# Using Behavior and Text Analysis to Detect Propagandists and Misinformers on Twitter

Michael Orlov[1] and Marina Litvak[2]

[1] NoExec
Israel
orlovm@noexec.org
[2] Shamoon College of Engineering
Beer Sheva, Israel
marinal@ac.sce.ac.il

**Abstract.** There are organized groups that disseminate similar messages in online forums and social media; they respond to real-time events or as persistent policy, and operate with state-level or organizational funding. Identifying these groups is of vital importance for preventing distribution of sponsored propaganda and misinformation. This paper presents an unsupervised approach using behavioral and text analysis of users and messages to identify groups of users who abuse the Twitter micro-blogging service to disseminate propaganda and misinformation. Groups of users who frequently post strikingly similar content at different times are identified through repeated clustering and frequent itemset mining, with the lack of credibility of their content validated through human assessment. This paper introduces a case study into automatic identification of propagandists and misinformers in social media.

**Keywords:** Propaganda · Misinformation · Social networks.

## 1 Introduction

The ever-growing popularity of social networks influences everyday life, causing us to rely on other people's opinions when making large and small decisions, from the purchase of new products online to voting for a new government. It is not surprising that by spreading disinformation and misinformation social media became a weapon of choice for manipulating public opinion. Fake content and propaganda are rampant on social media and must be detected and filtered out. The problem of information validity in social media has gained significant traction in recent years, culminating in large-scale efforts by the research community to deal with "fake news" [7], clickbait [6], "fake reviews" [2], rumors [8], and other kinds of misinformation.

We are confident that detecting and blocking users who disseminate misinformation and propaganda is a much more effective way of dealing with fake content, as it enables prevention of its massive and consistent distribution in social media. Therefore, in this paper we deal with detection of propagandists.

We define propagandists as groups of people who intentionally spread misinformation or biased statements, typically receiving payment for this task, similarly to the definition of "fake reviews" disseminators in [2]. An article in the Russian-language Meduza media outlet [12] describes one example of paid propagandists performing their task on a social network[3] while neglecting to delete the task description and requirements, as illustrated in Figure 1.

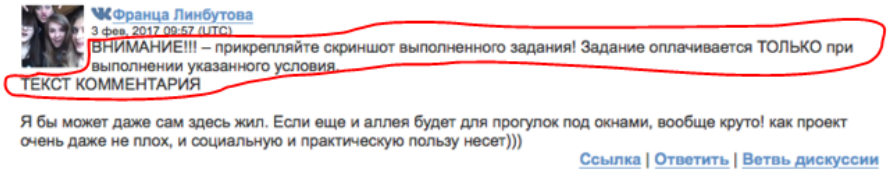


**Fig. 1.** A comment on VK social network that includes paid propaganda task description. The marked section translates as: *"ATTENTION!!! — attach a screenshot of the task performed! The task is paid ONLY when this condition is fulfilled. TEXT OF COMMENT."* The rest of the post promotes a municipal project.

Twitter is one of the most popular platforms for dissemination of information. We would expect Twitter to attract focused attention of propagandists — organized groups who disseminate similar messages in online forums and social media, in response to real-time events or as a persistent policy, operating with state-level or organizational funding. We explore, below, an unsupervised approach to identifying groups of users who abuse the Twitter micro-blogging service to disseminate propaganda and misinformation. This task is accomplished via behavioral analysis of users and text analysis of their content. Users who frequently post strikingly similar content at different times are identified through repeated clustering, and their groups are subsequently identified via frequent itemset mining. The lack of credibility of their content is validated manually. The most influential disseminators are detected by calculating their PageRank centrality in the social network and the results are visualized. Our purpose is to present a case study into automatic identification of propagandists in social media.

## 2 Related Work

The subject of credibility of information propagated on Twitter has been previously analyzed. Castillo et al. [5] observed that while most messages posted on Twitter are truthful, the service also facilitates spreading misinformation and false rumors. Dissemination of false rumors under critical circumstances was analyzed in [14], and the aggregation analysis on tweets was performed in order to

[3] VK is a social network popular in Russia, see https://vk.com.

differentiate between false rumors and confirmed news. Discussion about detecting rumors and misinformation in social networks remains very popular nowadays. Authors of [3] demonstrate the importance of social media for fake news suppliers by measuring the source of their web traffic. Hamidian and Diab [8] performed supervised rumors classification using the tweet latent vector feature. Large-scale datasets for rumor detection were built in [17] and [21].

However, not much attention has been paid to detection of *propagandists* in social media. Some works used the term *propaganda* in relation to spammers [13]. Metaxas [15]associated the theory of propaganda with the behavior of web spammers and applied social anti-propagandistic techniques to recognize trust graphs on the web. Lumezanu et al. [11] studied the tweeting behavior of assumed Twitter propagandists and identified tweeting patterns that characterize them as users who consistently express the same opinion or ideology. The first attempt to automatically detect propaganda on Twitter was made in [20], where linguistically-infused predictive models were built to classify news posts as suspicious or verified, and then to predict four subtypes of suspicious news, including propaganda.

In this paper, we address the problem of *automatically identifying paid propagandists*, who have an agenda, but do not necessarily spread false rumors, or even false information. This problem is principally different from what had been stated in other papers, classifying propaganda as rumor or equating it with spam, which is a much wider concept. Our approach is very intuitive and unsupervised.

## 3 Methodology

When using Twitter as an information source, we would like to detect tweets that contain propaganda[4], and users who disseminate it. We assume that propaganda is disseminated by professionals who are centrally managed and who have the following characteristics (partly supported by [11]): (1) They work in groups; (2) Disseminators from the same group write very similar (or even identical) posts within a short timeframe; (3) Each disseminator writes very frequently (within short intervals between posts and/or replies); (4) One disseminator may have multiple accounts; as such, a group of accounts with strikingly similar content may represent the same person; (5) We assume that propaganda posts are primarily political; (6) The content of tweets from one particular disseminator may vary according to the subject of an "assignment," and, as such, each subject is discussed in disseminator's accounts during some *temporal frame* of its relevance; (7) Propaganda carries content similar to an official governance "vision" depicted in mass media.

Based on the foregoing assumptions, we propose to perform the following analysis for detection of propagandists:

[4] Propaganda is defined as: *"posts that contain information, especially of a biased or misleading nature, that is used to promote or publicize a particular political cause or point of view"* (Oxford English Dictionary, 3rd Online Edition).

– Based on (1) and (2), given a time dimension, repeatedly cluster tweets posted during the same time interval (timeframe), based on their content. For each run, a group of users who posted similar posts (clustered together) can be obtained. Given $N$ runs for $N$ timeframes, we can obtain a group of users who consistently write similar content — these are users whose tweets were clustered together in most of the runs. The retrieved users can be considered good suspects for propaganda dissemination.

– Based on (3), the timeframes must be small, and clustering must be performed quite frequently.

– Based on (4), we do not distinguish between different individuals. Our purpose is to detect a set of accounts, where each individual (propagandist) can be represented by a single account or by a set of accounts.

– Based on (5), we can verify the final results of our analysis and see whether the posts published from the detected accounts indeed contain political content.

– Based on (6) and (7), we collect data that belongs to content that is discussed in mass media.

We outline, below, the main algorithm steps for the proposed methodology.

1. *Filtering and pre-processing tweets.* We consider only tweets in English and perform standard preprocessing using tokenization, stopword removal, and stemming. We also filter out numbers, non-textual content (like emoji symbols), and links.

2. *Split data set into timeframes.* We split the data set into $N$ timeframes, so that each split contains tweets posted at the same period of time (between two consecutive timeframes $n_i$ and $n_{i+1}$). The timeframes must be relatively short, according to assumption (3).

3. *Cluster tweets at each timeframe.* We cluster tweets at each timeframe $n_i$ in order to find a group of users who posted similar content (clustered together). K-means has been chosen as the unsupervised clustering method, using the elbow method to determine the optimal number of clusters. The simple vector space model [18] with adapted tf-idf weights[5] was used for tweets representation. We denote the clustering results (set of clusters) for timeframe $n_i$ by $C_i = \{c_{i_1}, c_{i_2}, \ldots, c_{i_k}\}$. The final clusters are composed of user IDs (after replacing tweet IDs by IDs of users who posted them), therefore the clusters are not disjointed.

4. *Calculate groups of users*[6] *frequently clustered together.* We scan the obtained clusters and, using adapted version of the AprioriTID algorithm [1, 10], compute groups of users whose posts were frequently clustered together. We start from generating a list $L_1$ from all single users $u_i$ appearing in at least $T$ (the minimum threshold specified by the user) timeframes. Then, we generate a list of pairs $L_2 = \{\langle u_l, u_m, i\rangle, u_l \in L_1, u_m \in L_1\}$ of users that are clustered together in at least $T$ timeframes. According to the Apriori algorithm, we then join pairs from $L_2$ in order to obtain $L_3$ and so forth. This step is necessary if we want to detect organized groups of propaganda disseminators.

[5] A tweet was considered as a document, and collection of all tweets as a corpus.

[6] By "user" we mean account and not individual, based on assumption (4).

5. *Identifying the most influential disseminators with PageRank centrality.* We construct an undirected graph, with nodes standing for users. We add an edge between two users if they have been clustered together at least once (in one timeframe). The weights on edges are proportional[7] to the number of times they were clustered together. As an option, edges having weights below the specified threshold $t$ can be removed from the graph. We calculate PageRank centrality on the resultant graph and keep the obtained scores for detected accounts as a disseminator's "influency" measure, as illustrated in Figure 2. Using an eigenvalue centrality metric for measuring influence in graph structure of a social network considers its "recursive" nature. For example, in [2] HITS algorithm [9] is adapted for computing the honesty of users and goodness of products.

6. *Visualize the "dissemination" network structure and analyze results.* We visualize the graph obtained in the prior step, where the PageRank centrality for each node affects its size. We also apply topic modeling in order to visualize main topics in the content that was detected as propaganda.

| Row ID | S Node | D PR |
|---|---|---|
| 0 | Col_Connaughton | 2.895 |
| 32 | syrializer | 2.277 |
| 3 | mrsn_34 | 2.087 |
| 26 | AmericanSyrians | 1.983 |
| 16 | AgendaOfEvil | 1.873 |
| 19 | awesomeseminars | 1.826 |
| 12 | Shababeeksouria | 1.794 |
| 28 | VitalAnon | 1.559 |
| 9 | mooredavid1970 | 0.967 |
| 17 | ferozwala | 0.917 |
| 23 | [illegible] | 0.905 |

**Fig. 2.** Partial example of a list of PageRank centrality values that were computed for the disseminators graph in step 5 above.

The algorithm's flow is shown in Figure 3.

## 4 Case Study

**Dataset**. Military airstrikes in Syria in September 2017 attracted worldwide criticism. Reflection of these events in Twitter can be tracked using the keyword *#syria*, determined via Hamilton 68 [19] as the most popular hashtag for 600 monitored Twitter accounts that were linked to Russian influence operations. Our case study was carried out on a dataset obtained from Twitter, collected using the Twitter Stream API with the *#syria* hashtag. The dataset covers 10,848 tweets posted by 3,847 users throughout September 9–12, 2017.

[7] Edge weights are normalized to be in range of $[0, 1]$.

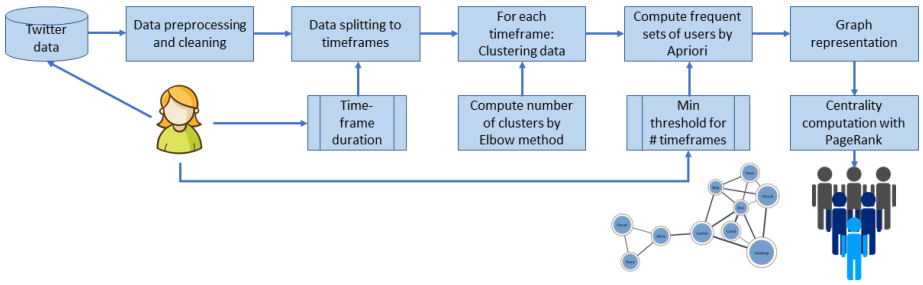


**Fig. 3.** Pipeline for detecting users who consistently post similar content.

**Parameters/Settings**. We performed clustering with 10 ($K = 10$) clusters, as an optimal clusters number according to the elbow method, 8 ($N = 8$) times (every 12 hours, according to our assumption that organized propagandists work regular hours), and looked for a group of accounts that consistently (all timeframes without exceptions, with $T = 100\%$) post similar content.

**Tools**. We have implemented the above-described process in KNIME, a data analytics, reporting, and integration platform [4].

**Results**. Our algorithm detected seven suspicious accounts. The content of messages posted by these accounts confirmed our suspicions of organized propaganda dissemination. Speaking formally, we manually approved 100% of precision. However, the recall was not measured due to the absence of manual annotation for all accounts in our data.

Topic modeling[8] results confirmed that most topics in the detected posts aligned well with political propaganda vocabulary. For example, the top topic words *attack, russia, report, isis, force, bomb, military* represent Russia's military operations in Syria, and *trump, attack, chemical, false, flag, weapons* represent an insinuated American undercover involvement in the area.

Activity analysis of the detected accounts confirmed assumption (3) about propagandists posting significantly more frequently than regular Twitter users. While regular users had 12.8 hours mean time between posts, propagandists featured 1.8 hours mean time. This assumption has been also confirmed by empirical analysis in [20].

## 5 Conclusions

This paper introduces initial stages in our research related to automatic detection of propagandists, based on analysis of users' behavior and messages. We propose an intuitive unsupervised approach for detecting Twitter accounts that disseminate propaganda. We intend to continue this research in several directions: (a) Extend our experiments with respect to other (baseline) methods, commercial domains, and various (standard) IR evaluation metrics; (b) Evaluate

[8] Topic modeling was performed using KNIME's LDA implementation.

in depth the contribution of each separate stage of our pipeline; (c) Incorporate additional (or alternative) techniques, like topic modeling, graph clustering, or analyzing web traffic of news sources, into our pipeline; (d) Adapt and apply our approach to tweets written in different languages, with focus on Russian, due to high popularity of Twitter among organized dissemination groups [16]; (e) Combine the proposed approach with authorship analysis to detect actual users that might use several accounts, according to assumption (4); (f) Perform geolocation prediction and analysis on the detected accounts to provide additional important information related to geographical distribution of organized propaganda dissemination activity; (g) Perform supervised classification of detected tweets for more accurate analysis; (h) Incorporate retweeting statistics into our network centrality analysis (step 6) to detect the most influential disseminators.

Our approach can be of great assistance in collecting a high quality dataset of propaganda and its disseminators, which then can be used for training supervised predictive models and for automatic evaluations. An automatic evaluation of our approach can be performed via verification of automatically detected accounts with accounts identified by public annotation tools, such as PropOrNot[9].

## References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I., et al.: Fast discovery of association rules. Advances in knowledge discovery and data mining **12**(1), 307–328 (1996)
2. Akoglu, L., Chandy, R., Faloutsos, C.: Opinion fraud detection in online reviews by network effects. In: ICWSM 2013. pp. 2–11 (2013)
3. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives **31**(2), 211–36 (2017)
4. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: Knime: The konstanz information miner. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.) Data Analysis, Machine Learning and Applications. pp. 319–326. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
5. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on Twitter. In: Proceedings of the 20th International Conference on World Wide Web. pp. 675–684. WWW '11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/1963405.1963500
6. Chen, Y., Conroy, N.J., Rubin, V.L.: Misleading online content: Recognizing clickbait as false news. In: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. pp. 15–19. ACM (2015)
7. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology **52**(1), 1–4 (2015)
8. Hamidian, S., Diab, M.T.: Rumor identification and belief investigation on Twitter. In: WASSA NAACL-HLT. pp. 3–8 (2016)
9. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: The web as a graph: Measurements, models, and methods. In: International Computing

[9] See http://www.propornot.com.

and Combinatorics Conference. pp. 1–17. Springer (1999). https://doi.org/3-540-48686-0_1
10. Li, Z.C., He, P.L., Lei, M.: A high efficient AprioriTid algorithm for mining association rule. In: 2005 International Conference on Machine Learning and Cybernetics. vol. 3, pp. 1812–1815. IEEE (Aug 2005). https://doi.org/10.1109/ICMLC.2005.1527239
11. Lumezanu, C., Feamster, N., Klein, H.: #bias: Measuring the tweeting behavior of propagandists. In: Sixth International AAAI Conference on Weblogs and Social Media (2012), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4588
12. Meduza: Authors of paid comments in support of Moscow authorities forgot to edit assignment (2017), https://meduza.io/shapito/2017/02/03/avtory-platnyh-kommentariev-v-podderzhku-moskovskih-vlastey-zabyli-otredaktirovat-zadanie
13. Mehta, B., Hofmann, T., Fankhauser, P.: Lies and propaganda: Detecting spam users in collaborative filtering. In: Proceedings of the 12th international conference on Intelligent user interfaces. pp. 14–21. ACM (2007). https://doi.org/10.1145/1216295.1216307
14. Mendoza, M., Poblete, B., Castillo, C.: Twitter under crisis: Can we trust what we RT? In: Proceedings of the First Workshop on Social Media Analytics. pp. 71–79. SOMA '10, ACM, New York, NY, USA (2010). https://doi.org/10.1145/1964858.1964869
15. Metaxas, P.: Using propagation of distrust to find untrustworthy web neighborhoods. In: Internet and Web Applications and Services, 2009. ICIW'09. Fourth International Conference on. pp. 516–521. IEEE (2009). https://doi.org/10.1109/ICIW.2009.83
16. Paul, C., Matthews, M.: The russian "firehose of falsehood" propaganda model. RAND Corporation (2016)
17. Qazvinian, V., Rosengren, E., Radev, D.R., Mei, Q.: Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1589–1599. EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), https://www.aclweb.org/anthology/D11-1147
18. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)
19. The Alliance for Securing Democracy: Hamilton 68 (2017), https://dashboard.securingdemocracy.org
20. Volkova, S., Shaffer, K., Jang, J.Y., Hodas, N.: Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 647–653 (2017)
21. Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K., Tolmie, P.: Towards detecting rumours in social media. In: AAAI Workshop: AI for Cities (2015)