

Detection of Promoted Social Media Campaigns

Emilio Ferrara,¹ Onur Varol,² Filippo Menczer,² Alessandro Flammini²

¹Information Sciences Institute, University of Southern California. Marina del Rey, CA (USA)

²Center for Complex Networks and Systems Research, Indiana University. Bloomington, IN (USA)

Abstract

Information spreading on social media contributes to the formation of collective opinions. Millions of social media users are exposed every day to popular memes — some generated organically by grassroots activity, others sustained by advertising, information campaigns or more or less transparent coordinated efforts. While most information campaigns are benign, some may have nefarious purposes, including terrorist propaganda, political astroturf, and financial market manipulation. This poses a crucial technological challenge with deep social implications: can we detect whether the spreading of a viral meme is being sustained by a promotional campaign? Here we study trending memes that attract attention either organically, or by means of advertisement. We designed a machine learning framework capable to detect promoted campaigns and separate them from organic ones in their early stages. Using a dataset of millions of posts associated with trending Twitter hashtags, we prove that remarkably accurate early detection is possible, achieving 95% AUC score. Feature selection analysis reveals that network diffusion patterns and content cues are powerful early detection signals.

Introduction

An increasing number of people rely, at least in part, on information shared on social media (SM) to form opinions and make choices on issues related to lifestyle, politics, health, and product purchases (Bakshy et al. 2011). This reliance provides motivation for a variety of parties (corporations, governments, etc.) to promote information and influence collective opinions through active participation in online conversations. Such a participation may be characterized by opaque methods to enhance both perceived and actual popularity of promoted information. Recent examples of abuse abound and include: (i) astroturf in political campaigns, or attempts to spread fake news under the pretense of grassroots conversations (Ratkiewicz et al. 2011); (ii) orchestrated boosting of perceived consensus on relevant social issues performed by some governments (Shearlaw 2015); (iii) propaganda and recruitment by terrorist organizations like ISIS (Berger and Morgan 2015); and (iv) actions involving SM and stock market manipulation (U.S. Securities and Exchange Commission 2015).

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

There are at least three important dimensions across which information campaigns deserve investigation. The first concerns the subtle notion of trustworthiness of information being spread, which may range from verified facts (Ciampaglia et al. 2015), to rumors and exaggerated, biased, unverified or fabricated news. The second concerns the strategies employed for the propaganda: from a known brand that openly promotes its products targeting users that have shown interest for the product, to the adoption of social bots, trolls and fake or manipulated accounts that pose as humans (Ferrara et al. 2014). The third dimension relates to the (possibly concealed) entities behind the promotion efforts and the transparency of their goals. Progress in any of these dimensions requires the availability of tools to automatically identify coordinated information campaigns in SM. But discriminating such campaigns from grassroots conversations poses both theoretical and practical challenges. Even the definition of “campaign” is challenging, as it may depend on determining strategies of dissemination, dynamics of user engagement, motivations, and more.

This paper takes a first step toward the establishment of reliable computational tools for the *detection* of promoted information campaigns. We focus on viral memes and on the task of discriminating between organic and promoted content. Advertisement is a type of campaign that is easy to define formally. Future efforts will aim at extending this framework to other types of information campaign.

The challenge of identifying promoted content

It is common to observe hashtags that enjoy a sudden burst in activity volume on Twitter. Such hashtags are labeled *trending* and highlighted. Hashtags may also be exposed prominently on Twitter for a fee. Such hashtags are called *promoted* and often enjoy bursts of popularity similar to those of trending hashtags. They are listed among trending topics, even though their popularity may be due to the promotional effort. Discriminating between promoted and organically trending topics is not trivial, as Table 1 shows: promoted and organic trending hashtags have similar characteristics. They may also exhibit similar volume patterns (Fig. 1). Promoted hashtags may preexist the moment they are given such status and may have originated in an entirely grassroots fashion, therefore displaying features largely indistinguishable from those of other grassroots hashtags on the same topic.

Table 1: Summary statistics of collected data about promoted and trending topics (hashtags) on Twitter.

	Promoted		Organic	
	mean	st. dev.	mean	st. dev.
Dates	1 Jan– 31 Apr 2013		1–15 Mar 2013	
No. campaigns	75		852	
Avg. no. of tweets	2,385	6,138	3,692	9,720
Avg. no. uniq. users	2,090	5,050	2,828	8,240
Avg. retweet ratio	42%	13.8%	33%	18.6%
Avg. reply ratio	7.5%	7.8%	20%	21.8%
Avg. no. urls	0.25	0.176	0.15	0.149
Avg. no. hashtags	1.7	0.33	1.7	0.78
Avg. no. mentions	0.8	0.28	0.9	0.35
Avg. no. words	13.5	2.21	12.2	2.74

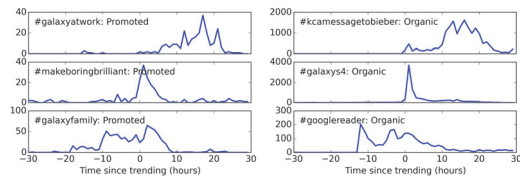


Figure 1: Time series of trending campaigns: volume (tweets/hour) relative to promoted (left) and organic (right) campaigns with similar temporal dynamics.

Data and methods

The dataset adopted in this study consists of Twitter posts. We collected in real time all trending hashtags (keywords identified by a ‘#’ prefix) relative to United States traffic that appeared from January to April 2013, labeling them as either promoted or organic, according to the information provided by Twitter itself. While Twitter allows at most one promoted hashtag per day, dozens of organic trends appear in the same period. Therefore we extracted organic trends observed during the first two weeks of March 2013 in our analysis. As a result, our dataset is highly imbalanced, with the promoted class more than ten times smaller than the organic one (cf. Table 1). Such an imbalance, however, reflects the actual data: we expect to observe a minority of engineered conversation blended in a majority of organic content. Thus, we did not balance the two classes by means of artificial resampling of the data: we studied the campaign prediction and detection problems under realistic class imbalance conditions.

For each of the campaigns, we retrieved all tweets containing the trending hashtag from an archive containing a 10% random sample of the public Twitter stream. The use of this large sample allows us to sidestep known issues of the Twitter’s streaming API (Morstatter et al. 2013). The collection period was hashtag-specific: for each hashtag we obtained all tweets produced in a four-day interval, starting two days before its trending point and extending to two days after that. This procedure provides an extensive coverage of the temporal history of each trending hashtag and its related tweets in our dataset, allowing us to study the characteristics of each campaign before, during and after the trending point.

Given that each trending campaign is described by a collection of tweets over time, we can aggregate data in sliding time windows $[t, t + \ell)$ of duration ℓ and compute features

on the subsets of tweets produced in these windows. A window can slide by time intervals of duration δ . To obtain the next window $[t + \delta, t + \ell + \delta)$, we capture new tweets produced in the interval $[t + \ell, t + \ell + \delta)$ and “forget” older tweets produced in the interval $[t, t + \delta)$. We experimented with various time window lengths and sliding parameters, and the optimal performance is often obtained with windows of duration $\ell = 6$ hours sliding by $\delta = 20$ minutes.

Features. Our framework computes features from a collection of tweets. The system generates a broad number of features (423) in the following five different classes:

(1) **Network and diffusion features.** We reconstruct three types of networks: (i) retweet, (ii) mention, and (iii) hashtag co-occurrence networks. Retweet and mention networks have users as nodes, with a directed link between a pair of users that follows the direction of information spreading — toward the user retweeting or being mentioned. Hashtag co-occurrence networks have undirected links between hashtag nodes when two hashtags occur together in a tweet. All networks are weighted according to the frequency of interactions and co-occurrences. For each network a set of features is computed, including in- and out-strength (weighted degree) distribution, density, shortest-path distribution, etc.

(2) **User account features.** We extract user-based features from the details provided by the Twitter API about the author of each tweet and the originator of each retweet. Such features include the distribution of follower and followee numbers, the number of tweets produced by the users, etc.

(3) **Timing features.** The most basic time-related feature we considered is the number of tweets produced in a given time interval. Other timing features describe the distributions of the intervals between two consecutive events, like two tweets or retweets.

(4) **Content and language features.** Our system extracts language features by applying a *Part-of-Speech* (POS) tagging technique, which identifies different types of natural language components, or *POS tags*. Phrases or tweets can be therefore analyzed to study how such POS tags are distributed. Other content features include statistics such as the length and entropy of the tweet content.

(5) **Sentiment features.** Our framework leverages several sentiment extraction techniques to generate various sentiment features, including *happiness score* (Kloumann et al. 2012), *arousal, valence and dominance scores* (Wariner, Kuperman, and Brysbaert 2013), *polarization and strength* (Wilson, Wiebe, and Hoffmann 2005), and *emotion score* (Agarwal et al. 2011).

Learning algorithms. We built upon a method called *k-nearest neighbor with dynamic time warping* (KNN-DTW) capable of dealing with multi-dimensional signal classification. *Random forests*, used as a baseline for comparison, treats each value of a time series as an independent feature.

KNN-DTW classifier. Dynamic time warping (DTW) is a method designed to measure the similarity between time series (Berndt and Clifford 1994). For classification purposes, our system calculates the similarity between two time series (for each feature) using DTW and then feeds these scores

into a k-nearest neighbor (KNN) algorithm (Cover and Hart 1967). KNN-DTW combines the ability of DTW to measure time series similarity and that of KNN to predict labels from training data. We explored various other strategies to compute the time series similarity in combination with KNN, and DTW greatly outperformed the alternatives. Unfortunately, the computation of similarity between time series using DTW is a computationally expensive task which requires $O(L^2)$ operations, where L is the length of the time series. Therefore, we propose the adoption of a *piece-wise aggregation strategy* to re-sample the original time series and reduce their length to increase efficiency with marginal classification accuracy deterioration. We split the original time series into p equally long parts and average the values in each part to represent the elements of the new time series. In our experiments, we split the L elements of a time series into $p = 5$ equal bins, and consider $k = 5$ nearest neighbors in KNN. The diagram in Fig. 2 summarizes the steps of KNN-DTW.

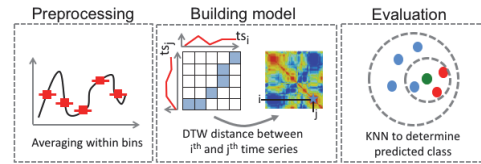


Figure 2: Diagram of KNN-DTW detection algorithm.

Baseline. We use an off-the-shelf implementation of *Random Forests* (Breiman 2001). This naive approach does not take time into consideration: each time step is considered as independent; therefore, we expect its performance to be poor. Nevertheless, it serves an illustrative purpose to underscore the complexity of the tasks, justify the more sophisticated approach outlined above, and gauge its benefit. Other traditional models we tried (SVM, SGD, Decision Trees, Naive Bayes, etc.) did not perform better.

Feature selection. Our system generates a set I of 423 features. We implemented a greedy forward feature selection method. This simple algorithm is summarized as follows: (i) initialize the set of selected features $S = \emptyset$; (ii) for each feature $i \in I - S$, consider the union set $U = S \cup i$; (iii) train the classifier using the feature set U ; (iv) test the average performance of the classifier trained on this set; (v) add to the set of selected features S the feature whose addition to S provides the best performance; (vi) repeat the feature selection process as long as a significant increase in performance is obtained. Other methods (simulated annealing, genetic algorithms) yielded inferior performance.

Results

Our framework takes multi-dimensional time series as input, which represent the longitudinal evolution of the set of features describing the campaigns. We consider a time period of four days worth of data that extends from two days before each campaign’s trending point to two days after. For all experiments, the system generated real-valued time series for each feature i represented by a feature vector \vec{f}^i consisting of 120 data points equally divided before and after the trending point. Time series are therefore encoded using the settings described above (windows of length $\ell = 6$ hours sliding every $\delta = 20$ minutes). Accuracy is evaluated by measuring the Area Under the ROC Curve (AUC) (Fawcett 2006) with 10-fold cross validation, and obtaining an average AUC score across the folds. We adopt AUC to measure accuracy because it is not biased by our class imbalance, discussed earlier (75 promoted vs. 852 organic hashtags).

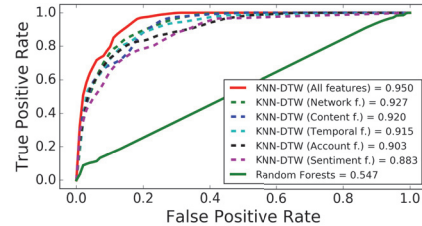


Figure 3: ROC curves showing our system’s performance. KNN-DTW obtains an AUC scores of 95% by using all features, and scores are provided for individuals feature classes as well. Random Forests scores only slightly above chance.

Method comparison. We carried out an extensive benchmark of several configurations of our system for campaign detection. The performance of the algorithms is plotted in Fig. 3. The average detection accuracy (measured by AUC) of KNN-DTW is above 95%: we deem it an exceptional result given the complexity of the task. The random forests baseline performs only slightly better than chance.

Our experiments suggest that time series encoding is a crucial ingredient for successful campaign classification. Encoding reduces the dimensionality of the signal by averaging time series. More importantly, encoding preserves (most) information about temporal trends. Accounting for temporal ordering is also critical. *Random forests* ignore long-term temporal ordering: data-points are treated as independent rather than as temporal sequences. KNN-DTW, on the other hand, computes similarities using a time series representation that preserves the long-term temporal order, even as time warping may alter short-term trends. This turns out to be a crucial advantage in the classification task.

Feature analysis. We use the greedy selection algorithm to identify the significant features, and group them by the five classes (user meta-data, content, network, sentiment, and timing) previously defined. Network structure and information diffusion patterns (AUC=92.7%), along with content features (AUC=92%), exhibit significant discriminating power and are most valuable to separate organic from promoted campaigns (*cf.* Fig. 3). Combining all feature classes boosts the system’s performance to AUC=95%.

Analysis of false negative and false positive cases. We conclude our analysis by discussing when our system fails. We are especially interested in false negatives, namely promoted campaigns that are mistakenly classified as organic. These types of mistakes are the most costly for a detection system: in the presence of class imbalance (that is, observing

significantly more organic than promoted campaigns), false positives (namely, organic campaigns mistakenly labeled as promoted) can be manually filtered out in post-processing. Otherwise, a promoted campaign mistakenly labeled as organic would easily go unchecked along with the many other correctly labeled organic ones.

Focusing our attention on a few specific instances of false negatives generated by our system, we gained some insights into the factors triggering the mistakes. First of all, it is conceivable that promoted campaigns are sustained by organic activity before promotion and therefore they are essentially indistinguishable from organic ones until the promotion triggers the trending behavior. It is also reasonable to expect a decline in performance for long delays: as more users join the conversation, promoted campaigns become harder to distinguish from organic ones.

The analysis of false positives provided us with some unforeseen insights as well. Some campaigns in our dataset, such as *#AmericanIdol* or *#MorningJoe*, were promoted via alternative communication channels (e.g., television, radio, etc.), rather than via Twitter. This has become a common practice in recent years, as more and more Twitter campaigns are mentioned or advertised externally to trigger organic-looking responses in the audience. Our system recognized such instances as promoted, whereas their ground-truth labels did not: this peculiarity distorted the evaluation to our detriment (in fact, those campaigns were unjustly counted as false positives by the AUC scores). However, we deem it remarkable that our system is capable of learning the signature of promoted campaigns irrespective of the mean(s) used for promotion (i.e., within the social media itself, or via external media promotion).

Conclusions

In this paper, we posed the problem of campaign detection and discussed the challenges it presents. We also proposed a solution based on supervised learning. Our system leverages time series representing the evolution of different features characterizing trending campaigns discussed on Twitter. These include network structure and diffusion patterns, sentiment, language and content features, timing, and user meta-data. We demonstrated the crucial advantages of encoding temporal sequences.

We achieved very high accuracy in campaign detection (AUC=95%), a remarkable feat if one considers the challenging nature of the problem and the high volume of data to analyze. One of the advantages of our framework is that of providing interpretable features and feature classes. We explored how the use of the various features affects detection performance. Feature analysis revealed that signatures of campaigns can be detected accurately, especially by leveraging information diffusion and content features.

Future research will unveil the characteristics (i.e., specific feature patterns) exhibited by promoted content and help understand how campaigns spread in SM. Further work is also needed to study whether different classes of campaigns (say, legitimate advertising vs. terrorist propaganda)

may exhibit distinct characteristics: classifying such campaigns after the trending point could be hard, due to the fact that the potential signature of artificial inception might get diluted after the campaign acquires the attention of the general audience. Therefore our final goal will be that of predicting the nature of a campaign before its trending point.

Acknowledgments. We thank M. JafariAsbagh, Q. Mei, Z. Zhao, and S. Malinchik for helpful discussions. This work was supported in part by ONR (N15A-020-0053), DARPA (W911NF-12-1-0037), NSF (CCF-1101743), and the McDonnell Foundation.

References

- Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; and Passonneau, R. 2011. Sentiment analysis of Twitter data. In *Proc. ACL Workshop on Languages in Social Media*, 30–38.
- Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone’s an influencer: quantifying influence on Twitter. In *Proc. WSDM*, 65–74.
- Berger, J., and Morgan, J. 2015. The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter. *The Brookings Project on US Relations with the Islamic World* 3:20.
- Berndt, D. J., and Clifford, J. 1994. Using dynamic time warping to find patterns in time series. In *Proc. of AAAI Workshop on Knowledge Discovery in Databases*, 359–370.
- Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.
- Ciampaglia, G. L.; Shiralkar, P.; Rocha, L. M.; Bollen, J.; Menczer, F.; and Flammini, A. 2015. Computational fact checking from knowledge networks. *PLoS ONE* 10(6):e0128193.
- Cover, T. M., and Hart, P. E. 1967. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on* 13(1):21–27.
- Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27(8):861–874.
- Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2014. The rise of social bots. *arXiv preprint arXiv:1407.5225*.
- Kloumann, I. M.; Danforth, C. M.; Harris, K. D.; Bliss, C. A.; and Dodds, P. S. 2012. Positivity of the English language. *PLoS One* 7(1):e29484.
- Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the sample good enough? Comparing data from Twitter’s streaming API with Twitter’s firehose. In *Proc. ICWSM*.
- Ratkiewicz, J.; Conover, M.; Meiss, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and tracking political abuse in social media. In *Proc. ICWSM*, 297–304.
- Shearlaw, M. 2015. From Britain to Beijing: how governments manipulate the Internet. Accessed online at <http://www.theguardian.com/world/2015/apr/02/russia-troll-factory-kremlin-cyber-army-comparisons>.
- U.S. Securities and Exchange Commission. 2015. Updated investor alert: Social media and investing — stock rumors. Accessed online at http://www.sec.gov/oiea/investor-alerts-bulletins/ia_rumors.html.
- Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 1–17.
- Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. ACL HLT/EMNLP*.