f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Feeding the troll detection algorithm: Informal flags used as labels in classification models to identify perceived computational propaganda

by Vlad Achimescu and Dan Sultănescu

Abstract

The authenticity of public debate is challenged by the emergence of networks of non-genuine users (such as political bots and trolls) employed and maintained by governments to influence public opinion. To tackle this issue, researchers have developed algorithms to automatically detect non-genuine users, but it is not clear how to identify relevant content, what features to use and how often to retrain classifiers. Users of online discussion boards who informally flag other users by calling them out as paid trolls provide potential labels of perceived propaganda in real time. Against this background, we test the performance of supervised machine learning models (regularized regression and random forests) to predict discussion board comments perceived as propaganda by users of a major Romanian online newspaper. Results show that precision and recall are relatively high and stable, and re-training the model on new labels does not improve prediction diagnostics. Overall, metadata (particularly a low comment rating) are more predictive of perceived propaganda than textual features. The method can be extended to monitor suspicious activity in other online environments, but the results should not be interpreted as detecting actual propaganda.

Contents

Introduction

The emergence of the Internet as a medium of communication has blurred the lines between producers and consumers of content (Thurman, 2008). In one view, the Internet has enabled people to freely express their opinions, as *netizens* (MacKinnon, 2012) in an online version of the *public sphere* (Habermas, 1989), where public opinion is created through horizontal interactions between peers, giving everyone a voice, rather than a top-down communication process from mass media to consumers. Public debate forums such as Reddit or comment sections of online newspapers encourage two-step flows of communication (Lazarsfeld, *et al.*, 1948), in which news is shared and reinterpreted through dialogue.

This positive view of a digital democracy tends to ignore the different motivations users have for posting online comments and assumes that there are no bad actors. Some users, referred informally as trolls and formally as *non-genuine actors* (Paletz, *et al.*, 2019), are motivated by intentions to manipulate public opinion, sending targeted political messages while pretending to be ordinary forum users. When their activity is not independent, but guided by an interested third party organization, foreign or domestic, they are engaging in *computational propaganda* (Bolsover and Howard, 2017), *political astroturfing* (Zhang, *et al.*, 2013) or *coordinated inauthentic behavior* (Weedon, *et al.*, 2017).

Recently, lists of non-genuine users linked to Russian sources were made public and banned from Twitter, Reddit, Facebook and Instagram in the United States [1]. Studies on computational propaganda show that that the presence of Russian bots and trolls on Eastern European social media and newspaper forums is non-negligible (Mihaylov, *et al.*, 2018; Zelenkauskaite and Niezgoda, 2017). Romania is a country vulnerable to anti-Western narratives, particularly its young population (Sultănescu, 2019), and there is increasing evidence of networks of Russian non-genuine user accounts operating on social networks in Romania (Dawson and Innes, 2019; Gleicher, 2019; Ioniță, 2019).

Coordinated networks of non-genuine actors corrupt the online public sphere by flooding it with disinformation (Ehrenfeld and Barton, 2019; Starbird, *et al.*, 2019), and false information tends to spread faster than truthful information (Vosoughi, *et al.*, 2018). Identifying non-genuine users is an important step in combating computational propaganda, which represents a real threat to deliberative democracy. Previous research on computational propaganda (Crothers, *et al.*, 2019; Grimme, *et al.*, 2017; Im, *et al.*, 2019; Stukal, *et al.*, 2017; Zannettou, *et al.*, 2019; Zannettou, *et al.*, 2018) has focused on detecting bots or trolls on social media and describing their behavior. Much of this research uses machine learning for classification, which is more efficient than manual classification. But the process cannot be fully automated, as every supervised learning algorithm for text classification requires an initial labeled set (Grimmer and Stewart, 2013). Various techniques to label users as trolls have been proposed (researcher, institutional and user labeling), but there is no consensus regarding the best labeling method. Some studies use original coding schemes directly on the texts (Stukal, *et al.*, 2017), while others use pre-existing labels, provided by online platform administrators (Im, *et al.*, 2019) or users (Mihaylov, *et al.*, 2018).

In this paper, the merits of a form of user labeling are investigated. Trolling is often sanctioned by forum users either formally or informally, and these sanctions, particularly informal flags (replies such as "you sound like a Russian troll"), can be used as a starting point to identify suspicious content posted by non-genuine users. We use a corpus of comments scraped from the message board of an online Romanian newspaper. We test two types of models, regularized regression and random forests, in a two-step supervised machine learning approach: first classifying messages as "informal flags" or "non-flags", then using the identified flags to classify new messages as "flagged as propaganda" or "not-flagged". We rely on two types of predictors as inputs to our classification models: *textual features* (*i.e.*, the words in the comments, tokenized and stemmed) and *metadata* (*i.e.*, timestamps, comment position or rating).

Our main goal is to build a model for predicting perceived propaganda with the lowest degree of prediction error, where flagged messages are identified with high precision and recall. Our setup permits checking for concept drift and feature importance. To analyze concept drift, we check whether the two-step classification, which requires updating the labels, improves accuracy compared to the one-step classification where a model is trained only on the initial manually classified flags. Finally, we investigate the types of features that influence the prediction of flags (step 1) and flagged messages (step 2), to check whether metadata or textual features are more important for the classification, and whether using both improves the prediction performance.

Theory and previous research

The online public sphere

The public sphere was conceptualized in the early 1960s by the German sociologist Jürgen Habermas. He defines the public sphere as "private people gathered together as a public and articulating the needs of society with the state" [2]. It manifests itself both through discussion and common action. It is "public" in the sense that it is non-exclusive (private), but it is also autonomous from the authority of the state (that encompasses a different understanding of "public").

It is tempting to see the rise of the Internet as reviving the public sphere in the twenty-first century, after the domination of the mass media gatekeepers in the previous century. Authors inspired by the writings of Habermas adapted the concept to the new technological realities of the Internet, by imagining an *online public sphere* as the emerging space for deliberative democracy (Dahlberg, 2001). By losing its spatiality, the public sphere is conceived as a more inclusive discursive space, giving voice to the voiceless (Mitra and Watts, 2002). With the rise of blogs, social media and online activism, the idea of the *netizens* — Internet users that can hold governments and corporate actors accountable (MacKinnon, 2012) — started to take shape. However, Geiger (2009) notes that Habermas himself did not hold such optimistic views, arguing that it did more to increase fragmentation and polarization, rather than inclusiveness.

Decentralized interactions on the Internet produce independent micro-publics instead of a public sphere (Geiger, 2009). Indeed, if we look at the large Internet platforms, it is hard to disagree with the fragmentation hypothesis. Facebook and Twitter users choose who to follow, locking themselves into partisan groups. Reddit is split into more than 100,000 subreddits, each with its own posting rules. Large newspapers, on the other hand, maintain a broader and more diverse audience, and the comments sections from some online newspapers look more like *communities of debate* (Ruiz, *et al.*, 2011). But the quality of these debate platforms can be affected by non-sincere debaters, such as *trolls*.

Trolling and informal flagging

Trolling is an umbrella term for a wide range of norm-defying patterns of behavior in online communities. Trolls are (often anonymous) forum users who "act repeatedly and intentionally to cause disruption or trigger conflict among community members" [3]. By promoting incivility and divisiveness, trolling is challenging the idea and functioning of authentic public debate (Munger, 2017; Starbird, *et al.*, 2019). Furthermore, trolling is contagious: a person can become a troll by exposure to antisocial messages (Cheng, *et al.*, 2017).

Most online communities have a formal and/or informal set of injunctive norms referring to unacceptable behaviors that may include how trolling is sanctioned by the community (Álvarez-Benjumea and Winter, 2018). There are *preventative* and *remedial* interventions in case of detecting deviance (Fichman and Sanfilippo, 2015; Sanfilippo, *et al.*, 2017). Preventative interventions refer to defining public standards and restricting membership to online communities. Moderators can apply remedial interventions after they consider that rules were broken. Remedial interventions include both *formal technological sanctions*, such as banning, filtering offensive language, flagging and censoring specific messages, and *informal sanctions*, such as polite private discussions, conversion attempts and confrontations.

In this paper, we refer to the strategy of users calling out or denouncing trolls by simply replying to their posts as *informal flagging*. This contrasts with a more formal form of flagging, such as downvoting or reporting a comment to the moderators for breaching the norms of an online community. Informal flagging can have other functions as well, such as warning the community not to trust specific users, but also attempting to discredit a user's opinion and marginalize them by associating them with a non-desirable entity (*e.g.*, intelligence agencies, political parties, or multinational corporations).

Computational propaganda and political astroturfing

The concept of trolling needs to go through a disambiguation process because it conflates intrinsically motivated trolls who enjoy creating controversy for their own amusement (Phillips, 2015) with extrinsically motivated trolls who are hired by a political group or foreign country to engage in online astroturfing. In astroturfing, deception and coordination are key features, as the appearance of popular support is used to legitimate the organization's interests, and political campaigns are masked as citizen initiatives (Zhang, *et al.*, 2013). If some users are not expressing their own

spontaneous opinions, but their participation is paid for and their messages are scripted, than the assumed person-to-person messaging in online communities becomes organization-to-people messaging resembling targeted advertisements (Sobkowicz and Sobkowicz, 2012).

Along with social bots (Grimme, *et al.*, 2017), extrinsically motivated trolls are non-genuine users (Paletz, *et al.*, 2019) employed by organizations for the purpose of spreading computational propaganda. Collaboration between non-genuine users has also been called coordinated inauthentic behavior (Weedon, *et al.*, 2017). Computational propaganda is "the assemblage of social media platforms, autonomous agents and big data tasked with the manipulation of public opinion" [4], and it can be employed both locally and abroad. A recent study finds that in 2019 at least 70 countries were targets of social media campaigns using cyber troops to "manufacture consensus, automate suppression, and undermine trust in the liberal international order" [5]. Sometimes foreign governments engage in computational propaganda to influence election outcomes or policy changes in other countries. Martin, *et al.*(2020) identified 53 foreign influence efforts in the last 10 years, of which 85 percent employed human non-genuine actors (trolls), most of them originating from Russia, targeting 20 countries.

Van Herpen (2016) describes the effort Russia takes to increase its soft power, by employing methods to achieve foreign policy objectives without using force, including the employment of inauthentic users. The new model of cross-border Russian propaganda has been called a "firehose of falsehood" [6], perpetrated by bots, bloggers and employees of troll farms, that produce Web content with high volume and velocity on multiple channels. These pseudo-users send repetitive but sometimes contradictory messages, lacking in objectivity and consistency, discrediting official sources without much concern for the truth. At home and in neighboring countries, they induce a climate of fear that can deter regular posters from expressing their opinion online (Aro, 2016).

The most famous case of institutionalized Russian propaganda is the Internet Research Agency, the "troll farm" (Chen, 2015) alleged to have attempted to influence the 2016 U.S. elections (DiResta, *et al.*, 2018). Troll and bot networks have been identified in Russia and its neighboring countries (Mihaylov, *et al.*, 2018; Stukal, *et al.*, 2017; Zelenkauskaite and Niezgoda, 2017). The European Parliament warns about foreign influence operations criticizing democratic institutions, human rights, or freedom of movement, and promoting populist parties (Bentzen, 2018), and anti-EU and anti-NATO master narratives abound in Russian state sponsored propaganda efforts in central and eastern Europe (Sultănescu, 2019). Facebook recently removed non-genuine pro-Russian accounts from Romania (Gleicher, 2019), and some IRA Twitter accounts were also posting in the Romanian language (Dawson and Innes, 2019).

In this paper, we use the term "trolls" to refer strictly to suspected non-genuine users associated with political astroturfing, particularly with spreading Russian propaganda on online newspaper forums.

We choose to analyze online newspaper comments on discussion boards (forums) rather than social media posts for several reasons. Interaction on these forums resemble more closely the two-step flow of communication model (Lazarsfeld, *et al.*,

1948), since comments are reactions to articles posted in the media rather than standalone posts on blogs or social media. Discussion boards are prone to trolling as they are more often anonymous, and access is not restricted to friends or followers. Discussion boards of mainstream newspapers are not ideologically homogenous places, and many discussions are highly emotional inter-faction disagreements (Sobkowicz and Sobkowicz, 2012). Exposure to discussion board comments can be high in some countries: a survey shows that 37 percent of likely voters in presidential elections in Romania say they read comments at the end of online newspapers at least once a week (Sultanescu, *et al.*, 2019). Data from newspaper discussion boards are publicly accessible, and all comments can be made available through Web scraping (Munzert, *et al.*, 2015), unlike in some social networks (*e.g.*, Facebook).

Detecting online propaganda

Manually identifying deviations from civil online debates, such as hate speech, trolling or propaganda is not scalable in the online environment. Machine learning can be a solution to this problem, but it requires an initial labeled set, and there is no standard way to obtain one. Several methods were used in previous research: *researcher labeling, institutional labeling* and *user labeling*. Each of them is briefly described below.

Content analysis traditionally requires that messages are coded directly by *researchers* employing a coding scheme (Grimmer and Stewart, 2013). Labeling a message as propaganda assumes a high level of expertise from the coders, and consensus is difficult to reach. *Researcher labeling* is more often applied to other forms of trolling, for example, when building an incivility and intolerance scale (Álvarez-Benjumea and Winter, 2018) or labeling sexist and racist discourse (Badjatiya, *et al.*, 2017). These cases are more suited to dictionary-based approaches. Russian bots on Twitter have been identified through a supervised machine learning algorithm (Sanovich, *et al.*, 2018; Stukal, *et al.*, 2017) that selects all articles with political keywords and uses a large number of human coders that manually classify users as "bots" or "humans" based on a series of guidelines.

To reduce temporal and monetary costs, pre-existing labels can be used, when available. *Institutional labeling* occurs when official sources, such as the owners of certain online platforms, publish lists of accounts suspected of engaging in spreading propaganda. There is limited transparency regarding the classification process for labels obtained this way, and data are static: rarely updated after release, only covering a particular time period. Institutional labeling was used to identify Russian trolls by training machine learning models on posts made by accounts closed by Twitter or Reddit as belonging to the Internet Research Agency (Crothers, *et al.*, 2019; Im, *et al.*, 2019).

Finally, there is *user labeling*. Labels produced by users of a public discussion forum can be repurposed for training models. In some cases, there are formalized ways in which users label comments (downvoting, flagging for inappropriate content). A study about the automated detection of hate speech (Cheng, *et al.*, 2017) relied on formal flagging (downvoting) to label norm breaking users. On most platforms, there is, however, no "flag for propaganda" button, since it is more unstable than hate speech, and it is harder to establish objective characteristics.

Informal flagging occurs when a user accuses another user of being a non-genuine actor. Employing these informal flags as labels requires an initial step of identifying informal flags (such as "You sound like a Russian troll") through a set of keywords, and then labeling the messages that the flags are replying to as perceived propaganda. To our knowledge, there are only two instances where informal flagging for propaganda (*i.e.*, calling out trolls) was analyzed. In a study on Russian trolls in Lithuania (Zelenkauskaite and Niezgoda 2017), classification was done manually. In another study, informal flags were used to identify perceived propaganda in the Bulgarian public sphere (Mihaylov, *et al.*, 2018). The authors used a one-step approach, searching for the word "troll" within comments and manually coding the informal flags, using regularized logistic regression.

In contrast, we propose a two-step approach, where we first use machine learning to identify new, unlabeled flags and then use those flags to identify messages flagged for propaganda. Compared to other labeling methods, user labels are a dynamic dataset, in the sense that users create new labels in real time, without researcher intervention. This can be an advantage for propaganda detection. It is important to know whether it is necessary to incorporate new labels into the dataset or if it is enough to train the classification model on old labels. Given the volatile nature of propaganda, with shifting topics and tactics (Paul and Matthews, 2016), we expect that periodically re-training the dataset would bring improvements in prediction accuracy. To our knowledge, no other study has used this strategy until now.

Data and methods

The data for this study are extracted from one of the largest online quality newspapers in Romania, www.hotnews.ro. The site had over 250,000 unique visitors per day in 2019 [7]. Their distinctive features are a strong pro-NATO and pro-EU orientation and the promotion of liberal democratic and free market values (Punti-Brun and Lozano, 2017). The comments section requires registration, but pseudonyms can be used, offering a degree of anonymity that may encourage trolling behavior. Moderation is mostly left to the users, who can police the forums by downvoting or upvoting other comments. Comments with a negative net rating (upvotes minus downvotes) are automatically hidden, but they can be revealed with a click.

Sometimes, users will reply to other users' posts, accusing them of being trolls, shills, sock puppets or spreading propaganda for monetary gains. These replies represent informal flags. In many cases, the accusations are explicitly about Russian propaganda. Figure 1 offers a graphical representation of the social network of comments and replies to comments from two articles. Figure 2 provides a few examples of comments flagged as Russian propaganda.

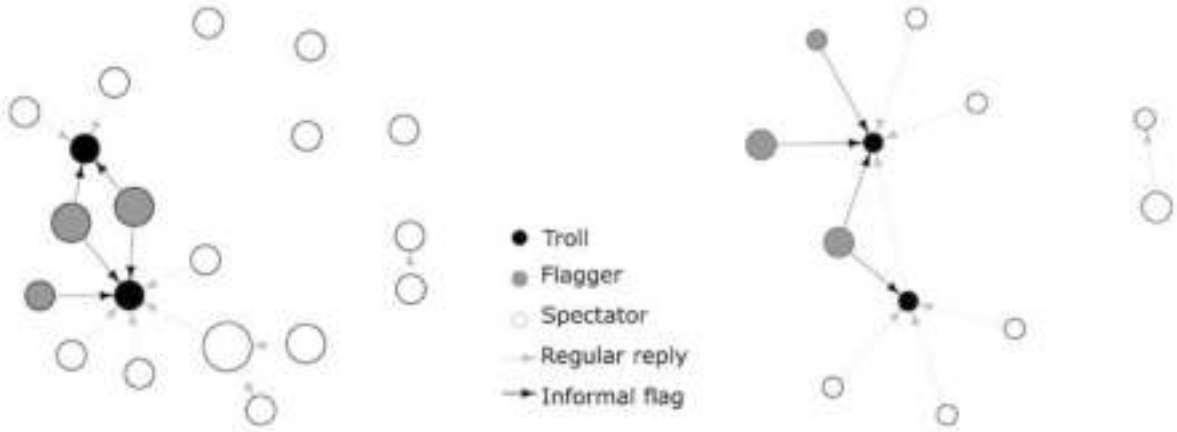


Figure 1: Social network of users' comments and replies under two online newspaper articles. Flaggers accuse trolls of spreading Russian propaganda. Area of circle is proportional to the overall number of comments posted by each user.

ARTICLE 1. Donald Trump says blaming Russia for interference in US elections is wrong (Jan 1 2017)
FLAGGED COMMENT: "Go Trump, stomp on these lackeys with their anti-Russia hysteria" (-2)

- INFORMAL FLAG: "Does Putin pay you in transferable or convertible rubles?" (+2)

ARTICLE 2. US ambassador Klemm: we do not agree to proposed changes in the justice system, we support Kovesi [leader of the Anti-Corruption Agency in Romania] (Jan 19 2017)
FLAGGED COMMENT: "Klemm supports Kovesi, but who is still supporting Klemm?"(-30)

- INFORMAL FLAG / REPLY: "The Romanian People, and also Prince Charles. You however are supported only financially by Vladimir Putin" (+5)

ARTICLE 3. Trump administration is developing an anti-missile defense system against Iran and North Korea (Jan 20, 2017)
FLAGGED COMMENT: "He can use the existing one, currently directed against Russia"(-10)

- INFORMAL FLAG / REPLY: "You (Russians) are so funny, Nikolai. So the shield is directed against Russia? Damn shield, I didn't know Russian missiles from Kaliningrad pass through Romania on their way to the US" (+5)

ARTICLE 4. Conflict between Ursula von der Leyen and Donald Trump (Mar 19 2017)
FLAGGED COMMENT: "EU without the UK army is so anemic. It was fed with vegetables. Merkel is a friend of Putin and will not be afraid of war. We must defend the West from the Ottomans!! Hahaha! The West is doing its part!"(-2)

- INFORMAL FLAG / REPLY: "If Merkel is afraid of Putin, then it means that Trump is Putin's stepfather. Where do you come up with this, you trolls?" (+2)

Figure 2: Examples of comments that were informally flagged as Russian propaganda on hotnews.ro (user rating of comment in parentheses). Translated from Romanian.

We collected the data through Web scraping. Each comment to each article from the first 10 months of 2017 was extracted along with relevant metadata: author name, date and time of posting, position of comment, rating of comment, article title, date and category.

We filtered the scraped comments using a set of 1,232 initial keywords in order to find informal flags. The keywords are broader than those used in previous research (Mihaylov, *et al.*, 2018), and contain all references to Russia, Putin, Kremlin or trolling in all possible declinations and conjugations (Appendix 2 shows the dictionary used to obtain the keywords). Two coders examined each comment from January–March 2017 that contained any keyword, and manually labeled it as "flag", if it was a reply accusing a poster of being a Russian troll, or "non-flag" otherwise.

In previous studies detecting computational propaganda, models were trained only on manually classified labels, and tested on a dataset from the same period. However, text content of propaganda and user reactions to propaganda could gradually change over time, thus the link between label *Y* and predictor set *X* may also shift. The inconsistency between present and past conditional distribution of *Y* given *X* is known in machine learning literature as *concept drift* (Tsymbal, 2004). Adaptive learning refers to "updating predictive models online during their operation to react to concept drifts" [8].

We contrast the one-step approach, in which we only use the initial manual labels, with the two-step approach, in which we first try to identify new informal flags from the future period, and then we use those as labels (with or without the initial flags). The goal is to check for concept drift: whether the prediction accuracy decreases unless the new labels are used. Both approaches are schematically presented in Figure 3.

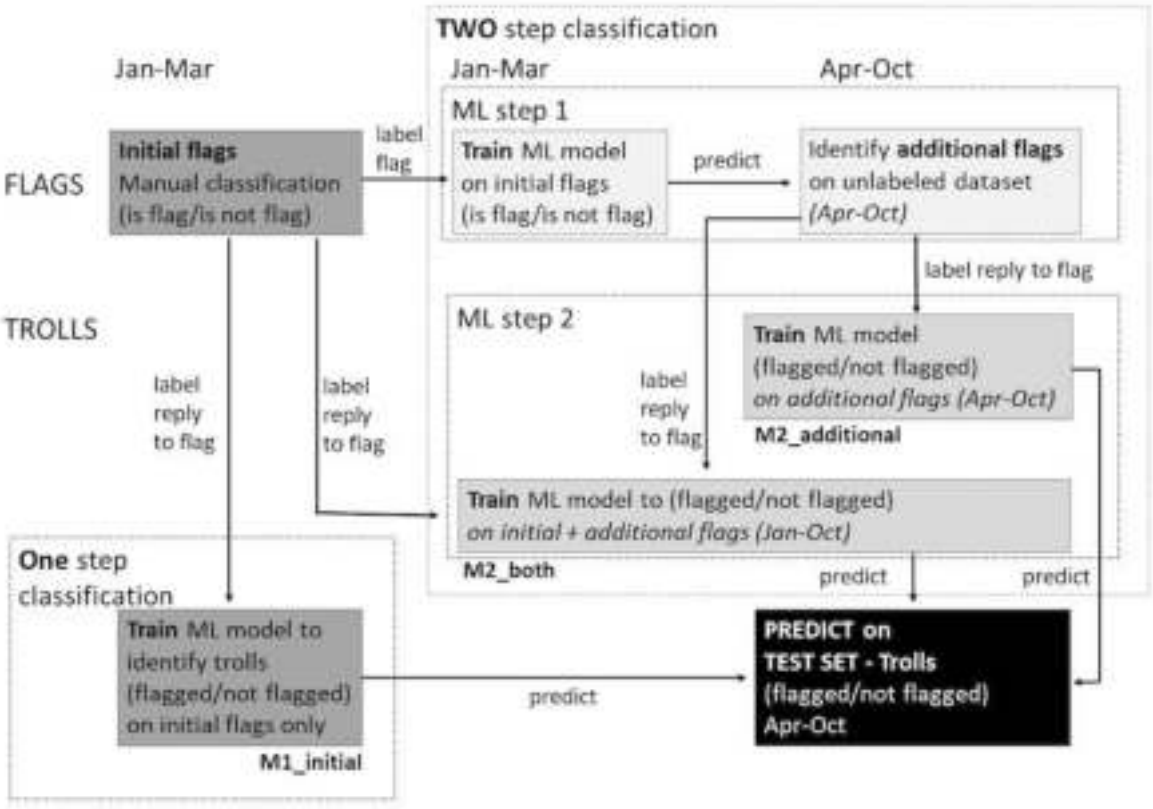


Figure 3: Flowchart of one-step and two-step classification of perceived Russian trolls on online newspaper forums.

In the one-step approach (Figure 3, bottom-left side), we label the comments that were informally flagged using only the initial flags obtained through manual classification. We train supervised machine learning (ML) models with five-fold cross-validation on data from January–March 2017 and select the one that has the best accuracy. We then use this model (M1_initial) to predict on a test set of later comments (April–October 2017) and calculate prediction diagnostics.

In the two-step approach (Figure 3, right side), we first use supervised machine learning to identify new flags (ML step 1) from the second period (April–October 2017). Then, we use these additional labels to predict whether a comment is flagged as Russian propaganda (ML step 2) in the same type of models and on the same test set used in the one-step approach. The advantage of the two-step approach is that it offers updated labels.

The first ML task in the two-step approach is *identifying an informal flag*. We trained the models on the labeled set of replies to comments from January to March 2017 that we manually classified as informal flags. We used a training set containing 70 percent of flags where we experimented with different values for hyper-parameters. Because of imbalanced classes, the training set oversamples flags. We checked the prediction diagnostics on the test set containing 30 percent of cases. We then predict new flags on a non-labeled set of comments made in a later period, from April to October 2017, and manually check if positive predictions are indeed flags, removing the others.

The second ML task in the two-step approach is *identifying perceived Russian propaganda*. The models are trained on messages that have at least one informal flag. We use a training set of additional flags from April–October 2017 (M2_additional) and another training set of both initial manually labeled flags and additional flags from January–October (M2_both). We check prediction diagnostics on the same test set of comments from April–October 2017 used in the one-step approach. If the two-step method is an improvement, prediction diagnostics for the model trained on more recent labels should be improved compared to the model trained on the initial labels. If the one-step and two-step methods provide similar results, there is no evidence for concept drift and re-training is not necessary.

In all the models, we randomly extracted comments from trusted users (at least 100 posts, were never flagged) and gave them the "not flagged" label. For each "flagged" case we had as labels, we extracted three "not flagged" cases. The choice of three was not random, since about one in four comments are posted by an informally flagged user.

Informal flagging is a reaction to another comment, where the flagger may be triggered by specific words or phrases in that comment, which is why we use the word content as textual features. We used a bag-of-words approach, a common practice in computational text analysis (Grimmer and Stewart, 2013). For each comment, we count the number of times each word (token) appears (0 if not at all). A document-term frequency matrix (comments in rows, words in columns, counts in cells) contains all the textual features. All words were set to lowercase and stemmed to reduce the sparsity (number of zeroes) of the matrix. Term frequency–inverse document frequency (tf–idf) weighting was applied. We also included textual features such as punctuation, number of links, hashtags, numbers or years posted, and percentage of words written in capital letters.

As additional features, we used metadata referring to the time the comment was posted, the popularity of the comment and the position of the comment within the thread (see Appendix 2for descriptives). We tested three configurations of features: textual features only (the document-term matrix), metadata only and mixed (textual features plus metadata). This was done to check if the content or other characteristics of the comment drive the prediction, and whether it is best to include both.

We chose regularized regression (with elastic net) and random forests as classifiers. These methods are commonly used in text classification (Gentzkow, *et al.*, 2019). The dataset contains many features, but relatively few labeled cases, which can lead to overfitting. Our chosen methods have feature selection built-in to avoid this situation [9].

Each method has advantages and disadvantages. Logistic regression converges faster but assumes linear relationships. We used it as a baseline model. Random forests are non-parametric and more computationally intensive but allow for complex interactions and nonlinear relationships. To avoid overfitting, we used multiple five-fold cross-validation (three times) and tuned the models with different hyperparameters and feature sets [10]. Then, we calculated precision, recall and F1 score to compare models in the test set [11], because they are more suitable for unbalanced datasets. We were particularly interested in how well the metadata affects prediction compared to the textual features.

Results

All 209,742 comments belonging to 22,349 articles posted between January and October 2017 on www.hotnews.ro were scraped from the site, along with metadata attached to each comment. Between January and March, we selected 2,118 comments that contained the keywords from Appendix 1. These comments were manually classified in two categories: "informal flags" and "not flags" by two raters, with 82 percent agreement. The cases where there was disagreement between raters were not included in the classification. In the end, 354 comments classified by both raters as flags were labelled as "informal flags" and 1,398 as "not informal flags".

In the next part, we present the classification diagnostics used to predict informal flags in the first step of the classification, also looking at variable importance partial dependence plots related to the flag-prediction model. We show some descriptive statistics regarding the prevalence of messages belonging to informally flagged users. Then we present the classification diagnostics used to predict flagged messages in the second step of the classification. We compare the one-step and two-step prediction diagnostics to answer our research question about concept drift. We compare metadata-only and textual features only diagnostics to answer our research question about feature importance in models predicting flagged comments. Finally, we look at feature importance and partial dependence plots to see how the most important features are related to predicting whether a message is flagged for propaganda or not.

Step 1: Identifying flags

We used the informal flags as labels in the first step of the supervised machine learning models, predicting whether a message is a flag or not. The classification diagnostics in Table 1show that overall, random forests perform similarly to regularized regressions (maximum F1 score is 0.54).

Comparing metadata and textual features

Regarding the feature sets tested, using only the word content of messages shows an increased precision (0.50 vs 0.42) and recall (0.47 vs 0.28) compared to using only metadata in the random forest models. It seems that textual features are more useful for

prediction, as flags can more easily be identified using a common set of keywords. Combining both metadata and word content shows a slight increase in precision for random forests (from 0.50 to 0.58), but not for regularized regression.

Table 1: Classification diagnostics for predicting flags in step 1, by method and feature set (rows).

	Random forests			Regularized regression		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Textual features	0.50	0.47	0.48	0.58	0.51	0.54
Metadata	0.42	0.28	0.33	0.33	0.67	0.44
Mixed	0.58	0.47	0.52	0.47	0.60	0.53

To find the most relevant features retained by the classification, we calculate variable importance scores and create partial dependence plots on the random forests model with mixed features.

Figure 4 shows predicted probabilities of a message being a flag depending on the comment rating and number of words in the comment in the metadata only model, all else set to average. These features are the most influential in separating flags from non-flags. Comments with fewer words (less than 50) and positively rated have a higher probability of being predicted as flags.

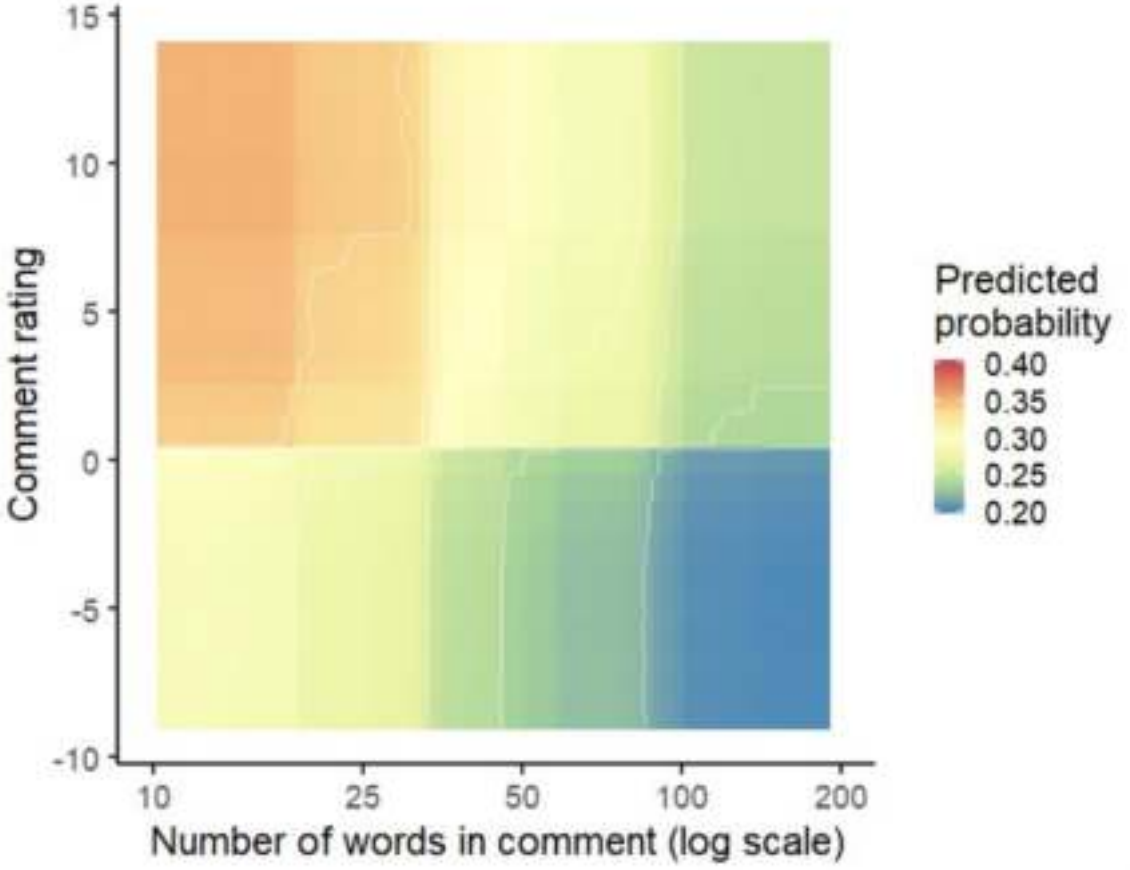


Figure 4: Heatmap of predicted probabilities of a post being classified as an informal flag.

Variable importance plots (Biau and Scornet, 2015) applied to the textual feature only model classify all features in terms of relative predictive power [12]. The predictor with the highest importance is set at 100, and all others are proportionally scaled in comparison to the highest. Appendix 3 (left) contains the top 50 features. The words that are most distinctive of informal flags are, unsurprisingly "esti" ("you are"), "rus" (Russian), and "postac" ("paid poster/troll"). Most of the other highly predictive textual features refer to Russia, Putin or Russian themed words (*e.g.*, "rubles", "Soviet", "USSR"), along with "Trump" and "platit" (paid).

Finding new informal flags

To obtain new labels, we predicted whether a post is an informal flag, using an ensemble of the top three most predictive models on an unlabeled dataset from a different period. Cases predicted as flags by at least two of the three models were investigated. From the 4,100 posts in the period April to October 2017 that contain the

initial keywords, 767 were classified as informal flags by the prediction model. Since the expected precision of the model is not very high, we manually inspected each of them. Two coders manually classified the predicted flags (85 percent in agreement) and found that 55 percent were indeed informal flags, close to the expected precision level of 0.58. This is much higher than the initial detection rate of 20 percent. A total of 425 additional flags were thus recorded for the second period.

Descriptive statistics for flagged users

Flagged users are among the most active on the forum, with an average of 1.1 comments/day, compared to 0.06 comments/day for non-flagged users (see Table 2). In total, 215 users (three percent of active users) were flagged at least once as trolls; in total, these users posted 35 percent of all comments on the forum. 78 users were flagged at least twice as trolls; in total, they posted 21 percent of all comments on the forum. 21 out of these 78 users were not labeled as trolls in the initial period (January–March 2017), showing that the pool of potential non-genuine users is increasing over time.

Table 2: Classification diagnostics for predicting flags in step 1, by method and feature set (rows).

	Number of users		**Comments of users**		**Comments per day**
	n	percent	*n*	percent	mean
never flagged	7101	0.971	135825	0.648	0.064
flagged once or more	215	0.029	73917	0.352	1.150
flagged twice or more	78	0.011	44773	0.213	1.910

Step 2: Identifying perceived propaganda

We labelled all “informally flagged” messages as perceived propaganda. Then we randomly selected control cases from users that were never flagged and posted at least 100 comments to label as non-propaganda. For every flagged message, we selected three non-flagged messages. After labeling the cases, we ran machine learning models predicting whether a message was flagged as propaganda or not.

Three different models were tested. Model M1_initial is trained on the initial flags (January–March 2017). This is the one-step approach. The other models represent the two-step approach. Model M2_additional is trained on the additional flags from the later period (April–October 2017), obtained through the algorithm described in step 1 (detecting more flags); Model M2_both uses both initial and additional flags. All models are tested on the later period.

Comparing performance of the two-step model to the one-step model

When using metadata in the feature set, the one-step and two-step models perform similarly (0.81–0.86 F1 score). This happens even though the one-step model (M1_initial) is trained on January to March data only while the two-step models (M2_additional and M2_both) are updated with more recent labels (April to October, same period as the comments in the test set). Precision and recall do not drop if the model is not re-trained for the mixed model. For the textual-features-only model, the F1 score increases from 0.32 to 0.42 in the random forests model that only uses the additional labels (M2_additional). Therefore, there is a hint of a concept drift in what words are associated with Russian propaganda over time, though adaptive training does not improve overall prediction when using full feature set.

Table 3: Classification diagnostics (F1-score) for models predicting messages flagged as Russian propaganda on test set from April–October 2017.

	M1_initial Trained on initial flags (Jan–Mar 2017)	**M2_additional** Trained on additional flags (Apr–Oct 2017)	**M2_both** Trained on initial+additional flags (Jan–Oct 2017)
	F1 score	*F1 score*	*F1 score*
Regularized regression:			
Textual features	0.28	0.31	0.39
Metadata	0.84	0.83	0.84
Mixed	0.81	0.82	0.84
Random forests:			
Textual features	0.32	0.42	0.37
Metadata	0.85	0.85	0.85
Mixed	0.86	0.85	0.86

Comparing metadata and textual features

Classification diagnostics in Table 3 show that, regardless of model, metadata are much more important than the textual features for predicting messages flagged as propaganda. The F1 score in the metadata only model is 0.8–0.9 compared to 0.3–0.4 for the textual-feature-only model. Precision is similar to recall. Random forests and regularized regression perform similarly for the metadata and mixed models (F1 score of over 0.8).

The most important feature in separating between potential propaganda and other messages is the comment rating. Informally flagged messages tend to have more negative ratings, and this is visible in the partial dependence plot in Figure 5. Posts with negative ratings have a higher chance of being informally flagged as potential propaganda. Additionally, flagged messages appear more frequently in threads with a higher number of comments.

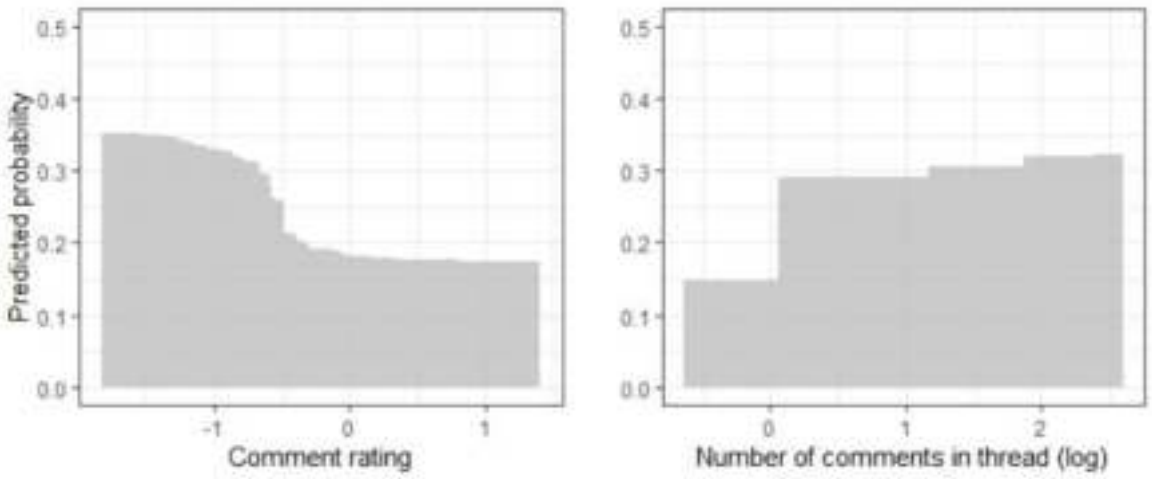


Figure 5: Partial dependence plot relating the comment rating with the probability of predicting a message flagged as potential propaganda.

Textual features do not appear as important in the classification compared to metadata. However, there are some associations: comments that contain more words referring to geopolitical entities (Russia, European Union, United States) are more often classified as propaganda (See Appendix 3 for the 50 most relevant features).

Discussion and conclusion

There is an increasing interest in studying the behavior of non-genuine actors on the Internet. As time passes, more influence campaigns that use social bots or trolls are uncovered (Bradshaw and Howard, 2019; Martin, *et al.*, 2020; Woolley and Howard, 2019), and there are concerns about their effects on eroding democratic values (Bentzen, 2018; Paul and Matthews, 2016; Sultănescu, 2019). On online discussion boards, users are sanctioning what they perceive to be foreign propaganda, by informally flagging comments. Our study revealed that on a popular Romanian online newspaper, more than one in five comments posted below an article belongs to a user flagged as a 'Russian troll' by other forum users.

Informal flags are a useful labeling method to detect perceived political astroturfing and propaganda on Romanian online discussion boards using supervised machine learning. The results of this study are comparable to results from previous research from Bulgaria (Mihaylov, *et al.*, 2015; Mihaylov, *et al.*, 2018; Mihaylov and Preslav, 2016), with over 80 percent of cases correctly predicted, and a large part of the prediction driven by metadata, particularly comment rating.

We compared a static design (one-step classification) with an adaptive design (two-step classification) to check for potential concept drift. Both approaches produce similar results, suggesting that the perceived propaganda detection model produces stable predictions over time on the news platform. We were expecting that periodically updating the labels in the model would improve accuracy, given the unstable nature of propaganda, but found weak evidence for that. Adaptive design does not improve overall prediction diagnostics in our setup. Concept drift may however occur over longer periods than the one tested here (one to seven months later), especially if flagging behavior patterns would change.

In our study, the textual features of the comments were more relevant for predicting informal flags (accusations of being Russian trolls), while metadata were more relevant for predicting perceived Russian propaganda. The low impact of word content on the predictive power of the model shows that prediction is highly driven by the action that most often accompanies an informal flag: downvoting. Thus, we could not find evidence that flaggers are triggered by specific keywords, and other mechanisms might be at play, such as flagging all messages from users previously labeled as trolls, with little regard for the content of the current message. Future research should focus on identifying factors that influence the informal flagging process.

The low impact of textual features on prediction raises concerns about external validity. If exported to a different forum with different rules, the model might not be able to predict with the same level of accuracy. Generalizing the exact model to detect trolling on other Web sites, or to other forms of trolling or harassment is not recommended. However, the procedure itself (combining metadata and word content in a one-step or two-step model) could be replicated, but success is highly dependent on the number and validity of informal flags on each Web site and the amount of metadata that can be

extracted. The model obtained in the first step can be applied to find more flags on other Romanian Web sites, since the textual features are driving prediction.

Externalization of labelling to users reduces costs and speeds up research. Human coders have less work to do because flags are easier to identify than trolls. New instruments can be created to help moderators identify political astroturfing and propaganda in real time. But before developing such a tool, more classification methods (including deep learning), different labeling strategies, larger samples and alternative feature sets (such as using part of speech tags or word embeddings) need to be explored, to further improve the classification diagnostics.

One of the main drawbacks of using informal flagging as labels is the risk of relying on false positives. In the absence of ground truth, we predict only what users think is Russian propaganda, and dishonest or uninformed labelers might bias results. The algorithm, even with a perfect prediction power cannot be more accurate than the human flaggers that labeled it. False flags might either demobilize regular users or increase the temptation of trolling. Also, trolls can adapt and find ways to thwart the instrument: they can change their posting behavior, or accuse random users of being Russian trolls themselves, which would affect the algorithm. We advise using additional methods of identifying propaganda if available for triangulation, rather than relying purely on user labels.

About the authors

Vlad Achimescu is a Ph.D. student at the Professorship for Statistics and Methodology at the University of Mannheim.
E-mail: v [dot] achimescu [at] uni-mannheim [dot] de

Dan Sultănescu is the research director of the Center for Civic Participation and Democracy at the National University of Political Studies and Public Administration, Bucharest.
E-mail: dan [dot] sultanescu [at] snspa [dot] ro

Acknowledgements

We would like to thank Prof. Florian Keusch, the members of the CDSS graduate school, and the entire FK2 group from the University of Mannheim for reviewing earlier drafts of this paper. We are grateful to Dana Sultănescu for helping us prepare and give the BigSurv 2018 presentation that provided the backbone of this paper. We also thank Prof. Susan Banducci, from the University of Exeter, for commenting on the presentation, and giving us valuable suggestions. Finally, we extend our gratitude to the Web masters of www.hotnews.ro, for keeping the comment section open and JavaScript free, making it easy for researchers to scrape user comments and metadata from their well-organized Web archive.

Data availability

The data was scraped from the archive of the Romanian online newspaper *Hotnews*, Web site: https://www.hotnews.ro/arhiva. The syntax for scraping can be provided by the authors by request. The dictionary applied to extract potential informal flags is presented in Appendix 1. Appendix 2 contains descriptives for the metadata used.

Software information

For data collection, we used Python 3.7.0 (https://www.python.org/downloads/release/python-370/). The *BeautifulSoup* library (Richardson, 2019) was used for Web scraping the comments. The syntax and outputs (.html generated from ipynb file) for scraping is presented in Appendix 5.

For data analysis, we used R version 3.6.1 (https://cran.r-project.org/src/base/R-3/). Words were tokenized and stemmed using the package *quanteda* (Benoit, 2019). For the machine learning models, we used the packages *caret* (Kuhn, *et al.*, 2019), *glmnet* (Friedman, *et al.*, 2019; Hastie and Junyang, 2016) and *randomForest* (Liaw and Wiener, 2018).

The R syntax and outputs for cleaning the data as well as training and testing the one-step and two-step models can be provided by the authors by request.

Notes

1. Lists and description of non-genuine users associated with the Russian Internet Research Agency can be obtained from the following lists:
Twitter: https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf.
Reddit: https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/.
Facebook and Instagram: https://democrats-intelligence.house.gov/social-media-content/social-media-advertisements.htm
https://about.fb.com/news/2019/10/removing-more-coordinated-inauthentic-behavior-from-iran-and-russia/.

2. Habermas, 1989, p. 140.

3. Fichman and Sanfilippo, 2015, p. 163.

4. Bolsover and Howard, 2017, p. 273.

5. Bradshaw and Howard, 2019, p. 1.

6. Paul and Matthews, 2016, p. 1.

7. Traffic statistics for all Romanian Web sites can be inspected at this site: https://www.brat.ro/sati/site/hotnews-ro-1/trafic-total/period_type/day/period_year/2019/period_filter/30days/grafic_type/clients#charts.

8. Gama, *et al.*, 2014, p. 2.

9. Regularized logistic regression (Hastie, *et al.*, 2009) penalizes weak coefficients, using either the L1 or L2 norm (ratio of L1/L2 penalization is controlled by the hyperparameter alpha). We also varied the strength of the penalty as hyperparameter lambda. Random forests (Breiman, 2001) refer to a tree-based classifier where random subsets of variables are chosen to build a large number of decision trees, and the average prediction over all trees is used for the final classification. We varied the number of variables randomly sampled at each split (hyperparameter *mtry*).

10. Hyperparameter tuning grids.
1) For regularized logistic regression, we used elastic net, with alpha {0,0.5,1} and 10 lambda values, from 10−4 to 104.
2) For random forests, we used number of variables available for splitting at each tree node: mtry of 10, 15, 20, 30 and 50.

11. We used precision, recall and F1 score because they show the performance of predicting the positive category (informal flags in the first ML task, messages that were informally flagged in the second ML task), and are not artificially inflated by the number of true negatives (not flag, not flagged), and can handle unbalanced datasets where the negative category has a much larger number of cases than the positive category.

12. The variable importance is calculated in the following manner, as described in the randomForest package documentation: “For each tree, the prediction error on the out-of-bag portion of the data is recorded (error rate for classification, MSE for regression). Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees, and normalized by the standard deviation of the differences.” (Liaw and Wiener, 2018, importance).

References

A. Álvarez-Benjumea and F. Winter, 2018. “Normative change and culture of hate: An experiment in online environments,” *European Sociological Review*, volume 34, number 3, pp. 223–237.
doi: https://doi.org/10.1093/esr/jcy005, accessed 15 August 2020.

J. Aro, 2016. “The cyberspace war: Propaganda and trolling as warfare tools,” *European View*, volume 15, number 1, pp. 121–132.
doi: https://doi.org/10.1007/s12290-016-0395-5, accessed 15 August 2020.

P. Badjatiya, S. Gupta, M. Gupta and V. Varma, 2017. "Deep learning for hate speech detection in tweets," *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760.
doi: https://doi.org/10.1145/3041021.3054223, accessed 15 August 2020.

K. Benoit, 2019. "quanteda: Quantitative analysis of textual data," at https://cran.r-project.org/web/packages/quanteda/index.html, accessed 15 August 2020.

N. Bentzen, 2018. "Foreign influence operations in the EU," *European Parliament* (10 July),
at https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2018)625123, accessed 15 August 2020.

G. Biau and E. Scornet, 2015. "A random forest guided tour," *arXiv*:1511.05741 (18 November), at https://arxiv.org/abs/1511.05741, accessed 15 August 2020.

G. Bolsover and P. Howard, 2017. "Computational propaganda and political big data: Moving toward a more critical research agenda," *Big Data*, volume 5, number 4, pp. 273–276.
doi: https://doi.org/10.1089/big.2017.29024.cpr, accessed 15 August 2020.

S. Bradshaw and P.N. Howard, 2019. "The global disinformation order: 2019 global inventory of organised social media manipulation," *Oxford Internet Institute*,
at https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf, accessed 15 August 2020.

L. Breiman, 2001. "Random forests," *Machine Learning*, volume 45, number 1, pp. 5–32.
doi: https://doi.org/10.1023/A:1010933404324, accessed 15 August 2020.

A. Chen, 2015. "The agency," *New York Times* (2 June),
at https://www.nytimes.com/2015/06/07/magazine/the-agency.html, accessed 15 August 2020.

J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil and J. Leskovec, 2017. "Anyone can become a troll: Causes of trolling behavior in online discussions," *>CSCW '17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1,217–1,230.
doi: https://doi.org/10.1145/2998181.2998213, accessed 15 August 2020.

E. Crothers, N. Japkowicz and H. Viktor, 2019. "Towards ethical content-based detection of online influence campaigns," *arXiv*:1908.11030 (29 August),
at https://arxiv.org/abs/1908.11030, accessed 15 August 2020.

L. Dahlberg, 2001. "The Internet and democratic discourse: Exploring The prospects of online deliberative forums extending the public sphere," *Information, Communication & Society*, volume 4, number 4, pp. 615–633.
doi: https://doi.org/10.1080/13691180110097030, accessed 15 August 2020.

A. Dawson and M. Innes, 2019. "The Internet Research Agency in Europe 2014-2016," *Cardiff University Crime & Security Research Institute*, and at https://www.cardiff.ac.uk/crime-security-research-institute/publications, accessed 15 August 2020.

R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R.B. Fox, J. Albright and B. Johnson, 2018. "The tactics & tropes of the Internet Research Agency," *New Knowledge* (17 December), at https://apo.org.au/node/211296, accessed 15 August 2020.

D. Ehrenfeld and M. Barton, 2019. "Online public spheres in the era of fake mews: Implications for the composition classroom," *Computers and Composition*, volume 54, 102525.
doi: https://doi.org/10.1016/j.compcom.2019.102525, accessed 15 August 2020.

P. Fichman and M.R. Sanfilippo, 2015. "The bad boys and girls of cyberspace: How gender and context impact perception of and reaction to trolling," *Social Science Computer Review*, volume 33, number 2, pp. 163–180.
doi: https://doi.org/10.1177/0894439314533169, accessed 15 August 2020.

J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon and J. Qian, 2019. "glmnet: Lasso and elastic-net regularized generalized linear models," at https://CRAN.R-project.org/package=glmnet, accessed 15 August 2020.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia, 2014. "A survey on concept drift adaptation," *ACM Computing Surveys*, article number 44.
doi: https://doi.org/10.1145/2523813, accessed 15 August 2020.

R.S. Geiger, 2009. "Does Habermas understand the Internet? The algorithmic construction of the blogo/public sphere," *Gnovis*>, volume 10, number 1, pp. 1–29, at http://www.gnovisjournal.org/2009/12/22/does-habermas-understand-internet-algorithmic-construction-blogopublic-sphere/, accessed 15 August 2020.

M. Gentzkow, B. Kelly, and M. Taddy, 2019. "Text as data," *Journal of Economic Literature*, volume 57, number 3, pp. 535–574.
doi: https://doi.org/10.1257/jel.20181020, accessed 15 August 2020.

N. Gleicher, 2019. "Removing coordinated inauthentic behavior from Russia," *Facebook* (17 January), at https://about.fb.com/news/2019/01/removing-cib-from-russia/, accessed 15 August 2020.

C. Grimme, M. Preuss, L. Adam and H. Trautmann, 2017. "Social bots: Human-like by means of human control?" *Big Data*, volume 5, number 4, pp. 279–293.
doi: https://doi.org/10.1089/big.2017.0044, accessed 15 August 2020.

J. Grimmer and B.M. Stewart, 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis*, volume 21, number 3, pp. 267–297.
doi: https://doi.org/10.1093/pan/mps028, accessed 15 August 2020.

J. Habermas, 1989. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Translated by T. Burger with the assistance of F. Lawrence. Cambridge, Mass.: MIT Press.

T. Hastie and Q. Junyang, 2016. "Glmnet Vignette" (13 September), at http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf, accessed 15 August 2020.

T. Hastie, R. Tibshirani and J. Friedman, 2009. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer. doi: https://doi.org/10.1007/978-0-387-84858-7, accessed 15 August 2020.

J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens and E. Gilbert, 2019. "Still out there: Modeling and identifying Russian troll accounts on Twitter," *arXiv*:1901.11162 (31 January), at https://arxiv.org/abs/1901.11162, accessed 15 August 2020.

S. Ioniță, 2019. Influența rusă "Influența rusaă in regiunea Europei de Est şi dincolo de ea [Russian influence in eastern Europe and beyond]," *Expert Forum*, at https://expertforum.ro/wp-content/uploads/2019/07/PB-Influenta_RU.pdf, accessed 15 August 2020.

M. Kuhn, J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan and T. Hunt, 2019. "caret: Classification and regression training," at https://CRAN.R-project.org/package=caret, accessed 15 August 2020.

P.F. Lazarsfeld, B.R. Berelson and H. Gaudet, 1948. *The people's choice: How the voter makes up his mind in a presidential campaign*. Second edition. New York: Columbia University Press.

A. Liaw and M. Wiener, 2018. "randomForest: Breiman and Cutler's random forests for classification and regression," at https://CRAN.R-project.org/package=randomForest, accessed 15 August 2020.

R. MacKinnon, 2012. "The netizen," *Development*, volume 55, number 2, pp. 201–204. doi: https://doi.org/10.1057/dev.2012.5, accessed 15 August 2020.

D.A. Martin, J.N. Shapiro and J. Ilhardt, 2020. "Trends in online foreign influence efforts," *Empirical Studies of Conflict Project, Princeton University*, at https://esoc.princeton.edu/publications/trends-online-influence-efforts, accessed 15 August 2020.

T. Mihaylov and N. Preslav, 2016. "Hunting for troll comments in news community forums," *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics*, at https://www.cl.uni-heidelberg.de/~mihaylov/publication/2016-mihaylovnakov-b/, accessed 15 August 2020.

T. Mihaylov, G. Georgiev and P. Nakov, 2015. “Finding opinion manipulation trolls in news community forums,” *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, at https://www.aclweb.org/anthology/K15-1032, accessed 15 August 2020.
doi: https://doi.org/10.18653/v1/K15-1032, accessed 15 August 2020.

T. Mihaylov, T. Mihaylova, P. Nakov, L. Màrquez, G.D. Georgiev and I.K. Koychev, 2018. “The dark side of news community forums: Opinion manipulation trolls,” *Internet Research*, volume 28, number 5, pp. 1,292–1,312.
doi: https://doi.org/10.1108/IntR-03-2017-0118, accessed 15 August 2020.

A. Mitra and E. Watts, 2002. “Theorizing cyberspace: The idea of voice applied to the Internet discourse,” *New Media & Society*, volume 4, number 4, pp. 479–498.
doi: https://doi.org/10.1177/146144402321466778, accessed 15 August 2020.

K. Munger, 2017. “Experimentally reducing partisan incivility on Twitter” (7 September), at https://kmunger.github.io/pdfs/jmp.pdf, accessed 15 August 2020.

S. Munzert, C. Rubba, P. Meißner and D. Nyhuis, 2015. *Automated data collection with R: A practical guide to Web scraping and text mining*. Chichester, West Sussex: Wiley.

S.B.F. Paletz, B. Auxier and E.M. Golonka, 2019. *A multidisciplinary framework of information propagation online*. Cham, Switzerland: Springer International.
doi: https://doi.org/10.1007/978-3-030-16413-3, accessed 15 August 2020.

C. Paul and M. Matthews, 2016. “The Russian ‘firehose of falsehood’ propaganda model: Why it might work and options to counter it,” *RAND Perspectives*,
at https://www.rand.org/pubs/perspectives/PE198.html, accessed 15 August 2020.

W. Phillips, 2015. *This is why we can’t have nice things: Mapping the relationship between online trolling and mainstream culture*. Cambridge, Mass.: MIT Press.

M. Punti-Brun and J.B. Lozano, 2017. “Online media in Romania: The case study of hotnews.ro,” In: C. Daba-Buzoianu, M. Bîră, A. Duduciuc and G. Tudorie (editors). *Exploring communication through qualitative research*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 41–58.

L. Richardson, 2019. “Beautiful Soup documentation,”
at https://www.crummy.com/software/BeautifulSoup/bs4/doc/, accessed 15 August 2020.

C. Ruiz, D. Domingo, J.L. Micó, J. Díaz-Noci, K. Meso and P. Masip, 2011. “Public sphere 2.0? The democratic qualities of citizen debates in online newspapers,” *International Journal of Press/Politics*, volume 16, number 4, pp. 463–487.
doi: https://doi.org/10.1177/1940161211415849, accessed 15 August 2020.

M.R. Sanfilippo, S. Yang and P. Fichman, 2017. “Managing online trolling: From deviant to social and political trolls,” *Proceedings of the 50th Hawaii International Conference on System Sciences*, at

doi: https://scholarspace.manoa.hawaii.edu/bitstream/10125/41373/1/paper0224.pdf, accessed 15 August 2020.

S. Sanovich, D. Stukal and J.A. Tucker, 2018. "Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia," *Comparative Politics*, volume 50, number 3, pp. 435–482.
doi: https://doi.org/10.5129/001041518822704890, accessed 15 August 2020.

P. Sobkowicz and A. Sobkowicz, 2012. "Two-year study of emotion and communication patterns in a highly polarized political discussion forum," *Social Science Computer Review*, volume 30, number 4, pp. 448–469.
doi: https://doi.org/10.1177/0894439312436512, accessed 15 August 2020.

K. Starbird, A. Arif and T. Wilson, 2019. "Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations," *Proceedings of the ACM on Human-Computer Interaction*, article number 127.
doi: https://doi.org/10.1145/3359229, accessed 15 August 2020.

D. Stukal, S. Sanovich, R. Bonneau and J.A. Tucker, 2017. "Detecting bots on Russian political Twitter," *Big Data*, volume 5, number 4, pp. 310–324.
doi: https://doi.org/10.1089/big.2017.0038, accessed 15 August 2020.

D. Sultănescu (editor). 2019. *Challenges in strategic communication and fighting propaganda in eastern Europe: Solutions for a future common project*. *NATO science for peace and security series, E, Human and societal dynamics*, volume 142. Amsterdam: IOS Press.

D. Sultanescu, V. Achimescu and D. Sultanescu, 2019. "Romania — New trends in civic participation," at http://civicparticipation.ro/wp-content/uploads/2019/11/Report-CPD-SNSPA-New-trends-in-civic-participation-2019.pdf, accessed 15 August 2020.

N. Thurman, 2008. "Forums for citizen journalists? Adoption of user generated content initiatives by online news media," *New Media & Society*, volume 10, number 1, pp. 139–157.
doi: https://doi.org/10.1177/1461444807085325, accessed 15 August 2020.

A. Tsymbal, 2004. "The problem of concept drift: Definitions and related work," *School of Computer Science and Statistics, Trinity College Dublin, Technical Reports*, TCD-CS-2004-15 (29 April), at https://www.scss.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-15.pdf, accessed 15 August 2020.

M.H. van Herpen, 2016. *Putin's propaganda machine: Soft power and Russian foreign policy*. London: Rowman & Littlefield.

S. Vosoughi, D. Roy and S. Aral, 2018. "The spread of true and false news online," *Science*, volume 359, number 6380 (9 March), pp. 1,146–1,151.
doi: https://doi.org/10.1126/science.aap9559, accessed 15 August 2020.

J. Weedon, W. Nuland and A. Stamos, 2017. "Information operations and Facebook," version 1.0 (27 April), at https://www.mm.dk/wp-content/uploads/2017/05/facebook-and-information-operations-v1.pdf, accessed 15 August 2020.

S. Woolley and P.N Howard (editors), 2019. *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford: Oxford University Press. doi: https://doi.org/10.1093/oso/9780190931407.001.0001, accessed 15 August 2020.

S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, 2019. "Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the Web," *arXiv*: 1801.09288 (4 March), at https://arxiv.org/abs/1801.09288, accessed 3 November 2018.

S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini and J. Blackburn, 2018. "Who let the trolls out? Towards understanding state-sponsored trolls," *arXiv*: 1811.03130 (7 November), https://arxiv.org/abs/1811.03130, accessed 15 August 2020.

A. Zelenkauskaite and B. Niezgoda, 2017. "'Stop Kremlin trolls': Ideological trolling as calling out, rebuttal, and reactions on online news portal commenting," *First Monday*, volume 22, number 5, at https://firstmonday.org/article/view/7795/6225, accessed 15 August 2020.
doi: https://doi.org/10.5210/fm.v22i5.7795, accessed 15 August 2020.

J. Zhang, D. Carpenter and M. Ko, 2013. "Online astroturfing: A theoretical perspective," *Proceedings of the Nineteenth Americas Conference on Information Systems*, at https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1620&context=amcis2013, accessed 15 August 2020.

Appendices

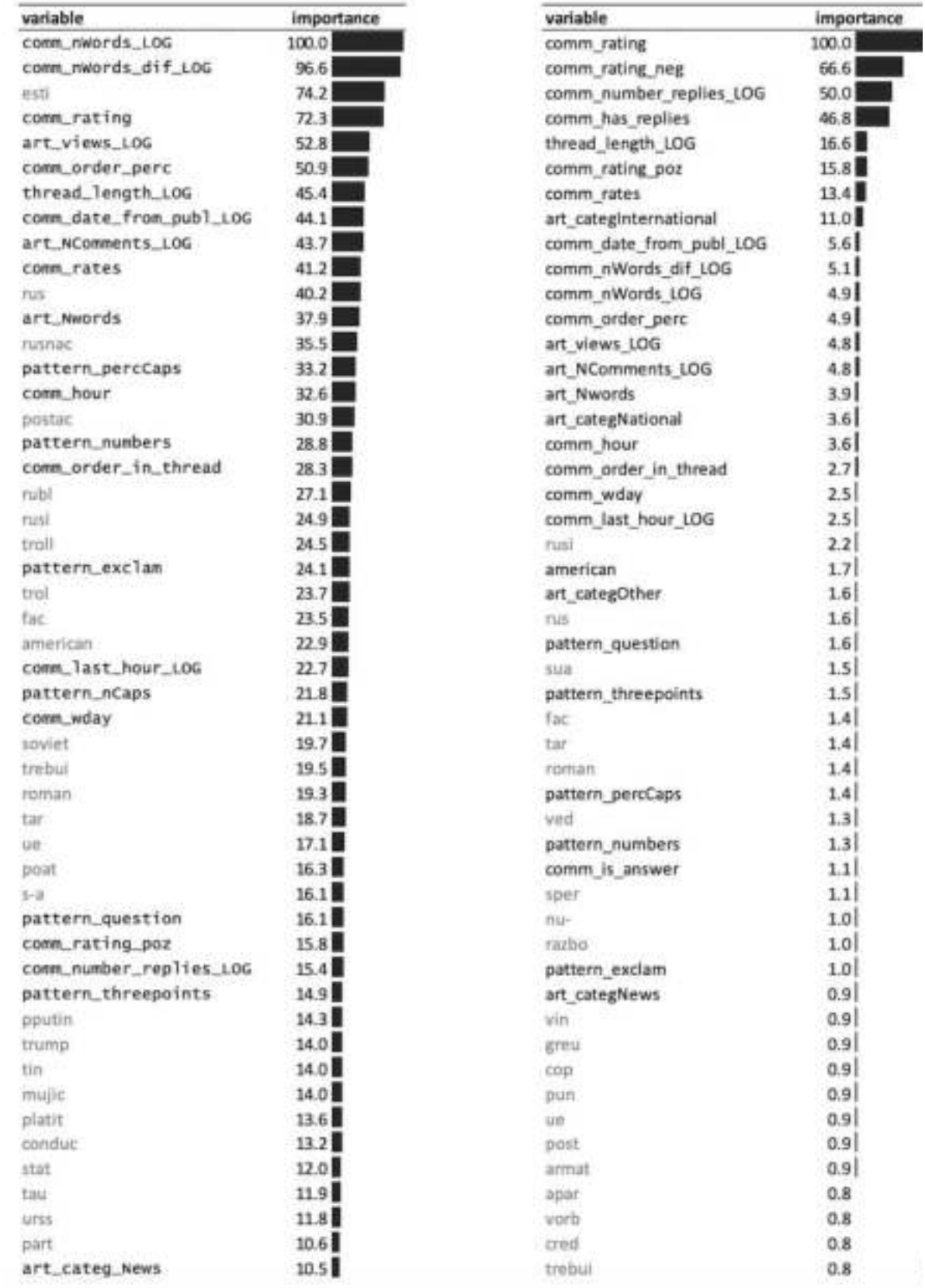
Appendix 1: Dictionary to generate keywords for manual classification of informal flags.

Dictionary object with 4 key entries.
— [russia]:
  — rus, rusia, russia*, rusesc, rusa, rusil, rusie*, ruso*, rusna*, filorus*, prorus*
— [ussr]:
  — urss, soviet*, bolsevi*, bolshevi*
— [putin]:
  — putin
— [other]:
  — rubl*, mujik*, mujic*

Appendix 2: Descriptives of metadata collected from each comment.						
	n	**min**	**max**	**median**	**mean**	**std.dev**
Article level						
number of words in article	209742	13	150	97	97.68	11.04
number of views	209742	125	265251	8235	14987	20303.27
number of comments	209742	0	460	28	45.35	52.73
number of pictures in article	209742	0	5	0	0.21	0.86
Comment level — length						
number of words in comment	209742	0	525	32	48.1	49.32
number of words minus average in article	209742	-176.71	471.82	-9.29	-0.01	45.39
Comment level — position						
comment order in comment list (percent)	209742	0	1	0.44	0.44	0.29
comment order in comment list (number)	209742	0	200	10	21.03	29.8

comment order in thread	209742	1	148	1	3.3	6.05
comment is reply to other comment	209742	0	1	0	0.46	0.5
number of replies the comment had	209742	0	10	0	0.61	1.23
Comment level — rating						
comment rating	209742	-100	100	3	4.79	10.79
number of raters	209742	0	269	7	11.2	13.05
Comment level — time posted						
day of week comment was posted	209742	1	7	4	3.96	1.82
hour comment was posted	209742	0	23	15	14.33	5.24
time spent from article publication to comment	209742	0	60	2.53	6.85	10.18
number of comments posted in the last hour by same user	209742	1	23	1	2.15	1.8
Thread level						

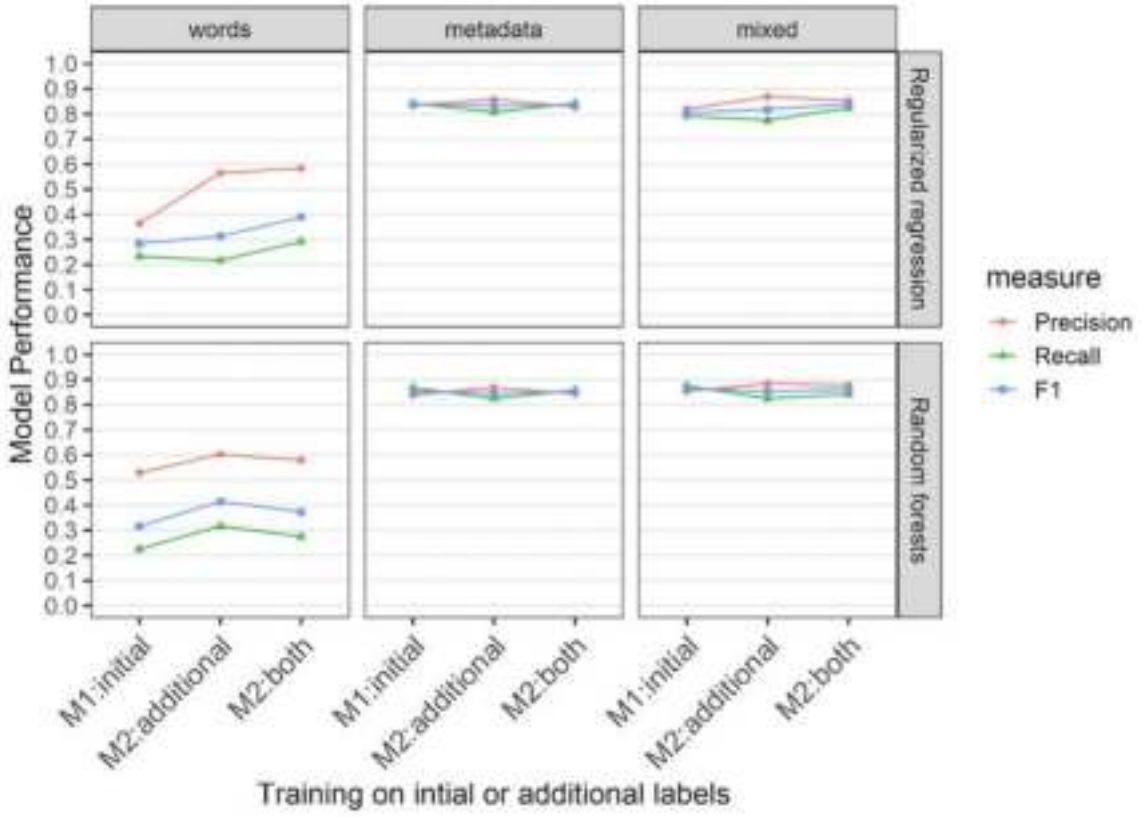
length of thread	209742	1	148	2	5.59	10.11



variable	importance
comm_nWords_LOG	100.0
comm_nWords_dif_LOG	96.6
esti	74.2
comm_rating	72.3
art_views_LOG	52.8
comm_order_perc	50.9
thread_length_LOG	45.4
comm_date_from_publ_LOG	44.1
art_NComments_LOG	43.7
comm_rates	41.2
rus	40.2
art_Nwords	37.9
rusnac	35.5
pattern_percCaps	33.2
comm_hour	32.6
postac	30.9
pattern_numbers	28.8
comm_order_in_thread	28.3
rubl	27.1
rusi	24.9
troll	24.5
pattern_exclam	24.1
trol	23.7
fac	23.5
american	22.9
comm_last_hour_LOG	22.7
pattern_nCaps	21.8
comm_wday	21.1
soviet	19.7
trebui	19.5
roman	19.3
tar	18.7
ue	17.1
poat	16.3
s-a	16.1
pattern_question	16.1
comm_rating_poz	15.8
comm_number_replies_LOG	15.4
pattern_threepoints	14.9
pputin	14.3
trump	14.0
tin	14.0
mujic	14.0
platit	13.6
conduc	13.2
stat	12.0
tau	11.9
urss	11.8
part	10.6
art_categ_News	10.5

variable	importance
comm_rating	100.0
comm_rating_neg	66.6
comm_number_replies_LOG	50.0
comm_has_replies	46.8
thread_length_LOG	16.6
comm_rating_poz	15.8
comm_rates	13.4
art_categInternational	11.0
comm_date_from_publ_LOG	5.6
comm_nWords_dif_LOG	5.1
comm_nWords_LOG	4.9
comm_order_perc	4.9
art_views_LOG	4.8
art_NComments_LOG	4.8
art_Nwords	3.9
art_categNational	3.6
comm_hour	3.6
comm_order_in_thread	2.7
comm_wday	2.5
comm_last_hour_LOG	2.5
rusi	2.2
american	1.7
art_categOther	1.6
rus	1.6
pattern_question	1.6
sua	1.5
pattern_threepoints	1.5
fac	1.4
tar	1.4
roman	1.4
pattern_percCaps	1.4
ved	1.3
pattern_numbers	1.3
comm_is_answer	1.1
sper	1.1
nu-	1.0
razbo	1.0
pattern_exclam	1.0
art_categNews	0.9
vin	0.9
greu	0.9
cop	0.9
pun	0.9
ue	0.9
post	0.9
armat	0.9
apar	0.8
vorb	0.8
cred	0.8
trebui	0.8

Appendix 3: Variable importance in random forest models (textual features in grey).



Appendix 4: Classification diagnostics for models predicting messages flagged as Russian propaganda on test set from April–October 2019, for models trained with different methods, feature sets and labels.

Editorial history

Received 14 April 2020; revised 11 August 2020; accepted 11 August 2020.

Feeding the troll detection algorithm: Informal flags used as labels in classification models to identify perceived computational propaganda
by Vlad Achimescu and Dan Sultănescu.
First Monday, Volume 25, Number 9 - 7 September 2020
https://firstmonday.org/ojs/index.php/fm/article/download/10604/9724
doi: http://dx.doi.org/10.5210/fm.v25i9.10604