social media
Begin Performance Period: 15-Apr-2016 **End Performance Period:** 14-Oct-2019
Report Term: 0-Other
Submitted By: Anna Rumshisky Email: anna_rumshisky@uml.edu
Phone: (978) 934-3619

STEM Degrees: 5 **STEM Participants:** 4

Major Goals: 1. Major goals (specific aims) as stated in the proposal

Specific Aim A. Develop a composite index of conflict intensity. Develop methods for detecting several conflict indicators in social media, to be combined into a composite index of conflict intensity. Use several measures of verbal and non-verbal user behavior and the associated network-scale effects, including:
(1) a measure of divergence between polarized user clusters, in the network induced by what users share, repost, or like; this will be based on novel methods we propose for for polarized community detection, as well as on existing measures of modularity and boundary-based polarization; (2) a measure of alignment of the user communities induced by different polarizing issues; (3) a measure of the intensity of mutual aggression and hostile person-to-person sentiment (flame wars), as indicated both by sentiment and user commenting patterns in mixed-commentary situations, i.e., when the members of polarized groups interact; (4) a measure of the vocabulary discrepancies between the opposing groups, including lexical meaning shifts and
neologisms, represented by distributional patterns captured by vector space embeddings. (5) a measure of presence and propagation of opposing sentiment towards trigger topics and events in user-generated text.

Specific Aim B. Analyze dynamic trends of the networks induced by user behavior. Detect and analyze the dynamics of polarized user networks, including user clique and cluster formation, their stability over time, and the changes in cluster modularity and density; track the formation and dynamics of user groups related to the conflict and individual user connections.
B1. Polarized community detection. Test and compare existing measures such as modularity and polarity-at-boundary, against novel methods to be developed in the scope of the project, including methods based on the multi-view graph representation of the user network and the adaptation of Bayesian topic modeling techniques representing users as probability distributions over topics, and topics as probability distributions over the content the users share, mention or like.
B2. Assessment of ideological issue alignment. Analyze the dynamics of the user networks induced by the issues related to the particular conflict, representing the resulting space as a set of partially overlapping user clusters. Develop a measure of issue alignment using the overlap in user clusters.

Specific Aim C. Flame war and political sentiment detection.
C1. Measuring verbal aggression. The analysis of verbal user behavior, including flame war detection and identifying instances of inciting to action, including directed and person-to-person sentiment.
C2. Measuring topic dominance. Analyze the trends in the number of users involved, their posting frequency, and topic distribution.

C3. Detecting directed sentiment towards salient topics. Track polarity and intensity of the sentiment expressed within the opposing user clusters towards some salient topics, as represented by common hashtags and common collocates extracted from news provided by the news aggregators. We will also track the propagation of sentiment towards new events through the user network.
C4. Tracking vocabulary shifts. Detect and measure vocabulary discrepancies between the opposing groups, including lexical shifts in meaning and introduction of neologisms.

Specific Aim D. Detecting information biases.
Develop methods that combine language-based and reaction-based bias detection. Develop the methods that use (1) collaborative filtering framework, (2) deep learning techniques using reaction-based data.

2. Changes / additions to the proposed goals
The original proposal included the analysis of two case studies: the Russian/Ukrainian conflict of 2014 and the 2016 US Elections. During the project period, we modified the scope of the proposed work to include the following case study: the 2011-2012 Bolotnaya protests in Russia. The Bolotnaya data has not received much attention, and it proved to be a good source for studying linguistic features that correlate with manifestations of conflict on the ground, such as the size and attendance of individual protest rallies. In the second half of the project, we have placed more emphasis on the Bolotnaya case study, suspending some of the work on the 2016 US Elections case study.

3. Milestones with % completion
• Data collection and preliminary analysis for Russian/Ukrainian conflict of 2014 [100%]
• Data collection and preliminary analysis for Bolotnaya conflict of 2012 [100%] (*new task)
• Data collection and preliminary analysis for the 2016 US elections [50%]
• Mechanical Turk annotation of the English language data [0%] (task suspended)
• Recruiting Russian language annotators [100%]
• Manual annotation of the Russian language data -- bias annotation [100%]
• Manual annotation of the Russian language data -- sentiment annotation [100%]
• Develop, implement, and test algorithms for polarization detection [75%]
• Evaluate polarization detection methods [75%]
• Develop, implement, and test the algorithms for sentiment analysis [100%] and bias detection [100%]
• Development of data and algorithms for political sentiment detection [100%]
• Development and evaluation of flame war detection and lexical change tracking algorithms [75%]
• Begin integrating the methods developed within the scope of the project to define a composite index for conflict intensity [75%]
• Development of algorithms for argument structure analysis [100%]
• Data and computational methods for detecting calls to political action [100%]
• Publication and dissemination of results [100%]

Accomplishments: Major accomplishments under the goals during the project period are listed below.

*** Data collection ***

The following data was collected/obtained by the project team:
• Data collected from VKontakte (the largest Russian social network, also popular in Ukraine): over 3.2B text of public posts of users who were members of pro-Russian and pro-Ukrainian groups during 2014 Maidan conflict in Ukraine.
• Data collected from VKontakte relevant to the 2011-2013 Bolotnaya anti-government protests in Russia (the most successful and largest anti-Putin protests in Russia until 2017) received via a collaborative relationship with New Economic School (NES). The dataset includes 100M posts from over 476 Bolotnaya protest-related groups and their user members.

*** Relationship between verbal and non-verbal conflict indicators ***

Using the Russian/Ukrainian Maidan data, we showed correlation patterns between the Random Walk Controversy (RWC) measure on user like-behavior pattern graphs and the mean and standard deviation of the sentiment expressed in politically-themed posts [Rumshisky et al 2017]:
• As the conflict intensifies, the RWC measure on user like-behavior pattern graphs and the standard deviation of

overall sentiment expressed by the opposing groups in politically-themed posts (SenSTD) increased in unison. RWC and SenSTD were positively correlated (Pearson and Spearman correlation values of 0.674 and 0.745, respectively).
• RWC and the average of the absolute value of the overall sentiment correlate negatively, confirming that negative sentiment accompanies the intensification of conflict (Pearson and Spearman correlation values of -0.598 and -0.291, respectively).

*** Argument structure ***

Using existing datasets for English, we have developed several algorithms for identifying the structure of arguments, as well as what makes an argument (a) persuasive, and (b) convincing:
• a novel attention-based neural network model for argument structure parsing [Potash et al 2017a].
• investigated several memory-based neural network models and other methods for integrating knowledge into models that identify more convincing arguments [Potash et al 2017b].
• a novel neural network model for predicting debate winners, which leveraged external signal from the audience for regularization at training time [Potash et al 2017c].

*** Bias detection ***

• We created 2 annotated datasets of biased news articles on Russia/Ukraine Maidan conflict: a gold standard resource with 347 articles, and silver standard resource with 10,000 articles pulled from highly polarized pro-Ukrainian and pro-Russian communities of users in VKontakte social network during the 2014 Maidan conflict.
• We trained classifier models to detect biased texts using the above data. A feed-forward neural network model using link address and article content tokens achieves 93.5% accuracy [Potash et al 2017d].

*** Sentiment analysis for Russian social media ***

• We developed a set of comprehensive sentiment annotation guidelines for annotation of social media posts in Russian. The posts are annotated on 3-point scale (negative, neutral, positive posts), with some formulaic speech act posts such as congratulations or greetings annotated as a separate subcategory, and unclear cases skipped.
• We recruited and trained Ukrainian/Russian bilingual annotators, and developed an annotation web-interface.
• We developed RuSentiment [Rogers et al 2018], a large high-quality dataset for analysis of sentiment in Russian social media - the largest and the most consistent of its kind that is currently available (18,453 posts and counting from VKontakte social network, processed by 3 annotators with kappa 0.637).
• Experiments on the new dataset with 3 baseline classifiers (logistic regression, RidgeClassifier, LinearSVC) yielded 0.65-0.7 accuracy when classifying into 5 classes (negative, positive, neutral, skip, speech acts), ≈0.9 for just the negative class against the rest, and ≈0.84 for just the positive class against the rest. The classifiers were trained on 13,764 annotated posts and tested on 3441. Each post was represented as the average of 300-dimensional word vectors trained on the 3.2B VKontakte corpus with FastText algorithm.
• Development of a novel sentiment analysis system based on neural memory network architecture, using the enriched key-value memory. Preliminary experiments with SentiRuEval dataset yielded 3% improvement over baselines.

*** Tracking controversy via vocabulary shifts ***

We have developed a novel technique for tracking the dynamics of a controversy via shifts in vocabulary used by the opposing sides. We applied the methodology of training word embeddings on temporal slices of the Maidan conflict data to show that cosine similarity between re-trained word vectors is a reliable measure of lexical drift [Rumshisky et al 2017]. Such drifts include:
• change in connotations of such words as as “Crimea”, “referendum”, and “Putin”;
• new meanings of some common words: e.g. the word “vata” (“cotton”) is now used by pro-Ukrainian speakers to refer to pro-Russian speakers, expressing a combination of “redneck” and “unquestioning support for Putin policies”.

*** Predicting real-life protest events from social media ***

Using the Bolotnaya case study, we showed that several methods can be used to predict the timing and magnitude of real-life protest events based on social media text. Specifically, we analyzed topic spikes and frequency of calls to action in user posts as predictors. For calls to action:

• We examined in-depth linguistic features that mark calls to action (CTA), such as imperatives, modal verb + infinitive and zero-infinitive constructions, and developed detailed annotation guidelines that included 14 core and borderline types of political CTAs.
• We annotated 1000 VKontakte posts from Russian Bolotnaya data and showed that they are relatively easy both to annotate (with IAA 0.78) and to classify (F1 of 0.77) [Rogers et al 2019].

*** Publications, Software, Annotated Datasets ***

Publications: 7 papers published (4 major conferences, two workshops)

Software:
• Sentiment detection baseline models, trained on the Russian sentiment dataset (RuSentiment).
• Bias detection model, trained on silver bias dataset for the Russian/Ukrainian Maidan conflict).
• Sentiment detection model using a neural network with key-value memory for sentiment labels, trained on SentiRuEval dataset of tweets from telecom/banking domain.
• Baseline models for detecting calls to action in Russian social media text.
• Neural models for extracting argument structure from English text.
• Data collection software.

Datasets:
• Russian sentiment dataset (RuSentiment) + annotation guidelines in English and in Russian.
• Word2vec embeddings using fastText trained on 300,000,000 posts from VK.
• Balanced list of political keywords for the Russian/Ukrainian Maidan conflict.
• Dataset and annotation guidelines for identifying calls to action in Russian social media text.
• Bias dataset (silver, gold).

Training Opportunities: • Graduate students involved in the project co-authored five major conference papers and two workshop papers produced by the project team, with a graduate student first author on four of the conference papers and one workshop paper.

• Graduate students involved in the project attended several professional conferences, including (1) The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies held in San Diego, California, June 12–17, 2016 (2) The Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), held in Copenhagen, Denmark, September 7–11, 2017 (3) Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017 (4) the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, held in Taipei, Taiwan, November 27 - December 1, 2017.

• The postdoctoral fellow involved in the project attended a professional conference to present a paper, the 27th International Conference on Computational Linguistics held in Santa Fe, New Mexico, August 20 - August 24, 2018.

• One of the Ph.D. students defended his Ph.D. dissertation in the fall of 2018.

• Four undergraduate students involved in the project had an opportunity to collaborate with our interdisciplinary project team, and one of them became a co-author on one of the publications produced by the team.

RPPR Final Report
as of 21-Apr-2022

Results Dissemination:
• The team published five conference papers and two workshop papers, which were presented at conferences by team members. PI discussed these papers with a team of linguistics researchers from University of Maryland exploring a possible collaboration.

• PI established a collaboration with Perceptronics Solutions, and they have collaborated on several SBIR contracts (PAIT N68335-19-075, ABDOC N68335-19-0838, ARTISAN N68335-20-C-0950).

• PI traveled to the Aberdeen Proving Ground to present the team's work to the researchers at the Army Research Lab. The PI and postdoctoral fellow further discussed collaboration possibilities with some of the ARL researchers at the COLING 2018 conference where the sentiment analysis paper was presented.

• One PhD dissertation was defended on the project work.

• One of our team members, professor Mikhail Gronas co-organized a interdisciplinary conference on the worsening relations between the U.S. and Russia in the last decade and its representation in the social media ("Mediating the New Cold War In The Digital Age', Dartmouth College, May 6, 2016) . Our team presented our data on detecting and quantifying conflict in social media, using the "new cold war" as a case study. We used this opportunity to acquaint political scientists, sociologists, historians, and journalists with the quantitative approaches to the conflict analysis and the prototypes for the conflict visualization tools we have began developing.
A. Rumshisky, M. Gronas, P. Potash, A. Romanov. 2016. Detecting and measuring the new cold war in social media.

Honors and Awards: Nothing to Report

Protocol Activity Status:

Technology Transfer: • PI traveled to the Aberdeen Proving Ground to present the team's work to the researchers at the Army Research Lab. The PI and postdoctoral fellow further discussed collaboration possibilities with some of the ARL researchers at the COLING 2018 conference where the sentiment analysis paper was presented.

• PI established a collaboration with Perceptronics Solutions, and they have collaborated on several SBIR contracts (PAIT N68335-19-075, ABDOC N68335-19-0838, ARTISAN N68335-20-C-0950).

PARTICIPANTS:

Participant Type: PD/PI
Participant: Anna Rumshisky
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Faculty
Participant: Mikhail Gronas
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Postdoctoral (scholar, fellow or other postdoctoral position)
Participant: Anna Rogers
Person Months Worked: 6.00 **Funding Support:**

Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Peter Potash
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Alexey Romanov
Person Months Worked: 15.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Alexey Romanov
Person Months Worked: 4.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Non-Student Research Assistant
Participant: Nikita Prianichnikov
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: Y

Participant Type: Non-Student Research Assistant
Participant: Mikhail Dubov
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Undergraduate Student
Participant: Ivan Blaskic
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Undergraduate Student
Participant: Alexander Gribov
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Consultant
Participant: Alex Potash
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Undergraduate Student
Participant: Gregory Smelkov
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Undergraduate Student
Participant: Patrick Kyoyetera
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Olga Kovaleva
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Consultant
Participant: Timothy Messen
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Consultant
Participant: Aleksandra Chashchina
Person Months Worked: 5.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Consultant
Participant: Olesia Shyrobokova
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Consultant
Participant: Oleksandra Tonachova
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Consultant
Participant: Roman Rieznyk
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

International Collaboration:

RUS

GBR

RUS

UKR

UKR

UKR

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: EMNLP 2017 Workshop "NLP Meets Journalism"
Date Received: 31-Aug-2018 Conference Date: 07-Sep-2017 Date Published:
Conference Location: Copenhagen, Denmark
Paper Title: Tracking Bias in News Sources Using Social Media: the Russia-Ukraine Maidan Crisis of 2013–2014
Authors: Peter Potash, Alexey Romanov, Anna Rumshisky, Mikhail Gronas
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: EMNLP 2017
Date Received: 31-Aug-2018 Conference Date: 09-Sep-2017 Date Published:
Conference Location: Copenhagen, Denmark
Paper Title: Here's My Point: Joint Pointer Architecture for Argument Mining
Authors: Peter Potash, Alexey Romanov, Anna Rumshisky
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: EMNLP 2017
Date Received: 31-Aug-2018 Conference Date: 09-Sep-2017 Date Published:
Conference Location: Copenhagen, Denmark
Paper Title: Towards Debate Automation: a Recurrent Model for Predicting Debate Winners.
Authors: Peter Potash, Anna Rumshisky
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: IJCNLP 2017
Date Received: 31-Aug-2018 Conference Date: 27-Nov-2017 Date Published:
Conference Location: Taipei, Taiwan
Paper Title: Length, Interchangeability, and External Knowledge: Observations from Predicting Argument Convincingness.
Authors: Peter Potash, Robin Bhattacharya, Anna Rumshisky
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: SocInfo 2017
Date Received: 31-Aug-2018 Conference Date: 13-Sep-2017 Date Published:
Conference Location: Oxford, UK
Paper Title: Combining Network and Language Indicators for Tracking Conflict Intensity
Authors: Anna Rumshisky, Mikhail Gronas, Peter Potash, Mikhail Dubov, Alexey Romanov, Saurabh Kulshreshth
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: COLING 2018
Date Received: 31-Aug-2018 Conference Date: 21-Aug-2018 Date Published: 21-Aug-2018
Conference Location: Santa Fe, New Mexico
Paper Title: RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian
Authors: Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, Alex Gribov.
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda
Date Received: 03-Mar-2022 Conference Date: 04-Nov-2019 Date Published: 04-Nov-2019
Conference Location: Hong Kong, China
Paper Title: Calls to Action on Social Media: Detection, Social Impact, and Censorship Potential
Authors: Anna Rogers, Olga Kovaleva, Mikhail Gronas, Anna Rumshisky
Acknowledged Federal Support: **Y**

DISSERTATIONS:

RPPR Final Report
as of 21-Apr-2022

Publication Type: Thesis or Dissertation
Institution: University of Massachusetts Lowell
Date Received: 31-Aug-2018 Completion Date: 8/29/17 10:57PM
Title: Neural argumentation: Structure and persuasion
Authors: Peter Potash
Acknowledged Federal Support: **N**

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Anna Rumshisky
Signature Date: 3/3/22 10:04PM

Title: Detecting civil conflict and information biases in polarized environments in social media
Contract Number: W911NF1610174
Period of Performance: **Start:** Apr 15, 2016 **End:** Oct 14, 2019

Final Report

This report describes our efforts and results of project work over the project period of Apr 15, 2016 – Oct 14, 2019. During the project period, the team made significant contributions towards several lines of research, which included

1. Developing datasets and methods for sentiment analysis it Russian,
2. Detection of bias in news sources using community-based citation patterns,
3. Computational analysis of argument structure and persuasiveness,
4. Evaluation of community detection methods,
5. Computational modeling of the relationship between verbal and non-verbal conflict indicators,
6. Prediction and tracking of live protest events via analysis of social media texts.

This document is structured as follows. We first list scientific publications and products produced under the grant. We then (1) describe the datasets we used in our studies, including our collection methods, and (2) detail the results of our work in the above efforts.

Grant Products

Publications

As a result of work on this grant, the project team published five papers in major conferences and two workshop papers. The publications are listed below.

1. A. Rumshisky, M. Gronas, P. Potash, M. Dubov, A. Romanov, S. Kulshreshtha, A. Gribov. Combining Network and Language Indicators for Tracking Conflict Intensity. *Proceedings of SocInfo 2017*. Oxford, United Kingdom.
2. P. Potash, A. Rumshisky. Towards Debate Automation: a Recurrent Model for Predicting Debate Winners *Proceedings of EMNLP 2017.* Denmark, Copenhagen.
3. P. Potash, A. Romanov, A. Rumshisky, M. Gronas. Tracking Bias in News Sources Using Social Media: the Russia-Ukraine Maidan Crisis of 2013–2014. *EMNLP 2017 Workshop "NLP Meets Journalism".* Denmark, Copenhagen.
4. P. Potash, A. Romanov, A. Rumshisky. Here's My Point: Joint Pointer Architecture for Argument Mining *Proceedings of EMNLP 2017.* Denmark, Copenhagen.
5. P. Potash, R. Bhattacharya, A. Rumshisky. Length, Interchangeability, and External Knowledge: Observations from Predicting Argument Convincingness *Proceedings of IJCNLP 2017.* Taipei, Taiwan.
6. A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, A. Gribov. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. *Proceedings of COLING 2018.* Santa Fe, New Mexico. [project page] [data]
7. A. Rogers, O. Kovaleva, M. Gronas, A. Rumshisky. Calls to Action on Social Media: Detection, Social Impact, and Censorship Potential. *Proceedings of The 2nd Workshop on NLP for Internet*

Freedom: Censorship, Disinformation, and Propaganda, pages 36–44, EMNLP 2019. Hong Kong, China.

Datasets and other resources:

- Russian sentiment dataset (RuSentiment) + annotation guidelines in English and in Russian.
- Word2vec embeddings using fastText trained on 300,000,000 posts from VK.
- Balanced list of political keywords for the Russian/Ukrainian Maidan conflict.
- Dataset and annotation guidelines for identifying calls to action in Russian social media text.
- Bias dataset (silver, gold).

1. Data collection efforts

In this section, we describe the data that was collected for the project, including the data collection systems developed for different data sources. A summary of all the data collected to date can be found in Table 1.

1.1 VKontakte data for the 2014 Russian/Ukrainian (Maidan) conflict.

For this project, we built a new scalable system for data collection from VKontakte social network (VK). Preliminary data collection for this project was done with our older data collection system that was built for Microsoft Azure platform, using Azure Queue, Azure Storage and Azure Cloud Service. In order to overcome the vendor lock-in, we developed a new version of this system that uses an open-source software stack under Linux OS. The system is implemented in Python using a PostgreSQL database and Redis-based message queue. Using this system, we have collected more than 440 million posts with meta-information from more than 1,350,000 previously identified "active" users from both pro-Maidan("evromaidan", pro-Western) and anti-Maidan (pro-Russian) groups. We defined *active* users as those who averaged 2 or more posts per 3 months at least once over the target time frame (Oct 1, 2013 - Oct 1, 2014).

General architecture of the system is shown in Figure 1. The system implements the following workflow:

(1) The user generates a data collection task, specifying (a) the method of VKontakte API, and (b) an SQL query which provides the parameters for the specified API method. This is depicted with the dashed arrow from the user to the database.
(2) The user invokes the data collection process on one or more data collection servers. Each server queries the VK API server with the specified method, using the information retrieved by the SQL query. Due to the API restrictions, each API key is limited to no more than three requests per second. Therefore, each server uses its own API key to speed up the data collection process. This process is depicted with the dashed arrows from the data collection servers to the VK API servers.
(3) When the data collection server receives a response from the VK API, it saves the results, including meta-information about the target method and parameters, to a Redis-based message queue. This process is depicted with the bold arrows from the data collection servers.
(4) A separate server reads from the message queue and saves the information to a database. Because the rate of data collection on each data collection server is limited, a single server that saves the

results into a database can handle many data collection servers. Efficient data insertion is achieved by using optimized bulk copy of data rows. In this mode, indices are disabled and the insertions are faster than with individual SQL "insert" statements. However, this mode fails to insert all the data rows if there is a violation of uniqueness of primary keys for even a single row. Therefore, the server first ensures that no constraints are violated.

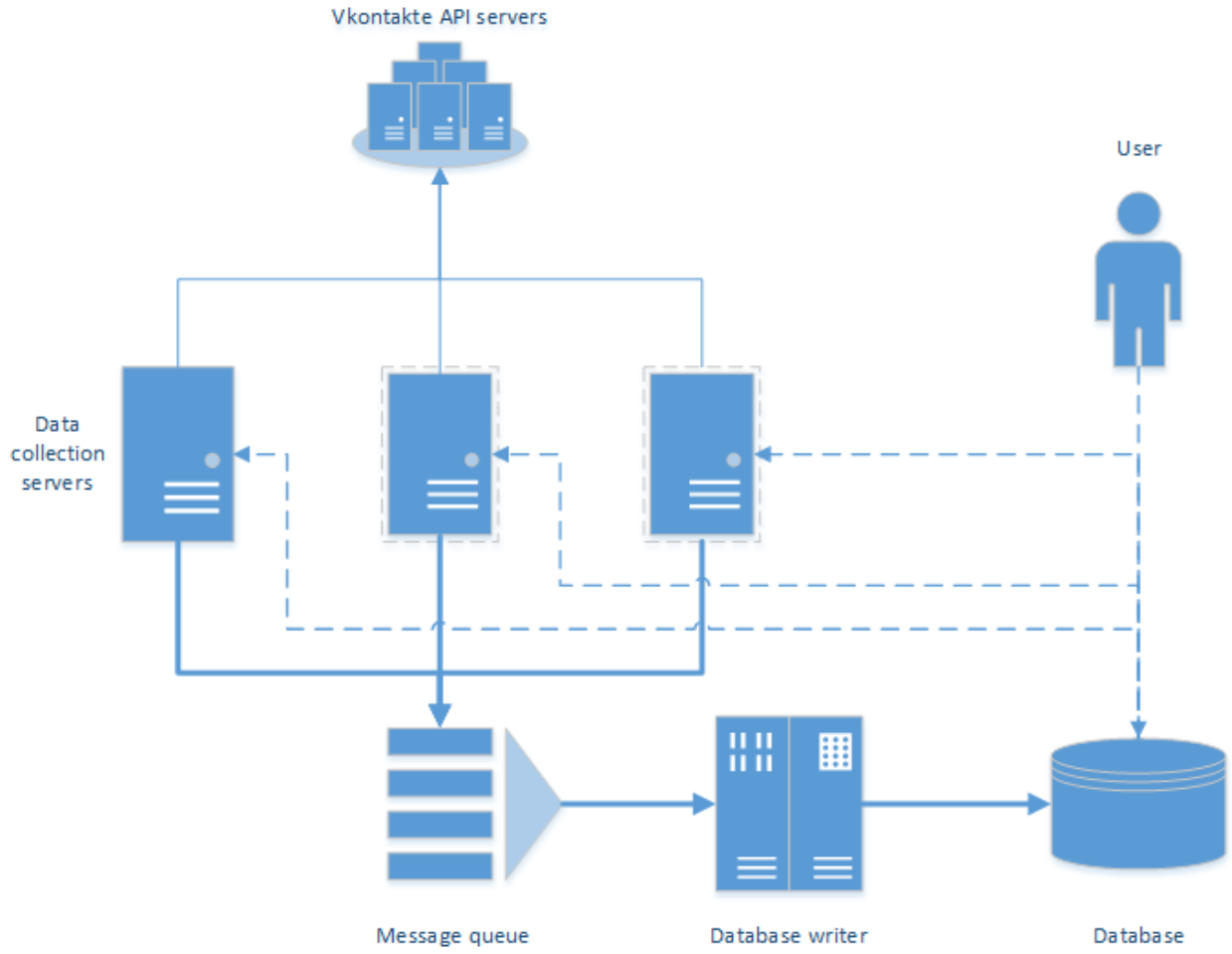


Fig. 1. General architecture of the VKontakte data collection system

This architecture allows us to perform large-scale continuous data collection. Specifically, this architecture easily supports hundreds of data collection servers saving results into a cluster of Redis servers as a message queue. The following methods of VKontake API are currently supported:

(1) `wall.get` - retrieves posts from a VK user's or group's wall
(2) `groups.getMembers` - retrieves members of the specified user group
(3) `users.get` - retrieves basic information about the VK user/users.

We have also implemented a method `execute` which allows us to execute up to 25 requests to the VK API in a single batch, which significantly reduces network-related delays.

Data statistics

Using the list of users identified during preliminary data collection for this project, we have collected over 443,550,000 posts from more than 1,354,000 unique "active" users from both the anti-Maidan and the pro-Maidan ("evromaidan") user groups. For each post, we collect the text, the date of the post, the number of likes/reposts/comments, and attached links, if any.

2.2 Vkontakte data for the 2011-2013 anti-government protests (Bolotnaya)

In 2013–2015, several members of our team created infrastructure for collecting Bolotnaya protest data for a collaborative study conducted by the New Economic School (NES) in Moscow. We have an agreement with NES to use this dataset. This dataset includes user information, friend lists, group membership, as well as posts and comments for the members of the "protest" groups. Overall, it comprises 61 million records in the "friends" table and 14 million records in "groups" table, as well as 100 million posts from the "protest" groups and users.

2.3 Twitter and political blog data for the 2016 U.S. presidential elections

Twitter data collection

We have developed a system for data collection from Twitter that has a similar architecture to the VK data collection system described above. The system currently supports the method `statuses/filter` from Twitter's streaming API with parameters `follow` and `track` for capturing either tweets from specific users or by filtering on keywords. When each tweet is retrieved, it is placed in a Redis queue. Another server reads from this queue, and when the size of the queue exceeds a threshold, the result is saved into a database using the batching technique described above. The systems automatically maintains the optimum threshold value so that the server is neither constantly busy because the queue is too small, nor idle; and also so that the size of the queue does not exceed the upper bound. It allows the system to dynamically adapt to the current volume of data it is getting, so that when it receives a large amount of tweets per second, the queue size grows and when only few tweets per second are received (for example, during the night), the queue shrinks. For each tweet, we save the text of the tweet, the information about hashtags, links, mentions that were present in the tweet, and meta-information of the user who posted it.

Using this system, we collected about 9.5 million tweets from a list of politicians and political journalists on Twitter, which we extracted from DBpedia. We also collected tweets related to different issues from regular user accounts based on a manually-constructed list of keywords and hashtags. Overall, we have over 24 million tweets for a variety of political keywords and hashtags, and several million tweets on a variety of polarizing issues (including 1.6M tweets for abortion-related keywords and hashtags and 660 thousands tweets for tax-related keywords and hashtags).

Blog data collection

We have also implemented an extensible system for scraping the data from online blogs and comment feeds. The system uses XPath specification to follow the HTML formatting and pagination of the specific website to collect the posts from the beginning till the latest post in the blog. It is designed to collect comments to a post even for websites that use a dynamic JavaScript comment system, such as Disqus or Facebook Comments Plugin. All information is saved in an SQL database. Thus far, we have collected

about 700,000 text units (articles/comments) from the manually created list of most popular blogs that take polarized positions on political and social issues.

Case Study	Source	Category	Type	Number of items	Notes
The Russian / Ukrainian conflict of 2014 (Maidan)	VKontakte	"Active" users	Posts	443,550,000	1,354,000 unique users
2011-2012 Bolotnaya protests in Russia	VKontakte	"Protest" users and groups	Friends	61,000,000	
	VKontakte	"Protest" users groups	Groups	14,000,000	
	VKontakte	"Protest" users and groups	Posts	100,000,000	
The U.S. Elections of 2016	Twitter	Politicians	Tweets	9,411,320	We also have collected related information, such as hashtags and user mentions.
		Political issues		24,159,710	
		Abortion		1,673,131	
		Taxes		687,598	
	Blogs		Articles	622,763	
	Blogs		Comments	13,487	

Table 1. Summary statistics of the data collected to date from different sources.

2. Russian sentiment analysis tools for conflict tracking

In this section, we describe our efforts to develop sentiment analysis datasets and tools for Russian language and conflict tracking analyses we conducted using these tools.

2.1 Existing Sentiment Analysis Data and Tools for Russian

At the beginning of the project, we conducted a thorough review of available systems and resources for Russian, which revealed a complete lack of freely available Russian sentiment analysis systems, despite the recently conducted shared tasks such as SentiRuEval 2016. Moreover, despite the recent success of deep learning methods in sentiment analysis [Yang2016Hierarchical], the proprietary sentiment analysis systems available for Russian tended to rely heavily on manually constructed rules and dictionaries. We therefore decided to develop a new, state-of-the-art open-source system for sentiment analysis for

Russian. To that end, we have conducted a comprehensive analysis of both available resources and annotated corpora for Russian, which we describe below. A summary of all resources is shown in Table 2.

1) ***SentiRuEval-2016*** – This corpus contains manually labeled tweets from "banks" and "telecom" domains. The tweets were labeled with respect to particular entities. For example, a tweet can be "positive" about one bank and "negative" about another and this situation is reflected in the annotation of the tweet. This dataset was used in a shared task SentiRuEval 2016 and, therefore, results of this task could be used as a baseline for our system. The characteristics of the dataset are presented in Table 3.

2) ***RuSentiLex*** – RuSentiLex is a Russian sentiment lexicon that contains 10,467 words and phrases with four sentiment categories (positive, negative, neutral, or positive/negative) and three sources of sentiment (opinion, emotion, fact)
3) ***Rubtsova corpus*** – Since manual annotation is expensive and labor-intensive, there were several impressive attempts to create silver annotation for a large corpus of tweets, using emoticons. Thus, the winning team in SemEval 2016 Task 4 used 90 million tweets labeled in this manner to pretrain the network [Deriu2016SwissCheese]. This approach of automated annotation was used in [Rubtsova2015Postroenie] in order to create a large corpus of tweets with 114,991 positive and 111,923 negative sentiment.
4) ***PNNL datasets*** – Our collaborators from the Pacific Northwest National Laboratory (PNNL) provided us with several silver datasets built at PNNL, including
    a) A dataset of 597,247 VKontakte posts, filtered by a manually created list of political keywords. These posts were annotated with sentiment using the POLARNIK system [Kuznetsova2013Testing].
    b) A corpus of tweets annotated with emotions, using hashtags corresponding to emotion names (such as #радость [joy] and #страх [fear]) and their synonyms. Native speakers manually validated the annotation. A system, trained on these tweets (5,717 emotional and 3,947 neutral) was used to annotate the posts from VKontatke with emotion labels.

Source	Category	Type	Number of items	Notes
SentiRuEval 2016	Banks domain	Tweets	10,890	Labeled with sentiment tweets
	Telecom domain	Tweets	12,705	
RuSentiLex		Lexicon	10,467	Labeled with sentiment words and phrases
Rubtsova corpus		Tweets	226,914	Automatically labeled with sentiment tweets

VKontakte	Contrasting public opinion dynamics and emotional response during crisis	Posts	597,247	Filtered by manually created list of political keywords, automatically annotated with sentiment and emotions
Twitter	Contrasting public opinion dynamics and emotional response during crisis	Tweets	5,717 + 3,947 neutral	Automatically annotated and validated with emotions tweets

Table 2. Russian sentiment resources and corpora.

domain	collection	neutral	positive	negative	total
telecom	training	4870	1354	2550	8643
	test	1016	226	1054	2247
banks	training	6977	704	1734	9392
	test	2240	312	722	3313

Table 3. Statistics for the SentiRuEval 2016 dataset

2.2 Neural Sentiment Analysis for Russian

We implemented several sentiment analysis systems to test out different deep learning architectures on existing Russian data. The first system used an Long Short-Term Memory (LSTM) recurrent neural network and two consecutive fully-connected layers with an exponential linear unit (ELU) [Clevert2015Fast] activation function after the first layer and a softmax activation function function to get a valid probability distribution over the sentiment classes (negative, neutral, or positive) in the last layer. The architecture of the system with an unrolled LSTM network, along with the model parameters, is shown in Figure 2.

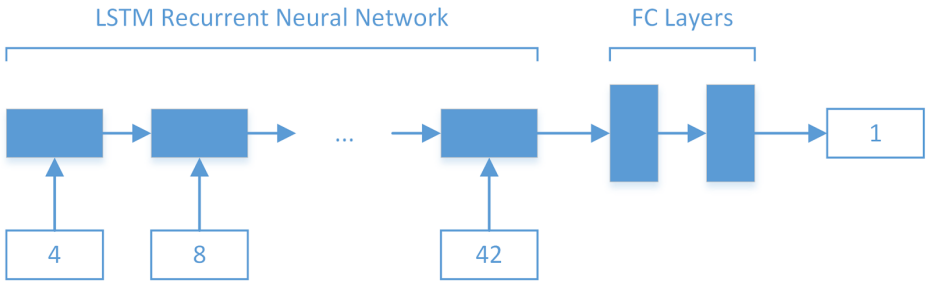


Fig. 2. Model architecture for the prototype sentiment analysis system. Layer size (given in parentheses) is as follows: LSTM (150), FC_1 (100), FC_2 (3).

To represent input sentences in a vector format, we trained **word2vec embeddings using the fastText utility [Bojanowski2016Enriching] on 300,000,000 posts from VK**. Using the derived embeddings, we trained a sentiment analysis model for 10 epochs with 0.6 dropout applied to the LSTM layer. The model achieved a 0.502 F1 score in the "banks" domain using the official SentiRuEval 2016 training and testing data, which would place it third in the official results. We also noticed that 300-dimensional vectors worked better than 100-dimensional. We believe this performance is encouraging, given the basic architecture used for the model.

Since available datasets and resources for Russian language are significantly smaller and not as diverse as English resources, we also decided to explore memory network-based architectures, a relatively new approach that is capable of integrating knowledge directly into the system. For example, [Prakash2017Condensed] successfully integrated medical knowledge in the form of Wikipedia articles to infer clinical diagnoses. Similarly, we intended to use linguistic knowledge to infer sentiment contained in social networks posts.

This model was based on [Sukhbaatar2015End], but instead of using a separate memory for each training sample, it used a global shared memory that was initialized with lexical knowledge from the available lexicons. Usually, such lexicons consist of a list of words with associated sentiment scores. For example, the word "bad" would have a score of -5, the word "good" would have a score of 5, and the word "ok" would have a score of 1. These scores can be directly incorporated into the model using the shared key-value memory. During the processing of tokens in the input sentence, the model produces attention over the memory keys and looks up the corresponding sentiment scores. Using these scores, as well as the tokens in the sentences themselves, the model produces the final sentiment score. As in the original architecture [Sukhbaatar2015End], the model could perform several "hops" of attending to the memory, as this enables better handling of harder cases that involve, for example, negation ("The movie was not bad at all"). An overview of the model architecture is presented in Fig. 3.

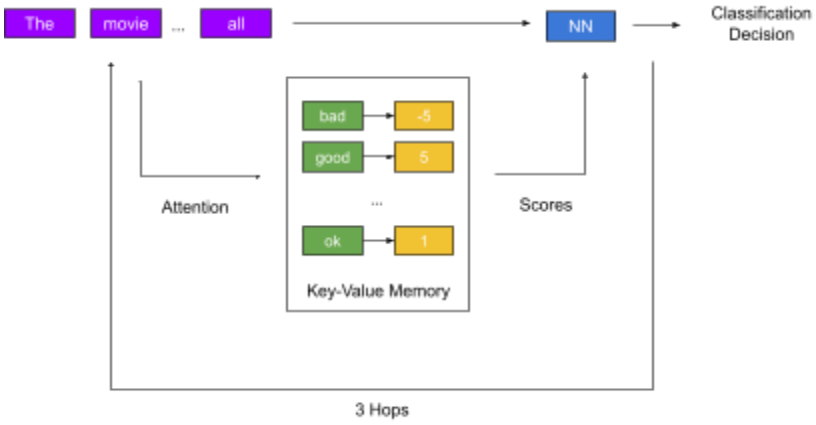


Fig. 3. Model architecture.

We examined several different configurations of the model shown in Fig. 3. Tested on data from the SentiRuEval challenge [Lukashevich2015SentiRuEval], the proposed models achieved results close to state of the art. The results of these experiments are shown in Table 4. For a relatively small SentiRuEval dataset, the enriched key-value memory network improves upon simpler baselines by 3%.

Model	Accuracy
Recurrent Neural Network	0.490
Dynamic Memory Network	0.508
Key-Value Memory Network	0.490
Enriched Key-Value Memory Network	0.530

Table 4. Accuracy on the SentiRuEval test set for different models.

2.3 RuSentiment Dataset: General Domain Sentiment for Russian

A thorough review of available datasets for Russian language showed a lack of resources sufficient for application to real-world social media data. The scarcity of high-quality sentiment analysis datasets is in part due to the difficulty and high cost of obtaining consistent manual annotation, given the numerous flavors in which sentiment can be expressed. For example, it can be expressed explicitly or implicitly; may be directed or not (i.e. the utterance can express how the speaker feels about something or merely express the subjective emotional state of the speaker); mixed sentiment can be expressed in the same utterance; sentiment may be expressed with irony or sarcasm, and so on.

Prior to this work, only two gold-standard sentiment resources were openly available for Russian: RuSentiLex and SentiRuEval. RuSentiLex is a lexicon, rather than a collection of sentences; containing roughly 10,000 entries. SentiRuEval focuses on two specific domains: telecom and banks. In order to develop sentiment analysis models that could be usefully applied to tracking the intensity of online conflict in Russian, we needed a large gold-standard resource for social media posts in Russian. We developed such a resource, the ***RuSentiment dataset*** for sentiment analysis in Russian, making use of the VKontakte data that we had collected in the initial stages of the current project.

RuSentiment, published in [Rogers2018RuSentiment], is a general-domain dataset for sentiment analysis of Russian social media.

2.3.1 Data statistics and annotation guidelines

RuSentiment includes a random selection of VKontakte data that was originally collected for research on political bias. The data contained the posts from the personal “walls” (i.e., posts on personal pages) of the users that were members of anti-Maidan leand pro-Maidan communities during the 2014 Maidan conflict in Ukraine. RuSentiment only includes the posts that were posted outside these communities, and do not contain political keywords. In total RuSentiment includes **31,185** posts, which makes it the largest openly available Russian dataset. Each post was labeled by 3 annotators, with the agreement producing the Fleiss’ kappa of 0.58.

RuSentiment contains labels for 5 classes:

1) the traditional negative/neutral/positive sentiment classes (both implicit and explicit sentiment);
2) formulaic speech acts classes, which include congratulatory posts, thank-you posts and greetings;
3) "skip" class that includes noisy posts (e.g. in other languages), posts that are not clear without context, and posts that were likely not created by the users themselves (e.g. reposts of song lyrics, jokes etc.)

The guidelines provided detailed instructions for 6 frequent cases of mixed sentiment, which helped to increase agreement. Smileys and hashtags were *not* treated as automatic markers of sentiment, as is often done in silver-standard datasets: for example, posts with "hedging" smileys that were the only markers of sentiment were considered neutral.

The annotation guidelines are released in both Russian and English versions, and could be quickly adopted for other languages. The annotation was conducted via the custom web-interface (Fig. 4), with each annotator processing 250-350 posts per hour on average. There were six annotators from Russia and Ukraine, all of them native speakers with linguistics background. Before starting the actual work, all of them completed training with gold-standard annotations and feedback.

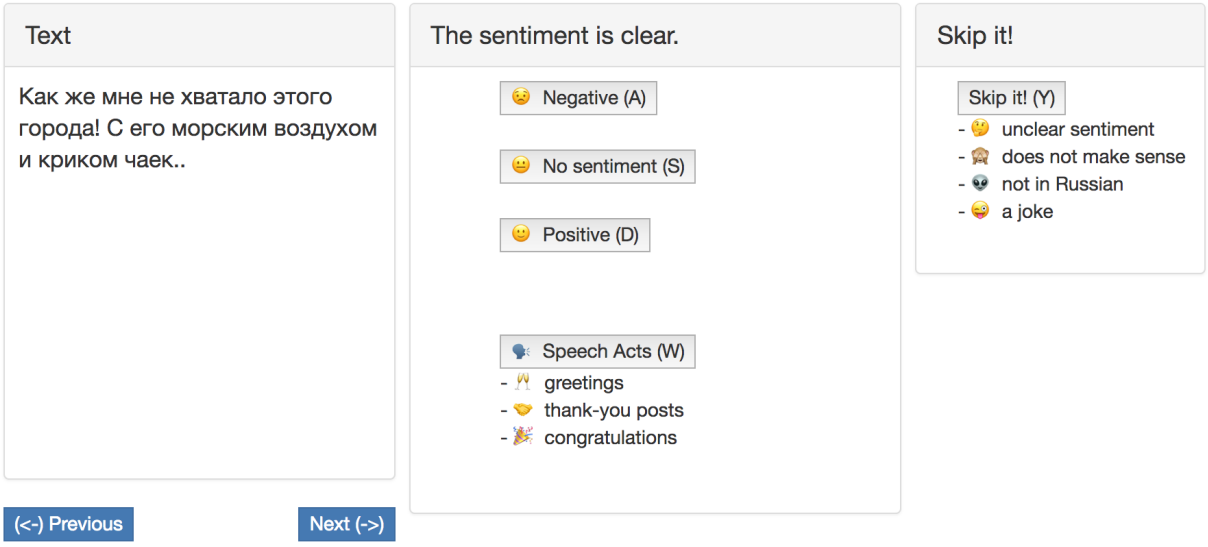


Fig. 4. Annotation web-interface for RuSentiment

2.3.2. Diversifying sentiment data with active learning

One major problem for sentiment analysis in general is that most speech is not sentiment-laden, and the classes of most interest (positive and negative sentiment) are comparatively more rare. Our RuSentiment data [Rogers2018RuSentiment] is not an exception: roughly 40% of the posts are neutral, 20% are positive and 10% are negative, with the rest divided between speech act and "skipped"posts.

The problem is exacerbated with the fact that sentiment is a very diverse phenomenon. As mentioned in section 1.1, the annotation guidelines for RuSentiment dataset included different types of

positive/negative sentiment: explicit (e.g. *I am sad*) and implicit (e.g. *My cat has died*), emotion (e.g. *I am sad*) and evaluation (e.g. *This is a bad cat*). In conjunction with the fact that the neutral class usually far outnumbers the positive/negative sentiment, this means that the automatic classification is not easy: the classes of interest are both comparatively rare and highly diverse.

We attempted to alleviate this problem by creating a part of RuSentiment dataset with active learning. After completing the initial annotation round (20,412) posts, we have experimented with 5-class classification by 4 baseline classifiers of different types, as summarized in Table 5. The posts were represented with TF-IDF [Schütze2008Introduction] scores or as average of FastText embeddings trained on Wikipedia, Common Crawl and our in-domain VKontakte corpus. The latter performed the best for all classifiers, and the best system was a neural-net-based classifier with 4 fully connected layers and non-linear activation patterns.

Model	**Feat.**	**F1**	**Prec.**	**Rec.**
Logistic Regression	CC	0.526	0.619	0.513
	VK	0.622	0.691	0.611
	Wiki	0.574	0.652	0.559
	TF-IDF	0.626	0.654	0.615
Linear SVM	CC	0.586	0.589	0.610
	VK	0.687	0.690	0.691
	Wiki	0.632	0.628	0.646
	TF-IDF	0.664	0.660	0.670
Gradient Boosting	CC	0.527	0.619	0.509
	VK	0.624	0.692	0.611
	Wiki	0.577	0.646	0.557
	TF-IDF	0.587	0.605	0.588
Neural Net Classifier	CC	0.604	0.603	0.623
	VK	**0.717**	**0.718**	**0.717**
	Wiki	0.661	0.658	0.666
	TF-IDF	0.593	0.599	0.589

Table 5. Baseline classifier performance (20,896 posts)

Fig 5. Distribution of true labels for binary NNC: "negative vs. other classes" and "positive vs. other classes"

Table 5 reports the average recall in the range of 0.7, but we found this to not be the case for binary classification for positive and negative sentiment. Fig. 5 shows the distribution of true labels in the output of binary classification by NNC, which suggests that a post was nearly equally likely to be placed in any probability bin. Thus we attempted to alleviate the problem by annotating additional 3,500 "negative" and 2,500 "positive" posts with certainty sampling (Koncz and Paralic, 2013; Fu et al., 2013): drawing an equal number of samples from the probability bins 0.3-0.7. 6,950 posts were annotated, and Fig. 6 shows that active learning did significantly change the class distribution.

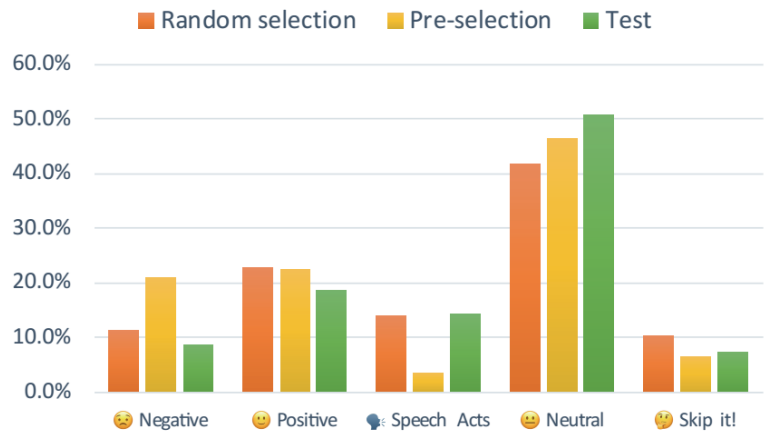


Fig.6. Distribution of classes in randomly selected and pre-selected parts of RuSentiment

However, we found that that effect came at the cost of inter-annotator agreement which went down from Fleiss' kappa 0.654 (for random posts only) to 0.449 (for the pre-selected posts only). Both our analysis of a sample of pre-selected posts and the annotators' comments suggested that the pre-selected posts had a higher ratio of middle-ground cases that were ambiguous between neutral and positive/negative sentiment.

Still, the overall effect of inclusion of posts selected with active learning was unambiguously positive for all classifiers (Table 6). Additional experiments with varying training set size (Fig. 7) suggest that the active learning effect is also muffled by the plateau effect, and it is possible that if active learning was employed at an earlier stage of the project the same accuracy range would be achievable with a smaller, but more diverse dataset.

Classifier	Data	F1	Precision	Recall
Logistic Regression	base	.679	.681	.683
	full	.688	.695	.695
Linear SVM	base	.679	.683	.684
	full	.681	.690	.687
Gradient boosting	base	.683	.684	.687
	full	.685	.693	.692
Neural Net Classifier	base	.685	.696	.692
	full	.716	.720	.721

Table 6. Results on the base (random) and full (random+pre-selected) posts

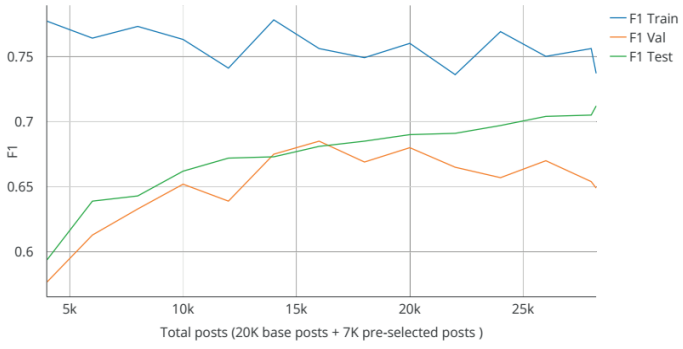


Fig. 7. Effect of increasing training set size in RuSentiment

2.4. Political Sentiment Analysis for Pro-Russian and Pro-Western Ukrainians.

While general domain sentiment analysis datasets such as RuSentiment [Rogers2018RuSentiment] should remain relevant in the foreseeable future, political datasets are mostly case-specific. Our 2014 Maidan conflict data from VKontakte social network is a case in point. First of all, many politicians who were

active and influential in 2014 are no longer active (e.g. Yanukovich, Yatsenyuk). Second, Ukraine has since banned VKontakte and other Russian social networks, increasing the segregation of users. Last but not the least, since Maidan Ukrainians have increasingly been speaking Ukrainian rather than Russian, with the choice of language also being a political statement. Any further studies on Ukrainian social media conducted solely on Russian data would be extremely one-sided.

The aspect of this data that is still relevant and likely to remain so in the near future is the linguistic separation of the pro-Western Russian-speaking Ukrainians and the supporters of the "Russian world". Qualitative analysis of data samples suggests that their "languages" differ significantly in two aspects:

1) **Vocabulary**. There is a growing number of diverging referring expressions for the same phenomena which simultaneously function as evaluative terms, and which tend to be used exclusively by one of the opposing sides of the conflict. For example, the pro-Russian media in Donetsk refer to the Kyiv government as the "junta", while the pro-Kyiv side refers to the current Donetsk area as "Donbabwe" (a blend of "Donetsk" and "Zimbabwe").

2) **Typical rhetorical structure and topics**. The general sentiment patterns and rhetorical structure of posts by the users on the two sides were different initially, and seemed to remain so. The pro-Western side of the conflict generally tended to discuss a wider range of topics: they do not focus exclusively on the conflict with the "Russian world", also talking about e.g. next steps for the country, how to fight corruption, etc. The pro-Russian posts, however, mostly follow the "whataboutist" agenda of official Russian media [Headley2015challenging]: the negative news about Ukraine/West and the Ukrainian nationalists that are portrayed as wannabe-killers of everybody Russian. Both sides emotionally accuse each other of war crimes.

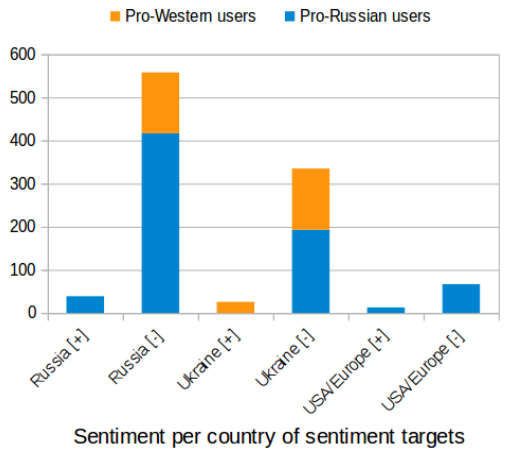
As a follow up to the work on RuSentiment, we conducted a pilot of aspect-based annotation of political posts that focused on different targets of sentiment in pro-Russian and pro-Western posts, and stance towards them. The posts were pre-selected by a list of over 120 political keywords, which improved the class imbalance problem we observed while constructing RuSentiment [Rogers2018RuSentiment] (since political posts of unhappy citizens are more rarely neutral than social media posts in general). The pilot annotation covered the following:

1) The stance of the poster (pro-Russian, pro-Western);
2) The target of the sentiment (country, people, army, politician, etc.);
3) The affiliation of the target of sentiment (e.g. Russian or Ukrainian politicians).

We annotated the walls of 50 users that were members of the Maidan community, and 50 anti-Maidan user walls. The Maidan users had 528 sentiment targets in 339 posts, while anti-Maidan users had 763 sentiment targets in 304 posts.

The breakdown by types of sentiment targets suggests that on both sides it is the people on the opposing side that are most frequent sentiment targets: 91.3% for pro-Western, 77.9% for pro-Russian users. The bulk of the remaining posts was devoted to politicians, mostly of the opposing side. Fig. 8 shows that a lot of posts have unexpected sentiment value, e.g. negative sentiment for Russia-associated targets in the

posts of pro-Russian users, or positive sentiment towards the West also in the pro-Russian user posts. This is due to the posts describing a negative socioeconomic situation. Neither side has much positive to say about "their" preferred country, which forces the argumentation on picking the lesser of two evils.

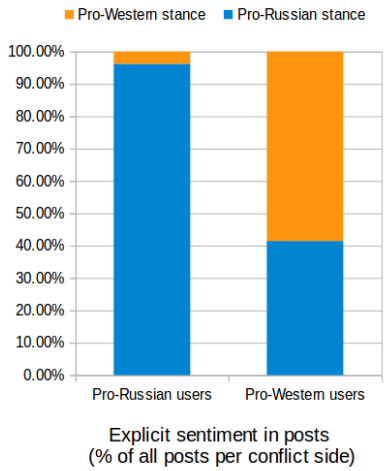


Example of positive sentiment (with regards to Russian government) followed by negative:
While **our leadership continues to score goals** in foreign policy and deal with the danger of war in our territory, the **saboteurs in the government** hit the fifth goal of General Management Means.
Example of negative sentiment (towards Ukrainian state) by a user not supporting annexation of Crimea:
I'm ok with living in a unified Ukraine, but not according to the criminal laws, not the Bandera way, these are not our heroes.
pro-Russian, anti-Western post example:
Brother slavs, brother Russians. If we die - make the Bandera Nazis pay for it, go after them all the way to Lwow, Latvia and Litva! Better even, all the way to Washington, the cradle of all fascism on the planet!

Fig. 8. Breakdown of sentiment targets (non-parodies) for pro-Western and pro-Russian users, per country with which the target is associated. About 80% of targets are the Russians/Ukrainians in general, with most of the remainder being politicians.

Interestingly, only pro-Russian users seem to be talking about the West at all, while the pro-Western Ukrainians are focusing on their own country and Russia. This is consistent with another general strategy of Soviet and now Russian propaganda: the higher ratings at home are achieved through the creating the impression of the homeland being under siege (by the West), and the necessity to unite in the face of the enemy. This requires the West to be always on the radar in the media.

Fig. 8 only includes data for the "matching" stance of the users, e.g. pro-Russian users expressing pro-Russian stance. Fig. 9 presents the statistics on the types of stance, which unexpectedly often does not match the political affiliation of the user – at least on the surface form. This unexpected finding is explained by a large number of *posts parodying the reasoning and/or style of the other side*. The pro-Western Ukrainians used this rhetorical structure significantly more often, as shown in Fig. 9.



Parody of pro-Western argument:
Are they f*cking kidding me? I came to see Europe. I got to the passport control. I'm all revolutionary, wearing a helmet, EU flag on my bag. I'm holding a ticket and my passport. But they are not letting me in! They say - yes, it's true, you don't need a visa, but show us your health insurance, your hotel booking confirmation, your bank account, your health certificate proving that you don't have AIDS, lice, or helminths, a police certificate approved by the consul and a consul certificate approved by the police, employment certificate, and your family records. Well f*ck off Europe, I'm going skiing in Carpathian mountains. We're already Europe anyway. Glory to Ukraine!

Parody of pro-Russian argument: Why, do you think people in the West do not eat shit??? Just look at the US debt!!! And in Norway they dress giraffes as children and rape them!!! And with Putin Russia is no longer on its knees!!! Crimea is ours!!! Why am I even talking to you, America is paying you in burgers, you're traitors, even though your grandfathers have fought in the great war!!! And also shit is actually good for you, vegetables grow well with it!!!

Fig. 9. Percentage of explicit sentiment targets in posts of pro-Western and pro-Russian users. The large number of posts with unexpected affiliation (pro-Western for pro-Russian users and vice versa) is due to parodies, such as shown on the right.

The distribution of posts in this pilot sample may not be exact, but it is overall consistent with our qualitative analysis. Due to the difficulties with consistent selection of spans by multiple annotators, the pilot annotation was conducted in two stages: first, the evaluative target spans were selected by a single annotator, and then two more annotators confirmed or rejected the sentiment labels selected by the first annotator.

We also compiled a pilot "Donetsk-Kyiv" glossary that was relevant at the time, and contained about 50 entries. Table 7 lists a few examples of this data. There are often symmetrical entries for how both sides frame an entity: e.g. the Ukrainians refer to the Donbas militants as *терористы* ("terrorists"), while the pro-Russian media describes them exclusively as *ополченцы* ("citizen soldiers / home guards"). Sometimes there is no such symmetry: for example, pro-Western Ukrainians refer to the Russian soldiers in Donetsk as *ихтамнеты* (literally "they who are not there"), while pro-Russian media for obvious reasons do not mention them.

Referent	Stance	Term	Notes
Putin		хуйло	A colloquialism for "dickhead", used by Ukrainians nearly exclusively to refer to Putin since 2014.
		Путин, В.В. Путин, Владимир Владимирович	Capitalization and/or spelling full name seem to also mark respect & political stance by itself.
Ukrainians		майдауны	A pro-Western Ukrainian who supported the Maidan protests (a blend of "Maindan" and "Down syndrome")

Russians		хохлы, каклы	Traditional Russian ethnic slur for Ukrainians that references the traditional Cossack haircut. The latter term is a new derogatory attempt to imitate the Ukrainian pronunciation of the former.
		москали, кацапы	Traditional Ukrainian ethnic slurs for Russians that reference Moscow (москаль) and goatee beard (кацап).
		россияне (as opposed to “русские”)	There are two terms for “Russian” in Russian: русский refers to language/culture/nationality, while российский refers to the Russian state. The latter is now used more often by the Russian-speaking Ukrainians to disassociate the language from the Putin policy.
		русские (as opposed to “россияне”)	The Russian foreign policy attempts to implement the idea that “Russian language” extends to “Russian culture” and then “Russian world” (“русский мир”), which according to them should include Ukraine. It is key for the “all Slavs are brothers” rhetoric to use the adjective that technically denotes Russian language/culture rather than the Russian citizenship.

Table 7. Framing glossary of Russian-Ukrainian conflict (sample)

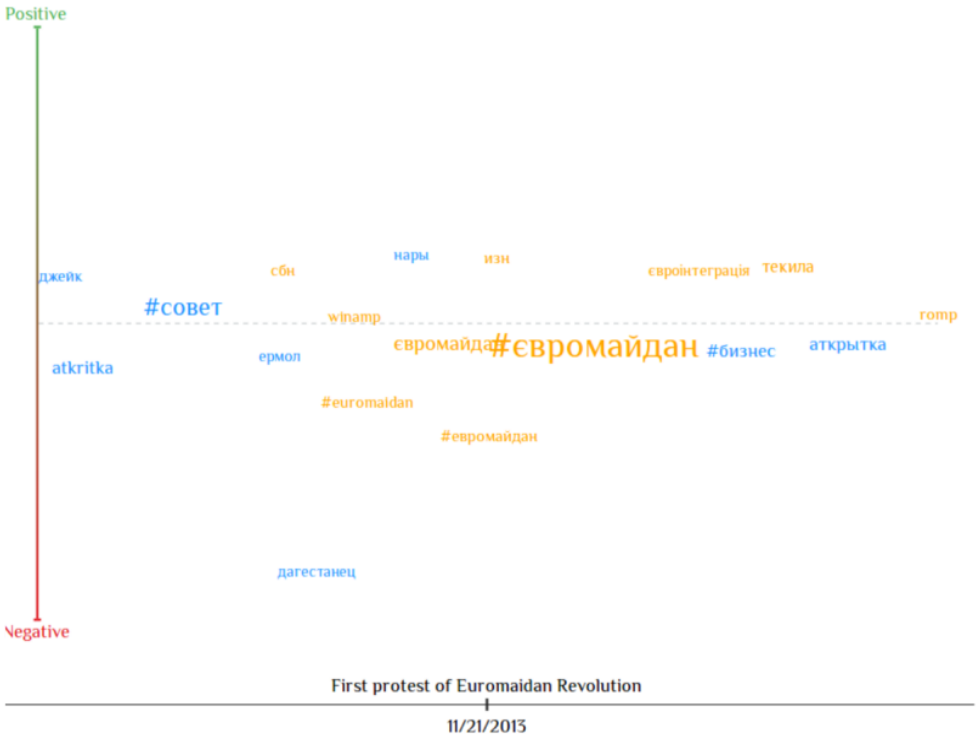
2.5 Using General Domain Sentiment Analysis for Conflict Tracking

In order to establish the most sentiment-laden topics discussed throughout the conflict, we divided a one year time frame into week-long slices and collected and analyzed user posts for each slice. Our goal was to visualize intuitively how different sides of the conflict saw and discussed the same events and political activists and how the difference in their views evolved with time.

We performed lemmatization and also transliteration of Ukrainian words that were directly mappable to Russian words, whenever possible. For each of the two communities (“Euromaidan” and “Antimaidan”), we computed term frequency-inverse document frequency (TF-IDF) values which evaluate how important every word in a weekly collection of posts of a given community is with respect to the whole time frame. The benefit of TF-IDF is that it discards the words that are equally popular among the entire dataset and gives higher scores to the words that only appear in specific time slices.

Using TF-IDF scores produced by the method, we extracted the top 10 most important (i.e. most discussed) words per week for each of the two camps. Next, we built a neural network-based sentiment analysis model with one layer of GRU units [Chung2014Empirical] and computed the average sentiment score over all the posts containing these words. This enabled the analysis of trends for each of the groups to mention a certain concept in a predominantly positive or negative way. For clarity of visualization, we used a force-driven words layout, where two words would attract or repulse depending on how close the contexts they appear in were. To quantitatively estimate this similarity, we used the Euclidean distance between word embeddings, derived through training a fastText [Bojanowski2016Enriching] model on the whole corpus of users’ posts. We also manually annotated each of the time slices with the actual events occurring within a given slice, in order to be able to overlay the sentiment towards specific concepts with the actual events occurring on the ground.

Fig. 10 shows screenshots from our demo with two different time segments as examples of the resulting layout. The interface has a lever that allows the user to select the desired time period. In the layout, the font size of every word reflects its TF-IDF score (i.e. importance). This visualization makes it easy to confirm, for example, that as the conflict evolved, the pro-Ukrainian side tended to mention the conflict more frequently and in a substantially more positive way than the pro-Russian side. Ukrainian orthography began to appear, showing evidence of the pro-Ukrainian side gradually becoming further separated from the pro-Russian side.



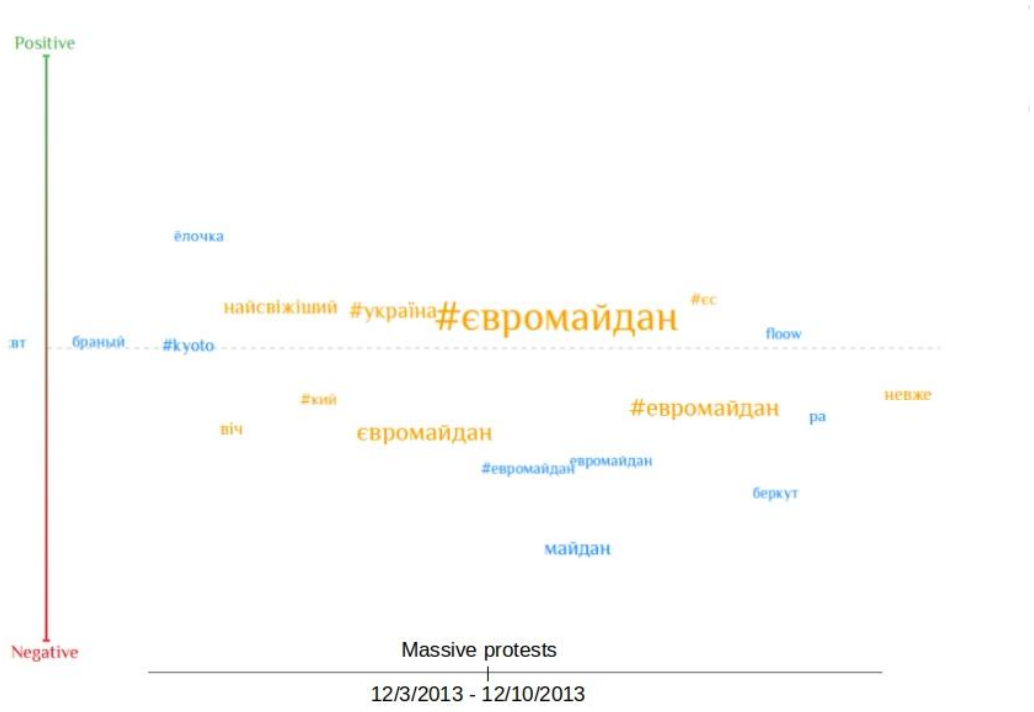


Fig. 10. Sentiment visualization of the topics discussed weekly by Euromaidan (orange) and Antimaidan (blue) communities. The most important keywords shown for late November 2013 (top) and early December 2013 (bottom).

3. Predicting protest events by analyzing social media

We conducted several studies evaluating our ability to predict and/or identify the timing and magnitude of real-life protest events that take place in the streets from the social media data stream. Specifically, we were interested in identifying the linguistic features that correlate with intense manifestations of conflict. Below we describe the *Bolotnaya* protest dataset and the outcomes of the studies we conducted, the last of which was published in [Rogers2019Calls].

3.3.1 "Bolotnaya" dataset

The "Bolotnaya" dataset contains posts, likes and groups of users from VKontakte, the largest Russian social network. It was created on the basis of a list of 476 protest groups, which was compiled by Yandex (the largest Russian search engine) and used by the New Media Center (Moscow, Russia) in 2014 to collect the data for these user groups. The dataset includes the posts from these user groups (218386 records), the users (221812 records), and their likes (837906 records) and friends (11343216 records). The data is used by an agreement with New Media Center. Table 8 provides the basic statistics for the dataset.

Riot groups	476
Riot groups' members	1 459 632
Riot groups' posts	91 553
Time frame covered	Aug 2010 - Oct 2014
Users liking riot groups' posts	57754
Total links between users who liked the same posts	395 636 801
Average likes-based graph degree	6850
Riot communities examples	❤ Краснодар ("❤ Krasnodar"): https://vk.com/public32822100 Голос Кемерова ("Voice of Kemerovo"): https://vk.com/public33048737

Table 8. Bolotnaya dataset statistics.

3.3.2. Topic spikes as predictors for protest rallies

This part of our work focused on developing analytical tools for tracking the most important trends and events in social media. Specifically, we developed a filtering algorithm that is based on the changes in word frequencies over time. The general flow of the analysis is inspired by traditional statistical natural language processing [Manning1999Foundations] and is as follows:

1) We group the posts by the year and month and sort the posts from each month by the number of likes they received, retaining only the top *k* most liked posts;
2) We then combine the posts from the same month into a single document and compute TF-IDF scores, retaining only the top *n* most important words according to the TF-IDF score.
3) Finally, we filter out the words that were in the top *n* for only one month.

The intuition for this is that it will filter out the noisy entities that are most likely coming from the locally popular posts that quickly gain popularity and are immediately forgotten. For each remaining token, we calculate the maximum change in the TF-IDF score across the target months, sort them in the decreasing order, and leave only top *m* tokens that have the largest changes in score. Finally, we calculate the pairwise correlation in TF-IDF scores between every token in this list and create "clusters" of most correlated tokens.

The resulting plots for several word clusters are shown in Figures 11-12, which suggest that this method is reasonably effective at aligning the social media text with prominent political events. For example, the words shown in Fig. 11a had a strong spike in June of 2012, which corresponds to the so-called "March of Millions" that happened in Moscow on June 12th; it was a large protest rally which reportedly had over 100,000 participants. Fig. 11b shows that during the protest rallies in Moscow (including May 2012 and January 2013) the posts on the group wall frequently mentioned such words as "demand" and "march".

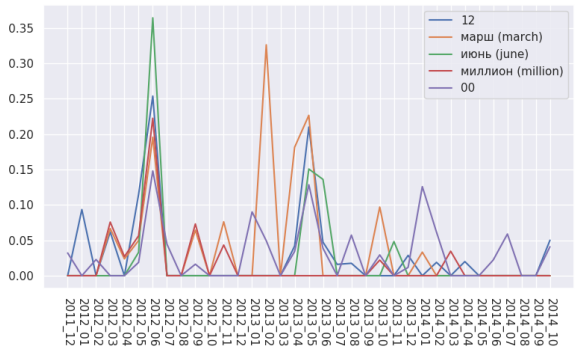


Fig. 11a. Use frequency trends for the words which correlate the most with the token “12”. June 12th is the date of the “March of Millions” on June 12th in Moscow (“00” came from the fact that the announcements of the march contained the time).

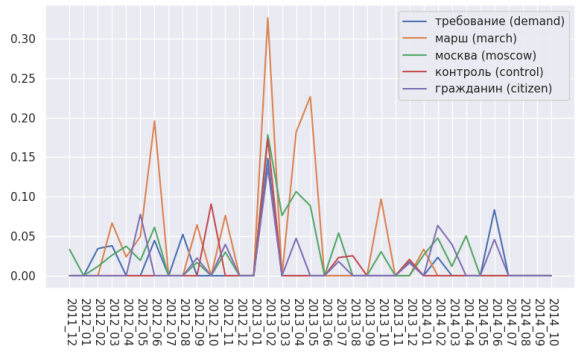


Fig. 11b. Use frequency trends for the words which correlate the most with the token *требование* (“demand”). Note the similarities to the graph on the left.

Similarly, the spikes in Fig. 12 corresponds to the Crimea crisis of 2014. Note that the words “наш” (“our”) and “крым” (“Crimea”) constitute the phrase “крым наш” which has come to denote the common Russian attitude towards the annexation. This phrase is typically used with either extremely negative or extremely positive connotations, depending on the user’s political views.

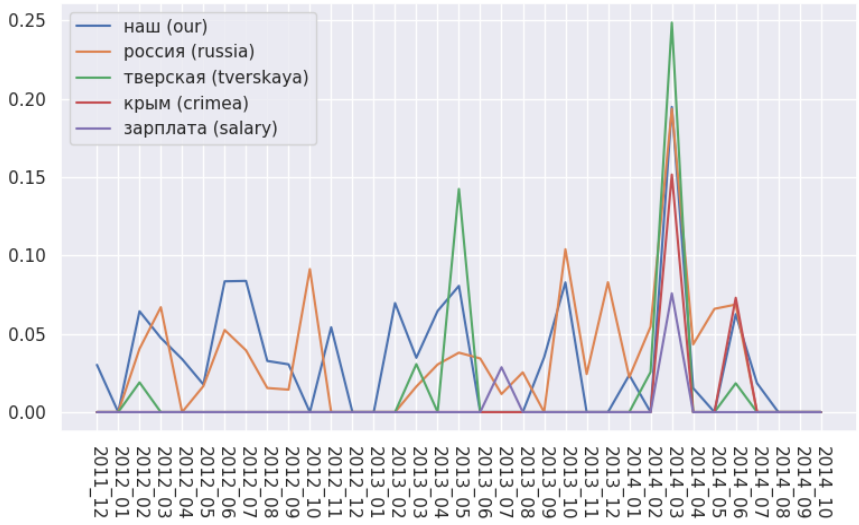


Fig. 12. The words which correlate the most with the word “our” (top). The plot shows the importance of the following tokens (top to bottom): “our, “Russia”, “Tverskaya”, and “Crimea”. Note that “our” and “Crimea” (“крым наш” in Russian) is a widely used neologism which has come to carry either extremely negative or extremely positive connotations.

Topic frequency spikes in social media streams have been used to detect and track natural disasters and other events [Atefeh2015survey], and our present analysis confirms the same holds for significant political events. A qualitative study of these results suggests that the present term filtering method identifies reasonably effective indicators for significant political events.

3.3.3. Calls to action as predictors for protest rallies

We hypothesized that protest rallies that occur in the streets will coincide with increased agitation, dissatisfaction, and *calls to action* in the social media. Calls to action on social media are known to be effective means of mobilization in social movements, and a frequent target of censorship.

We limited our consideration to the time frame between December 2011 and July 2013, which corresponded to the period of active political protests in Russia. Throughout this timeframe, a number of rallies occurred, which were accompanied by numerous discussions in social networks. We extracted the posts on "riot" groups' walls, grouped them by date and left the top-100 most liked posts to reduce lexical noise irrelevant to political topics. Our assumption was that any protests were likely to be preceded by a growing amount of social dissatisfaction that is expressed in the posts.

Preliminary study

In a preliminary study, we focused on calls to actions specifically, which in a simplified conception, we operationalized simply as the number of sentences which ended with an exclamation point in the posts for each month. An initial qualitative analysis suggested that such sentences in these groups tended to include general expressions of dissatisfaction and direct calls to action. Typical examples would be "6-7 мая сотворим историю своими руками!" ("Let's make history ourselves on the 6-7th of May!"), "За Россию без Путина!" ("For Russia without Putin!") and "15 сентября вся Россия выходит на Марш миллионов!" ("Entire Russia is going to the March of Millions on the 15th of September!"). We summed the counts of calls-to-action for every month to represent the overall dynamic of social attitude over time. Using Wikipedia data about the attendance of individual rallies [Wikipedia2018Russian], we took an average of the participants counts and used it to describe the rallies size/importance. We plotted both the count for calls-to-action and the size of the rallies over time in order to check if there was any correlation between them (see Fig. 13). If there were no notable rallies during certain months, we manually set size of the rallies to 0. This simple approach produced a Pearson correlation coefficient of 0.51 between the number of monthly rally attendees and the number of exclamatory sentences in the protest groups. The two series are shown in Fig. 13 to illustrate the observed correspondence.

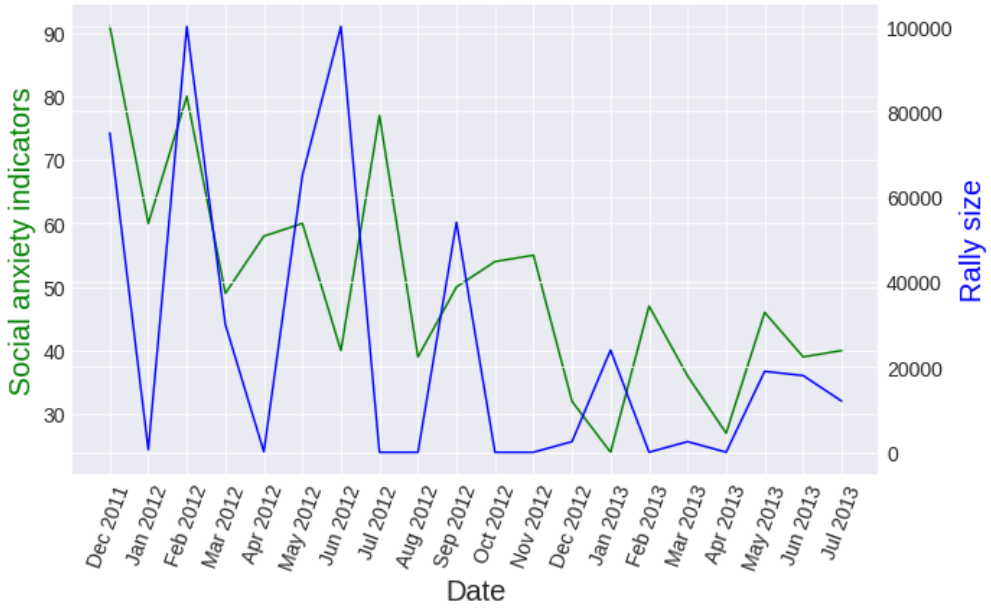


Fig. 13. The correlation between the linguistic features showing social anxiety and the attendance of the individual protests in the time frame between December 2011 and July 2013. For the months when no meetings occurred, the corresponding rally size is set to 0.

This initial study demonstrated the potential for using verbal indicators in social media to track and/or predict size and attendance of individual protests and other civil unrest events.

Calls-to-action dataset

In a follow-up study, published in [Rogers2019Calls], we examined in more depth linguistic features that mark *calls to action* (CTA), such as imperatives, modal verb + infinitive and zero-infinitive constructions, and others. Specifically, we developed a dataset that included 14 core and borderline types of political CTAs, and and showed that they are relatively easy both to annotate (with IAA 0.78) and to classify (F1 of 0.77, even with a small amount of annotated data). We also showed that in Bolotnaya data, the volume of CTAs on social media had a moderate positive correlation with actual rally attendance.

Prototypical CTAs are imperatives prompting the addressee to perform some action, such as “Don’t let the government tell you what to think!”. This seems like a straightforward category to annotate, but in reality CTAs may be expressed in various ways, including both direct and indirect speech acts. There are many borderline cases that would in the absence of clear guidelines decrease inter-annotator agreement (IAA). To the best of our knowledge, our work was the first to create a detailed schema for CTA annotation in the context of a political protest. Note that we were concerned not so much with CTAs in particular, but with a broader category of “material with collective action potential”. Figure 14 below shows different categories of CTAs in the detailed annotation guidelines we developed.

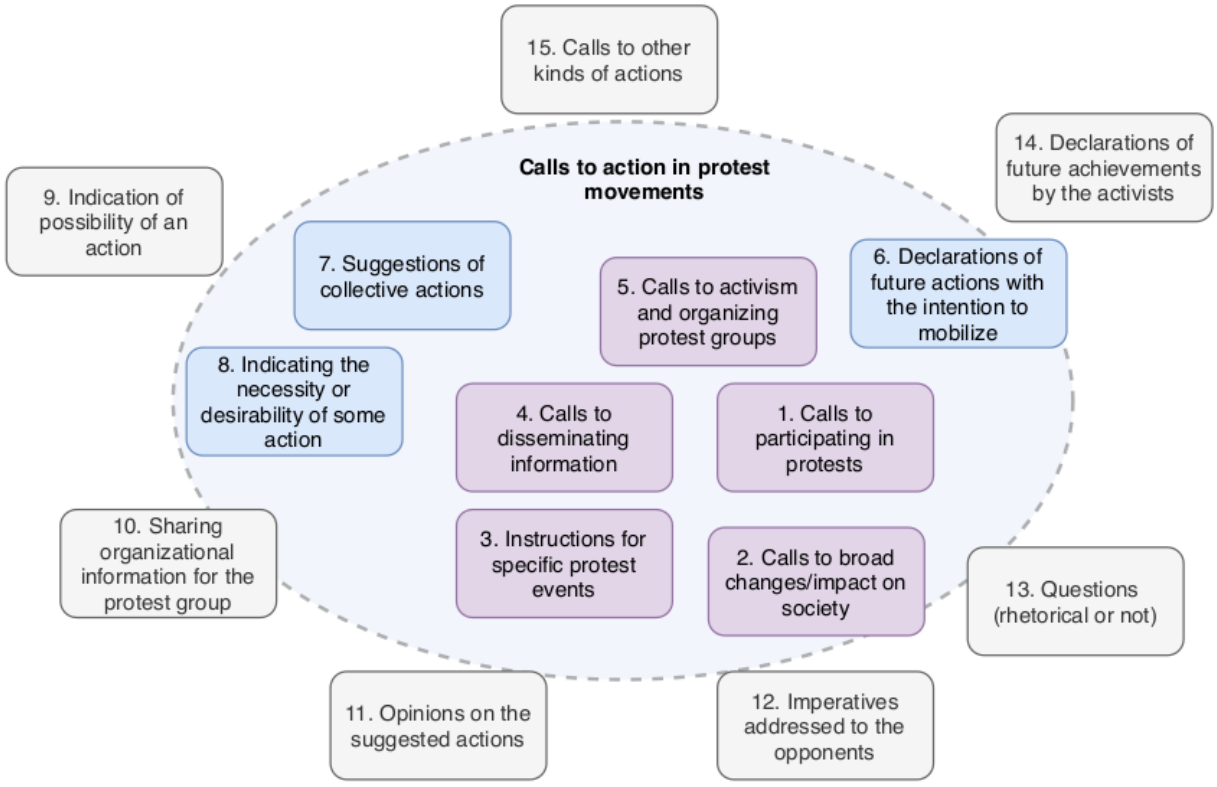


Fig. 14. The core and peripheral cases of political CTAs. This study focuses on types 1-8. Examples for each type:

1. Everybody, join us tomorrow in Sakharov square!
2. If you love Russia, if you love your home city of Smolensk, start the fight with the crooks and thieves!
3. Do not form a line or arrange to meet in a specific place.
4. Invite foreign press and TV – let them see what is going on in our capital!
5. Observers in Kaluga, please respond!
6. That's ok, we will tell them what we think of them even in the square in front of the Central market!
7. I suggest we put on white stripes on our arms as a symbol of honest elections. That's easy to do!
8. On the 10th of March we should come in large numbers!
9. You can download the leaflet with the invitation here.
10. This is the beginning! We will start activities when we will have 50 members. We repeat, participation in this group can
11. only be active.
12. I do like the idea of the government's resignation, but I think your slogans are too emotional. Furthermore, I'm against
13. calling an early election.
14. Out with you, McFaul! And take Putin and Medvedev with you, together with Nemtsov and Chirikova!
15. Is THAT really our choice? (rhetorical)
16. Today at 10 pm Vlad and I are going to post the leaflets around the city. Who wants to help us? (factual)
17. Together we will get rid of Putin's lies and dictatorship!
18. Everybody, come to my birthday party on Saturday!

Since CTAs overall constitute a small portion of all posts, we pre-selected the data for annotation using a manually created seed list of 155 protest-related keywords and phrases, such as "participate", "share", "join", "fair elections", etc. We annotated **1000 VKontakte posts** from **Russian Bolotnaya data**. The annotation was performed on the level of full post, not individual sentences. We considered a post as CTA if it included even one instance of a political CTA as defined above. Ambiguous cases were treated as political CTA, as long as they could function as such: for example, "Join us tomorrow!" could refer to both a protest or a birthday party. Each post was annotated by 3 native Russian speakers, using the

classification interface of Prodigy 1 annotation tool. The inter-annotator agreement as estimated by Krippendorf's alpha was .78. In the end, we obtained 871 posts on which at least 2 annotators agreed. 300 of them were identified as CTAs, and 571 - as non-CTAs.

Automatic detection of CTAs

We randomly split the collected CTA dataset into the train and test parts in the 80/20% ratio. We trained several models to perform post classification, with Logistic Regression (LR) Support Vector Machine classifier with a linear kernel (SVC) used as baselines. We used TF-IDF representations of both original posts and posts lemmatized with pymorphy 3 library (Korobov, 2015). We picked the best regularization hyperparameters for each model through cross-validation based on the average F1 score over 5 folds.

Our state-of-the-art models included two versions of BERT (Devlin et al., 2019): the multilingual model released in the PyTorch repository of BERT 4 , and the Russian version (RuBERT) released by DeepPavlov 5. The latter model is initialized as multilingual BERT and further fine-tuned on Russian Wikipedia and news corpora (Kuratov and Arkhipov, 2019). Both models have 12 layers and 180M parameters. We trained both models for 40 epochs with the batch size of 32 and the learning rate of 5e −5 . We also included a contextual embedder of the ELMo model (Peters et al., 2018) pre-trained for Russian and released by DeepPavlov 6 . The posts were split into sentences using the NLTK library 7 and each sentence token was encoded by the ELMo embedder into a 1024-dimensional vector. The classification was performed by a standard LSTM network (Hochreiter and Schmidhuber, 1997) with a hidden size of 256 units followed by a linear layer. We trained the network for 25 epochs and with the learning rate of 0.001. The results of all the classification experiments are shown in Table 9 below. The best performance was achieved by RuBERT, with LSTM on ELMo close second.

Classifier	Acc.	F1
LR (no lemmatization)	0.78	0.67
LR (lemmatization)	0.82	0.71
SVC (no lemmatization)	0.80	0.68
SVC (lemmatization)	0.78	0.65
BERT multilingual	0.8	0.73
RuBERT	0.86	0.78
LSTM on ELMo	0.83	0.75

Table 9. CTA classification performance

CTAs for predicting social unrest

We ran the trained RuBERT CTA classifier over 91K posts falling in the date range between Dec 2011 through Jul 2013 from the Bolotnaya dataset. Figure 15 shows the volume of posts identified as CTAs, plotted against the Wikipedia data about attendance of individual rallies. When no attendance data is available, we assume that there were 0 protest events. The two green lines correspond to upper and lower attendance estimates. The blue line shows the detected CTAs. Despite the noisiness and incompleteness of the available protest data, Pearson's correlation between attendance estimates and the number of detected CTAs is about 0.4, which is considered to be "moderate". This could make CTAs a useful additional factor to systems based on spatiotemporal, demographic, and/or network activity features.

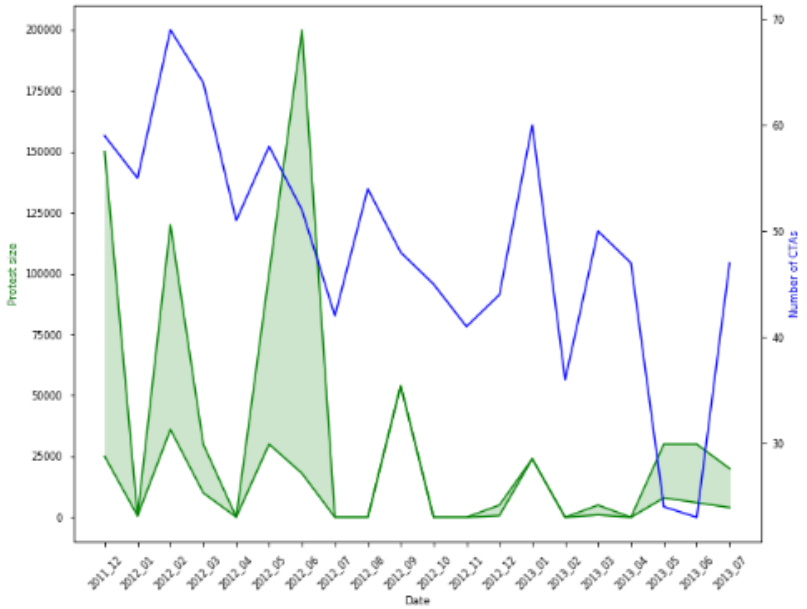


Fig. 15. The correlation between the detected CTAs (blue) and the rally attendance (green) per month. The two green lines reflect the upper and lower attendance estimates depending on the source of data used

We also conducted experiments to estimate the real-world effect of likes and reposts of CTA posts. Intuitively, one would expect that a higher number of likes and reposts of CTA posts should result in higher attendance for protest rallies. To see whether that was the case for Bolotnaya data we calculated the number of shares and likes on posts detected as CTAs by our classifier, and all other posts in the sample. Figure 16 shows these numbers plotted against the attendance of the protest events. The pattern we actually observed in Bolotnaya data is different: before the March of the Millions the average number of both reposts and likes is spiking before a protest event, and going down after it. This corresponds to preparation and the aftermath of a major event. Interestingly, after the March of the Millions there was much like/repost activity which did not result in any larger events. This can be attributed to the introduction of the anti-protest laws that effectively stifled the movement: the link between social media and real world activity clearly becomes weaker.

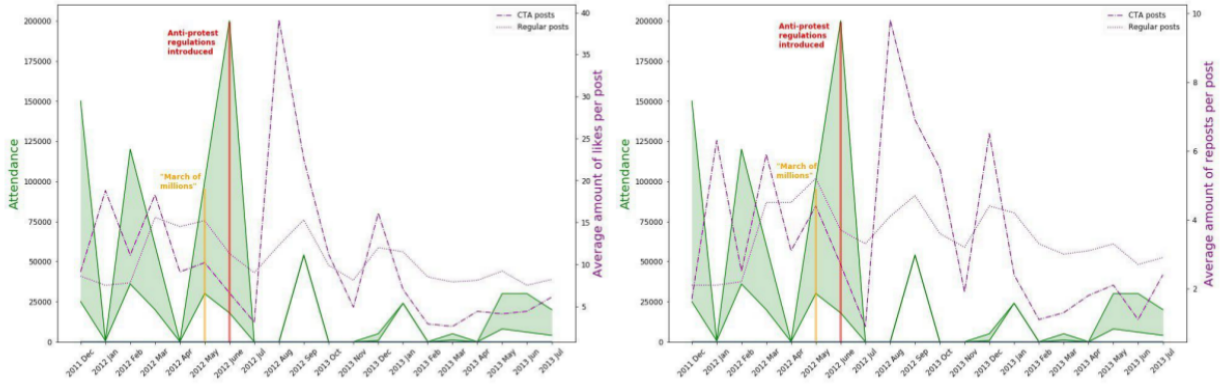


Fig. 16. Average number of likes and reposts on CTA and non-CTA posts vs rally attendance.

4. Argument analysis

One interesting side of the conflict is that the representations of the same events by the opposing sides begin to diverge. We believe that identifying and extracting opposing arguments, as well as the sentiment intensity around them is crucial in understanding how opposing representations diverge. Therefore elucidating the logical structure of argument helps to explain and predict the dynamics of controversy. With this goal in mind, we worked on developing methodologies for argument mining. We focused on the methods that require minimal processing of raw text and dependence on external tools, in order to enable easy adaptation of such methods to new low-resource languages.

Specifically, we developed computational methods for predicting **argument structure**, as well as **argument persuasiveness** and **convincingness**. We worked with several neural network architectures to tackle these problems, including sequence-to-sequence modeling, attention mechanisms, and memory networks. We used English language data for algorithm development and evaluation, since argument mining corpora did not exist for Russian. This work was published in [Potash2017Here], [Potash2017Towards], and [Potash2017Length], respectively.

In order to determine argument structure in text, one must understand how different individual components of the overall argument are linked. We developed the first neural network-based approach to link extraction in argument mining. Specifically, we proposed a novel architecture that applies Pointer Network sequence-to-sequence attention modeling to structural prediction in discourse parsing tasks. We then developed a joint model that extends this architecture to simultaneously address the link extraction task and the classification of argument components. The proposed joint model achieved state-of-the-art results on two separate evaluation corpora, showing far superior performance than the previously proposed corpus-specific and heavily feature-engineered models. Our results demonstrated that jointly optimizing for both tasks is crucial for high performance. Figure 17 below shows an example of argument structure form one of the corpora used in this study. Figure 18 shows the joint pointer network architecture that jointly identified argument component spans and relations between them. Table 10 shows results of the joint architecture against several baselines.

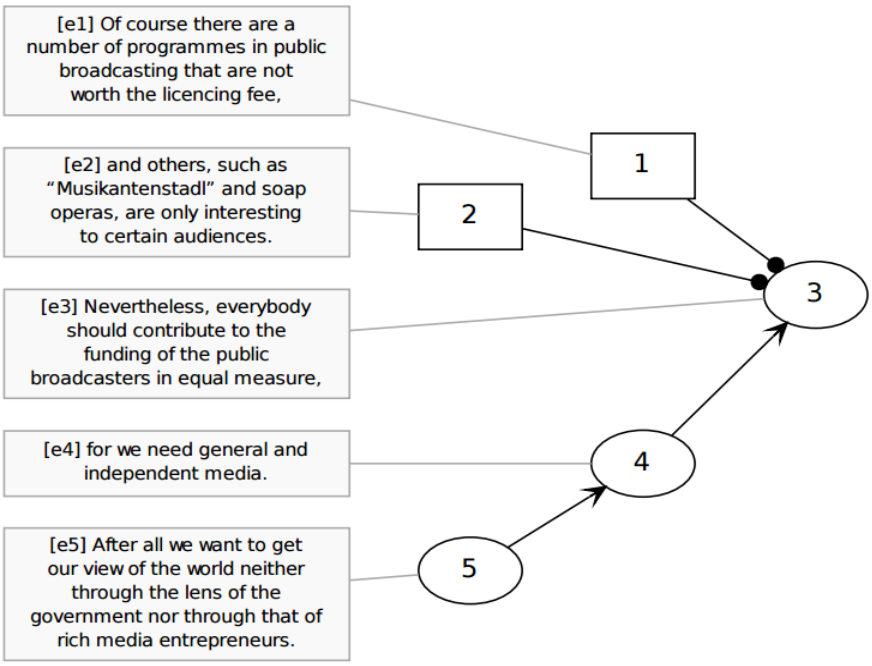


Fig. 17. Argument structure showing the supporting and attacking statements for the debated position from [Peldszus2015Joint].

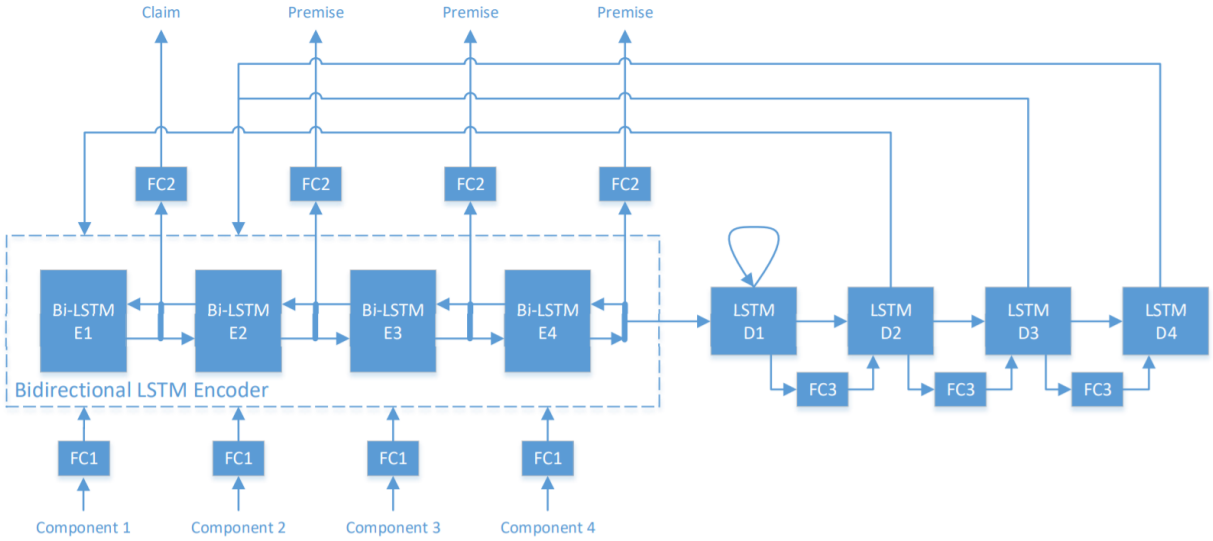


Fig. 18. Pointer network architecture for joint prediction of argument structure and argument components [Potash2017Here].

	Type prediction			Link prediction		
Model	Macro f1	Cl f1	Pr f1	Macro f1	Link f1	No Link f1
Simple	.817	-	-	.663	.478	.848
Best EG	**.869**	-	-	.693	.502	.884
MP+p	.831	-	-	.720	.546	.894
Base Classifier	.830	.712	.937	.650	.446	.841
ILP Joint Model	.857	.770	.943	.683	.486	.881
Joint Model	.813	.692	.934	**.740**	**.577**	**.903**

Table 10. Results of joint argument structure / component prediction

For argument persuasiveness, we introduced a state-of-the-art recurrent predictive model for predicting debate winners. By having an accurate predictive model, we were able to objectively rate the quality of a statement made at a specific turn in a debate. The model was based on a recurrent neural network architecture with attention, which allowed the model to effectively account for the entire debate when making its prediction. Our model achieved state-of-the-art accuracy on a dataset of debate transcripts annotated with audience favorability of the debate teams. Figure 19 shows the joint training objective that leveraged audience favorability for prediction of debate winners.

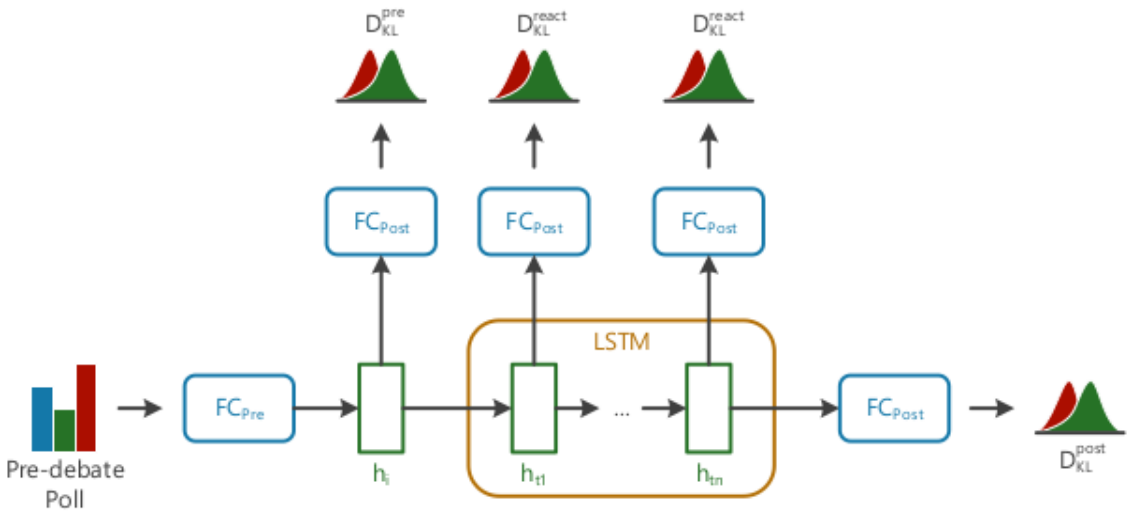


Fig. 19. Illustration of the training objective that leverages per-turn audience favorability to predict debate winners [Potash2017Towards]. Kullbach-Leibler divergence between actual and predicted audience response was used during training.

Table 11 below shows that audience favorability regularization (LSTM+reg) produces a substantial improvement over both LSTM model with attention (LSTM+Att) and over logistic regression models using argument flow (LR Flow, LF Flow*) and audience reaction (LR React) features:

Model	Accuracy
LR BOW	0.50
LR React	0.60
LR Flow	0.63
LR Flow*	0.65
LSTM	0.55
LSTM + Att	0.57
LSTM + Reg	0.64
LSTM + Att, Reg	**0.71**
LSTM + Att, Drpt	0.60

Table 11. Results on the debate winner prediction task

For argument convincingness, our study provided insight into three key aspects related to predicting argument convincingness. First, we explicitly displayed the power that text length possesses for predicting convincingness in an unsupervised setting. Second, we showed that a bag-of-words embedding model posts state-of-the-art on a dataset of arguments annotated for convincingness, outperforming an SVM with numerous hand-crafted features as well as recurrent neural network models that attempt to capture semantic composition. Finally, we assessed the feasibility of integrating external knowledge when predicting convincingness, as arguments are often more convincing when they contain abundant information and facts. We experimented with different methods, from bag-of-words and simple similarity heuristics to using memory networks to compute the weighting for the relevant Wikipedia articles [Potash2017Length].

5. Bias classification

We created **a silver dataset of biased news articles** pulled from VKontakte, the biggest Russian social network which at the time of data collection had 320 million registered users and is the most popular social network in both Russia and Ukraine. During the 2014 Maidan conflict in Ukraine, both pro-Russian (also known as “Antimaidan”) and pro-Ukrainian side (also known as “Pro-” or “Evromaidan”) were represented online by large numbers of Russian-speaking users. We had originally pulled 10,000 hyperlinks to outside sources (such as news articles and blog posts) posted by Promaidan and Antimaidan communities, each of them having over 2 million members. An overwhelming majority of these links were posted only by one side of the conflict, with less than 1% posted by both, indicating a very high degree of polarization. We collected the articles mentioned by the two opposing communities, creating a silver dataset of online news preferred by each side.

In order to assess whether this silver data could be used to predict the bias in the news articles available from news aggregators, we also created a small **manually annotated gold standard dataset of biased articles**. We collected links from the first 5 pages of Google News Russia by using “maidan” and “Ukraine” query words. This resulted in a total of 1,039 links, out of which only 219 were present in our silver dataset. Out of these 1,039 links, 675 were active at the time of the annotation. Two annotators labeled the articles on a scale from - 2 to 2, where -2 was strongly Antimaidan, -1 was weakly

Antimaidan, 0 was neutral, 1 was weakly Promaidan, and 2 was strongly Promaidan. The annotators could also assign the label "N/A" if the article was not related to the Maidan crisis. After merging non-zero labels, there were 40 Anti, 92 Pro, and 215 neutral articles on which the annotators agreed.

We show that this silver data can be used to detect high-accuracy models for bias detection. In particular, we show that a Naive Bayes classifier, using just the domain name as a feature, trained on the silver data derived from the user linking patterns, can achieve (1) 90.3% accuracy predicting which user community posted a link (2) 82.6% accuracy predicting whether the article would contain Pro- or Antimaidan viewpoint. This conclusively confirms the fact that user citation patterns are very strong predictors of source bias. This work has been published as a workshop paper [Potash2017Tracking], please refer to it for further details on both this data and the classification results.

6. Relationship between verbal and non-verbal conflict indicators

As part of our work on creating a composite index of conflict intensity, we investigated the feasibility of tracking controversy over time by looking at the relationship between different controversy measures, including the quantitative controversy measures proposed by Garimella et al. [Garimella2016Quantifying]. We explored three methods: (1) a network-based measure that analyzes the separation of the high-degree nodes of two non-overlapping connected communities, the Random Walk Controversy Measure (RWC), (2) a language-based method, which uses a dictionary-based sentiment analysis tool to analyze the mean and standard deviation of sentiment, and (3) a qualitative language-based methodology that leverages continuous word representations to analyze change in word meanings, similar to the methodology of Kim et al. [Kim2014Temporal]. We applied these measures in a comprehensive case study using the Maidan Crisis in Ukraine [Rumshisky2017combining] as well as three smaller case studies in U.S. politics during the 2016 presidential election cycle.

Using the Russian/Ukrainian Maidan data, we showed correlation patterns between the Random Walk Controversy (RWC) measure on user like-behavior pattern graphs and the mean and standard deviation of the sentiment expressed in politically-themed posts [Rumshisky2017combining]:

- As the conflict intensifies, the RWC measure on user like-behavior pattern graphs and the standard deviation of overall sentiment expressed by the opposing groups in politically-themed posts (SenSTD) increased in unison. RWC and SenSTD were positively correlated (Pearson and Spearman correlation values of 0.674 and 0.745, respectively).
- RWC and the average of the absolute value of the overall sentiment correlate negatively, confirming that negative sentiment accompanies the intensification of conflict (Pearson and Spearman correlation values of -0.598 and -0.291, respectively).

Figure 20 shows the correlation patterns between the three measures. Figure 21 shows user like-behavior graphs, as polarization increases between the opposing communities. October and November 2013 graphs show a random arrangement of nodes in the absence of conflict. This is followed by a clear division of communities, which are still connected by a bridge of users attending to both sides of the discourse. As polarization increases and the conflict drives out the neutrals, the bridge begins to thin out around May 2014 and disappears completely in June 2014. As the bridge no longer exists, there is virtually no way for

the representatives of the opposing sides to experience the discourse from the other side. This corresponds to the plateau observed in the graph of the random walk controversy measure shown in Figure 20.

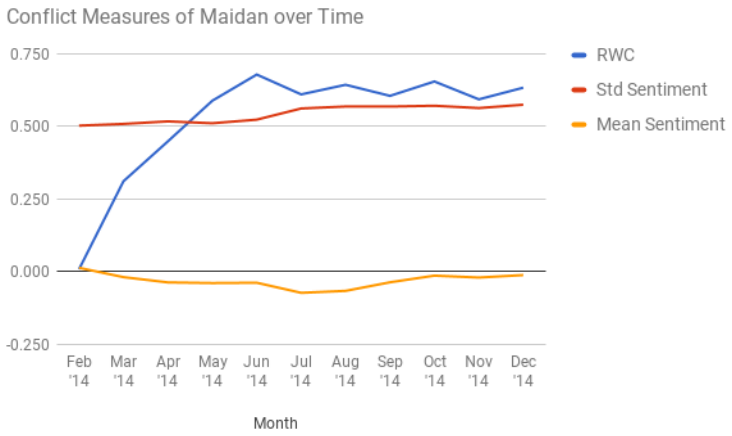


Fig. 20. Conflict indicator correlation graph for Russian/Ukrainian Maidan conflict.

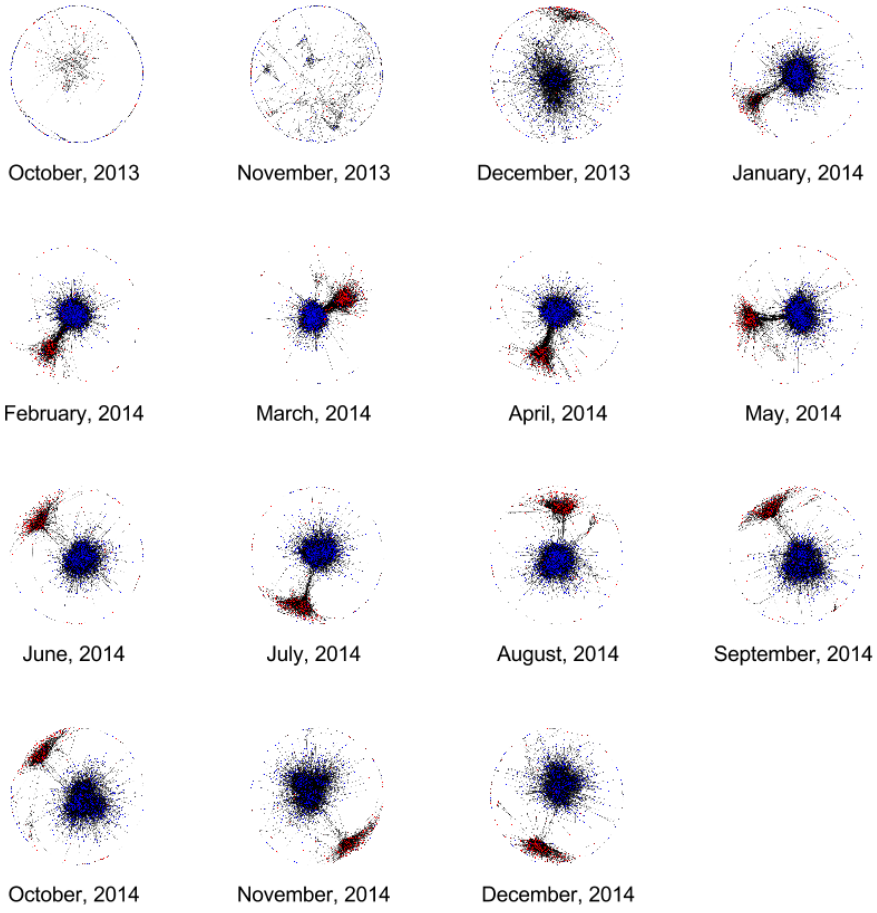


Fig. 21. Pro- and anti-Maidan VKontakte user like-behavior networks as conflict develops (Fruchterman-Reingold force-directed graph). Blue nodes represent Evromaidan (pro-Ukrainian) users, red nodes represent Antimaidan (pro-Russian) users.

7. Community detection

Using the Maidan dataset of user and group wall posts from the VKontakte, we experimented with community detection for the users involved in the Russian/Ukrainian conflict of 2014. We built a graph of active users, where the nodes represented individual users and an edge between two nodes appeared if the corresponding users liked the same post.

Following the methodology we described in [Rumshisky2017conflicting], we only considered "active users", which were defined as users who averaged two posts per three months at least once over the target time frame (Oct 1, 2013 - Oct 1, 2014). The resulting graph contained 9.2K nodes and 2.5M edges. We extracted the largest strongly connected component, which produced 8.9K nodes and 1M edges. Every node was labeled as either "EURO" or "ANTI", depending on which set of Maidan-related user groups the user belonged to. We refer to these labels as ground-truth classes to be recovered automatically by applying community detection algorithms to the induced graph.

We evaluated several widely used state-of-the-art community detection algorithms, implemented in an open-source igraph library [Csardi2006the] designed for handling computation-intensive operations on graphs. Specifically, we used Walktrap, Fast-greedy, Label Propagation, Leading eigenvector, Infomap and Multilevel algorithms. These algorithms rely on different principles for community detection, and as a result, they identify different user clusters in the graph structure.

Fastgreedy [Clauset2004finding] and Multilevel [Blondel2008fast] methods optimize modularity scores. Leading Eigenvector is based on the spectral optimization of modularity by using the eigenvalues and eigenvectors of the modularity matrix [Newman2006finding]. Infomap [Rosval2007information] and Walktrap [Pons2005computing] both use random walks to analyze the information flow through a network. Label Propagation assumes that each node in the network is assigned to the same community as the majority of its neighbors [Raghavan2007near]. A more detailed review on all of these algorithms can be found in [Yang2016comparative].

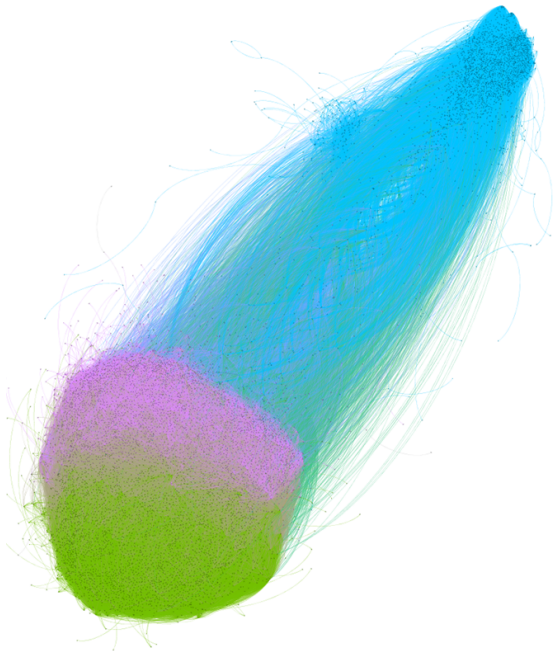
We performed qualitative analysis of the output for each algorithm. We sorted the communities by size and annotated the top-10 largest communities (when available) with the top-10 most liked the posts within each community. As expected, the most liked posts tended to be highly polarized in terms of the expressed political opinion. Top most liked posts for each community are shown in Figure 22.

Fastgreedy

Надо быть умственно отсталым, чтобы проживая в одной стране, скандировать имя другой	You gotta be retarded to promote one country while living in another
ПРОШУ МАКСИМАЛЬНЫЙ РЕПОСТ! ВСЁ, КРЫМЧАНЕ! ВЫ меня окончательно разозлили... Тогда вот вам правда! Новости сегодня очень смешные. ВР Крыма приняла решение о входе в Россию и присоединении к рублёвой зоне.	PLEASE SHARE! THAT'S IT, CRIMEANS! YOU finally drove me crazy... Here's the truth! Today's news is hilarious. The Verkhovna Rada of Crimea decided to join Russia and the ruble zone.

ПРОСТО ЛАЙК И РЕПОСТ. #Осознание@anti_usa_news	JUST LIKE AND SHARE. #Awareness@anti_usa_news
Вступите в группу РУССКОЕ ОБОЗРЕНИЕ и расскажите друзьям!	Join the club RUSSIAN SURVEY and tell your friends!

Значить так. Ми подумали. І вирішили: 1. Межигір"я - або лікарня майбутнього, або Охмадит, або дитячий табір. 2. Донести ЮЛІ голосно і хором ПОГОВОРИЛА І ФАТИТЬ. На лікування. Будь ласка. ПОЧИНАЄМО ІНФОРМАЦІЙНУ ХВИЛЮ!	So, we've thought about all this and this is what we decided: 1) Mezhigirya should be either a hospital of the future, or a Ohmadit, or a children's camp. 2) The whole [Maidan] crowd should let Yulia [Timoshenko] know loud and clear that her time has passed, it's about time she retired. Let's start the information wave!
Українці, всі по лайку та репосту ;)	Ukranians, let's all like and share this ;)



Infomap

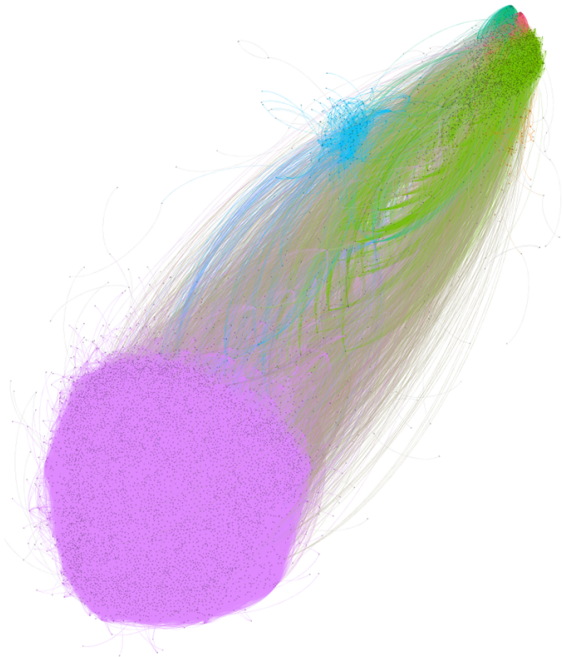
Значить так. Ми подумали. І вирішили: 1. Межигір"я - або лікарня майбутнього, або Охмадит, або дитячий табір. 2. Донести ЮЛІ голосно і хором ПОГОВОРИЛА І ФАТИТЬ. На лікування. Будь ласка. ПОЧИНАЄМО ІНФОРМАЦІЙНУ ХВИЛЮ!	So, we've thought about all this and this is what we decided: 1) Mezhigirya should be either a hospital of the future, or a Ohmadit, or a children's camp. 2) The whole [Maidan] crowd should let Yulia [Timoshenko] know loud and clear that her time has passed, it's about time she retired. Let's start the information wave!

ПРОСТО ЛАЙК И РЕПОСТ. #Осознание@anti_usa_news	JUST LIKE AND SHARE. #Awareness@anti_usa_news

http://hobosti.ru	http://hobosti.ru

'РУССКОЕ ОБОЗРЕНИЕ приближается к отметке в 100 000 подписчиков. Мы за великую Россию! Мы за нравственность! Мы за великий русский народ!	RUSSIAN SURVEY is reaching 100 000 members. We stand for Great Russia! We stand for morals! We stand for the great Russian nation!

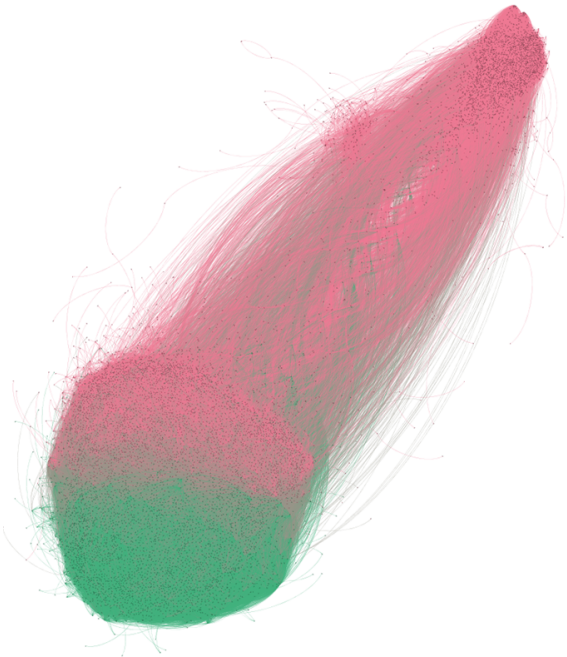
Русские плечо к плечу с Кавказцами,Казахами и Белорусами vs Нацисты из Киева" В одном из турецких отелей в курортном городке Кемер произошёл настоящий бой между украинскими и российскими туристами	Russians are back to back with Kazakhs, Caucasians, Belarusians vs Kiev nazis! There was a real fight between Russian and Ukranian tourists in a hotel in Kemer resort town.



Leading eigenvector

Надо быть умственно отсталым, чтобы проживая в одной стране, скандировать имя другой	You gotta be retarded to promote one country while living in another
ПРОШУ МАКСИМАЛЬНЫЙ РЕПОСТ! ВСЁ, КРЫМЧАНЕ! ВЫ меня окончательно разозлили... Тогда вот вам правда! Новости сегодня очень смешные. ВР Крыма приняла решение о входе в Россию и присоединении к рублёвой зоне.	PLEASE SHARE! THAT'S IT, CRIMEANS! YOU finally drove me crazy... Here's the truth! Today's news is hilarious. The Verkhovna Rada of Crimea decided to join Russia and the ruble zone.

Значить так. Ми подумали. І вирішили: 1. Межигір"я - або лікарня майбутнього, або Охмадит, або дитячий табір. 2. Донести ЮЛІ голосно і хором ПОГОВОРИЛА І ФАТИТЬ. На лікування. Будь ласка. ПОЧИНАЄМО ІНФОРМАЦІЙНУ ХВИЛЮ!	So, we've thought about all this and this is what we decided: 1) Mezhigirya should be either a hospital of the future, or a Ohmadit, or a children's camp. 2) The whole [Maidan] crowd should let Yulia [Timoshenko] know loud and clear that her time has passed, it's about time she retired. Let's start the information wave!
Українці, всі по лайку та репосту ;)	Ukranians, let's all like and share this ;)

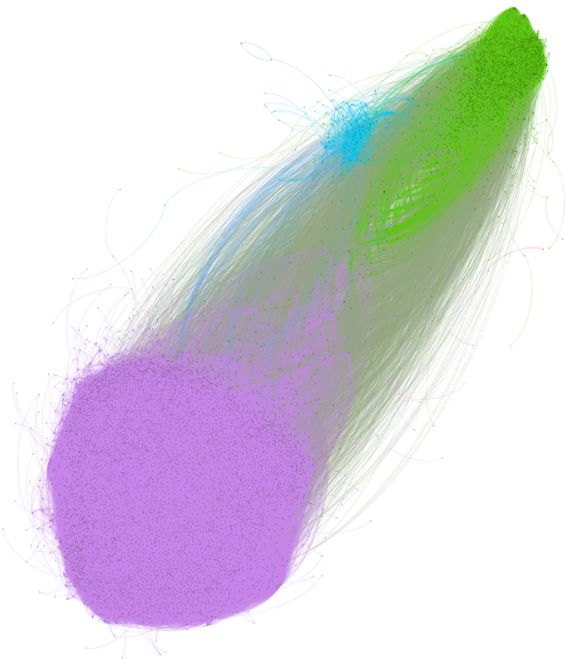


Label propagation

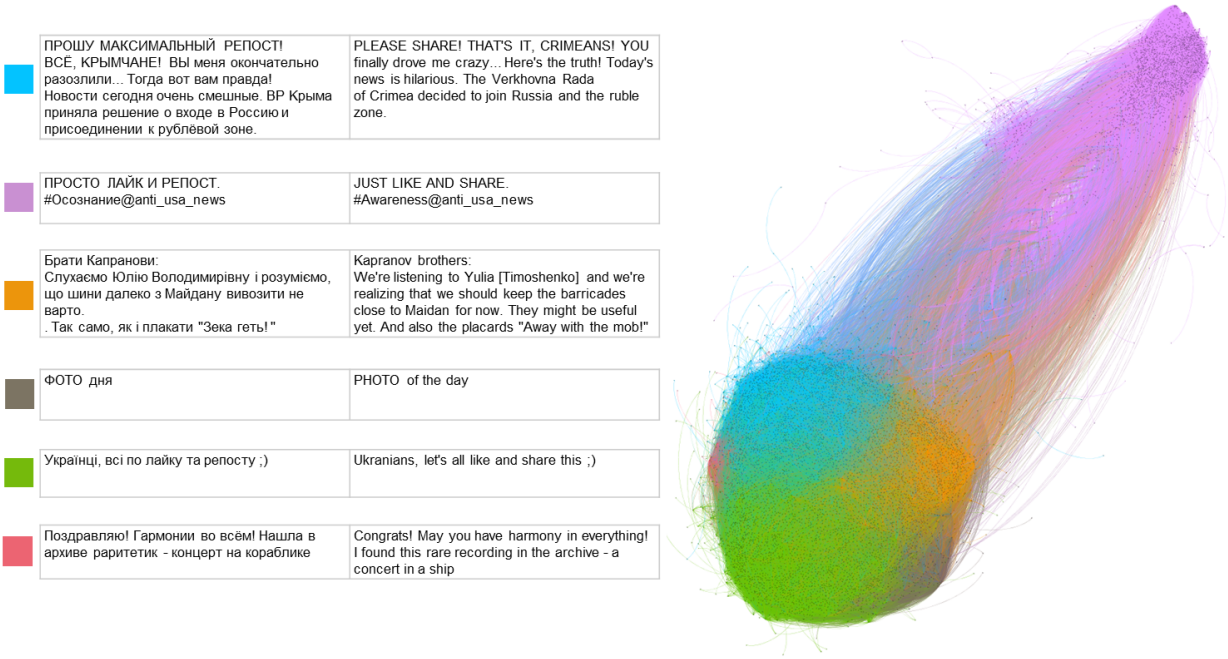
Значить так. Ми подумали. І вирішили: 1. Межигір"я - або лікарня майбутнього, або Охмадит, або дитячий табір. 2. Донести ЮЛІ голосно і хором ПОГОВОРИЛА І ФАТИТЬ. На лікування. Будь ласка. ПОЧИНАЄМО ІНФОРМАЦІЙНУ ХВИЛЮ!	So, we've thought about all this and this is what we decided: 1) Mezhigirya should be either a hospital of the future, or a Ohmadit, or a children's camp. 2) The whole [Maidan] crowd should let Yulia [Timoshenko] know loud and clear that her time has passed, it's about time she retired. Let's start the information wave!
Українці, всі по лайку та репосту ;)	Ukranians, let's all like and share this ;)

ПРОСТО ЛАЙК И РЕПОСТ. #Осознание@anti_usa_news	JUST LIKE AND SHARE. #Awareness@anti_usa_news
'РУССКОЕ ОБОЗРЕНИЕ приближается к отметке в 100 000 подписчиков. Мы за великую Россию! Мы за нравственность! Мы за великий русский народ!	RUSSIAN SURVEY is reaching 100 000 members. We stand for Great Russia! We stand for morals! We stand for the great Russian nation!

http://hobosti.ru	http://hobosti.ru
Пенсионерам пропишут галлюциногенные грибы	Pensioners are prescribed with hallucinogenic mushrooms



Multilevel



Walktrap

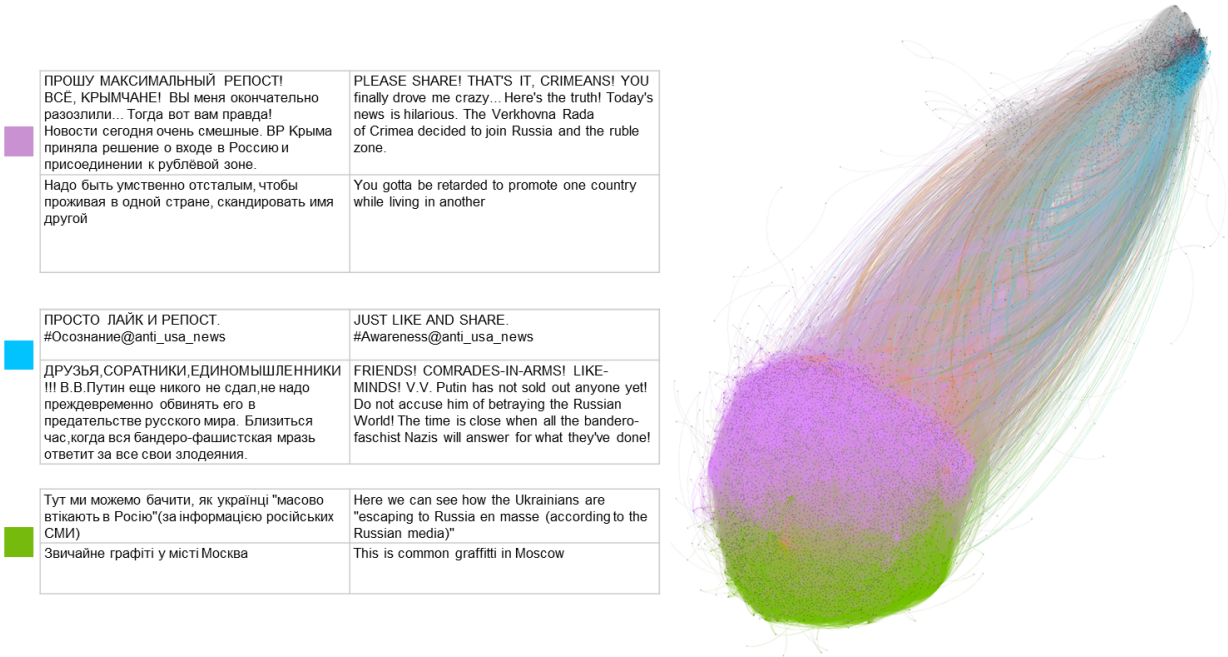


Fig. 22. Top most shared posts for the communities detected by the six surveyed algorithms.

Visualizations shown in Figure 22 use the force-driven ‘Forced Atlas’ layout implemented in the Gephi library [Bastian2009gephi]. Table 12 shows the number of communities detected by each algorithm, as well as the amount of time needed to run it on the user graph we used in these experiments (cf. the statistics given above).

	Walktrap	Fast-greedy	Label propa-gation	Leading eigen-vector	Infomap	Multilevel
Elapsed time	~300 sec	~200 sec	~0.6 sec	~6 sec	~907 sec	~3 sec
Number of detected communities	1517	13	6	2	169	7

Table 12. Number of communities and runtime for the surveyed algorithms.

Based on the results obtained by the surveyed algorithms, our observations are as follows:

- Walktrap algorithm is too granular, tending to form communities consisting of a single node. Given that our data came from only two major sources, detection of 1517 communities seems to be unreasonable.
- Though Leading Eigenvector correctly reproduces 2 communities, it can't properly handle the intermediate zone between the two, which is possibly explained by the fact that the initial graph is unbalanced. We conclude that this algorithm is not the best choice for community detection tasks on real social-network data.
- Both Multilevel and Walktrap correctly reconstruct one of the communities while artificially splitting the other into too many subdivisions. Again, this might be because of the class imbalance.
- Out of all the tested algorithms, we found Label Propagation and Infomap to be most promising in terms of adequately reconstructing ground-truth classes. An interesting note here is that both of the algorithms detect a separate community of users (shown in blue in Figure 22) who have a sarcastic attitude to the Russian media, while using the same lexicon as the Antimaidan group does. This tends to be a useful complement to linguistic methods analyzing the semantics of the posts.

Since 5704 of the users belong to the Evromaidan community and 3266 to the Antimaidan community, the graph is unbalanced. Depending on how the connectivity between the nodes is handled in each case, different algorithms produce user communities of different granularity. Tables 13 and 14 below provide, for each of the algorithms, raw statistics needed to compute any of the standard criteria of clustering quality (such as Purity, Normalized Mutual Information, BCubed F-measure, etc.) Table 13 shows the number of detected communities corresponding to each of the two ground truth classes. Table 14 shows the number of users in EURO / ANTI ground truth classes for the top 5 largest communities detected by each algorithm. The splits between the ground truth classes in each of the top communities shows how well each algorithm is able to capture the ground truth, i.e. how well the largest communities represent the overall user division. For the algorithms that detected fewer than 5 communities, the data for all the detected communities is shown.

Algorithm	Communities corresponding to the *Antimaidan* ground-truth class	Communities corresponding to the *Evromaidan* ground-truth class
Walktrap	1208	337
Fastgreedy	10	10
Label propagation	6	4
Leading eigenvector	2	2
Infomap	157	41
Multilevel	7	6

Table 13. Number of communities detected for each ground truth class.

	1 (ANTI/EURO)	2 (ANTI/EURO)	2 (ANTI/EURO)	4 (ANTI/EURO)	5 (ANTI/EURO)
Walktrap	282/3777	106/1488	407/2	213/2	66/61
Fastgreedy	270/2972	224/632	2759/78	2/11	2/1
Label propagation	694/5676	2395/25	163/1	8/2	3/0
Leading eigenvector	2987/3097	279/2607	-	-	-
Infomap	533/5589	1392/18	166/4	73/0	68/0
Multilevel	2703/56	190/2085	197/2069	123/1123	34/229

Table 14. Split between the ANTI and EURO ground truth classes in the detected communities.

References

[Rogers2018Rusentiment] A. Rogers, A. Romanov, A.Rumshisky, S. Volkova, M. Gronas, A. Gribov. RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 755–763). 2018. http://www.aclweb.org/anthology/C18-1064

[Rogers2019Calls] A. Rogers A, O. Kovaleva, A. Rumshisky. Calls to action on social media: Detection, social impact, and censorship potential. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, EMNLP 2019* Nov (pp. 36-44).

[Headley2015Challenging] Headley, J. Challenging the EU's claim to moral authority: Russian talk of 'double standards.' *Asia Europe Journal*, *13*(3), 297–307. 2015. https://doi.org/10.1007/s10308-015-0417-y

[Rumshisky2017combining] A. Rumshisky, M. Gronas, P. Potash, M. Dubov, A. Romanov, S. Kulshreshtha, A. Gribov. "Combining Network and Language Indicators for Tracking Conflict Intensity." *Proceedings of SocInfo 2017*. Oxford, United Kingdom.

[Bastian2009gephi] M. Bastian, S. Heymann, M. Jacomy. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. 2009.

[Yang2016comparative] Z. Yang, R. Algesheimer, C.J. Tessone. A comparative analysis of community detection algorithms on artificial networks. Scientific Reports. 2016

[Csardi2006the] G. Csardi, T. Nepusz. The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org

[Clauset2004finding] A. Clauset, M.E. Newman, C. Moore. Finding community structure in very large networks. Physical Review E 70, 2004

[Rosvall2007information] M. Rosvall, C.T. Bergstrom. An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences 104, 7327–7331 (2007)

[Raghavan2007near] U.N. Raghavan, R. Albert, S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76, 036106 (2007).

[Newman2006finding] M.E. Newman. Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74, 036104 (2006)

[Blondel2009fast] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008.

[Pons2005computing] P. Pons, M. Latapy. Computing communities in large networks using random walks. In Computer and Information Sciences-ISCIS 2005, 284–293 2005.

[Wikipedia2018Russian] Wikipedia contributors. "2011–2013 Russian protests." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 6 Aug. 2018. Web. 31 Aug. 2018.

[Bojanowski2016Enriching] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

[Schütze2008Introduction] Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.

[Chung2014Empirical] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

[Manning1999Foundations] Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

[Atefeh2015survey] Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence*, *31*(1), 132-164

[Yang2016Hierarchical] Yang, Zichao, et al. "Hierarchical attention networks for document classification." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.

[Deriu2016SwissCheese] Deriu, Jan, et al. "SwissCheese at SemEval2016 Task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision." *Proceedings of SemEval (2016)*: 11241128.

[Rubtsova2015Postroenie] Рубцова, Ю. В. "Построение корпуса текстов для настройки тонового классификатора." *Программные продукты и системы 1* (109) (2015).

[Kuznetsova2013Testing] Kuznetsova, Ekaterina S., Natalia V. Loukachevitch, and Ilia I. Chetviorkin. "Testing rules for a sentiment analysis system." Proceedings of International Conference Dialog. 2013.

[Clevert2015Fast] Clevert, DjorkArné, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)." arXiv preprint arXiv:1511.07289 (2015).

[Vinyals2015Pointer] Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks."Advances in Neural Information Processing Systems. 2015.

[Stab2016Parsing] Stab, Christian, and Iryna Gurevych. "Parsing Argumentation Structures in Persuasive Essays."arXiv preprint arXiv:1604.07370 (2016).

[Peldszus2015Joint] Peldszus, Andreas, and Manfred Stede. "Joint prediction in MSTstyle discourse parsing for argumentation mining." Proc. of the Conference on Empirical Methods in Natural Language Processing. 2015.

[Prakash2017Condensed] Prakash, Aaditya, et al. "Condensed Memory Networks for Clinical Diagnostic Inferencing." *AAAI*. 2017.

[Sukhbaatar2015End] Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. "End-to-end memory networks." *Advances in neural information processing systems*. 2015.

[Lukashevich2015SentiRuEval] Н. В. Лукашевич, Ю. В. Рубцова. "SentiRuEval-2016: преодоление временных различий и разреженности данных для задачи анализа репутации по сообщениям

твиттера." *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции Диалог*. 2015.

[Potash2017Here] P. Potash, A. Romanov, A. Rumshisky. " Here's My Point: Joint Pointer Architecture for Argument Mining . " Proceedings of EMNLP 2017. Denmark, Copenhagen.

[Potash2017Towards] P. Potash, A. Rumshisky. " Towards Debate Automation: a Recurrent Model for Predicting Debate Winners." Proceedings of EMNLP 2017. Denmark, Copenhagen.

[Potash2017Length] P. Potash, R. Bhattacharya, A. Rumshisky. " Length, Interchangeability, and External Knowledge: Observations from Predicting Argument Convincingness. " IJCNLP 2017. Tapei, Taiwan.

[Potash2017Tracking] P. Potash, A. Romanov, A. Rumshisky, M. Gronas." Tracking Bias in News Sources Using Social Media: the Russia-Ukraine Maidan Crisis of 2013–2014 ." EMNLP 2017 Workshop "NLP Meets Journalism". Denmark, Copenhagen

[Peldszus2015Joint] Peldszus, A. and Stede, M. (2015). Joint prediction in mst-style discourse parsing for argumentation mining. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 938–948.

[Garimella2016Quantifying] Garimella, Kiran, et al. "Quantifying controversy in social media." Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM, 2016.

[Kim2014Temporal] Kim, Yoon, et al. “Temporal Analysis of Language through Neural Language Models”. In *Proceedings of ACL 2014*.

Political calls to action on Social Media

Annotation guidelines
v.0.2

Anna Rogers, Viktoriya Khaichuk,
Anna Rumshisky, Mikhail Gronas

Despite considerable interest towards detection of calls to political action in social media, to the best of our knowledge, at this point there is no clear definition of such calls, and no annotated datasets that are publicly accessible. At the same time, the number of commercial systems for detecting dissenting and activist social media posts in Russia is growing: https://meduza.io/feature/2018/10/16/politsiya-po-vsey-rossii-pokupaet-sistemy-monitoringa-sotssetey-oni-pomogayut-iskat-ekstremizm-ne-vyhodya-iz-rabochego-kabineta

One of the goals of this project is to operationalize political calls to action (CTA) in social media and provide a publicly accessible data that would level the playing field for researchers on social media. The pilot annotation and experimental results are described in our paper (Rogers et al., 2019).

We define “calls to action” (CTA) as calls to some action that would induce a change in society that the speaker considers desirable. There may be calls to action with other, non-political content, such as invitations to birthday parties. In the scope of this project, we distinguish between the following post categories:

- CTA vs non-CTA
- Political vs non-political posts
- Skip (unclear)

The guidelines were originally developed with data from Russian VKontakte network that was collected in the wake of the Bolotnaya protests in Russia (2010-2012). This version of the guidelines contains Russian examples translated and/or adapted to English.

We perform post-level annotation. In cases where CTAs are embedded in longer posts, mixed with other content, we still consider the post as a CTA.

CTA categories

Prototypical CTAs are imperatives prompting the addressee to perform some action, such as *"Don't let the government tell you what to think!"*. This seems like a straightforward category to annotate, but in reality CTAs may be expressed in various ways, including both direct and indirect speech acts (Brown, 1980). There are many borderline cases that would in the absence of clear guidelines decrease inter-annotator agreement (IAA). There is relevant work on the task of identification of requests in emails (Lampert et al., 2010) and intention classification for dialogue agents (Quinn & Zaiane, 2014), but, to the best of our knowledge, this work is the first to create a detailed schema for CTA annotation in the context of a political protest.

The current work on censorship is concerned not so much with CTAs in particular, but with a broader category of "*material with collective action potential".* (King et al., 2013) defines such materials as those that '(a) involve protest or organized crowd formation outside the Internet; (b) related to individuals who have organized or incited collective action on the ground in the past; or (c) relate to nationalism or nationalist sentiment that have incited protest or collective action in the past.' In other words, this definition only concerns offline events, and does not include various forms of "crowd protesting" such as calls to share information critical of the government.

Based on extensive manual analysis of samples from Bolotnaya dataset, we identified 5 core and 9 borderline cases for political CTAs, shown in Fig. 1. Since we were interested in CTAs for social movements, we excluded any other CTAs that would formally fit the criteria, such as invitations, marketing CTAs etc. We also excluded any other protest-related posts, such as reports of protest events. Of the core and borderline CTA cases, we chose to consider 8 as CTAs.

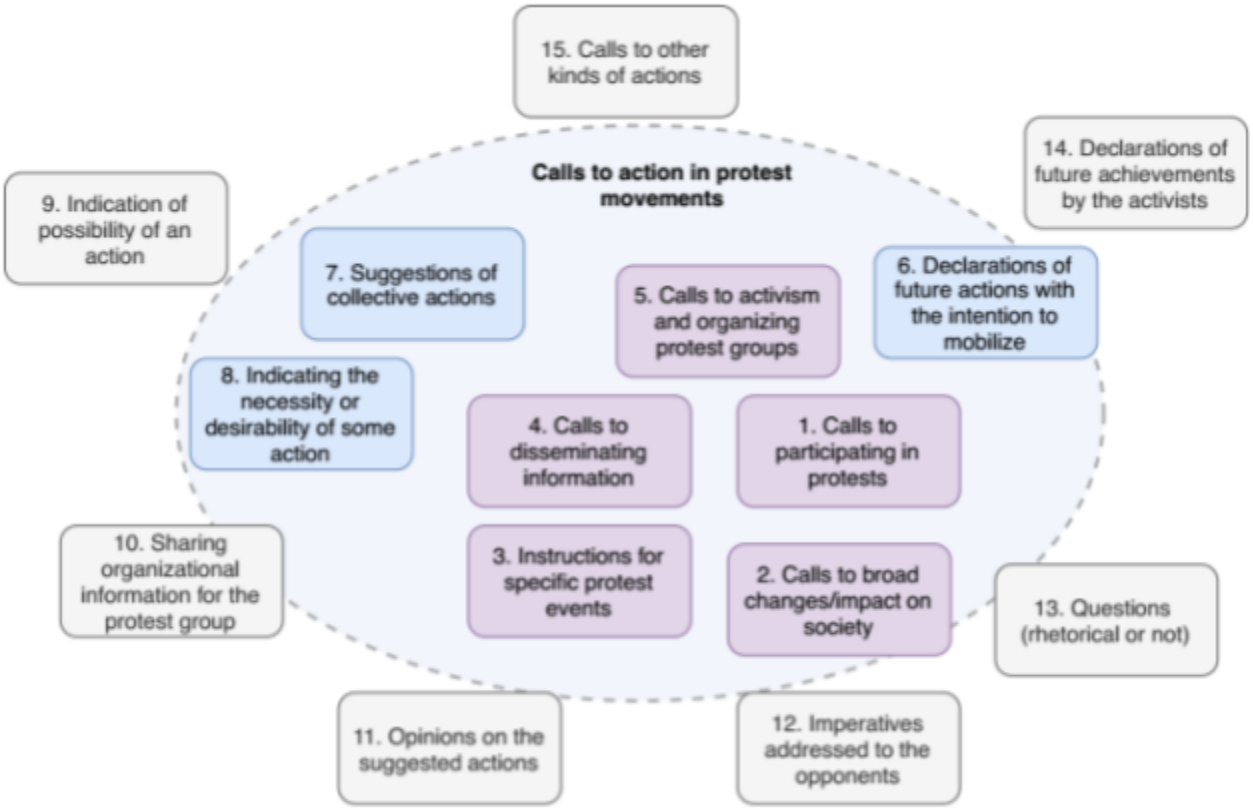


Fig. 1. The core and peripheral cases of political CTAs.

The choice to exclude or include certain edge cases does not have a firm theoretical underpinning and would vary depending on the researcher's perspective and the case study. For example, in our Bolotnaya data we opted to not include broad rhetorical questions like "*For how much longer shall we put up with this?*", but in a different context (especially in a different culture) they could be key. Inter-annotator agreement depends on how the guidelines' describing the chosen policy explicitly.

Political calls to action: core cases

1. Calls to participating in protests

This is the most straightforward case: imperatives inviting the addressee to take part in protest events, or notifications of such events that are accompanied with the implicit "Everybody, please come" message.

- *Everybody, join us tomorrow in Sakharov square!*
- *Everybody, on Sunday (Mar 18), 12 pm there will be a rally in the city esplanade.*

2. Calls to broad changes/impact on society

These CTAs express the poster's desire to make a positive impact on society or prevent a negative one. The proposed actions tends to be rather broad, and the target audience abstract.

- *Dear people of Smolensk and guests of our city! If you still love Russia, if you still love your mother hero city of Smolensk, start the fight against the crooks and thieves!*
- *Let's not put our city to shame!*
- *Let's make the authorities think!*
- *Wake up, our mother region of Smolensk! We are not slaves!*

3. Instructions for specific protest events

Unlike the previous category, this one includes calls to specific actions that are desirable/undesirable for specific rallies.

- *We meet at the Marsovo pole, 6.30-7pm. Then we will leave [for the rally]. Bring white ribbons, balloons and anything else that can be used to identify the movement.*
- *Do not form a line or agree on a specific place to meet.*
- *The rally is approved! 2pm, Feb 4th we will meet by the city council building.*

4. Calls to disseminating information

This category includes both the calls to spreading information through any media, and to recruiting more people into the movement.

- *Please, help us spread the information about the rally as much as possible!*
- *Invite your friends!*
- *Invite foreign press and journalists - let them see what's going on in our capital!*
- *Don't forget to invite your friends! There's a lot of us, we need to stand together.*
- *We ask you to print the leaflets and spread the information in any (legal) ways you can.*
- *Share your versions of leaflets to post around the city!*

5. Calls to activism and organizing protest groups

First of all, this category includes various CTAs coming from the leaders and organizers of the protest events:

- *Observers in Kaluga, please respond!*

We also include in this category various instructions on how to interact with the organizers/activists:

- *Everybody who wishes to go to Moscow on Constitution day with the STAL organization, write a direct message to the organizers to learn the details.*
- *The rally is approved, since we expect a lot of people we ask you to let us know beforehand if you'd like to speak there.*

6. Declarations of future actions with the intention to mobilize

The future tense by itself is not a marker of a call to action, but combined with "we all" subject it may indicate an implicit CTA: something will definitely happen because the people will so, and the addressee is invited to ride that wave, so to speak.

- *We will assemble on June 12 2012 to discuss again the results of their activity, and if there are none we will not leave until something is done!*
- *That's ok, we will tell them what we think of them even in the square in front of the Central market!*

7. Suggestions of collective actions

Here we include grassroot suggestions to perform some protest-related action. These may come from the current or future organizers.

- *We made a group "For honest elections and honest media", I suggest we move all discussions there.*
- *I suggest we put on white stripes on our arms as a symbol of honest elections. That's easy to do!*

8. Indicating the necessity/desirability of some action

This category includes all kinds of modal expressions and their equivalents that are not direct imperatives, but express the speaker's belief that some action should be performed, and thus implicitly calls for that action.

- ***It's time** we stop fighting in the internet and start fighting in the streets. Especially after such insolence [on the part of authorities].*
- *On March 10th there **must** be a lot of us!*
- *You **can** participate in the rally and later join us!*
- *We **need** a different Russia!!!!!*
- *We **demand** that the authorities: (1) investigate and make public the results of the investigation of the actions of the ex-mayor of Atrakhan…*
- *We **want** our country to be honest!*

Calls to action: peripheral cases

9. Indication of the possibility of an action

While we include modal expressions of necessity, obligation, and desirability (category #8), we opted to exclude posts that only indicate the possibility that the addressee performs some action. However they could also be considered implicit CTAs.

- *You **can** download the leaflet here.*
- *The leaflet has 2 pages. If printed 2-sided, **it is possible** to hand it out easily.*
- *It **would be nice** to make the leaflets by 13.04.2013, and to distribute them in the area.*

10. Sharing organizational information for the protest group

Many posts share information about a group that is not directly actionable.

- *This is the group for rally organizers in Kaluga.*

However, sometimes they too could be treated as implicit CTAs:

- *This is the beginning! We will start activities when we will have 50 members. We repeat, participation in this group can only be active.*

This could be viewed as an implicit invitation to join the group, so that its activities could finally begin.

11. Sharing opinions on the suggested actions

These posts are typically discussions of proposed actions, which themselves may be CTAs.

- *I do like the idea of the government's resignation, but I think your slogans are too emotional. Furthermore, I'm against calling an early election.*

With a more direct language (imperatives, modals) they could also be viewed as instances of categories #7 or #8:

- *I **suggest** that we stick to slogans that are not so emotional.*
- *I did not vote for United Russia, I voted for other scum. I **demand** the recount!*

Still, when the message is personal it sounds like there's less potential for sharing. The CTA case becomes stronger if the mode switches to "we" instead of "I":

- *We **have to** keep the situation stable, and early election would upset that.*

12. Imperatives addressed to the political opponents.

Formally these are imperatives, but we exclude them because the speakers hardly expect the opponents to follow the suggestions, unlike in the core CTA cases.

- *Out with you, McFaul! And take Putin and Medvedev with you, together with Nemtsov and Chirikova!*
- *Putin, don't chicken out!*

13. Questions (rhetorical or not)

Organizational questions that elicit real information are not political CTAs.

- *Today at 10 pm Vlad and I are going to post the leaflets around the city. Who wants to help us?*

Rhetorical questions are a harder case, since they may imply a call to a broad change in sociery (category #2). However, in the current project we chose to not annotate them.

- *Is THAT really our choice?*
- *Why is the United Russia allowed to do that, and we are not?*

14. Declarations of future achievements by the activists

This category is different from #6 in that the intention to mobilize is less clear, one perceives the message more as a statement and less as an invitation to perform the action together. Admittedly, the border is thin.

- *Together we will get rid of Putin's lies and dictatorship!*
- *We will make it happen that something real, something honest actually happens in our country.*

15. Other kinds of political posts

This is a category for all other political posts. We do not count them as CTA, even if they are political and written by protesters.

- *There aren't many people here.*
- *The protest rally against rigged election will take place on December 24th at 2pm in the Aloye Pole. The rally is approved by the city administration.*
- *No to covering up the fraud!*

Non-political calls to action

This category includes calls to all other kinds of actions, not relevant to politics. They include a wide variety of events and cover all the above modes of expressions, from imperatives to rhetorical questions.

- *Everybody, come to my birthday party on Saturday!*
- *Share this song with a friend.*
- *Let the holiday be called by its original name.*

Sometimes it is not clear without more context whether the post was political. For example, the following post could be either about a rally or about a school event:

- *Invite your friends.*

In such cases, our policy was to treat them as political. However, if a long descriptive political post contains at least one call to action of any of the core CTA categories, it did count.

Bibliography

Brown, G. P. (1980). Characterizing Indirect Speech Acts. *American Journal of Computational Linguistics*, *6*(3). http://www.aclweb.org/anthology/J80-3002

King, G., Pan, J., & Roberts, M. E. (2013). How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review*, *107*(2), 326–343. https://doi.org/10.1017/S0003055413000014

Lampert, A., Dale, R., & Paris, C. (2010). Detecting Emails Containing Requests for Action. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 984–992. http://www.aclweb.org/anthology/N10-1142

Quinn, K., & Zaiane, O. (2014). Identifying questions & requests in conversation. *Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering*, 10.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2019). Calls to Action on Social Media: Potential for Censorship and Social Impact. *EMNLP-IJCNLP 2019 Second Workshop on Natural Language Processing for Internet Freedom (Accepted)*.