Towards Measuring Content Coordination in Microblogs

Dmitri Roussinov[1]

[1] University of Strathclyde, 16 Richmond Street, Glasgow G11XQ, United Kingdom
dmitri.roussinov@strath.ac.uk

Abstract. The value of microblogging services (such as Twitter) and social networks (such as Facebook) in disseminating and discussing important events is currently under serious threat from automated or human contributors employed to distort information. While detecting coordinated attacks by their behaviour (e.g. different accounts posting the same images or links, fake profiles, etc.) has been already explored, here we look at detecting coordination in the content (words, phrases, sentences). We are proposing a metric capable of capturing the differences between organic and coordinated posts, which is based on the estimated probability of coincidentally repeating a word sequence. Our simulation results support our conjecture that only when the metric takes the context and the properties of the repeated sequence into consideration, it is capable of separating organic and coordinated content. We also demonstrate how those context-specific adjustments can be obtained using existing resources.

Keywords: Language Models, Simulating Text, Online Bots And Trolls.

1 INTRODUCTION

Recent media reports discovered massive efforts by various political groups worldwide to over-represent their support by employing automated or paid-human contributors [3]. For example, the 2016 US presidential election witnessed use of automated bots on both sides, with 5:1 ratio for the winner [7]. The Islamic State of Iraq and the Levant (ISIL or ISIS) has been noted to use coordinated bots [11]. Twenty (*20*) percent of all the internet comments in China are believed to be made by paid pro-government trolls [12]. Russian government spends millions of dollars every year on similar activities [13].

As we further elaborate in our "Related Work" section, several methods to detect coordination in microblogging activities have been proposed. However, they are so far based only on troll's behaviour and profile characteristics. Meanwhile, several studies noted the occurrence of identical word sequences that can potentially serve as tell-tales of ongoing coordination in the content, for example the use of the same *6* word sequence ("Ukrainians killed him out of jealousy") in Twitter right after a Russian opposition leader's assassination [5] or *14*-word sequence ("How Chris Coons budget works- uses tax $ 2 attend dinners and fashion shows") to smear a US democratic senator Chris Coons [10]. While repeating those sequences indeed looks suspicious, we still don't know *what are the properties (e.g. the minimum length, rarity of the words*

used, number of repetitions, etc.) of the repeated sequence to be suspicious since repetitions happen in organic (not-coordinated) communication as well. For example, several tweets wrote "Earthquake hits central Alaska", when such event indeed occurred on May 7th, 2017. Is this suspicious? While we do not claim to provide complete answers to those questions here, we still make several important steps towards it by providing a framework for future work. Our contributions are the following: 1) we propose to model the classes of repetitions rather than individual suspicious sequences. 2) By using simulation and counter-examples, we demonstrate that without taking the context of the post (topic) into consideration, repetitions from organic communication may look unjustifiably suspicious (false positives). 3) We propose the necessary context adjustments that allow separating organic coincidences from coordinated ones.

The next section presents the related work, followed by the description of our framework. The "Conclusions…" section summarizes our findings and possible future directions.

2 RELATED WORK

Distinguishing automated from human accounts has been successfully tackled by several research projects, e.g. [1][2][4][10], which offered various successful machine learning detection methods that are based on the *behaviour* of the coordinated bots (trolls, users, actors, etc.), such as posting the same links or same digital object or using fake profiles or being somehow associated with other, already detected trolling accounts. However, the behaviour-only methods would not solve the problem for the following reason: as those methods become known through their publications, the bots' coordinators will simply modify their behaviour to avoid being caught. As an alternative, here we focus on detecting coordination in the content, since it is intrinsically inseparable from the bots' purpose: to amplify a certain message by artificially inflating the presence of certain content. A closely related problem of *plagiarism detection* has been receiving significant scrutiny resulting in a number of useful tools [9], however they are not known to involve quantitative estimates, but rather treat any repetitions as suspicious. Also, potentially relevant to the task here are the algorithms on review-spam detection, e.g. [6], but it is still a different task since the review spamming accounts are typically short lived, the reviews themselves are much longer than the microblog posts, and the contributors are not connected into a network.

3 SIMULATING REPETITIONS IN TEXT

The task we are trying to solve here is formally the following: given a "suspicious" repeating sequence of words (*n*-gram), estimate the probability of occurrence of this sequence more than once in organic (not coordinated) set of short documents (tweets, posts, etc.). If this probability is very low (e.g. *<0.0001*), then we may claim with a high certainty that coordination is taking place. The suspect string is typically identified by a manual investigation [5] or by automatically applying a clustering algorithm [2][1].

Once the suspect sequence is identified, we can try to estimate the probability of generating it by applying a language model [8], e.g. using Microsoft's n-grams service (https://azure.microsoft.com/en-us/services/cognitive-services/), trained on the portion of WWW indexed by their search engine (Bing). For the suspicious sentence from [5] it gives us $4.8 \cdot 10^{-16}$, thus accidentally repeating it in entire Twitter even once is highly unlikely. But how can we generalize from this to the other suspect sequences? For better generalization, we suggest to model the classes of repetitions rather than specific sequences, so we can distinguish between types of repetitions that are suspects and those that do happen in organic posts. Thus, we define a *repetition class C(n, p)* as a repeated sub-sequence of *n* content-bearing words (n-gram), with *p* equal to their maximum probability of occurrence. We would intuitively expect the classes with large *n* and small *p* to be suspect, while the repetitions in the classes with small *n* and large *p* to be quite common. Ignoring the stopwords is justifiable since their use is determined by grammatical relationships between the content-bearing words around them. Based on the same Bing's n-grams statistics, the repetition class for the example sequence above will be $C(3, 3.4 \cdot 10^{-5})$, where $3.4 \cdot 10^{-5}$ is the probability of occurrence of the word *killed* as the most frequent out of the three context-bearing words (*ukranians, killed, jealousy*). Table 1 lists some of example sequences from real trolling attacks reported in the prior works and from organic communication, along with the parameters defining their repetition classes.

Table 1. Examples of repeated sequences from trolling attacks and organic communication along with their repetition class parameters.

	n	**p**
Coordinated:		
Ukrainians killed him out of jealousy	3	$3.4 \cdot 10^{-5}$
Ukrainians killed him out of jealousy. He stole a girlfriend from one of them.	5	$3.6 \cdot 10^{-5}$
How Chris Coons budget works- uses tax $ 2 attend dinners and fashion shows	10	$1.7 \cdot 10^{-4}$
Organic:		
Earthquake hits central Alaska	4	$1.6 \cdot 10^{-4}$
16 foreigners among 39 killed in Istanbul nightclub	4	$3.5 \cdot 10^{-5}$
Spanish prosecutors have charged Catalan cabinet	5	$7.1 \cdot 10^{-5}$

To estimate the probabilities of occurrences within those repetition classes, we run a simulation by sampling *n*-grams from a uniform distribution matching the class probabilities (*p*) and the lengths (*n*). We generated *1000* "tweets" of a typical size (*10* words), and looked for repetitions using a hash table. We also obtained similar results by using the Zipf distribution with several sets of typical parameters, but omitting them here due to space limitations. We did not observe occurrences in any of the classes defined by those examples in any of *10,000* simulation runs. This suggests that our metric based on a notion of a repetition class correctly identifies the examples from coordinated attacks as suspicious (probability of happening in organic posts < *1/10000*). But the metric also erroneously identifies all the sequences from organic

communication as suspect, thus underestimating the probability of repetition. In reality, the tweets are not random utterances as they are typically posted about certain events. Thus, their word distributions are strongly affected by the topic. For example, according to a search run over an indexed copy of Wikipedia, the probability of the word *jealousy* increases almost *200* fold when the document already has the word *killed*.

Table 2. Examples of repeated sequences from trolling attacks and organic communication along with their repetition class parameters adjusted for the context. The last column is the number of simulation runs *S*, out of *10000*, in which any repetitions in that class occurred.

	n	**p**	**S**
Coordinated:			
Ukrainians killed him out of jealousy	3	$7.2 \cdot 10^{-3}$	10000
Ukrainians killed him out of jealousy. He stole a girlfriend from one of them.	5	$7.2 \cdot 10^{-3}$	0
How Chris Coons budget works- uses tax $ 2 attend dinners and fashion shows	10	$1.23 \cdot 10^{-3}$	0
Organic:			
Earthquake hits central Alaska	4	$1.2 \cdot 10^{-2}$	3911
16 foreigners among 39 killed in Istanbul nightclub	4	$2.3 \cdot 10^{-2}$	5677
Spanish prosecutors have charged Catalan cabinet	5	$1.0 \cdot 10^{-2}$	38

This observation can be quantified by introducing the probability adjustment factors $a(w|T)$ for each word w estimated as the ratio of the probability of occurrence within a particular topic T to the probability of occurrence in the corpus (regardless of a topic):

$$a(w|T) = \frac{p(w|T)}{p(w)}, \quad (1)$$

where T is a topic defined by a boolean query (e.g. "*assassination AND russia*" here). The probability of a word occurrence conditional on the topic T is estimated as $p(w|T) = \frac{\#(w\ AND\ T)}{\#(T)}$, where $\#(q)$ is the number of documents in the corpus matching the query q. The probability of a document having the word w is estimated as $p(w) = \frac{\#(w)}{W}$, where W is the total number of documents in the corpus, or can be obtained from Bing's n-grams. Alternatively, for the words closely related to the topic, the adjustments can be estimated empirically by running the search queries defining the topic T (or the related hashtags) in Twitter and counting the occurrences of those words in the returned tweets. For the words defining the topic itself (e.g. *alaska*, *earthquake*), those probabilities typically range between *0.05* and *0.1*. Table 2 shows the same examples of repetitions with their classes adjusted for the context. The last column (S) shows the number of runs in which any repetitions within that class occurred out of all

10,000 simulation runs. The following can be observed: 1) The repetitions classes corresponding to the examples from the organic posts do indeed happen, and, as a result, those repetitions will not be flagged as suspicious. 2) The classes of repetitions corresponding to the first sentence from the coordinated attack ("Ukrainians…") also happen, and, thus this sentence alone may not serve as sufficient evidence of coordination contrary to the investigators' in [5] claim. 3) Only when combined with the sequence immediately following it in the posts under investigation (next line in the table), the entire sequence belongs to the classes of repetitions that do not happen in organic posts. 4) The sentence about Chris Coons falls into a class of repetitions which signals a coordinated attack.

Table 3 presents additional simulation runs for various repetition classes. It is possible to observe the following: 1) Repetitions with $n=3$ (repeating a sequence with *3* content bearing words) are normally not suspicious (happen in organic communication) unless all those words are rare ($p < 5 \cdot 10^{-5}$, which generally means not among *10000* most frequent words) 2) For $n=4$, any repetition is suspect (does not happen in organic communication) unless it includes the words highly associated with the topic ($p > .01$, which is often the case with the words defining the topic itself, e.g. *earthquake* and *alaska* here. 3) Repeating a sequence of *5* or more content bearing words is always suspicious, regardless of the magnitudes of the context adjustments. While the occurrence estimates we obtained using Zipf distribution are somewhat smaller, they support the same observations.

Table 3. Simulation results for various repetition classes: Number of trials out of 10000 in which repetitions happen.

n-gram length:	**n=3**	**n=4**	**n=5**	**n=6**	**n=7**
p:					
.00005	0	0	0	0	0
.0001	3	0	0	0	0
.0003	31	0	0	0	0
.0005	54	0	0	0	0
.001	126	0	0	0	0
.003	7861	27	0	0	0
.005	10000	47	4	0	0
.01	10000	3911	38	0	0

4 CONCLUSIONS AND FUTURE WORK

Our estimates and numeric simulations here demonstrate that it is possible to quantify coordination in the content, which potentially leads to exposing unwelcome activities in the microblogging posts (e.g. Twitter), and, thus, reducing the damage that it inflicts. We have proposed a metric that is based on modeling repetitions within a class rather than trying to model repeating individual sequences. This study also suggests that context-specific adjustments are necessary and demonstrates how they can be obtained based on a training corpus. We have illustrated this on several examples from past

works and selected real microblog posts, leaving room for future more powerful approaches, such as those based on machine learning models and more formal evaluation.

Acknowledgements: I would like to thank N. Puchnina for providing a number of valuable suggestions.

References

1. Beutel, A., Xu, W., Guruswami, V., Palow, C., Faloutsos, C.: Copy-Catch: Stopping Group Attacks By Spotting Lockstep Behavior In Social Networks. In.: 22nd International Conference on World Wide Web, pp. 119–130 (2013).
2. Cao, Q., Yang, X., Yu, J., Palow, C.: Uncovering large groups of active malicious accounts in online social networks. In.: ACM SIGSAC Conference on Computer and Communications Security. ACM, pp. 477–488 (2014).
3. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The Rise of Social Bots. Communications of the ACM, 59(7), 96-104 (2016).
4. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Detection of Promoted Social Media Campaigns. In.: Tenth International AAAI Conference on Web and Social Media (2016).
5. Global Voices: Social Network Analysis Reveals Full Scale of Kremlin's Twitter Bot Campaign. Global Voices, 02/04/2015 (2015).
6. Jindal, N., Liu, B.: Review spam detection. In.: WWW Conference 2007, pp. 1189–1190 (2007).
7. Howard, P.N., Kollanyi, B. and Woolley, S.: Bots and Automation over Twitter during the US Election. Zugriff am, 14 (2016).
8. Ponte, J.M., Croft, W.B.: A Language Modelling Approach to Information Retrieval. In.: Research and Development in Information Retrieval. pp. 275–281 (1998).
9. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection, Notebook Papers of CLEF LABs and Workshops (2011).
10. Ratkiewicz, J., Conover, M.D., Meiss, M., Gonçalves, B., Flammini, A., Menczer, F.: Detecting and Tracking Political Abuse in Social Media. In.: ICWSM (2011).
11. Shane, S., Hubbard, B.: ISIS displaying a deft command of varied media. New York Times (2014).
12. Simon, J.: The New Censorship: Inside the global battle for media freedom. New York: Columbia University Press (2015).
13. Sindelar, D.: The Kremlin's Troll Army. The Atlantic. August 12 (2014).