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)
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

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

Source: Survey data, CPD@SNSPA, June, July, August 2018,
Russian influence in Romania

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

- **RUSSIA**: 16%
  - EU: 57% (6% + 10%)
  - NATO: 57% (5% + 11%)
  - US: 48% (4% + 12%)

- **EU**: 57%
  - RUSSIA: 16%
  - NATO: 46%
  - US: 36%

- **NATO**: 57%
  - RUSSIA: 16%
  - EU: 46%
  - US: 36%

- **US**: 48%
  - RUSSIA: 16%
  - EU: 36%
  - NATO: 36%

---

Source: Survey data, CPD@SNSPA, June–August 2018, 
Sputnik network

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

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


**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**
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
Manual labeling

● Keywords to identify flags:

bolsevic bolsevica bolsevic bolsevic bolsevicul bolsevik mujic mujici mujici mujicul pro
putin putinu rrusia ruble rubele rusa rusasi ruseasca rusesc rusi rusia rusiaasa rusiacd rusiade
rusianu rusiapentru russiaa rusiei rusii rusilor rusilorun rusnac rusnacul ruso rusonicele rusofil 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
**STEP 1 - predicting new flags**

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

350 / 2100 (17%) manually classified as flags

**APRIL - OCTOBER**
- ~4100 comments
- All contain keywords
- Prediction, then manual classification

720 / 4100 predicted as flags
- 430 / 720 (60%) manually classified as flags
### STEP1. Flag/ Not Flag Classification diagnostics

<table>
<thead>
<tr>
<th></th>
<th>RAND. FORESTS</th>
<th>GLMNET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>METADATA</td>
<td>0.54</td>
<td>0.24</td>
</tr>
<tr>
<td>WORDS</td>
<td>0.63</td>
<td>0.54</td>
</tr>
<tr>
<td>MIXED</td>
<td><strong>0.69</strong></td>
<td>0.45</td>
</tr>
</tbody>
</table>

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

- Flaggers are people of few words
- However, they receive more positive ratings
**STEP 2. Troll / Not Troll classification diagnostics**

<table>
<thead>
<tr>
<th></th>
<th>Jan – Mar</th>
<th>Apr - Oct</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 score</td>
<td>METADATA</td>
<td>Initial Flags</td>
</tr>
<tr>
<td></td>
<td>0.85</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>WORDS</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>MIXED</td>
<td>0.85</td>
</tr>
</tbody>
</table>

- 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

IF YOU HAVEN’T BEEN CALLED A RUSSIAN BOT
YOU PROBABLY ARE NOT A PATRIOT.

NPC meme is dehumanizing

Trump supporters are russian bots
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!