Is Informal Flagging for Propaganda in User Comments Helpful to Identify Anti-Western Narratives?

The Benefits and Risks of Relying on User-Based Labeling

Vlad **Achimescu** (University of Mannheim, Germany)
Dan **Sultanescu** (CPD SNSPA, Bucharest, Romania)
Dana **Sultanescu** (CPD SNSPA, Bucharest, Romania)

SNSPA
Center for Civic Participation and Democracy

UNIVERSITY OF MANNHEIM

Premises

Online anti-Western propaganda - a persistent phenomenon with increasing levels of intensity.

Comments sections of online news articles: public sphere or fertile ground for opinion manipulation trolls ?

“Informal flagging” might serve as a form of identifying topics and narratives used by anti-Western propaganda

Online comments - two steps

EU without the UK army (Sunday, March 19 2017, 15:48) **0 (34 votes)**

soundtrack [user]

is so anemic. It was fed with curled dock!!!
Merkel is a friend of Putin and will not be afraid of war.
We must defend the west from the Ottomans!! Hahahha!
The US is doing its part.

reply send

The trolls are high (Monday, March 20 2017, 0:38) **+6 (12 votes)**

pro_bono [user] reply to **soundtrack**

If Merkel is a friend of Putin, then it means that Trump is Putin's stepfather ... Where do you get all this, you trolls?

reply send

Comment
potential
TROLL

Reply
potential
FLAGGER

Translation from comments to article with the topic “Conflict between Ursula von der Leyen and Donald Trump”2017-03-19

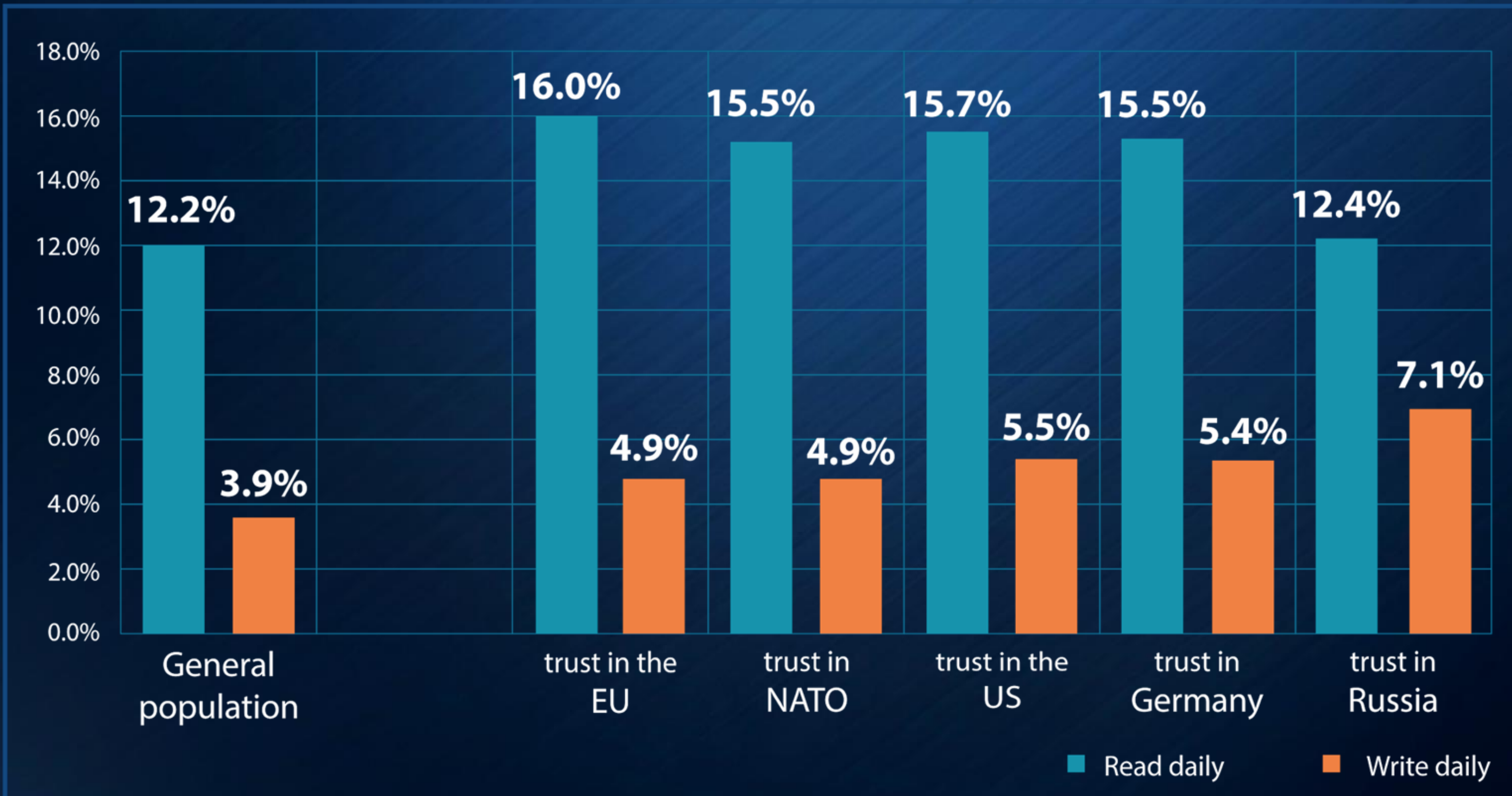
Research Questions

Is Informal Flagging for Propaganda in User Comments Helpful to Identify Anti-Western Narratives?

Does the two-step classification improve accuracy compared to the one-step classification?

Does combining metadata with text content improve prediction accuracy?

Active online media consumption in Romania



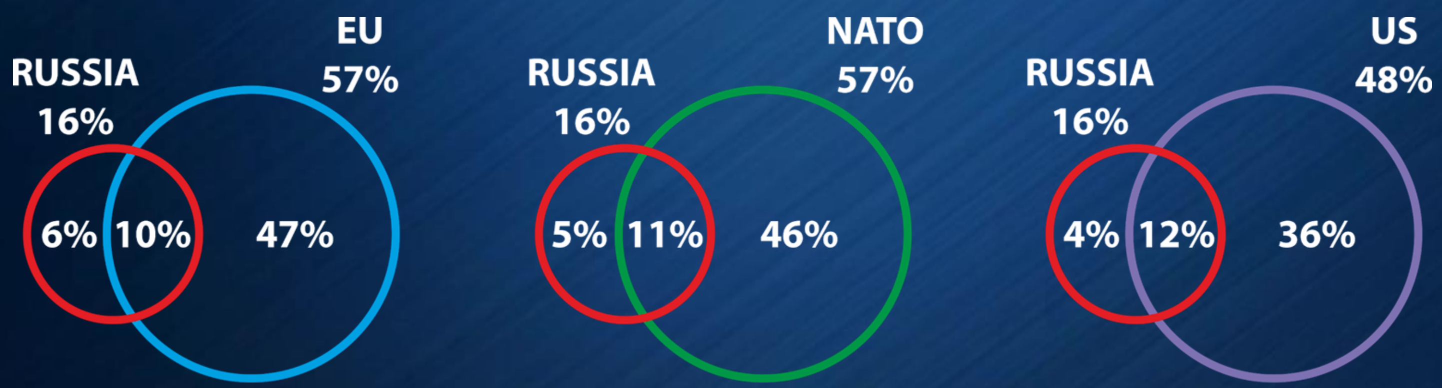
SNSPA Center for Civic Participation and Democracy

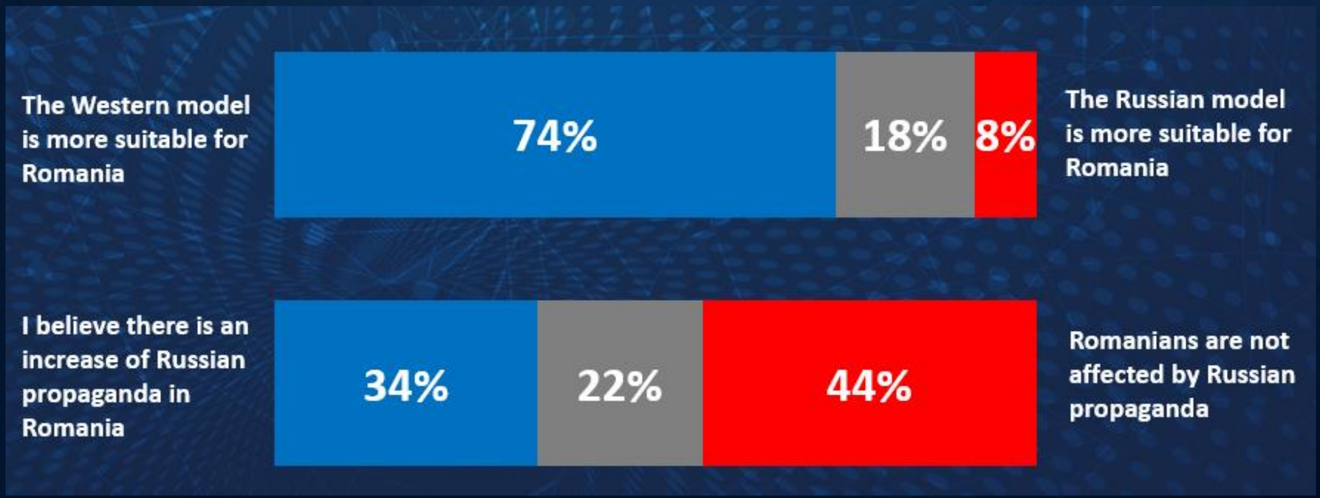
UNIVERSITY OF MANNHEIM

Source: Survey data, CPD@SNSPA, June, July, August 2018, http://civicparticipation.ro/uncategorized/anti-western-propaganda-in-romania-2/

Russian influence in Romania

Trust in countries or institutions / sum of “a great deal” and “quite a lot”



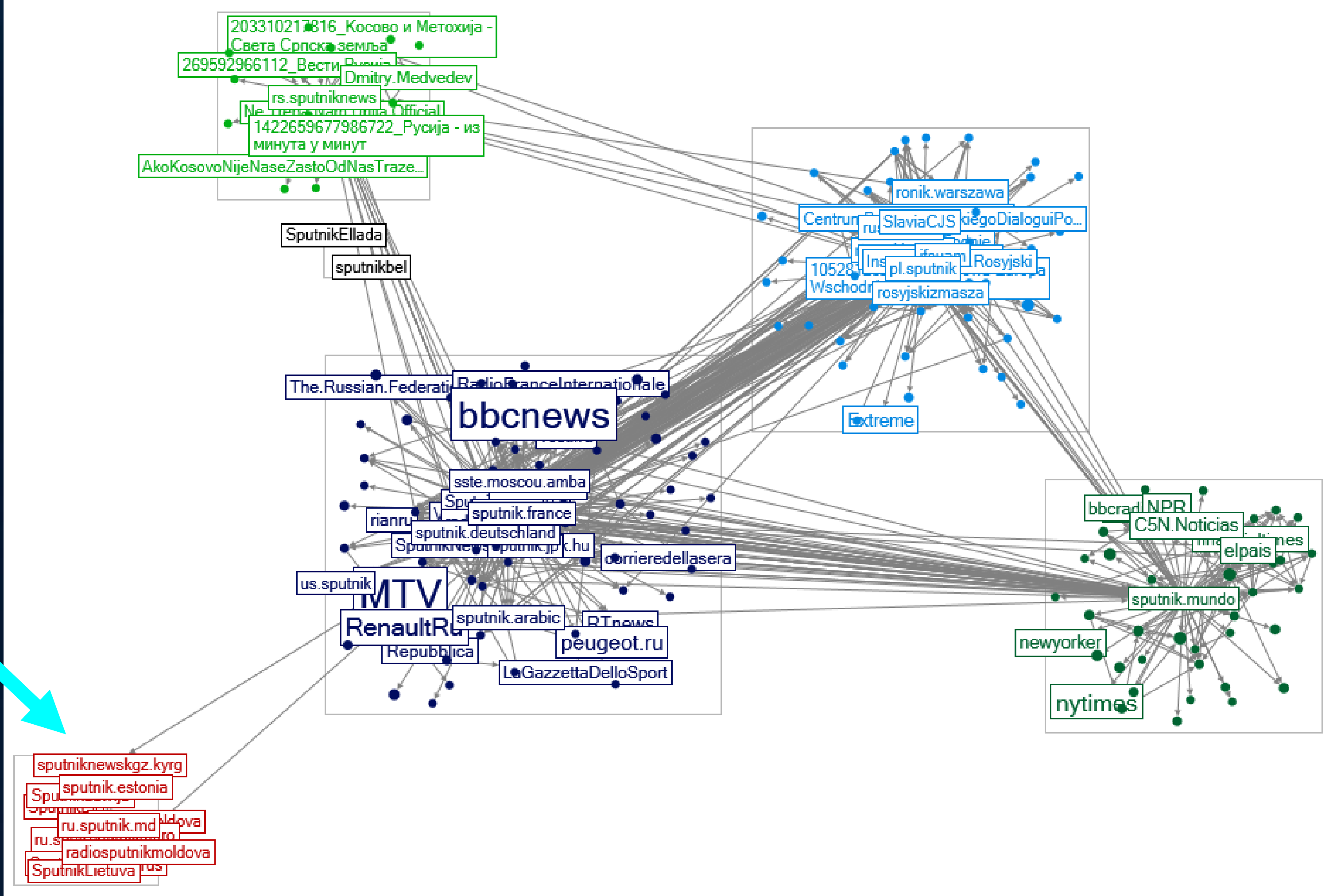


SNSPA
Center for Civic Participation and Democracy

UNIVERSITY OF MANNHEIM

Source: Survey data, CPD@SNSPA, June–August 2018,
http://civicparticipation.ro/uncategorized/anti-western-propaganda-in-romania-2/

Sputnik network



SNSPA Center for Civic Participation and Democracy

UNIVERSITY OF MANNHEIM

Source: NodeXL data, CPD@SNSPA, September 2018, civicparticipation.ro

Previous research

ONLINE **TROLLING** - *civil and uncivil public discourse, norms*

Munger 2017, Cheng et al 2017, Alvarez-Benjumea & Winter 2018

ONLINE RUSSIAN/ANTI-WESTERN **PROPAGANDA** - *esp. in Eastern Europe*

Paul and Matthews 2016; Chen 2015, Van Herpen 2016; Aro 2016; Franke 2015; Pomerantsev & Weiss, 2014

COMPUTATIONAL PROPAGANDA / ASTROTURFING - *focus on bots, less on trolls*

Bolsover and Howard 2017; Sanovich, Stukal and Tucker 2015

INFORMAL **FLAGGING** - *ML rarely applied to identify online trolls, one step*

Zannettou et al. 2018 - institutional flagging of Russian propaganda

Zelenkauskaite and Niezgoda 2017 - informal flagging of Russian propaganda

Mihaylov and Nakov 2016 - informal flagging, machine learning in one step

Research procedure

- Web **scraping** comments
- Selecting keywords and **labeling cases**
- **Machine Learning Models - STEP 1**
  TASK = identify informal flags
- **Machine Learning Models- STEP 2**
  TASK = identify perceived trolls

Comment
potential
TROLL

Predicted in **step 2**

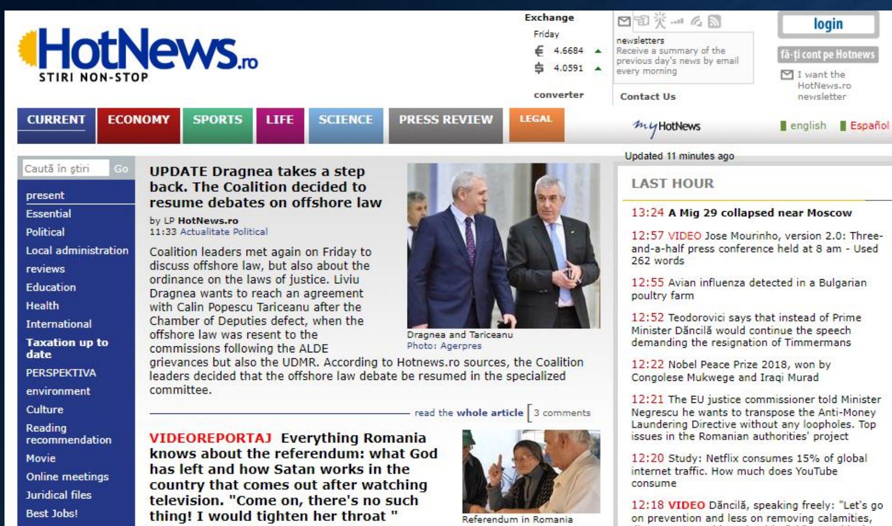
Reply
potential
FLAGGER

Predicted in **step 1**

Data

Data source: www.hotnews.ro

- January - October 2017
- **209.000 comments**
- 20.000 articles
- 7.300 registered users



Variables/Features:

- **Metadata** (25 features) - dense
  - **Article**: views, section, size
  - **Position of comment** : is reply, replies to other comments, order in thread
  - **Rating of comment** and number of raters
  - **Day and hour** comment was posted

- **Content** (350-1200 features)
  - bag of words / sparse
    - lowercase, stemmed, no stopwords
    - TF-IDF weights, ngrams
    - At least in 5/10/20 documents
  - Patterns : numbers, punctuation, hashtags, links, emojis, ALL CAPS

Methods

Supervised ML for classification ->

- STEP1: flags / non-flags ; STEP2: trolls / non-trolls

Methods:

- logistic regression (L1 & L2 regularization)
- random forests (5-100 features / tree)

Tuning

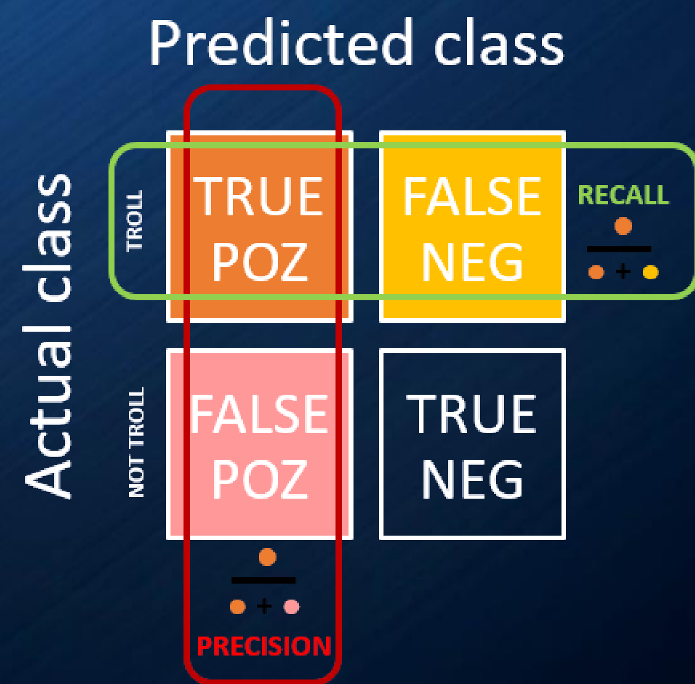
- **70% training set / 30% test set**
- **Cross-validation:** 5-fold, 3 times
- Different feature sets tested
- Oversampling flags and flagged comments

Performance measures for classification:

- Precision and Recall
- F1 score

SOFT: R 3.4.4 quanteda, glmnet, randomForests, caret

Confusion Matrix



SNSPA
Center for Civic Participation and Democracy

UNIVERSITY OF MANNHEIM

Manual labeling

- **Keywords to identify flags:**

bolsevic bolsevica bolsevici bolsevicii bolsevicul bolsevik mujic mujici mujiciei mujicul pro putin putinnu rrusia ruble rublele rusa rusasi ruseasca rusesc rusi rusia rusiaasa rusiacu rusiade rusianu rusiapentru rusiasa rusiei rusii rusilor rusilornu rusnac rusnacul ruso rusoaicele rusofil rusofili rusofilii rusofobia rusofobie russia russian soviet sovietelor sovietica sovietice sovieticul sovieticus urss

- **Period of search:** Jan - Mar 2017
- 2.100 / 82.000 **comments contain keywords**
- **Manual labeling:** 350 / 2.100 are flags

STEP1 - predicting new flags

Reply
potential
FLAGGER

Predicted in **step 1**

JANUARY - MARCH

- 2100 comments
- All contain keywords
- Manual classification (2 coders)

350 / 2100 (17%)
manually classified as flags

TRAINING SET (70%)

TEST SET (30%)

APRIL - OCTOBER

- ~4.100 comments
- All contain keywords
- Prediction, then manual classification

720 / 4100
predicted as flags

430 / 720 (60%)
manually classified as flags

VIRGIN SET

STEP1. Flag/ Not Flag
Classification diagnostics

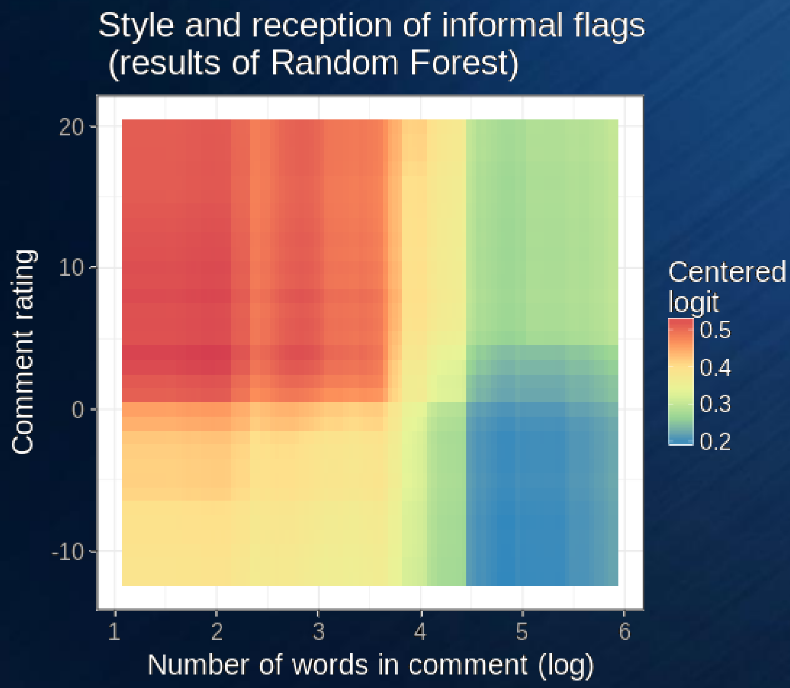
Reply
potential
FLAGGER

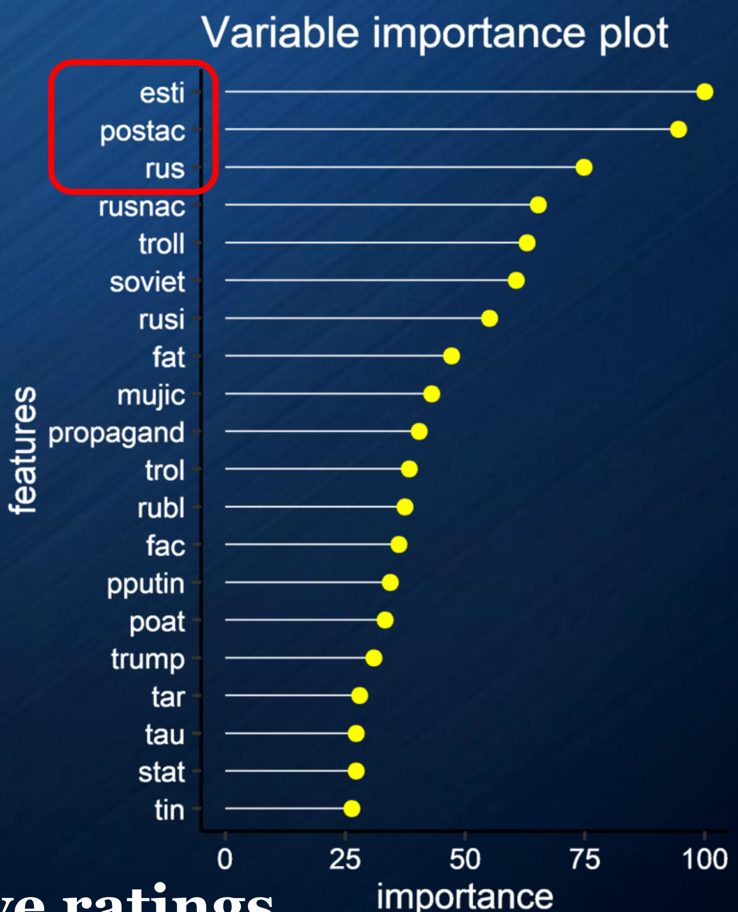
Predicted in **step 1**

	RAND. FORESTS		GLMNET	
	Precision	**Recall**	**Precision**	**Recall**
METADATA	0.54	0.24	0.32	0.60
WORDS	0.63	0.54	0.43	0.52
MIXED	**0.69**	0.45	0.45	0.53

- Random forests > Regularized regression
- Word tokens > Metadata , both increase precision but reduce recall
- Best configuration: RF / Mixed / no n-grams / normalized (F1 = 0.54)

STEP1. Flaggers - feature importance





- Flaggers are people of few words
- However, they receive **more positive ratings**

STEP 2. Troll / Not Troll classification diagnostics

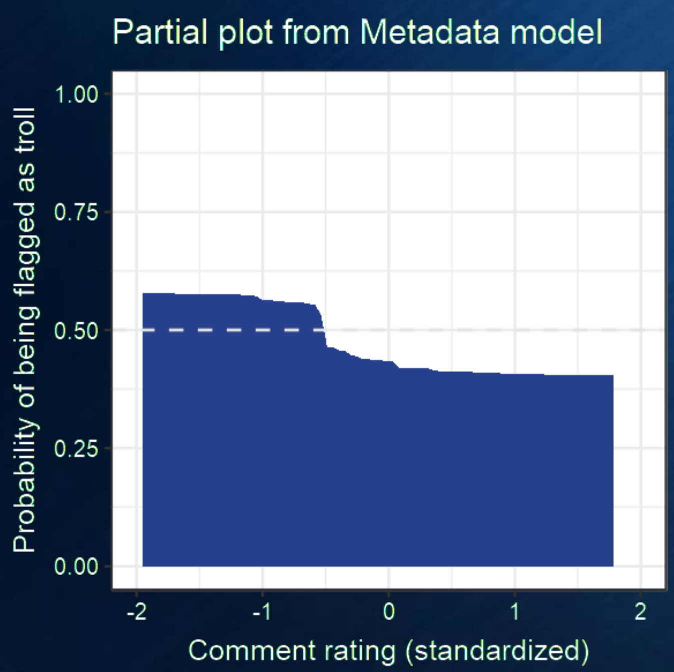
Comment
potential
TROLL

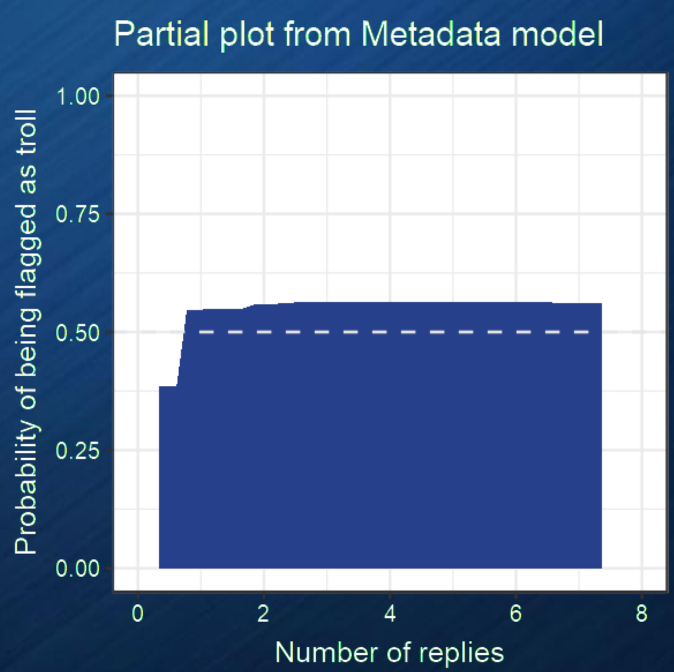
Predicted in **step 2**

F1 score	Jan – Mar	Apr - Oct		
	Initial Flags	**Initial Flags**		**All Flags**
METADATA	0.85	0.72	<	0.81
WORDS	0.41	0.31	<	0.41
MIXED	**0.85**	0.76	<	**0.85**

- Two test sets: one in Jan-Mar, one in Apr-Oct and two models
- Model 1 trained on **initial flags**, Model 2 - on **initial and additional flags**
- Metadata more informative than word tokens, combination no added value
- Is accuracy stable over time? Model updated with new training cases performs better in the second part of the year

STEP 2. Trolls - distinctive features





Comments flagged as propaganda are more **controversial**; they:

- Have lower ratings, more replies
- Use more words related to Russia, the EU or the US
- Use fewer words related to local politics and less punctuation

Summary, Benefits & Risks

- Higher accuracy prediction in two step procedure over time
- Content of comment for predicting flags, metadata for predicting trolls

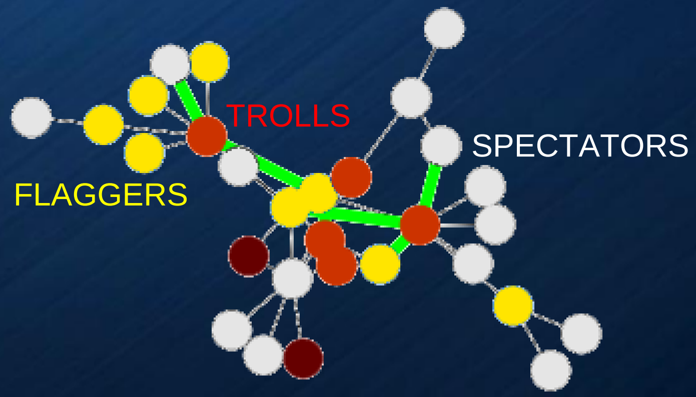
- Externalization of labelling reduces costs
- Instrument for moderators to identify anti-Western trolls in real time

- The risk of relying on false positives, dishonest or uninformed labelers
- Trolls adapting to thwart the instrument

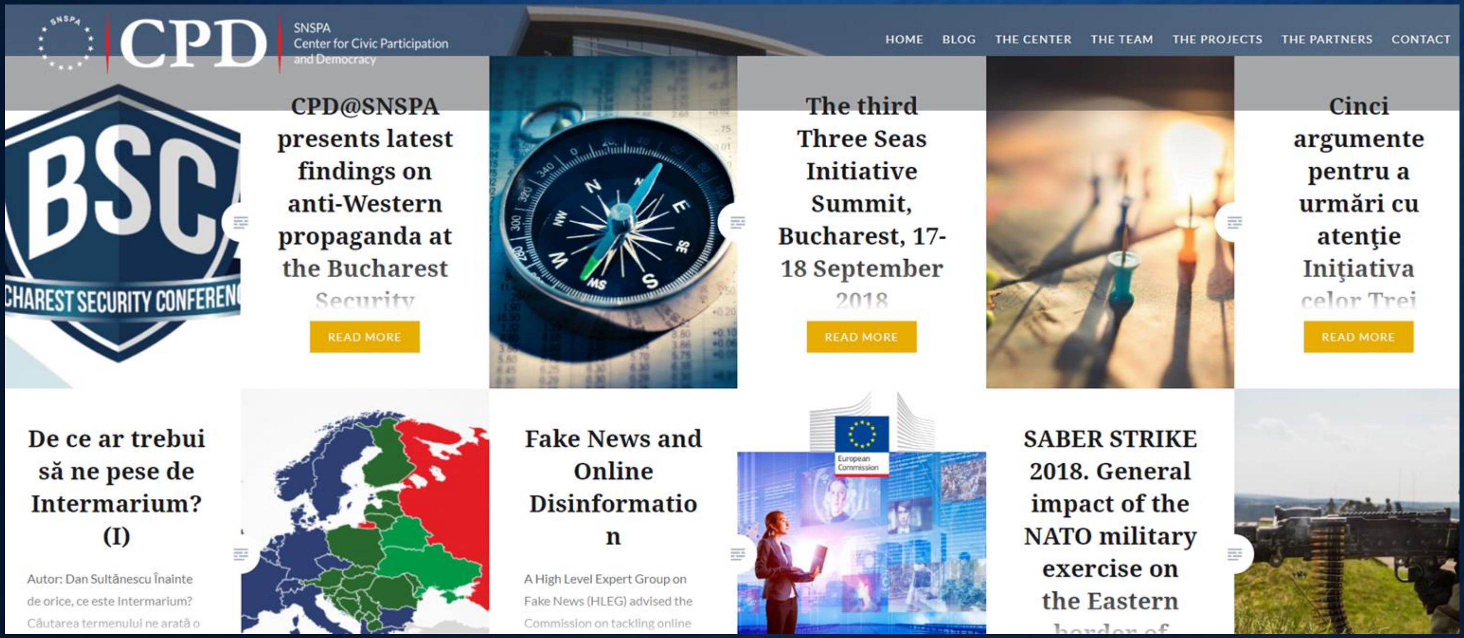
Meta-trolling

Next steps

- Estimate number of trolls on forum
- Reinforcement learning
- External validation
- Topic modeling
- Network analysis
- Experiments
- Survey of forum users



More about our work



Thank You!

SNSPA
Center for Civic Participation and Democracy

UNIVERSITY OF MANNHEIM