Varol et al. *EPJ Data Science* (2017) 6:13
DOI 10.1140/epjds/s13688-017-0111-y

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Early detection of promoted campaigns on social media

Onur Varol[1], Emilio Ferrara[1,2*], Filippo Menczer[1,3] and Alessandro Flammini[1,3]

*Correspondence:
emilio.ferrara@gmail.com
[1]School of Informatics and Computing, Indiana University, Bloomington, IN, USA
[2]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA
Full list of author information is available at the end of the article

Abstract

Social media expose millions of users every day to information campaigns - some emerging organically from grassroots activity, others sustained by advertising or other coordinated efforts. These campaigns contribute to the shaping of collective opinions. While most information campaigns are benign, some may be deployed for nefarious purposes, including terrorist propaganda, political astroturf, and financial market manipulation. It is therefore important to be able to detect whether a meme is being artificially promoted at the very moment it becomes wildly popular. This problem has important social implications and poses numerous technical challenges. As a first step, here we focus on discriminating between trending memes that are either organic or promoted by means of advertisement. The classification is not trivial: ads cause bursts of attention that can be easily mistaken for those of organic trends. We designed a machine learning framework to classify memes that have been labeled as trending on Twitter. After trending, we can rely on a large volume of activity data. Early detection, occurring immediately at trending time, is a more challenging problem due to the minimal volume of activity data that is available prior to trending. Our supervised learning framework exploits hundreds of time-varying features to capture changing network and diffusion patterns, content and sentiment information, timing signals, and user meta-data. We explore different methods for encoding feature time series. Using millions of tweets containing trending hashtags, we achieve 75% AUC score for early detection, increasing to above 95% after trending. We evaluate the robustness of the algorithms by introducing random temporal shifts on the trend time series. Feature selection analysis reveals that content cues provide consistently useful signals; user features are more informative for early detection, while network and timing features are more helpful once more data is available.

Keywords: social media; information campaigns; advertising; early detection

1 Introduction

An increasing number of people rely, at least in part, on information shared on social media to form opinions and make choices on issues related to lifestyle, politics, health, and products purchases [1–3]. Such reliance provides a variety of entities - from single users to corporations, interest groups, and governments - with motivation to influence collective opinions through active participation in online conversations. There are also obvious incentives for the adoption of covert methods that enhance both perceived and actual popularity of promoted information. There are abundant examples of recently reported abuse: astroturf in political campaigns, or attempts to spread fake news through

social bots under the pretense of grassroots conversations [4–6]; pervasive spreading of unsubstantiated rumors and conspiracy theories [7]; orchestrated boosting of perceived consensus on relevant social issues performed by governments [8]; propaganda and recruitment by terrorist organizations, like ISIS [9, 10]; and actions involving social media and stock market manipulation [11].

The situation is ripe with dangers as people are rarely equipped to recognize propaganda or promotional campaigns as such. It can be difficult to establish the origin of a piece of news, the reputation of its source, and the entity behind its promotion on social media, due both to the intrinsic mechanisms of sharing and to the high volume of information that competes for our attention. Even when the intentions of the promoter are benign, we easily interpret large (but possibly artificially enhanced) popularity as widespread endorsement of, or trust in, the promoted information.

There are at least three questions about information campaigns that present scientific challenges: what, how, and who. The first concerns the subtle notion of trustworthiness of information, ranging from verified facts [12], to rumors and exaggerated, biased, unverified or fabricated news [4, 7, 13]. The second considers the tools employed for the propaganda. Again, the spectrum is wide: from a known brand that openly promotes its products by targeting users who have shown interest, to the adoption of social bots, trolls and fake or manipulated accounts that pose as humans [5, 14–16]. The third question relates to the (possibly concealed) entities behind the promotion efforts and the transparency of their goals. Even before these question can be explored, one would need to be able to *identify* an information campaign in social media. But discriminating such campaigns from grassroots conversations poses both theoretical and practical challenges. Even the very definition of 'campaign' is conceptually difficult, as it entangles the nature of the content (e.g., product or news), purpose of the source (e.g., deception, recruiting), strategies of dissemination (e.g., promotion or orchestration), different dynamics of user engagement (e.g., the aforementioned social bots), and so on.

This paper takes a first step toward the development of computational methods for the *early detection* of information campaigns. In particular, we focus on trending memes and on a special case of promotion, namely advertisement, because they provide convenient operational definitions of social media campaigns. We formally define the task of discriminating between organic and promoted trending memes. Future efforts will aim at extending this framework to other types of information campaign.

1.1 The challenge of identifying promoted content

On Twitter, it is common to observe *hashtags* - keywords preceded by the # sign that identify messages about a specific topic - enjoying sudden bursts in activity volume due to intense posting by many users with an interest in the topic [17–19]. Such hashtags are labeled as *trending* and are highlighted on the Twitter platform. Twitter algorithmically identifies trending topics in a predetermined set of geographic locations. Although Twitter recently included personalized and clustered trends, the ones in the collection analyzed here correspond to single hashtags selected on the basis of their popularity. Unfortunately, detailed knowledge about the algorithm and criteria used to identify organic trends is not publicly available [20]. Other hashtags are exposed prominently after the payment of a fee by parties that have an interest in enhancing their popularity. Such hashtags are called *promoted* and often enjoy subsequent bursts of popularity similar to those of trending hashtags, therefore being listed among trending topics.

Table 1 Summary statistics of collected data about promoted and organic trends on Twitter

	Promoted		Organic	
Dates	**1 Jan-31 Apr 2013**		**1-15 Mar 2013**	
No. trends	**75**		**852**	
	Mean	**St. dev.**	**Mean**	**St. dev.**
Number of tweets	2,385	6,138	3,692	9,720
Number of unique users	2,090	5,050	2,828	8,240
Retweet ratio	42%	13.8%	33%	18.6%
Reply ratio	7.5%	7.8%	20%	21.8%
Number of URLs	0.25	0.176	0.15	0.149
Number of hashtags	1.7	0.33	1.7	0.78
Number of mentions	0.8	0.28	0.9	0.35
Number of words	13.5	2.21	12.2	2.74

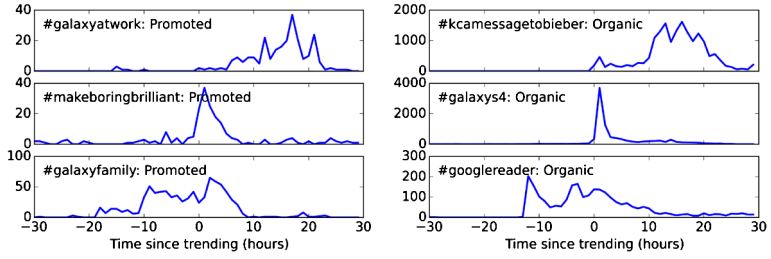

Figure 1 Time series of trending hashtags. Comparison of the time series of the volume (number of tweets per hour in our sample) relative to promoted (left) and organic (right) trends with similar temporal dynamics.

Of course, once Twitter labels a hashtag as trending, it is not necessary to detect whether or not it is promoted - this information is disclosed by Twitter. However, since it is difficult to manually annotate a sufficiently large datasets of campaigns, we use organic and promoted trending topics as a *proxy* for a broader set of campaigns, where promotion mechanisms may be hidden. Our data collection methodology provide us with a large source of reliable 'ground truth' labels about promotion, which represent an ideal testbed to evaluate detection algorithms. These algorithms have to determine whether or not a hashtag is promoted based on information that would be available even in cases where the nature of a trend is unknown. We stress that our goal of distinguishing mechanisms for promoting popular content is different from that of predicting viral topics, an interesting area of research in its own right [21–23].

Discriminating between promoted and organically trending topics is not trivial, as Table 1 illustrates - promoted and organic trending hashtags often have similar characteristics. One might assume that promoted trends display volume patterns characteristic of exogenous influence, with sudden bursts of activity, whereas organic trends would conform to more gradual volume growth patterns typical of endogenous processes [17, 24, 25]. However, Figure 1 shows that promoted and organic trends exhibit similar volume patterns over time. Furthermore, promoted hashtags may preexist the moment in which they are given the promoted status and may have originated in an entirely grassroots fashion. It is therefore conceivable for such hashtags to display features that are largely indistinguishable from those of other grassroots hashtags about the same topic, at least until the moment of promotion.

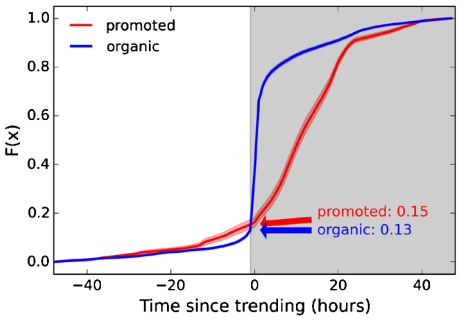

Figure 2 Cumulative fraction of tweets as a function of time. On average, only 13% of the tweets in the organic class and 15% of the tweets in the promoted class are produced prior to the trending point. The majority of tweets are observed after the trending point, with a rapid increase around trending time.

The analysis in this paper is motivated by the goal of identifying promoted campaigns at the earliest possible time. The early detection task addresses the difficulty of judging the nature of a hashtag using only the limited data available immediately before trending. Figure 2 illustrates the shortage of information available for early detection. It is also conceivable that once the promotion has triggered interest in a hashtag, the conversation is sustained by the same mechanisms that characterize organic diffusion. Such noise around popular conversations may present an added difficulty for the early detection task.

1.2 Contributions and outline

The major contribution of this paper, beyond formulating the problem of detection of campaigns in social media, is the development and validation of a supervised machine learning framework that takes into consideration the temporal sequence of messages associated with a trending hashtag on Twitter and successfully classifies it as either 'promoted' (advertised) or 'organic' (grassroots). The proposed framework adopts time-varying features built from network structure and diffusion patterns, language, content and sentiment information, timing signals, and user meta-data. In the following sections we discuss the data we collected and employed, the procedure for feature extraction and selection, the implementation of the learning framework, and the evaluation of our system.

2 Data and methods

2.1 Dataset description

The dataset adopted in this study consists of Twitter posts (*tweets*) that contain a trending hashtag and appeared during a defined observation period. Twitter provides an interface that lists trending topics, with clearly labeled *promoted* trends at the top (Figure 3). We crawled the Twitter webpage at regular intervals of 10 minutes to collect all organic and promoted hashtags trending in the United States between January and April 2013, for a total of $N = 927$ hashtags. This constitutes our ground-truth dataset of *promoted* and *organic* trends.

We extracted a sample of organic trends observed during the first two weeks of March 2013 for our analysis. While Twitter allows for at most one promoted hashtag per day, dozens of organic trends appear in the same period. 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 our expectation to observe in the Twitter stream a minority of promoted conversations blended in a majority of organic content. Therefore we did not balance the classes by resampling, to study the campaign detection problem under realistic conditions.

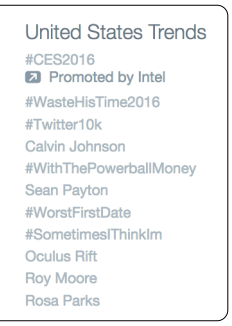

Figure 3 Screenshot of Twitter U.S. trends taken on Jan. 6, 2016. The hashtag #CES2016 was promoted on this date.

Hashtags may trend multiple times on Twitter. However, those in our collection only trended once during our observation period. For each trend, we retrieved all tweets containing the trending hashtag from an archive containing a 10% random sample of the public Twitter stream. 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 in our dataset and its related tweets, allowing us to study the characteristics of each trend before, during, and after the trending point.

Given that each trend 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. The next window therefore contains tweets produced in the interval $[t + \delta, t + \ell + \delta)$. 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.

We have made the IDs of all tweets involved in the trending hashtags analyzed in this paper available in a public dataset (carl.cs.indiana.edu/data/ovarol/trend-dataset.tar.gz).

2.2 Features

Our framework computes features from a collection of tweets in some time interval. The system generates 487 features in five different classes: network structure and information diffusion patterns, content and language, sentiment, timing, and user meta-data. The classes and types of features are reported in Table 2 and discussed next. All of the feature time series in this study are available in our public dataset.

2.2.1 Network and diffusion features

Twitter actively fosters interconnectivity. Users are linked by means of *follower/followee* relations. Content travels from person to person via *retweets*. Tweets themselves can be addressed to specific users via *mentions*. The network structure carries crucial information for the characterization of different types of communication. In fact, the usage of network features significantly helps in tasks like astroturf detection [4]. Our system reconstructs three types of networks: retweet, mention, and 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 have occurred together in a tweet. All networks are weighted

Table 2 List of 487 features extracted by our framework

Class	Feature description	No. of features
Network (†)	Number of nodes	1
	Number of edges	1
	(*) Strength distribution	8
	(*) In-strength distribution	8
	(*) Out-strength distribution	8
	(*) Distribution of number of nodes in the connected components	8
	Network density of whole and largest connected component	2
	Network assortativity of whole and largest connected component	2
	Mean shortest path length of the largest connected component	1
User	(*) Sender's follower count	8
	(*) Sender's followee count	8
	(*) Sender's number of favorite tweets	8
	(*) Sender's number of Twitter statuses posted	8
	(*) Sender's number of lists subscribed to	8
	(*) Originator's follower count	8
	(*) Originator's followee count	8
	(*) Originator's number of favorite tweets	8
	(*) Originator's number of Twitter statuses posted	8
	(*) Originator's number of lists subscribed to	8
Timing	Number of tweets appeared in a given window	1
	(*) Time between two consecutive tweets	8
	(*) Time between two consecutive retweets	8
	(*) Time between two consecutive mentions	8
Content	(*) Number of hashtags in a tweet	8
	(*) Number of mentions in a tweet	8
	(*) Number of URLs in a tweet	8
	(*,**) Frequency of POS tags in a tweet	64
	(*,**) Proportion of POS tags in a tweet	64
	(*) Number of words in a tweet	8
	(*) Entropy of words in a tweet	8
Sentiment	(***) Happiness scores of aggregated tweets	2
	(***) Valence scores of aggregated tweets	2
	(***) Arousal scores of aggregated tweets	2
	(***) Dominance scores of single tweets	2
	(*) Happiness score of single tweets	8
	(*) Valence score of single tweets	8
	(*) Arousal score of single tweets	8
	(*) Dominance score of single tweets	8
	(*) Polarization score of single tweets	8
	(*) Entropy of polarization scores of single tweets	8
	(*) Positive emoticons entropy of single tweets	8
	(*) Negative emoticons entropy of single tweets	8
	(*) Emoticons entropy of single tweets	8
	(*) Ratio between positive and negative score of single tweets	8
	(*) Number of positive emoticons in single tweets	8
	(*) Number of negative emoticons in single tweets	8
	(*) Total number of emoticons in single tweets	8
	Ratio of tweets that contain emoticons	1

† We consider three types of network: retweet, mention, and hashtag co-occurrence networks. The hashtag co-occurrence network is undirected. * Distribution types. For each distribution, the following eight statistics are computed and used as individual features: min, max, median, mean, std. deviation, skewness, kurtosis, and entropy. ** Part-of-Speech (POS) tag. There are eight POS tags: verbs, nuns, adjectives, modal auxiliaries, pre-determiners, interjections, adverbs, and pronouns. *** For each feature we compute mean and std. deviation.

according to the number 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, and so on (cf. Table 2).

2.2.2 User-based features

User meta-data is crucial to classify communication patterns in social media [5, 26]. 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, and the number of tweets produced by the users (cf. Table 2).

2.2.3 Timing features

The temporal dimension associated with the production and consumption of content may reveal important information about campaigns and their evolution [27]. 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 (cf. Table 2).

2.2.4 Content and language features

Many recent papers have demonstrated the importance of content and language features in revealing the nature of social media conversations [28–32]. For example, deceiving messages generally exhibit informal language and short sentences [33]. 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*. The following POS tags are extracted: verbs, nouns, adjectives, modal auxiliaries, predeterminers, interjections, adverbs, pronouns, and wh-pronouns (for details and examples see www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html). Tweets can be therefore analyzed to study how such POS tags are distributed. Other content features include the length and entropy of the tweet content (cf. Table 2).

2.2.5 Sentiment features

Sentiment analysis is a powerful tool to describe the attitude or mood of an online conversation. Sentiment extracted from social media conversations has been used to forecast offline events, including elections and financial market fluctuations [34, 35], and is known to affect information spreading [36, 37]. Our framework leverages several sentiment extraction techniques to generate various sentiment features, including *happiness score* [38], *arousal, valence and dominance scores* [39], *polarization and strength* [40], and *emotion score* [41] (cf. Table 2).

2.3 Feature selection

Our system generates a set I of $|I| = 487$ features (cf. Table 2) designed to extract signals from a collection of tweets and distinguish promoted trends from organic ones. Some features are more predictive than others; some are by definition correlated with each other due to temporal dependencies. Most of the correlations are related to the volume of data. For instance the two most correlated features immediately prior to the trending point are the size of the hashtag cooccurrence network and the size of its largest connected component (Pearson's $\rho = 0.75$). This is why it is important to perform feature selection

to eliminate redundant features and identify a combination of features that yield good classification performance.

There are several methods to select the most predictive features in a classification task [42]. We implemented a simple greedy forward feature selection method, 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 features in U; (iv) test the average performance of the classifier trained on this set; (v) add to S the feature that provides the best performance; (vi) repeat (ii)–(v). We terminate the feature selection procedure if the AUC (cf. Section 2.5) increases by less than 0.05 between two consecutive steps. Most of the experiments terminate after selecting fewer than 10 features. The time series for the selected features are passed as input to the learning algorithms. In the next subsections we provide details about our experimental setting and learning models.

2.4 Experimental setting

Our experimental setting follows a pipeline of feature selection, model building, and performance evaluation. We apply the *wrapper* approach to select features and evaluate performance iteratively [43]. During each iteration (Figure 4), we train and evaluate models using candidate subsets of features and expand the set of selected features using the greedy approach described in Section 2.3. Once we identify the set of features that performs best, we report results of experiments using only this set of features.

In each experiment and for each feature, an algorithm receives in input a time series with $L = 35$ data points to carry out its detection. The length of the time series and its delay D with respect to the trending point are discussed in Section 3; different experiments will consider different delays.

A set of feature time series is used to either train a learning model or evaluate its accuracy. The learning algorithms are discussed in the next subsection. For evaluation, we compute a Receiver Operating Characteristic (ROC) curve, which plots the true positive rate (TPR) versus the false positive rate (FPR) at various thresholds. Accuracy is evaluated by measuring the Area Under the ROC Curve (AUC) [44] with 10-fold cross validation, and averaging AUC scores across the folds. A random-guess classifier produces the diagonal line where TPR equals FPR, corresponding to a 50% AUC score. Classifiers with higher AUC scores perform better and the perfect classifier in this setting achieves a 100% AUC score. We adopt AUC to measure accuracy because it is not biased by the imbalance in our classes (75 promoted trends versus 852 organic ones, as discussed earlier).

2.5 Learning algorithms

Let us describe the learning systems for online campaign detection based on multidimensional time-series data from social media. We identified an algorithm, called *K-Nearest Neighbor with Dynamic Time Warping* (KNN-DTW), that is capable of dealing with multidimensional time series classification. For evaluation purposes, we compare the classification results against two baselines: SAX-VSM and KNN. These three methods are described next.

2.5.1 KNN-DTW classifier

KNN-DTW is a state-of-the-art algorithm to classify multidimensional time series, illustrated in Figure 4. During learning, we provide our model with training and testing sets generated by 10-fold cross validation. Time series for each feature are processed in paral-

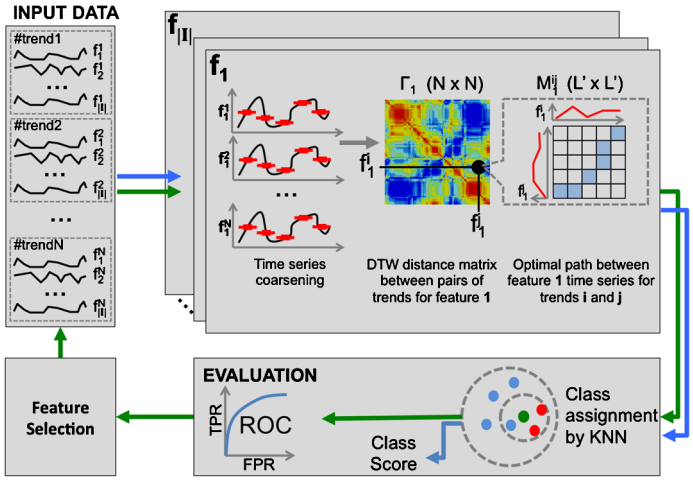

Figure 4 Wrapper method description for KNN-DTW. We present the pipeline of our complete system, including feature selection and model evaluation steps. Input data feed into the system for training (green arrow) and testing (blue arrow) steps.

lel using *dynamic time warping* (DTW), which measures the similarity between two time series after finding an optimal match between them by 'warping' the time axis [45]. This allows the method to absorb some non-linear variations in the time series, for example different speed or resolution of the data.

For efficiency, we initially apply a time series coarsening strategy called *piece-wise aggregation*. We split each original time series into p equally long sections and replace the time-series values by the section averages, reducing the dimensionality from L to $L' = L/p$. For trend i and feature k, we thus obtain a coarsened time series $f_k^i = \{f_{k,1}^i, f_{k,2}^i, \ldots, f_{k,L'}^i\}$. Then, DTW computes the distance between all pairs of points of two given trend time series f_k^i and f_k^j. Each element of the resulting $L' \times L'$ distance matrix is $M_k^{ij}(t,t') = (f_{k,t}^i - f_{k,t'}^j)^2$. Points closer to each other are more likely to be matched. To create a mapping between the two time series, an optimal path is computed over the time-series distance matrix. A path must start from the beginning of each time series and terminate at its end. The path between first and last points is then computed by minimizing the cumulative distance (γ) over alternative paths. This problem can be solved via dynamic programming [45] using the following recurrence: $\gamma(t,t') = M(t,t') + \min\{\gamma(t-1,t'-1), \gamma(t-1,t'), \gamma(t,t'-1)\}$ (indices i, j, k dropped for readability). The distance γ_k^{ij} is used as the ij-th element of the $N \times N$ trend similarity matrix Γ_k.

The computation of similarity between time series using DTW requires $O(L'^2)$ operations. Some heuristic strategies use lower-bounding techniques to reduce the computational complexity [46]. Another technique is to re-sample the data before adopting DTW. Our coarsening approach reduces the computational costs by a factor of p^2. We achieved a significant increase in efficiency with marginal classification accuracy deterioration by setting $p = 5$ ($L' = 7$).

In the evaluation step, we use the K-Nearest Neighbor (KNN) algorithm [47] to assign a class score to a test trend q. We compare q with each training trend i to obtain a DTW distance γ_k^{iq} for each feature k. We then find the $K = 5$ labeled trends with smallest DTW distance from q, and compute the fraction of promoted trends s_k^q among these nearest neighbors. We finally average across features to obtain the class score $\bar{s^q}$. Higher values

of $\bar{s^q}$ indicate a high probability that q is a promoted trend. Class scores, together with ground-truth labels, allow us to compute the AUC of a model, which is then averaged across folds according to cross validation.

2.5.2 SAX-VSM classifier

Our first baseline, called SAX-VSM, blends symbolic dimensionality reduction and vector space models [48]. Time series are encoded via Symbolic Aggregate approXimation (SAX), yielding a compact symbolic representation that has been used for time series anomaly and motif detection, time series clustering, indexing, and more [49, 50]. A symbolic representation encodes numerical features as words. A vector space model is then applied to treat time series as documents for classification purposes, similarly to what is done in information retrieval. In our implementation, we first apply piece-wise aggregation and then use SAX to represent the data points in input as a single word of L' letters from an alphabet $\aleph$. This choice and the parameters $|\aleph| = 5$ and $L' = 4$ are based on prior optimization [48], and variations to these settings only marginally affect performance. Each time-series value is mapped into a letter by dividing the range of the feature values into $|\aleph|$ regions in such a way as to obtain equiprobable intervals under an assumption of normality [50]. In the training phase, for each feature, we build two sets of words corresponding to organic and promoted trends, respectively. In the test phase, a new instance is assigned to the class with the majority of word matches across features. In case of a tie we assign a random class. For further details about this baseline and its implementation, we refer the reader to the SAX-VSM project website (github.com/jMotif/sax-vsm_classic).

2.5.3 K-nearest neighbors classifier

Our second baseline is an off-the-shelf implementation of the traditional *K-Nearest Neighbors* algorithm [47] for time-series classification. We used the Python scikit-learn package [51]. We selected KNN because it can capture and learn time-series patterns without requiring any pre-processing of the raw time-series data. We created the feature vectors for each trend by concatenating into a single vector the continuous-valued time series representing each feature. The nearest neighbor classifier computes the Euclidean distance between pairs of single-vector time series. For a test trend, the class score is given by the fraction of promoted trends among the $K = 5$ nearest neighbors.

3 Results

In this section, we present results of experiments designed to evaluate the ability of our machine learning framework to discriminate between organic and promoted trends. For all experiments, each feature time series consists of 120 real-valued data points equally divided before and after the trending point. Although in principle we could use the entire time series for classification, ex-post information would not serve our goal of early detection of social media campaigns in a streaming scenario that resembles a real setting, where information about the future evolution of a trend is obviously unavailable. For this reason, we consider only a subset consisting of L data points ending with delay D since the trending point; $D \leq 0$ for early detection, $D > 0$ for classification after trending. We evaluate the performance of our detection framework as a function of the delay parameter D. The case $D = 0$ involves detection immediately at trending time. However, we also consider $D < 0$ to examine the performance of our algorithms based on data preceding the trending point;

of course the detection would not occur until $D = 0$, when one would become aware of the trending hashtag. Time series are encoded using the settings described above ($L = 35$ windows of length $\ell = 6$ hours sliding every $\delta = 20$ minutes).

3.1 Method comparison

We carried out an extensive benchmark of several configurations of our system for campaign detection. The performance of the algorithms as a function of varying delays D is plotted in Figure 5.

In addition, we introduce random temporal shifts for each trend time series to test the robustness of the algorithms. In real-world scenarios, one would ideally expect to detect a promoted trend without knowing the trending point. To simulate such scenarios, we designed an experiment that introduces variations that randomly shift each time series around its trending point. The temporal shifts are sampled from Gaussian distributions with different variances. We present the results of this experiment in Figure 6.

KNN-DTW and KNN display the best detection accuracy (measured by AUC) in general. Their performance is comparable (Figure 5). The AUC score is on average around 95% for detecting promoted trends after trending. In the early detection task, we observe scores above 70%. This is quite remarkable given the small amount of data available before the trending point. KNN-DTW also displays a strong robustness to temporal shifts, pointing to the advantage of time warping (Figure 6). The KNN algorithm is less robust because it computes point-wise similarities between time series without any temporal alignment; as the variance of the temporal shifts increases, we observe a significant drop in accuracy. SAX-VSM benefits from the time series encoding and provides good detection performance (on average around 80% AUC) but early detection accuracy is poor, close to random for $D < 0$. A strong feature of SAX-VSM is its robustness to temporal shifts, similar to KNN-DTW.

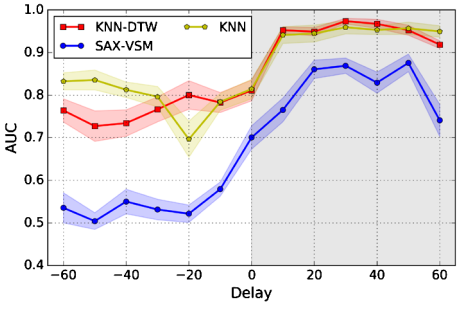

Figure 5 Methods comparison. Classification performance of different learning algorithms on encoded and raw time series. The AUC is measured for various delays D. Confidence intervals represent standard errors based on 10-fold cross validation.

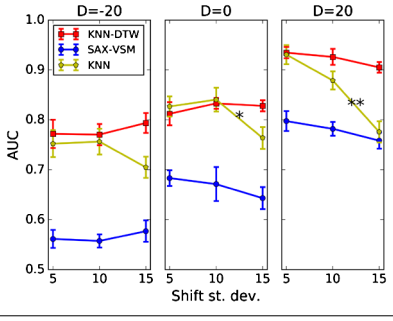

Figure 6 Temporal robustness. AUC of different learning algorithms with random temporal shifts versus the standard deviation of the shifts. We repeated the experiment for various delay values D. Significance levels of differences in consecutive experiments are marked as (*) $p < 0.05$ and (**) $p < 0.01$.

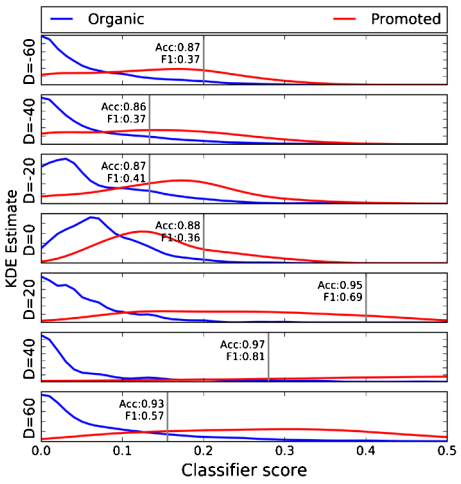

Figure 7 Distributions of KNN-DTW classifier scores. We use Kernel Density Estimation (KDE), a non-parametric smoothing method, to estimate the probability densities based on finite data samples. We also show the threshold values that separate the two classes yielding an optimal F1 score.

Our experiments suggest that temporal encoding is a crucial ingredient for successful classification of time-series data. Encoding reduces the dimensionality of the signal. More importantly, encoding preserves (most) information about temporal trends and makes an algorithm robust to random shifts, which is an importance advantage in real-world scenarios. SAX-VSM ignores long-term temporal ordering. 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 to achieve both high accuracy and robustness.

Using AUC as an evaluation metric has the advantage of not requiring discretization of scores into binary class labels. However, detection of promoted trends in real scenarios requires binary classification by a threshold. In this way we can measure accuracy, precision, recall, and identify misclassified accounts. Figure 7 illustrates the distribution of probabilistic scores produced by the KNN-DTW classifier as a function of the delay for the two classes of trends, organic and promoted. The scores are computed for leave-out test instances, across folds. An ideal classifier would separate these distributions completely, achieving perfect accuracy. Test instances in the intersection between two distributions either are misclassified or have low-confidence scores. Examples of misclassified instances are discussed in Section 3.3. For $D < 0$, KNN-DTW generates more conservative scores, and the separation between the organic and promoted class distributions is smaller. For $D > 0$, KNN-DTW scores separate the two classes well. To convert continuos scores into binary labels, we calculated the threshold values that maximize the F1 score of each experiment; this score combines precision and recall. Trends with scores above the threshold are labeled as promoted. The best accuracy and F1 score are obtained shortly after trending, at $D = 20$.

3.2 Feature analysis

Let us explore the roles and importance of different features for trend detection. To this end, we identify the significant features using the greedy selection algorithm described in Section 2.3, and group them by the five classes (user meta-data, content, network, sentiment, and timing) previously defined. We focus on KNN-DTW, our best performing method. After selecting the top 10 features for different delays D, we compute the frac-

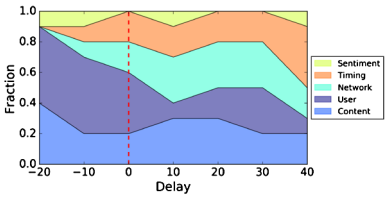

Figure 8 KNN-DTW feature analysis. Stacked plot showing how different feature classes are represented among the top 10 selected features.

Table 3 Top 10 features for experiments with different values of *D*

Delay	Features	Classes
40	Number of tweets	Timing
	Max. proportion of pronouns in a tweet	Content
	Entropy of hashtag cooccurrence network degree	Network
	Entropy of time between two consecutive mentions	Timing
	Mean time between two consecutive tweets	Timing
	Entropy of emoticon scores	Sentiment
	Median time between two consecutive tweets	Timing
	Max. originator's followers count	User
	Kurtosis of mention network degree distribution	Network
	Entropy of pre-determiner POS frequency in a tweet	Content
0	Max. hashtag cooccurrence network degree	Network
	Entropy of number of originator's friends count	User
	Max. originator's statuses count	User
	Median time between two consecutive tweets	Timing
	Skewness of time between two consecutive mentions	Timing
	Median of sender's lists count	User
	Min. originator's lists count	User
	Median of mention network out-degree	Network
	Min. frequency of adjective POS in a tweet	Content
	Mean frequency of noun POS in a tweet	Content

tions of top features in each class, as illustrated by Figure 8. We list the top features for experiments $D = 0$ (early detection) and $D = 40$ (classification) in Table 3.

The usefulness of content features does not appear to change significantly between early and late detection. In the early detection task, user features seem to contribute significantly more than any other class, possibly because early adopters reveal strong signals about the nature of trends. As we move past the trending point, signals from early adopters are flooded by increasing numbers of participants. Timing and network features become increasingly important as the involvement of more users allows to analyze group activity and network structure patterns.

3.3 Analysis of misclassifications

We conclude our analysis by discussing when our system fails. In Figure 9, we illustrate how some key features of misclassified trends diverge from the majority of the trends that are correctly classified. We observe that some misclassified trends follow the temporal characteristics of the other class. This is best illustrated in the case of volume (number of tweets).

An advantage of continuous class scores is that we can tune the classification threshold to achieve a desired balance between precision and recall, or between false positives and false negatives. False negative errors are the most costly for a detection system: a promoted trend mistakenly labeled as organic would easily go unchecked among the larger number of correctly labeled organic trends. Focusing our attention on a few specific in-

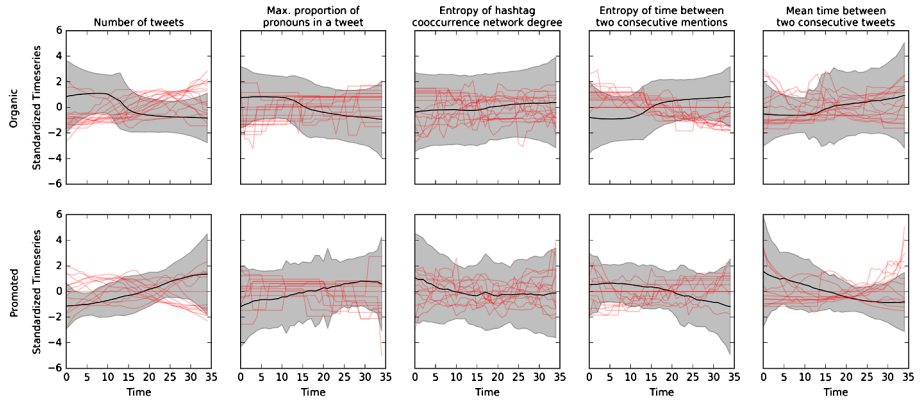

Figure 9 Comparison between feature time series of misclassified and correctly classified trends. Time series of the top five features (columns) for promoted (top) and organic (bottom) trends in the $D = 40$ detection task. The black lines and gray areas represent the average and 95% confidence intervals of time series for correctly classified trends. Time series of misclassified trends are shown in red. Misclassified organic trends (false positives) are: `#bobsantigoldlive`, `#evildead`, `#galaxyfamily`, `#gethappy`, `#madmen`, `#makeboringbrilliant`, `#nyias`, `#oneboston`, `#stingray`, `#thewalkingdead`, `#timeto365`, `#watchsuitstonight`, and `#whyiwatchsuits`. Misclassified promoted trends (false negatives) are: `#1dmemories`, `#20singersthatilike`, `#8thseed`, `#bnppo13`, `#ciaa`, `#expowest`, `#jaibrooksforpresident`, `#justintimberweek`, `#kobalt400`, `#mentionsomeonecuteandbeautiful`, `#nyc`, `#realestate`, `#stars`, `#sxsw`, `#wbc`, and `#wcw`.

stances of false negatives generated by our system, we gained some insight on the reasons triggering the mistakes. First of all, it is conceivable that promoted trends 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 trends become harder to distinguish from organic ones. This may explain the dip in accuracy observed for the longest delay (cf. Figure 5).

False positives (organic trends mistakenly labeled as promoted) can be manually filtered out in post-processing and are therefore less costly. However, analysis of false positives provides for some insight as well. Some trends in our dataset, such as `#watchsuitstonight` and `#madmen`, were promoted via alternative communication channels (television and radio), 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. Those campaigns were therefore wrongly counted as false positives, penalizing our algorithms in the evaluation. We find it remarkable that in these cases our system is capable of learning the signature of promoted trends, even though the promotion occurs outside of the social media platform.

4 Related work

Recent work on social media provides a better understanding of human communication dynamics such as collective attention and information diffusion [52], the emergence of trends [53, 54], social influence and political mobilization [2, 55–57].

Different information diffusion mechanisms may determine the trending dynamics of hashtags and other memes on social media. Exogenous and endogenous dynamics pro-

duce memes with distinctive characteristics [17, 24, 25, 37, 54]: external events occurring in the real world (e.g., a natural disaster or a terrorist attack) can generate chatter on the platform and therefore trigger the trending of a new, unforeseen hashtag; other topics (e.g., politics or entertainment) are continuously discussed and sometimes a particular conversation can accrue lots of attention and generate trending memes. The promotional campaigns studied here can be seen as a type of exogenous factor affecting the visibility of memes.

The present work, to the best of our knowledge, is the first to investigate the early detection of promoted content on social media. We focus our attention on advertisement, which can play an important role in information campaigns. Trending memes are considered an indicator of collective attention in social media [17, 58], and as such they have been used to predict real-world events, like the winner of a popular reality TV show [59]. Although emerging from collective attention, communication on social media can be manipulated, for example for political gain, as in the case of astroturf [4, 60].

Recent work analyzes emerging topics, memes, and conversations triggered by real world events [61–63]. Studies of information dissemination reveal mechanisms governing content production and consumption [64] as well as prediction of future content popularity. Cheng *et al.* study the prediction of photo-sharing cascade size [22] and recurrence [23] on Facebook. Machine learning models can predict future popularity of emerging hashtags and content on social media [65, 66]. Features extracted from content [67], sentiment [37, 68], community structure [69, 70], and temporal signatures [71–73] are commonly used to train such models. In this paper we leverage similar features, but for the novel task of campaign detection. Furthermore, our task is more challenging because we deal with dynamic features whose changes over time are captured in high-dimensional time series.

Another topic related to our research is rumor detection. Rumors may emerge organically as genuine conversation and spread out of control. They are characterized and sustained by ambiguous contexts, where correctness and completeness of information or the meaning of a situation is not obviously apparent [74]. Examples are situations of crisis or topics of public debate [75]. Existing systems to identify rumors are based mostly on content analysis [76, 77] and clustering techniques [78, 79]. An open question is to determine if rumor detection might benefit from the wide set of feature classes we propose here.

The proposed framework is based on a mixture of features common in social media data, including emotional and sentiment information. The literature has reported extensively on the use of social media content to describe emotional and demographic characteristics of users [26, 36, 37]. The use of language in online communities is the focus of two recent papers [28, 29]: the authors observe that the language of social media users evolves, and common patterns emerge over time. The language style of users adapts to achieve better fitness in the conversation [80]. These findings suggest that language contains strong signals, in particular if studied in conjunction with other dimensions of the data. Our study confirms the importance of content for campaign detection.

Finally, our system builds on network features and diffusion patterns of social media messages. Network structure and information diffusion in social media have been studied extensively [81, 82]. Network features are highly predictive of certain types of social media abuse, like astroturf, that attempt to simulate grassroots online conversations [4, 5, 14, 83]. Such artificial campaigns produce peculiar patterns of information diffusion: the topology

of retweet or mention networks is often a stronger signal than content or language. The present findings are consistent with this body of work, as network features are helpful in detecting promoted content after trending.

5 Conclusions

As we increasingly rely on social media to satisfy our information needs, it is important to recognize the dynamics behind online campaigns. In this paper, we posed the problem of early-detection of promoted trends on social media, discussed the challenges that this problem presents, and proposed a supervised computational framework to attack it. The proposed system leverages time series representing the evolution of different features characterizing trending campaigns. The list includes features relative to network structure and diffusion patterns, sentiment, language and content features, timing, and user metadata. We demonstrated the crucial advantages of encoding temporal sequences.

We achieved good accuracy in campaign detection. Our early detection performance is remarkable when one considers the challenging nature of the problem and the low volume of data available in the early stage of a campaign. We also studied the robustness of the proposed algorithms by introducing random temporal shifts around the trending point, simulating realistic scenarios in which the trending point can only be estimated with limited accuracy.

One of the advantages of our framework is that of providing interpretable feature classes. We explored how content, network, and user features affect detection performance. Extensive feature analysis revealed that signatures of campaigns can be detected early, especially by leveraging content and user features. After the trending point, network and temporal features become more useful.

The availability of data about organic and promoted trends is subject to Twitter's recipe for selecting trending hashtags. There is no certain way to know if and when social media platforms make any changes to such recipes. However, nothing in our approach assumes any knowledge of a particular platform's trending recipe. If the recipe changes, our system could be retrained accordingly.

This work represents an important step toward the automatic detection of campaigns. The problem is of paramount importance, since social media shape the opinions of millions of users in everyday life. Further work is needed to study whether different classes of campaigns (say, legitimate advertising vs. terrorist propaganda) may exhibit characteristics captured by distinct features. Many of the features leveraged in our model, such as those related to network structure and temporal attributes, capture activity patterns that could provide useful signals to detect astroturf [4]. Therefore, our framework could in principle be applied to astroturf detection, if longitudinal training data about astroturf campaigns were available.

Acknowledgements

We thank Mohsen JafariAsbagh, Qiaozhu Mei, Zhe Zhao, and Sergey Malinchik for helpful discussions in 2012 and 2013. We are also grateful to two anonymous reviewers, whose suggestions greatly improved this paper. This work was supported by ONR (N15A-020-0053), NSF (grant CCF-1101743), DARPA (grant W911NF-12-1-0037), and the McDonnell Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions
All authors conceived the system and the experiments. OV implemented the system and performed the experiments. All authors analyzed and discussed the results and contributed to the manuscript.

Author details
[1]School of Informatics and Computing, Indiana University, Bloomington, IN, USA. [2]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA. [3]Indiana University Network Science Institute, Bloomington, IN, USA.

Publisher's Note

Received: 18 March 2016 Accepted: 22 June 2017 Published online: 05 July 2017

References

1. Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on Twitter. In: Proc. of the 4th ACM international conference on web search and data mining, pp 65-74
2. Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489(7415):295-298
3. Olteanu A, Varol O, Kıcıman E (2017) Distilling the outcomes of personal experiences: a propensity-scored analysis of social media. In: Proc. of the 20th ACM conference on computer-supported cooperative work and social computing
4. Ratkiewicz J, Conover M, Meiss M, Goncalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. In: Proceedings of the 5th international AAAI conference on weblogs and social media, pp 297-304
5. Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59(7):96-104
6. Bessi A, Ferrara E (2016) Social bots distort the 2016 us presidential election online discussion. First Monday 21:11
7. Bessi A, Coletto M, Davidescu GA, Scala A, Caldarelli G, Quattrociocchi W (2015) Science vs conspiracy: collective narratives in the age of misinformation. PLoS ONE 10(2):0118093
8. Shearlaw M (2015) From Britain to Beijing: how governments manipulate the Internet. http://www.theguardian.com/world/2015/apr/02/russia-troll-factory-kremlin-cyber-army-comparisons
9. Berger J, 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
10. Ferrara E, Wang W-Q, Varol O, Flammini A, Galstyan A (2016) Predicting online extremism, content adopters, and interaction reciprocity. In: International conference on social informatics. Springer, Berlin, pp 22-39
11. U.S. Securities and Exchange Commission (2015) Updated investor alert: social media and investing - stock rumors. http://www.sec.gov/oiea/investor-alerts-bulletins/ia_rumors.html
12. Ciampaglia GL, Shiralkar P, Rocha LM, Bollen J, Menczer F, Flammini A (2015) Computational fact checking from knowledge networks. PLoS ONE 10(6):0128193
13. Zhao Z, Resnick P, Mei Q (2015) Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 1395-1405
14. Varol O, Ferrara E, Davis CA, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. arXiv:1703.03107
15. Clark EM, Jones CA, Williams JR, Kurti AN, Nortotsky MC, Danforth CM, Dodds PS (2015) Vaporous marketing: uncovering pervasive electronic cigarette advertisements on Twitter. arXiv:1508.01843
16. Haustein S, Bowman TD, Holmberg K, Tsou A, Sugimoto CR, Larivière V (2016) Tweets as impact indicators: examining the implications of automated "bot" accounts on Twitter. J Assoc Inf Sci Technol 67(1):232-238
17. Lehmann J, Gonçalves B, Ramasco JJ, Cattuto C (2012) Dynamical classes of collective attention in Twitter. In: Proc. the 21th international conference on world wide web, pp 251-260
18. Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on web search and data mining. ACM, New York, pp 177-186
19. Myers SA, Leskovec J (2014) The bursty dynamics of the Twitter information network. In: Proceedings of the 23rd international conference on world wide web. ACM, New York, pp 913-924
20. Twitter Inc. (2016) FAQs about trends on Twitter. https://support.twitter.com/articles/101125
21. Weng L, Menczer F, Ahn Y (2013) Virality prediction and community structure in social networks. Sci Rep 3:2522
22. Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J (2014) Can cascades be predicted? In: Proceedings of the 23rd international conference on world wide web. ACM, New York, pp 925-936
23. Cheng J, Adamic LA, Kleinberg JM, Leskovec J (2016) Do cascades recur? In: Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 671-681
24. Sornette D, Deschâtres F, Gilbert T, Ageon Y (2004) Endogenous versus exogenous shocks in complex networks: an empirical test using book sale rankings. Phys Rev Lett 93(22):228701
25. Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 33-41
26. Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, Rosenquist JN (2011) Understanding the demographics of Twitter users. In: Proceedings of the 5th international AAAI conference on weblogs and social media
27. Ghosh R, Surachawala T, Lerman K (2011) Entropy-based classification of retweeting activity on Twitter. In: Proceedings of KDD workshop on social network analysis (SNA-KDD)
28. Danescu-Niculescu-Mizil C, West R, Jurafsky D, Leskovec J, Potts C (2013) No country for old members: user lifecycle and linguistic change in online communities. In: Proceedings of the 22nd international conference on world wide web, pp 307-318
29. McAuley JJ, Leskovec J (2013) From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In: Proceedings of the 22nd international conference on world wide web. ACM, New York, pp 897-908

30. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A (2013) The Twitter of Babel: mapping world languages through microblogging platforms. PLoS ONE 8(4):61981
31. Botta F, Moat HS, Preis T (2015) Quantifying crowd size with mobile phone and Twitter data. R Soc Open Sci 2(5):150162
32. Letchford A, Moat HS, Preis T (2015) The advantage of short paper titles. R Soc Open Sci 2(8):150266
33. Briscoe E, Appling S, Hayes H (2014) Cues to deception in social media communications. In: Proceedings of the Hawaii international conference on system sciences
34. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: ICWSM, vol 10, pp 178-185
35. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1-8
36. Mitchell L, Harris KD, Frank MR, Dodds PS, Danforth CM (2013) The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place. PLoS ONE 8(5):64417
37. Ferrara E, Yang Z (2015) Quantifying the effect of sentiment on information diffusion in social media. PeerJ Comput Sci 1:26
38. Kloumann IM, Danforth CM, Harris KD, Bliss CA, Dodds PS (2012) Positivity of the English language. PLoS ONE 7(1):29484
39. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav Res Methods 45:1191-1207
40. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. ACL, pp 347-354
41. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of Twitter data. In: Proceedings of the workshop on languages in social media. ACL, pp 30-38
42. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157-1182
43. John GH, Kohavi R, Pfleger K et al (1994) Irrelevant features and the subset selection problem. In: Machine learning: proceedings of the eleventh international conference, pp 121-129
44. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861-874
45. Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. In: Proc. of AAAI workshop on knowledge discovery in databases. Seattle, pp 359-370
46. Keogh E, Ratanamahatana CA (2005) Exact indexing of dynamic time warping. Knowl Inf Syst 7(3):358-386
47. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21-27
48. Senin P, Malinchik S (2013) Sax-vsm: interpretable time series classification using sax and vector space model. In: Data mining (ICDM), 2013 IEEE 13th international conference on. IEEE Press, New York, pp 1175-1180
49. Lin J, Keogh E, Lonardi S, Chiu B (2003) A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery, pp 2-11
50. Lin J, Keogh E, Wei L, Lonardi S (2007) Experiencing sax: a novel symbolic representation of time series. Data Min Knowl Discov 15(2):107-144
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825-2830
52. Weng L, Flammini A, Vespignani A, Menczer F (2012) Competition among memes in a world with limited attention. Sci Rep 2:354
53. Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 497-506
54. Ferrara E, Varol O, Menczer F, Flammini A (2013) Traveling trends: social butterflies or frequent fliers? In: Proc. of the first ACM conference on online social networks, pp 213-222
55. Conover MD, Davis C, Ferrara E, McKelvey K, Menczer F, Flammini A (2013) The geospatial characteristics of a social movement communication network. PLoS ONE 8:55957
56. Conover MD, Ferrara E, Menczer F, Flammini A (2013) The digital evolution of Occupy Wall Street. PLoS ONE 8:64679
57. Varol O, Ferrara E, Ogan CL, Menczer F, Flammini A (2014) Evolution of online user behavior during a social upheaval. In: Proceedings of the 2014 ACM conference on web science. ACM, New York, pp 81-90
58. Wu F, Huberman BA (2007) Novelty and collective attention. Proc Natl Acad Sci 104:17599-17601
59. Ciulla F, Mocanu D, Baronchelli A, Gonçalves B, Perra N, Vespignani A (2012) Beating the news using social media: the case study of American Idol. EPJ Data Sci 1:8
60. Metaxas PT, Mustafaraj E (2012) Social media and the elections. Science 338(6106):472-473
61. Aggarwal CC, Subbian K (2012) Event detection in social streams. In: SDM, vol 12. SIAM, Philadelphia, pp 624-635
62. Becker H, Naaman M, Gravano L (2011) Beyond trending topics: real-world event identification on Twitter. In: ICWSM, vol 11, pp 438-441
63. Cataldi M, Di Caro L, Schifanella C (2010) Emerging topic detection on Twitter based on temporal and social terms evaluation. In: Proceedings of the tenth international workshop on multimedia data mining. ACM, New York, p 4
64. Ciampaglia GL, Flammini A, Menczer F (2015) The production of information in the attention economy. Sci Rep 5:9452
65. Tsur O, Rappoport A (2012) What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In: Proceedings of the fifth ACM international conference on web search and data mining. ACM, New York, pp 643-652
66. Ma Z, Sun A, Cong G (2013) On predicting the popularity of newly emerging hashtags in Twitter. J Am Soc Inf Sci Technol 64(7):1399-1410
67. Jamali S, Rangwala H (2009) Digging digg: comment mining, popularity prediction, and social network analysis. In: Web information systems and mining, 2009. WISM 2009. International conference on. IEEE Press, New York, pp 32-38
68. Krauss J, Nann S, Simon D, Gloor PA, Fischbach K (2008) Predicting movie success and academy awards through sentiment and social network analysis. In: ECIS, pp 2026-2037

69. Weng L, Menczer F, Ahn Y-Y (2014) Predicting successful memes using network and community structure. In: Proc. eighth international AAAI conference on weblogs and social media (ICWSM). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8081
70. Weng L, Ratkiewicz J, Perra N, Gonçalves B, Castillo C, Bonchi F, Schifanella R, Menczer F, Flammini A (2013) The role of information diffusion in the evolution of social networks. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 356-364
71. Pinto H, Almeida JM, Gonçalves MA (2013) Using early view patterns to predict the popularity of youtube videos. In: Proceedings of the sixth ACM international conference on web search and data mining. ACM, New York, pp 365-374
72. Figueiredo F, Benevenuto F, Almeida JM (2011) The tube over time: characterizing popularity growth of Youtube videos. In: Proceedings of the fourth ACM international conference on web search and data mining. ACM, New York, pp 745-754
73. Wang S, Yan Z, Hu X, Yu PS, Li Z (2015) Burst time prediction in cascades. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence. AAAI Press, Menlo Park, pp 325-331
74. DiFonzo N, Bordia P (2007) Rumor, gossip and urban legends. Diogenes 54(1):19-35
75. Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: can we trust what we RT? In: Proceedings of the first workshop on social media analytics. ACM, New York, pp 71-79
76. Qazvinian V, Rosengren E, Radev DR, Mei Q (2011) Rumor has it: identifying misinformation in microblogs. In: Proceedings of the conference on empirical methods in natural language processing. ACL, pp 1589-1599
77. Kwon S, Cha M, Jung K, Chen W, Wang Y (2013) Prominent features of rumor propagation in online social media. In: Proc. IEEE international conference on data mining series (ICDM)
78. Ferrara E, JafariAsbagh M, Varol O, Qazvinian V, Menczer F, Flammini A (2013) Clustering memes in social media. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. IEEE Press, New York, pp 548-555
79. JafariAsbagh M, Ferrara E, Varol O, Menczer F, Flammini A (2014) Clustering memes in social media streams. Soc Netw Anal Min 4(1):1-13
80. Das A, Gollapudi S, Kıcıman E, Varol O (2016) Information dissemination in heterogeneous-intent networks. In: Proceedings of the 8th ACM conference on web science. ACM, New York, pp 259-268
81. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: Proc. of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 44-54
82. Lerman K, Ghosh R (2010) Information contagion: an empirical study of the spread of news on Digg and Twitter social networks. In: Proceedings of the 4th international AAAI conference on weblogs and social media, pp 90-97
83. Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Patil S, Flammini A, Menczer F (2011) Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th international conference on world wide web, pp 249-252