Policy and Internet

Event Prediction With Learning Algorithms—A Study of Events Surrounding the Egyptian Revolution of 2011 on the Basis of Micro Blog Data

Benedikt Boecking, Margeret Hall, and Jeff Schneider

We aim to predict activities of political nature influencing or reflecting societal-scale behavior and beliefs by applying learning algorithms to Twitter data. This study focuses on capturing domestic events in Egypt from November 2009 to November 2013. To this extent we study underlying communication patterns by evaluating content and metadata of 1.3 million tweets through computationally supported classification, without targeting specific keywords or users from the Twitter stream. Support Vector Machine (SVM) and Support Distribution Machine (SDM) classification algorithms are applied to detect and predict societal-scale unrest. Latent Dirichlet Allocation (LDA) is used to create content-based input patterns for the SVM while the SDM is used to classify sets of features created from meta-data. The experiments reveal that user centric approaches based on meta-data outperform methods employing content-based input despite the use of well established natural language processing algorithms. The results show that distributions over user centric meta information provide an important signal when detecting and predicting events. Applying this approach can assist policymakers and stakeholders in their efforts toward proactive community management.

KEY WORDS: machine learning, event prediction, classification, Twitter data, modeling, politics, Arab Spring

Introduction

Near-to-real time measurement of societal indicators for proactive community management is an overarching policy goal. Until recently, the measurement of societal health and success has been conducted via economic indicators or indices concentrating on macro social indicators. Due to the methodology, such indicators highlight condition changes considerably after their occurrence. Thereby, current indicators are restricted; consequently, policymakers are unable to use them as a comprehensive, detailed, and prompt measurement of human welfare or public opinion. For instance, current indicators were unable to predict the shock waves accompanying the Arab Spring or crises in Venezuela and the Ukraine.

This is where recent innovations driven by the computational sciences can make a difference. Machine learning algorithms have evolved over the years to be successfully

Published by Wiley Periodicals, Inc., 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford, OX4 2DQ.

employed on structured and unstructured data, enabling advancements in many areas such as image and voice recognition or for the navigation of autonomous vehicles (Mitchell, 1997). The methods are invaluable as they help both machines and humans to gain a high-level understanding of large collections of data. Machine learning algorithms are popular because the techniques are designed to deal with noise, are highly scalable, and are able to work with almost arbitrary types of data. Machine learning offers the promise of analyzing large amounts of data and simultaneously tackling until-now nontrivial issues such as language, bias, automation, and generalization. With the advent of social media, scholars have applied machine learning approaches to social media data to enrich or even supplement traditional data sources. The general interest of policymakers, societal stakeholders, and researchers in social media stems from the rich variety of information contained within, as well as its often public nature, and ready availability. Using learning algorithms to detect or predict domain specific events can lead to generalizable approaches that may help to address a multitude of policy issues such as public health, safety, and economic issues, for example, through early warnings of violent protests or diseases.

Furthering this goal, this study targets domestic events of political nature in Egypt which can be associated with either the standing government or the opposition. We aim to predict events in a time frame of four years surrounding the Egyptian Revolution of 2011 on the basis of Twitter data to investigate whether conflicts on a national level can be monitored or even predicted using social media data. It is reasonable to believe that if our approach works for domain specific events, in this case conflict events in Egypt, the findings will generalize well to events in other domains. The case of the Egyptian Revolution proved opportune as it is widely acknowledged that the revolution was enhanced by online social media (Lotan et al., 2011; Starbird & Palen, 2011); that intermittent periods of peace and protest are recorded in this time span; and that the detection and prediction of social unrest is a key interest of policymakers and community managers. We define detection to be the task of identifying the occurrence of an event as it happens or shortly after on the basis of input data gathered within a short window surrounding the incident. We define prediction to be the task of forecasting the occurrence of an event based on input data only gathered prior to the incident. We further examine which input features provide the signal for event detection and prediction. Therefore we pose the research question: Which parameters are necessary to detect and predict civil unrest and conflict based on hidden patterns found in Twitter data?

The second section covers literature on Twitter data in relation to events and machine learning. Our setup for event detection and prediction is described in the third section. The results of our experiments are presented in the fourth section followed by a discussion of the findings in the fifth section and the conclusion in the sixth section.

Related Work

Numerous studies focus on the use of Twitter, resulting in a very large research corpus. This section presents related publications, focusing on learning algorithms, event detection, prediction tasks, and the use of Twitter during events and emergency situations as well as during the Egyptian Revolution of 2011.

Twitter and the Egyptian Revolution of 2011

Several articles address the role of social media within the Arab Spring in general and the Egyptian Revolution of 2011 in particular (e.g., Anderson, 2011; Choudhary, Hendrix, Lee, Palsetia, & Liao, 2012; Khondker, 2011; Lotan et al., 2011; Starbird & Palen, 2012; Wolfsfeld, Segev, & Sheafer, 2013). The studies provide evidence for the coverage of the Egyptian Revolution on Twitter making event prediction in a wider time frame surrounding the revolution an appealing task. Starbird and Palen (2012) examine information diffusion activity through a subset of tweets during the Egyptian Revolution of 2011 and find that the protesters were clearly using social media services to coordinate their actions and garner support. The authors extracted the 1,000 most highly retweeted users by use of popular hashtags related to the protests. They highlight that 30 percent of the users appear to have been in Cairo during the event, many of which were among the protesters out in the streets. With regard to the role of social media during the revolutions, most researchers tend to agree that it was used as tool supporting the cause but that other means of communication and organization would likely have substituted it had social media not been available (see e.g., Khondker, 2011).

Twitter and Learning Algorithms

With regard to learning algorithms, Twitter has been used in various problem settings (e.g., Asur & Huberman, 2010; Conover, Goncalves, Ratkiewicz, Flammini, & Menczer, 2011; Li, Lei, Khadiwala, & Chang, 2012; Puniyani, Eisenstein, Cohen, & Xing, 2010; Sakaki, Okazaki, & Matsuo, 2010; Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010). The articles include tasks such as the assignment of tweets and users to categories, the inference of network structures, and the detection of events, that is, determining the occurrence of an event as it occurs or shortly after. Popular approaches for the creation of input features in these studies include the use of text representations such as term frequencies and topic modeling, or information gathered through reconstructed networks such as retweet networks (see e.g., Conover et al., 2011), as well as sentiment analysis where a sentiment score may be assigned to each tweet (see e.g., Asur & Huberman, 2010).

Twitter and Crisis or Conflict Events

Several studies have investigated information flow and user behavior on microblogging platforms during common emergency and crisis situations (e.g., Hua et al., 2013; Mendoza, Poblete, & Castillo, 2010; Qu, Huang, Zhang, & Zhang, 2011; Starbird, Maddock, Orand, Achterman, & Mason, 2014; Starbird & Palen, 2010, 2011; Vieweg, Hughes, Starbird, & Palen, 2010). The studies point to common reporting behavior during events, low variance in vocabulary, the importance of retweets, and also message attributes such as their increased frequency and contraction in length during major events.

Twitter and Event Detection

Twitter data has been used for event detection in a different settings such as in first story detection tasks (e.g., Petrović, Osborne, & Lavrenko, 2010) where streaming algorithms are used to discover threads of similar tweets. Popular classification algorithms such as Naive Bayes, Support Vector Machines (SVMs), and Decision Trees have also previously been applied for event detection (e.g., Becker, Naaman, & Gravano, 2011; Popescu & Pennacchiotti, 2010; Popescu, Pennacchiotti, & Paranjpe, 2011; Sakaki et al., 2010; Sankaranarayanan, Samet, Teitler, Lieberman, & Sperling, 2009). For example, Sakaki et al. (2010) train a binary classifier using a SVM for the purpose of detecting specific target events such as typhoons and earthquakes. The authors additionally show that Twitter location information can be exploited to track and map the events. Atefeh and Khreich (2013) provide an extensive survey of research on Twitter event detection and categorize the literature by event type, detection task, and detection method.

Twitter and Prediction

Aside from event prediction, other prediction tasks on the basis of social media and Twitter data in particular have been studied in different settings (see Kalampokis, Tambouris, & Tarabanis, 2013) including the prediction of election outcomes, movie ratings, stock market movements, and disease outbreaks. Recently researchers have also investigated the task of predicting events using social media data (Xu, Lu, Compton, & Allen, 2014) including the use of Twitter data in particular (Compton, Lee, Lu, De Silva, & Macy, 2013, 2014; Ramakrishnan et al., 2014). Compton et al. (2013) and the extension of the work in Compton et al. (2014) create alerts for civil unrest in Latin America from social media. Their system identifies a number of informative posts by applying a pipeline of filters to large amounts of posts to produce a number of daily alerts that can be managed by a human auditor. The filters in Compton et al. (2014) extract Twitter and Tumblr posts first by keywords, then by mentions of future dates, to then apply a logistic regression classifier using text features and a classifier that labels the source as either individuals or organizations. The posts are additionally filtered by locations and repostings.

Ramakrishnan et al. (2014) forecast events with an alert system using a multitude of sources, among them Twitter, news, Internet traffic, and economic indicators. The system generates alerts for analyst consumption that provide detailed information on the who, when, where, why, and how of the forecasts. The prediction models used by Ramakrishnan et al. (2014) also rely on the application of several filters, among them keywords, date referencing, geolocation, reposts, and term frequencies. The authors also use a model that dynamically expands vocabulary on the basis of seed keywords to generate a larger vocabulary of interest.

To the best of our knowledge, previous studies on event detection or prediction have not avoided the use of keyword or lead user filters to extract input data, a potentially strong source of bias that our study attempts to avoid. We are also unaware of any studies to date that use distribution divergences over meta features from social media data in a classification setting.

Method

This section will cover the applied classifiers and the methods of creating features. The decision to use Twitter data was made due to the public nature of the service and corresponding data availability. In addition, tweets can be geotagged which allows the extraction of tweets that can be attributed to a place, making event detection within a state's borders possible. The time frame before, during, and after the Egyptian Revolution of 2011 was chosen because it covers a time where alterations of the social order occurred multiple times. Tweets are selected based on location if they are believed to originate in Egypt. However, no other restrictions are employed on input data selection as no keywords or specific groups are explicitly targeted. A simplified, high level overview of our approach is as follows:

(1) For all days of available Twitter data, use events extracted from an event database and determine if a given day was "eventful" or "eventless."

(2) Use Twitter data to create daily input features for classifiers, where a feature is defined as follows:

   (a) a daily text feature vector created using topic modeling to employ a classifier on individual patterns.

   (b) daily "bags" of feature vectors of metadata such as number of followers and followees to use a classifier on distributions.

(3) Assign labels from (1) to feature vectors from (2).

   (a) use daily label windows of up to three days, for example, to predict on the basis of one day of Twitter data if events will occur in the following three days.

(4) Classification with 10-fold cross-validation for performance evaluation.

A classifier assigns categorical labels to previously unseen observations by learning from a set of training examples for which the label is known. The classifiers we use are binary classifiers that process a numerical feature input based on Twitter data. A qualitative output signal called a label that is either positive or negative and determines the occurrence of an event is then assigned. We use these binary classifiers to predict and detect event occurrences on the basis of daily Twitter data. We create content-based features from the text of tweets as well as features based on metadata such as a user's number of followers. Previous research has already pointed to the importance of metadata such as by Garcia-Herranz, Moro, Cebrian, Christakis, and Fowler (2014), who use followees to find groups of users central to a network in a method which may be used for sampling from large volumes of social media data (Figure 1).

Classifiers

Two different but related algorithms for classification are used during the experiments. One is the SVM which is a nonlinear maximum margin classifier (Cortes & Vapnik, 1995) that operates on individual data points. The purpose of finding the maximum margin separator is to improve generalizability of the separation of the classes. SVMs are used with a similarity function called a kernel that implicitly maps the input data into some high dimensional feature space where it finds the maximum margin separating hyperplane. The result is that a linear separation in the larger feature space may correspond to a nonlinear separator in the original space. The other classifier we use is the Support Distribution Machine (SDM), an extension of the SVM that was chosen because it enables classification using an approach that estimates divergences of unknown distributions from samples of the data (Poczos, Xiong, Sutherland, & Schneider,

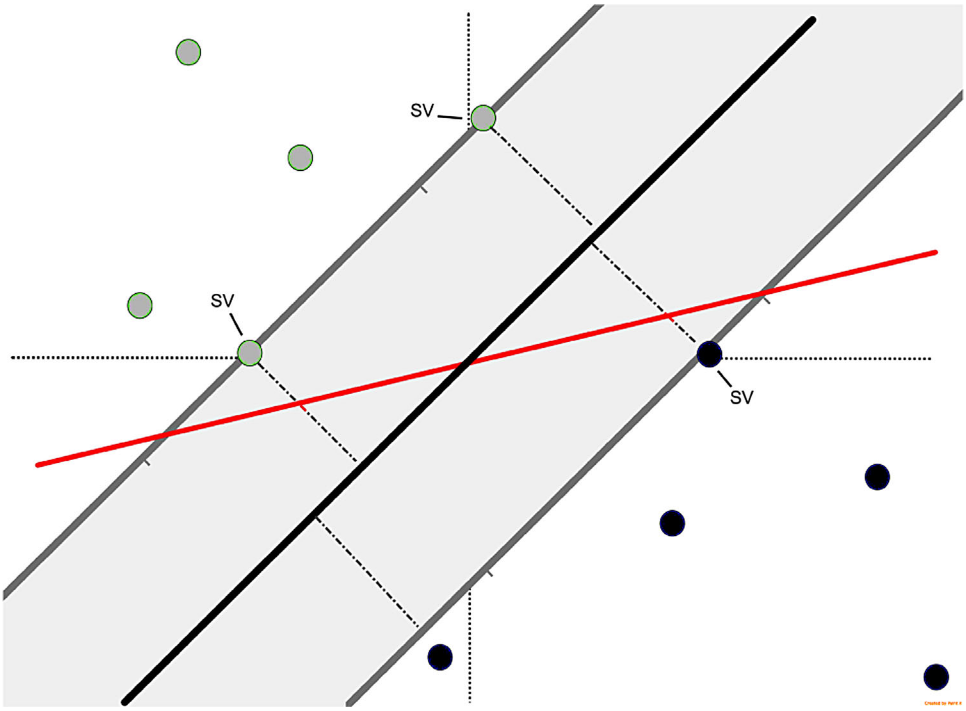

Figure 1. A Linear Maximum Margin Separating Decision Surface in $\mathbb{R}^2$. *Notes*: Although the red plane separates the patterns correctly, the black line does so while maximizing the margin to the closest patterns, the *Support Vectors* (SV).

2012). To establish a baseline comparison for classification accuracy, the values are compared to the performance achieved by a majority classifier. A majority classifier, sometimes called default classifier, provides a baseline accuracy by always assigning the label that belongs to the majority class of the training set. This is to say, if a data set has 90 percent negative labels, the majority classifier achieves 0.9 accuracy in 10-fold cross-validation by always predicting the negative label for every pattern.

Support Vector Machine. Given pairs of patterns and labels (x_1, y_1) ... (x_m, y_m) that constitute the set of training observations, the dual form of the nonlinear soft-margin SVM can be written as follows (Schölkopf & Smola, 2001):

$$\max_{\alpha \in \mathbb{R}^m} \widehat{\alpha} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j),$$

$$\text{subject to } 0 \leq \alpha_i \leq C, i = 1, ..., m,$$

$$\text{and } \sum_{i=1}^{m} \alpha_i y_i = 0.$$

where C is a penalty parameter, α are the Lagrange multipliers, and K is a Mercer kernel, that is, a function that quantifies similarity of two elements of a set by assigning a distance value and which fulfills Mercer's condition (see Schölkopf & Smola, 2001). K maps the data into feature space where the learning algorithm constructs the linear maximum margin separating hyperplane. This hyperplane may then correspond to a nonlinear separation of the patterns in the original input space.

Support Distribution Machine. In traditional classification tasks the individual data points are usually treated as the object of interest. As introduced above, the SVM operates on patterns from a finite-dimensional vector space. In contrast, the SDM[1] generalizes kernel machines so that classification can be done on groups of data points by treating the patterns within as independent and identically distributed (i.i.d.) samples of some unknown distribution. The SDM extends the SVM from vector space to the space of distributions by using kernel functions which estimate distance values between distributions. In the style of Poczos et al. (2012) the problem will be formally defined. Assume T sample sets $\{X_1, \ldots, X_T\} \in X$ with labels $\mathrm{Y_t} \in Y = \{\pm 1\}$.

Let tth input $\mathrm{X}_t = \{\mathrm{X}_{\mathrm{t}1}, \ldots, \mathrm{X}_t, m_t\}$ be m_t i.i.d. samples from density p_t.

The objective function of the SVM dual changes to:

$$\widehat{\alpha} = \arg\max_{\alpha \in \mathbb{R}^T} \sum_{i=1}^{T} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{T} \alpha_i \alpha_j y_i y_j K(p_i, p_j),$$

subject to $\sum_i \alpha_i y_i = 0.0 \le \alpha_i \le C$. Now kernels on i.i.d. sample sets need to be defined. Kernels can be defined on the basis of the following quantity:

$$D_{\alpha,\beta}(p\|q) \int p^{\alpha}(x) q^{\beta}(x) p(x) dx,$$

where $\alpha, \beta \in \mathbb{R}$. Kernels can for example be defined by using:

$$e^{-\frac{\mu^2(p,q)}{2\sigma^2}},$$

and setting $\mu(p,q)$ to the L_2 distance or the Rényi-α divergence:

$$\mu(p,q) = R_{\alpha}(p\|q) \frac{1}{\alpha - 1} log \int p^{\alpha}(x) q^{1-\alpha}(x) dx.$$

Poczos et al. (2012) create an estimator for $D_{\alpha,\beta}(p\|q)$ for some α,β. Let $X = (X_1,\ldots, X_n)$ be an i.i.d. sample of size n of a distribution with density p. Let $Y = (Y_1,\ldots,Y_m)$ be an i.i.d. sample of size m of a distribution with density q. Let ρ_k (i) denote the Euclidean distance of the kth nearest neighbor of X_i in X and let $v_k(i)$ denote the Euclidean distance of the kth nearest neighbor of X_i in Y. The following estimator defined by Poczos et al. (2012) is L_2 consistent under certain conditions:

$$\widehat{D}_{\alpha,\beta} = \frac{B_{k.\alpha,\beta}}{n(n-1)^{\alpha} m^{\beta}} \sum_{i=1}^{n} \rho_k{}^{-d\alpha}(i) \nu_k{}^{-d\beta}(i),$$

where

$$B_{k,\alpha,\beta} = \bar{c}^{-\alpha-\beta} \frac{\Gamma(k)^2}{\Gamma(k-\alpha)\Gamma(k-\beta)}.$$

Γ is the Gamma function and $\bar{c}$ is the volume of a d-dimensional unit ball. Note that a kernel used in an SVM-based classifier such as the one introduced here needs to fulfill Mercer's condition. Poczos et al. (2012) state that the Rényi-α divergence is not symmetric, the triangle inequality does not hold, and that it does not lead to positive semi-definite Gram matrices. However, these deficiencies can be corrected by symmetrizing the resulting matrices and then projecting to the cone of positive semi-definite matrices by discarding any negative eigenvalues from their spectrum (see Higham, 2002; Poczos et al., 2012).

Input Feature Creation

We create content-based feature vectors of daily Twitter activity by using topic modeling on the tweets' text as well as by creating low dimensional representations of unique hashtags. Metadata features are created on the basis of the additional information that accompanies the text gathered through Twitter's application programming interface (API), such as how many followees and followers a user had at the time a tweet was collected, or whether it is a retweet of another user's tweet.

Features From Text. Latent Dirichlet Allocation (LDA) is a widely adopted probabilistic generative topic model (Blei, Ng, & Jordan, 2003). The *MAchine Learning for LanguagE Toolkit* (MALLET; McCallum, 2002) and its implementation of LDA was used to create numerical feature vectors from texts. In the style of Blei and Lafferty (2009) the generative process of LDA is as follows. Assume documents in a corpus of size D are created from a fixed number of K topics, that is, distributions over a fixed vocabulary of terms of size V. Also, for each document there exists a distribution over the topics from which it is created. Let Dir_V (η) denote a V-dimensional symmetric Dirichlet distribution with scalar parameter η. Let Dir_K $(\vec{\alpha})$ denote a K-dimensional Dirichlet distribution with vector parameter $\vec{\alpha}$. Also, let Mult denote the Multinomial distribution. The generative process of LDA by which documents in a corpus are created can be described as follows (Blei & Lafferty, 2009):

(1) For each topic k,

  (a) draw a distribution over words $\overrightarrow{\beta_k} \sim \mathrm{Dir}_V(\eta)$.

(2) For each document d,

  (a) draw a vector of topic proportions $\vec{\theta}_d \sim \mathrm{Dir}_K(\vec{\alpha})$

  (b) For each word n,

    (i) draw a topic assignment $Z_{d,n} \sim \mathrm{Mult}(\vec{\theta}_d), Z_{d,n} \in \{1, ..., K\}$

    (ii) draw a word $W_{d,n} \sim \mathrm{Mult}(\overrightarrow{\beta_{Z_{d,n}}}), W_{d,n} \in \{1, ..., V\}$

The hidden variables in this model are the topics $\vec{\beta}$, the proportions of topics per document $\vec{\theta}$, and the topic assignments per word Z. $\vec{\alpha}$ controls the concentrations of topics per document, while η controls the topic-word concentrations. The multinomial parameters $\vec{\beta}$, the topics, are smoothed by being drawn from a symmetric Dirichlet conditioned on the data.

There are several methods that can be used for model fitting such as SparseLDA (Yao, Mimno, & McCallum, 2009) which is based on Gibbs sampling. Gibbs sampling for LDA is a Markov Chain Monte Carlo method for iterative sampling that can be used to estimate the distribution over topic assignment to word tokens (Griffiths & Steyvers, 2004). As Steyvers and Griffiths (2007) describe, posterior estimates of this distribution can then be used to approximate the hidden variables of the generative process such as the distributions of words in topics and the distributions of topics in the documents.

In LDA each input document is comprised of a set of latent topic proportions. However, Twitter messages are extremely short and sparse and may contain no more than one topic. Using LDA on short input documents degrades performance compared to longer documents as Tang, Meng, Nguyen, Mei, and Zhang (2014) describe. To improve LDA performance, tweets are thus aggregated to produce daily input documents.

The number of topics k is a user defined parameter. Naturally, too low a number of topics will not reflect the underlying structure and keep unrelated content in the same topic while too large a number of topics overfits the data that may result in loss of information. The number of topics in this study is evaluated on the performance of the learning algorithms on the data when using feature vectors of different numbers of topics. The assumption made here is that a number of topics close to ground truth will provide the clearest underlying hidden signal for event detection and prediction through the features. As such, classification performance is expected to stabilize for numbers of topics close to the ground truth.

Several text preprocessing steps are conducted on the Twitter messages. All unwanted characters and items in the messages are removed such as punctuation, emoticons, user-mentions, and URLs. The predominant language contained in a single tweet is guessed using heuristics based on characters and character trigrams as well as dictionary lookups. Topic modeling is then conducted separately on English and Arabic sets which are the languages that make up the majority of tweets. English tweets are converted to lowercase, and stop words are removed separately for the English and Arabic texts. Porter stemming (Porter, 2001) is used on the English tweets, while the Arabic tweets are stemmed as proposed by Taghva, Elkhoury, and Coombs (2005), using implementations provided by the Natural Language Toolkit open source library (Bird, Klein, & Loper, 2009). Input documents of daily tweets are created and passed on to LDA. The topic modeling output of document topic proportions for each language are combined *ex post*.

Features From Hashtags. We also use hashtags to create feature vectors. Twitter hashtags are user-created, and can be any arbitrary combination of characters preceded by the hashtag symbol "#" and separated by a whitespace. We count the use of unique hashtags for daily documents of tweets. This approach results in a very large unique hashtag matrix. To reduce the dimensionality, truncated *Singular Value Decomposition* (SVD) is used. SVD is a matrix factorization technique that can be used to attain a low-rank approximation of a given matrix. Efficient approaches exist to use truncated SVD on large matrices such as randomized algorithms that compute partial matrix decompositions as presented by Halko, Martinsson, and Tropp (2011). While the use of SVD reduces the number of columns, it retains the similarity structure between the rows. Thus, the use of SVD attempts to uncover latent relations of hashtags in the sparse original matrix, grouping the columns by preserving the most important uncorrelated factors. The rows of the matrix are then used as daily feature vectors.

Features From Metadata. We explore different ways of using metadata to create feature vectors. Analogous to the hashtag features, we count tweets sent per unique author by day, resulting in a sparse matrix of documents by unique authors. In an approach mirroring the one on the aforementioned hashtag matrix, the dimensionality of the resulting document by unique author matrix was reduced using truncated SVD. We expect the resulting low dimensional feature vectors to identify groups of authors with common tweeting behavior.

We also create bags of feature vectors for use with the SDM. On Twitter, users can follow other users by adding them as followees. User centric feature vectors per tweet are created containing any combination of the following attributes: the number of followees, number of followers, and the number of tweets sent by the respective user. Of these features, the combination of followees and followers represents degree centrality, that is, a measure of the importance of a node in the network based on its edges (see e.g., Russell, 2011). Using the SDM with these features amounts to estimating distribution divergences over degree centrality of active parts of a network. It provides a way to identify communication behavior of a network on a day to day basis without the need to first estimate the entire network.

Additional features were created using message length and message frequency. The attributes were studied alone and in combination with the meta information mentioned previously. To compare SDM performance on these features with the SVM, statistical measures such as the mean and variance as well as distribution parameters of a fitted lognormal distribution were used to aggregate the information on a daily basis. Here, the advantages of the SDM become clear. It is able to use the features "as is" on the basis of single tweets which compose the daily bags of sample data.

Label Data

To capture societal-scale events of political nature in Egypt a macro-level representation of daily events is needed. Further, a quantitative representation of events is required to facilitate an analytical approach for label creation, that is, we need a measure of magnitude for events to create binary labels. One of the few publicly accessible event databases covering the time period we focus on is the Global Database of Events, Language, and Tone (GDELT; Leetaru & Schrodt, 2013),[2] which contains geolocated political event data and monitors print, broadcast, and web media from all around the globe in over 100 languages. The data has for example been used in a prediction setting by Racette et al. (2014). A data source such as GDELT is no immaculate source of ground truth. Machine coded event databases are inherently noisy and no matter how large the number of news sources used to create event data, a certain media bias can carry over. For example Hammond and Weidmann (2014) compare GDELT data to hand coded events and caution on the accuracy of GDELT on a subnational level. Despite concerns, GDELT is applied here since our study currently only requires a counting function on a daily macro level to create binary labels; accuracy on a micro level is not as important as in other applications. For example, location information is only needed to be accurate to the location of Egypt's borders. Detailed information on actors and ethnicity is not needed.

Target categories of conflict events are selected from the different categories that GDELT captures on the basis that the actions are (1) domestically oriented and (2) can be associated with either the standing government or the opposition (see Supporting Information).[3] The event data set is used on a highly aggregated macro level and the recorded number of mentions across news sources is used as a measure of event importance. Due to noise in both the Twitter and the event data set we set out to detect and predict events in short windows of up to three days, for example, on the basis of one day of Twitter data we want to predict if an event occurred in the following three days or not. To detect the spikes the number of mentions that deviate from the noisy baseline of daily event mentions we determine reasonable thresholds (see Figure 2). Based on that measure, a day is labeled as positive on which one or more events occur that fit the target profile and surpass the threshold, while all other days are labeled as negative. Figure 2 shows the number of mentions for extracted events in Egypt as well as different levels of thresholds to visualize the spikes that are captured by the thresholds. In aggregating the number of mentions on a daily basis for the selected event categories and in selecting a high threshold value, it is assumed that the impact of noise present in the data set is marginalized and that most of the important spikes are captured. By using such a simple measure in assessing the importance of events on a particular day some events may be mistakenly omitted. However, this way the measure uses a minimal amount of assumptions. As mentioned above, labeling on the basis of these binary signals is done with two window sizes of either one or three days. Choosing a window of more than one day smoothens the label data and adjusts for noise in the input and label data set. A maximum window size of three days was chosen to prevent an overall excessive number of positive labels and blurring of events. If one or more days of event data within the window are positive, the day of Twitter data will be labeled positive and negative otherwise. For prediction, the label window concerns the day(s) immediately following the day of Twitter data. For detection, the day of Twitter data is considered in the center of the window. The choice of considering

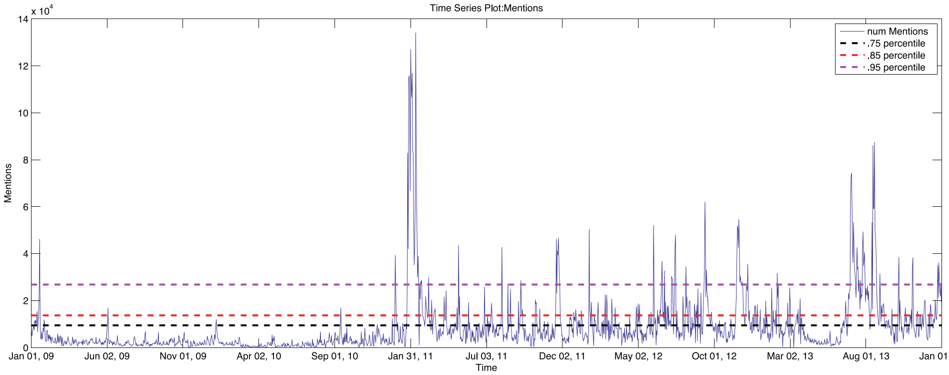

Figure 2. The Smoothed Number of Mentions for Extracted GDELT Events and Different Thresholds for Binary Label Creation.

a sliding window approach for label creation for the present classification task is made due to the noise present in both the input data as well as the label data.

To create labels the 0.75 percentile was selected as the threshold value of the importance measure, above which days were considered positive, that is, one or more important events occurred. This level captured most significant spikes above a base level of noise. Lower threshold values in combination with a label window greater than one day would eventually result in the majority of the labels in the data set being positive.

Twitter Data

Sampling Twitter data through target users, keywords, or hashtags can make sense in problem settings where a concise set of keywords or users can be anticipated. However, in the case of sampling by hashtags, González-Bailón, Wang, Rivero, Borge-Holthoefer, and Moreno (2014) show that this strategy introduces a bias and may artificially crop a periphery of activity. For the purpose of detecting and predicting events during times of social change, a predefined set of keywords or lead users could allow relevant tweets and tweeting behavior to escape the sample. It can also contribute to echo chamber effects, whereby the selection of a hashtag as both a data collection unit and the proxy of a phenomena cause the proxy and not the phenomena to be detected. Further, the possibility of choosing the wrong keywords introduces a source for additional errors, noise, and bias. Moreover, in a practical application concerning prediction, the relevant set of keywords may be impossible to anticipate. We thus extract tweets on the basis of geolocation to gather tweets that most likely stem from the target area, making no assumptions on their content or on user behavior. The Twitter data used stems from a data set established through sampling from the Twitter streaming API with Gardenhose streaming access at an approximate sample rate of 10 percent of the entire Twitter feed.[4] If location information for a tweet is available and determined to be in Egypt, the tweet is extracted. However, the percentage of geotagged tweets on Twitter is very low, generally between 1 and 2 percent. Hence, if the tweet's location is unavailable, the user's location associated with their profile is used as a proxy for their current location, an approach that other studies follow in a similar way (e.g., Ramakrishnan et al., 2014). This process resulted in 1.3 million tweets located in Egypt being extracted. Twitter data was gathered from November 2009 until the end of November 2013 as too little data was available before this time period. The extracted data set contains messages from 57,238 unique users and the messages contain 81,253 unique hashtags.

Experiments

All results quoted in this section are the averaged scores from 10-fold cross-validation (see e.g., Kohavi, 1995). For each training fold, parameter selection is done on the basis of threefold cross-validation within the training set. All SDM results quoted use the Rényi-α divergence where $\alpha = 0.9$ and a nearest neighbor setting of $k = 5$. The SVM uses a Gaussian kernel.

SVM Classification of Document Topic Feature Vectors

A total of 1,042,680 tweets in predominantly English or Arabic were used to create daily feature vectors. As explained in the previous section, models were first fitted to the corpora for different number of topics k. To choose a reasonable number of topics, classification performance was evaluated across several dimensions. The word cloud of Figure 3 illustrates that words related to the Egyptian Revolution of 2011 are reasonably well allocated within at least one of the topics of a topic model fitted to daily documents of tweets with 100 topics. Performance of features from Arabic and English tweets were used separately and in combination. Further, the performance in detection and prediction was compared across different label settings with regard to window size and the threshold value on number of mentions used to create the binary labels. Figure 4 shows an example of how classification performance changes along different numbers of topics. Most results from hereon will be quoted for features from a model with a reasonable $k = 100$ topics as the accuracy stabilized around this value in the different evaluation settings.

Table 1 shows how classification performance changes with respect to different label window sizes in a prediction setting. In the present case, increasing the label window to three days results in an increase in classification accuracy along with an increase in the number of positive labels. A possible explanation is that the performance increase results from a smoothing of the noise in both

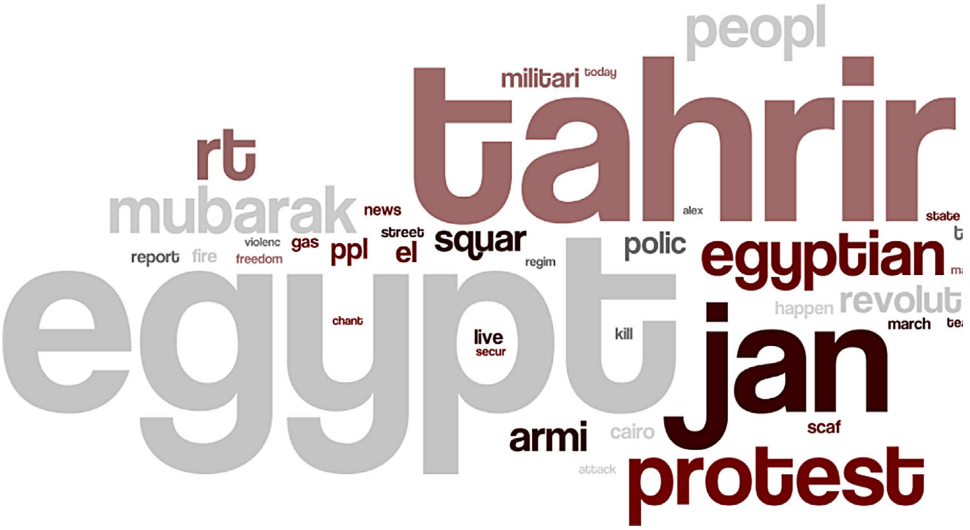

Figure 3. The Top 40 Stemmed Words of a Topic From a Model Trained With 100 Topics on Documents of Daily English Tweets. *Note*: The word size is proportionate to its probability in the topic, but not proportionate to its frequency in the corpus.

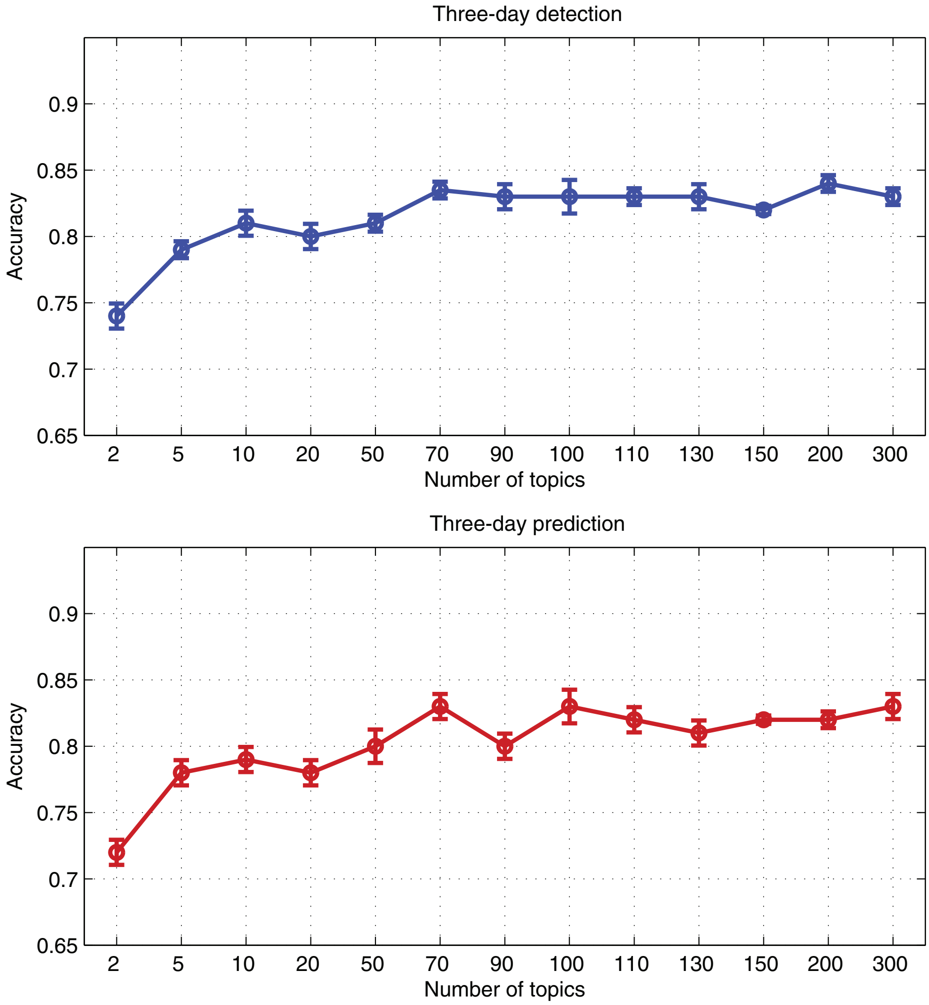

Figure 4. SVM Classification Accuracy of Topic Model Features With Varying Numbers of Topics. *Notes*: Bars denote standard error. The top plot shows a three-day detection window with the classified day in the center. The bottom plot shows a three-day prediction with the classified day preceding the label window.

Table 1. Ten-Fold Cross-Validation Prediction Accuracies of SVM Classification With Topic Modeling Features in Comparison to the Majority Classifier

Window Size	70 Topics	100 Topics	Majority Accuracy
1	0.790	0.799	0.695
2	0.800	0.795	0.604
3	0.828	0.826	0.509

Notes: The label creation threshold was set to the 0.75 percentile. The table indicates prediction results for different label window sizes.

Table 2. SVM Accuracies of Daily Topic Features, English and Arabic Tweets Separately and Combined

Task	100 Topics Arabic	100 Topics English	100 Topics Combined
Prediction	0.788 (0.008)	0.780 (0.013)	0.826 (0.013)
Detection	0.804 (0.007)	0.772 (0.010)	0.834 (0.012)

Notes: Standard error is shown in parentheses. The label window was set to three days, and the binary threshold on number of mentions was set to the 0.75 percentile.

Twitter and event label data. The classification task was also conducted using English and Arabic topic modeling output as shown in Table 2. The results suggest that a combination of both signals provides significantly better classification results in terms of cross-validated accuracy.

Hashtag Features

Classification results using vectors from reduced hashtag count matrices cannot match the classification performance of their topic modeling counterparts. In an effort to assess a common point of comparison between the SVM and SDM, we also created count vectors using unique hashtags per hour. SVM classification of feature vectors of daily counts of unique hashtags beat the classification accuracy achieved using the SDM hourly features. A possible explanation for this is that creating hourly bags results in bags of feature vectors that are too small and that these bags of hourly features additionally violate the SDM assumption of having an i.i.d. sample input. The results can be found in Table 3.

Metadata-Based Features

In the same approach followed to create hashtag-based features, matrices of counts of unique users are created and their size is then reduced using truncated SVD. This feature beats the performance achieved by using topic modeling-based features and hashtag-based features in a three-day label window setting. A combination of topic features and these user features leads to another small improvement in terms of accuracy, however standard errors for these results range between 0.006 and 0.012 do not allow for general conclusions to be drawn. Results for the three-day detection and prediction task are shown in Table 4.

Table 3. SVM Classification Accuracies of Hashtag Features Across Different Dimensions (Dim)

	SVM		SDM	
Task	# Daily 200 Dim	# Daily 100 Dim	# Hourly 200 Dim	# Hourly 100 Dim
Prediction	0.738	0.755	0.701	0.711
Detection	0.752	0.761	0.692	0.689

Notes: The label window was set to three days, and the label creation threshold was set to the 0.75 percentile.

Table 4. SVM Classification Accuracies of User ID-Based Features, Topic Model-Based Features, and a Combination of Topic Model and User Features Across Different Dimensions (Dim)

Task	50-Dim User	100-Dim User	100 Topics	100 Topics & 100-Dim User
Prediction	0.843	0.870	0.826	0.871
Detection	0.854	0.858	0.834	0.870

Notes: The label window was set to three days, and the label creation threshold was set to the 0.75 percentile.

Table 5. SDM Classification Accuracies of Different Bags of One-Dimensional Meta Features

Task	No. Followers	No. Followers	No. Tweets	Message Length
Detection	0.827	0.745	0.747	0.751
Prediction	0.819	0.739	0.743	0.728

Notes: The label window was set to three days, and the label creation threshold was set to the 0.75 percentile.

Classification of Metadata-Based Features With the Support Distribution Machine. The per-tweet meta information consistently available throughout our Twitter data set is the number of followees, the number of followers, the total number of tweets sent, and message length. These features were used to create daily bags of feature vectors for the SDM. classification tasks using the meta information separately revealed that the strongest classification performance with a one-dimensional feature vector was achieved by using the number of followees (see Table 5). Permutations of these features were also classified to explore their combined predictive power using the SDM classifier. The best classification results were obtained using a two-dimensional feature vector encompassing the number of followees and followers. The results achieved with this feature beat the classification performances of all other approaches (see Table 6). Information on precision and recall for this experiment can be found in Tables 7 and 8. A summary of the detection and prediction results in Figures 5 and 6 also visualizes the standard error, showing that the cross-validated accuracy using the followees and friends feature provides a significant improvement over content-based efforts. Attempts to use the SVM with daily metadata using distribution parameters such as the mean and variance as well as distribution parameters of a lognormal distribution did not provide competitive results.

Table 6. SDM Classification Accuracies of Daily Bags of a User Centric Feature of Followees and Followers

Task	No. Followees & Followers	100 Topics	Majority
Detection	0.884	0.834	0.509
Prediction	0.895	0.826	0.509

Notes: The label window was set to three days, and the label creation threshold was set to the 0.75 percentile.

Table 7. Precision and Recall for SDM Three-Day Prediction Window

Label	Precision	Recall	Support
0	0.94	0.86	775
1	0.85	0.93	657
Avg/Total	0.90	0.89	1432

Notes: Label threshold at 0.75. Feature: No. Followees and Followers.

Table 8. Precision and Recall for SDM Three-Day Detection Window

Label	Precision	Recall	Support
0	0.94	0.86	776
1	0.85	0.93	656
Avg/Total	0.90	0.89	1432

Notes: Label threshold at 0.75. Feature: No. Followees and Followers.

Discussion

The previous section showed results for binary classification tasks of predicting and detecting societal-scale events in Egypt using Twitter data extracted on the basis of geolocation information. The prediction and detection tasks are insofar successful as they comfortably outmatch a time-series feature and a majority classifier in terms of classification accuracy. Interestingly,

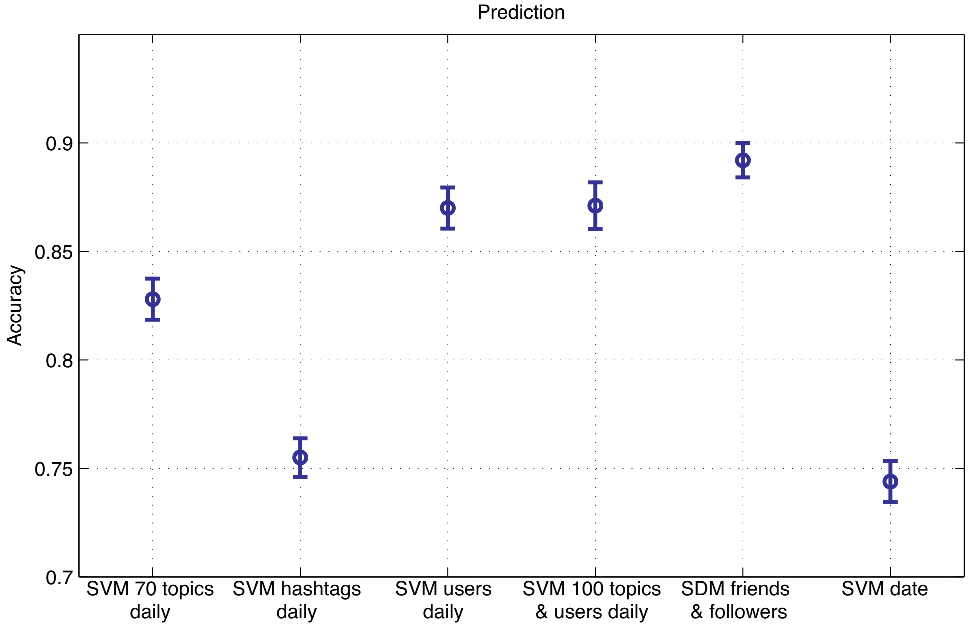

Figure 5. Summary of the Best Prediction Classification Accuracies for a Label Window of Size Three, Label Threshold at the 0.75 Percentile.

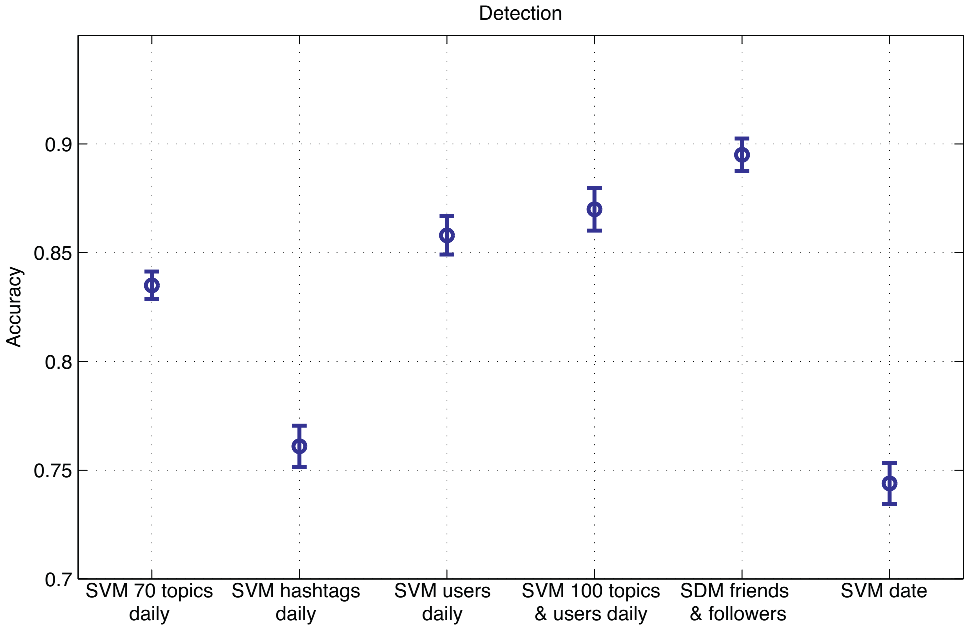

Figure 6. Summary of the Best Detection Classification Accuracies for a Label Window of Size Three, Label Threshold at the 0.75 Percentile.

distribution divergences over user centric features of metadata outperform content-based features in the prediction and detection settings. This indicates that not only the tweet content, but who is tweeting it contributes to forecasting and detection power of tweets. We have also seen that the noise in both the event data set used for label creation and the Twitter data set makes exact detection and prediction a difficult task. This goes hand in hand with the standard limitations of Big Data approaches, which often find that results are accurate but not precise (boyd and Crawford, 2011). With a label window spanning three days, the error reductions with respect to the majority classifier become more impressive as the window smoothes the periods of events. Judging from the extracted events displayed in Figure 2, there is a comparatively peaceful period in 2009 and 2010 in Egypt which contrasts the increasing times of unrest in 2011, 2012, and 2013. Thus, it is possible to achieve respectable classification performance by predicting negative labels for early days in the data set and positive labels later on. To check that the classifiers do not achieve the quoted classification results by simply detecting a time trend in the data, the best results were compared to the performance of a simple date-based feature. Our results show that the topic modeling-based approach, as well as the metadata approach, outperform this date feature comfortably. A major contribution of this work is its focus on and application of methods with limited bias in gathering input data, requiring a low amount of parameter selection. Specifically, by not culling the stream of messages as recommended by González-Bailón et al. (2014) and by giving performance and robustness information and time separated analyses as recommended by Ruths and Pfeffer (2014), common research biases of social media and network analyses are largely mitigated.

Classification Performance Using Content Features

Some valuable conclusions can be drawn on content-based features although they were outperformed by user centric features in classification performance. The classification performance for detection and prediction tasks improve notably when the Arabic and English topic modeling features were combined. The results show that although English is the *lingua franca* of Twitter, the inclusion of information conveyed in other languages can have a significantly positive performance impact. This has impact not only in community management fora, but in policy research.

Interestingly, an attempt to exploit the topical information provided by users through hashtags neither matched the accuracies achieved by the metadata features nor that of the topic modeling features. The approach relied on counting occurrences of unique hashtags and reducing dimensionality of resulting matrices with truncated SVD. The success of the same approach used on matrices counting tweets by unique authors suggests that truncated SVD can indeed be used to find latent structures in similar matrices. A possible explanation for the gap in performance may be found in the sparsity of hashtags. The vocabulary displayed in tweets concerning events may offer more common ground to uncover relations to similar events, which in turn makes it easier for text-based approaches to uncover the similarity. Many studies have focused on prediction tasks in shorter time frames that used popular hashtags or keywords for sampling or as a basis for prediction via frequency counts. Chung and Mustafaraj (2011) outline some of these shortcomings and our results support the conclusion that for general detection and prediction purposes it is necessary to go beyond keywords polarity alone lest the results lead to spurious relationships.

Metadata-Based Features

SVD was used to create low dimensional feature vectors on the basis of unique authors for use with the SVM. The process is expected to reduce noise and group authors by tweeting behavior. The approach provided good classification results and outperformed the content-based methods. The results suggest that the approach succeeded in identifying groups of users with similar tweeting behavior, which then helped the classifier to identify days when societal scale events occurred.

Even better performance was achieved using the SDM on meta information, which also provided a much simpler approach compared to all attempts previously described. The best SDM classification performance was achieved through daily bags of the number of followees and followers per tweet. This corresponds to estimating distributions of degree centrality for users tweeting per day and then calculating the divergences among these distributions. This method thereby indirectly measures weighted activity of nodes of a network without ever actually reconstructing a network. The performance of this degree centrality feature is closely followed in accuracy by combinations of the number of total tweets per user combined with their number of followers or followees. The good performance of the classifier in these cases may be caused by its ability to identify patterns of behavior of groups of authors identified by user centric features that only occur during or before target events. The feature vector of followees and followers does surely have its limits in the amount of information it can convey. But generated from the diverse input data set of the Twitter data, it stands out in its simplicity and performance. While the use of the feature with the SDM involves kernel estimations of distribution divergences, it is not subject to any prior modeling or preprocessing steps. It can be used in the same way for every user as opposed to content-based features where issues of language, slang, stop words, or unwanted characters need to be addressed individually. Another advantage is that the simple feature draws on the strengths of the SDM in working with samples of patterns. In the case of problems with data collection or bandwidth the SDM can still perform well, even if the sample size is significantly smaller. The experiments also showed that efforts in using metadata-based information with the SVM through statistical summarization could not match the superior classification achieved by the SDM. The experiments highlight the advantages of the SDM in scenarios were bags of features are present that can be used in the SDM without any preceding steps. Of use to both the researcher and the policymaker, SDM delivers robust results without extensive data preparation and handling. These results highlight the advantages of employing the latest advancements in machine learning methods on policy-related tasks.

Limitations

The methodology followed in our study does have drawbacks. We acknowledge that location filtering of input tweets is a potential source of bias that needs to be investigated further. In addition, the labels for our classification experiments were established on the basis of machine coded events from a variety of sources of international news sources. When working with a machine coded event data set issues of noise, validity, and bias arise. The creation of labels to detect events relies on the sources to capture the events accurately in the first place. Also, in creating binary labels through a threshold value, the method relies on the target events to be mentioned sufficiently across sources to generate peaks in the number of mentions, possibly carrying over a media bias.

There are also some issues that need to be considered in practical applications. The Twitter network evolves over time. Topics of conversations change rapidly and the number of followees and followers of users evolves. It is thus possible that the underlying ground truths of present topics and distributions over meta features evolve over time as the network changes. Hence, training data may have to be adjusted for such effects.

Future Work

Since periods of events were detected and predicted through the use of a label window, a more careful separation of the prediction and detection tasks is necessary as the windows slightly overlapped. In addition, a more rigorous approach to label creation is required to address questions of bias and verifiability as well as to increase the quality of the labels, for example, through a combination of machine coded events and human verification. It would also be interesting to investigate the present approach in a multi-class classification setting, for example, with regard to violence levels of events. Finally, Sakaki et al. (2010) have shown that by using spatial information they were able to estimate the location of earthquakes close to the actual center. Since all tweets used in this study contain spatial information, the prediction and detection tasks may be extended to mapping the center of the events, for example, through identifying key users *ex post* and to estimate the spatial accuracy achieved with such an approach.

Conclusion

In the efforts of policymakers and stakeholders to guarantee sustainable growth, stability, security, and progress, the struggle to *ex ante* predict events and detect their accompanying shock waves is a common issue. Predicting target events through social media data can be used to provide immediate, high-level feedback on important societal indicators in the form of event signals to policymakers. However, the research task of predicting societal-scale events using social media data still leaves room for many design choices and interpretations. The approach we followed by only selecting tweets according to geolocation and using a range of target events from a machine coded event database to create binary labels arguably retains more noise in both the input and the label data set. But we argue that the findings generalize well and that they may be adapted to more narrowly defined use cases. Care was taken to cross-validate performance, to limit sources of bias, to check the performance against a time-series trend, to compare accuracy values to a majority classifier, and to conduct experiments on an extended time period of four years.

Our findings show that it is possible to both detect and predict societal-scale events using Twitter data in a binary classification setting. In addition, our work shows that estimating distributions over samples of very simple user centric features derived from metadata can outmatch more elaborate content-based approaches in detecting and predicting events. The results suggest that users can be grouped in their event-related messaging behavior through meta-data. This relatively simple approach can be taken by even those with little to no experience in machine learning to produce reliable, robust results. In addition to being a straightforward approach, the simple methods based on distribution divergences presented in this work help to retain initial sample sizes which would otherwise be reduced when resorting to content-based approaches or the reconstruction of communication networks. Finally, minimized bias in the approach creates replicable research and a framework for future online social network and social media studies. Our model allows policymakers to apply cutting edge technologies in their assessments of social media and social network data with minimal bias and low data handling requirements. Thus enabled, policymakers should feel confident in applying machine learning tools to their Big Data analyses.

Benedikt Boecking, M.Sc., Research Programmer, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA [boecking@andrew.cmu.edu].
Margeret Hall, M.A., Dr.des., Ph.D. candidate, Karlsruhe Service Research Institute (KSRI), Karlsruhe Institute of Technology, Karlsruhe, Germany.
Jeff Schneider, Ph.D., Research Professor, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Notes

1. An implementation of the SDM can be found here: https://github.com/dougalsutherland/py-sdm.
2. Recently, a legal dispute regarding several data sources in GDELT has raised concerns (Racette et al., 2014). However, GDELT has been relocated and is again publicly available. The current GDELT project homepage states that the issues were resolved by an independent panel at the University of Illinois. Further, the data in question concerns entries from the historical backfiles which were not used in our study.
3. The list of selected categories is available as additional material on the journal homepage.
4. Brendan O'Connor, at Carnegie Mellon University's School of Computer Science, helped assemble the Twitter data set.

References

Anderson, L. 2011. "Demystifying the Arab Spring: Parsing the Differences Between Tunisia, Egypt, and Libya." *Foreign Affairs* 90 (3): 2–7.

Asur, S., and B.A. Huberman, 2010. "Predicting the Future With Social Media." In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* Vol. 1 of WI-IAT '10. Washington, DC: IEEE Computer Society, 492–99.

Atefeh, F., and W. Khreich. 2013. "A Survey of Techniques for Event Detection in Twitter." *Computational Intelligence* 31 (1): 132–64.

Becker, H., M. Naaman, and L. Gravano. 2011. "Beyond Trending Topics: Real-World Event Identification on Twitter." *In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2745Hila/3207.

Bird, S., E. Klein, and E. Loper, 2009. *Natural Language Processing With Python*, 1st ed. Sebastopol, CA: O'Reilly Media, Inc.

Blei, D.M., and J. Lafferty, 2009. "Topic Models." In *Text Mining: Classification, Clustering, and Applications*, eds. A. Srivastava and M. Sahami. London: Taylor & Francis.

Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. "Latent Dirichlet Allocation." The Journal of Machine Learning Research 3: 993–1022.

boyd, d., and K. Crawford, 2011. "Six Provocations for Big Data." In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. http://ssrn.com/abstract=1926431

Choudhary, A., W. Hendrix, K. Lee, D. Palsetia, and W.-K. Liao. 2012. "Social Media Evolution of the Egyptian Revolution." *Communications of the ACM* 55 (5): 74–80.

Chung, J., and E. Mustafaraj. 2011. "Can Collective Sentiment Expressed on Twitter Predict Political Elections?" *In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/viewFile/3549%26lt%3B/4126.

Compton, R., C. Lee, J. Xu, L. Artieda-Moncada, T.-C. Lu, L. Silva, and M. Macy. 2014. "Using Publicly Visible Social Media to Build Detailed Forecasts of Civil Unrest." *Security Informatics* 3 (4). http://link.springer.com/article/10.1186/s13388-014-0004-6.

Compton, R., T.-C. Lu, L. De Silva, and M. Macy. 2013. "Detecting Future Social Unrest in Unprocessed Twitter Data: Emerging Phenomena and Big Data." *In Proceedings of the 2013 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 56–60.

Conover, M.D., G. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer. 2011. "Predicting the Political Alignment of Twitter Users." *In Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*. IEEE, 192–99.

Cortes, C. and V. Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3): 273–97.

Garcia-Herranz, M., E. Moro, M. Cebrian, N.A. Christakis, and J.H. Fowler. 2014. "Using Friends as Sensors to Detect Global-Scale Contagious Outbreaks." *PLoS ONE* 9 (4): e92413.

González-Bailón, S., N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno. 2014. "Assessing the Bias in Samples of Large Online Networks." *Social Networks* 38: 16–27.

Griffiths, T.L., and M. Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (Suppl 1): 5228–35.

Halko, N., P.G. Martinsson, and J.A. Tropp. 2011. "Finding Structure With Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions." *SIAM Review* 53 (2): 217–88.

Hammond, J., and N.B. Weidmann. 2014. "Using Machine-Coded Event Data for the Micro-Level Study of Political Violence." *Research & Politics* 1 (2). http://rap.sagepub.com/content/1/2/2053168014539924.

Higham, N.J. 2002. "Computing the Nearest Correlation Matrixa Problem From Finance." *IMA Journal of Numerical Analysis* 22 (3): 329–43.

Hua, T., C.T. Lu, N. Ramakrishnan, F. Chen, J. Arredondo, D. Mares, and K. Summers. 2013. "Analyzing Civil Unrest Through Social Media." *Computer* 46 (12): 80–4.

Kalampokis, E., E. Tambouris, and K. Tarabanis. 2013. "Understanding the Predictive Power of Social Media." *Internet Research* 23 (5): 544–59.

Khondker, H.H. 2011. "Role of the New Media in the Arab Spring." *Globalizations* 8 (5): 675–79.

Kohavi, R. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Vol. 2, IJCAI'95*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1137–43.

Leetaru, K., and P. Schrodt. 2013. "GDELT: Global Data on Events, Language, and Tone, 1979-2012." Paper presented at the International Studies Association Annual Conference, April, San Francisco, CA.

Li, R., K.H. Lei, R. Khadiwala, and K.C.-C. Chang. 2012. "TEDAS: A Twitter-Based Event Detection and Analysis System." In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering (ICDE)*, 1273–76. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6228186.

Lotan, G., E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and d. boyd. 2011. "The Arab Spring—The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions." *International Journal of Communication* 5 (0). http://ijoc.org/index.php/ijoc/article/view/1246/643.

McCallum, A.K. 2002. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu.

Mendoza, M., B. Poblete, and C. Castillo. 2010. "Twitter Under Crisis: Can We Trust What We RT?" In *Proceedings of the First Workshop on Social Media Analytics, SOMA '10*, New York, NY: ACM, 71–9.

Mitchell, T.M. 1997. *Machine Learning*, 1 ed. New York, NY: McGraw-Hill, Inc.

Petrović, S., M. Osborne, and V. Lavrenko. 2010. "Streaming First Story Detection With Application to Twitter." In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*. Stroudsburg, PA: Association for Computational Linguistics, 181–89.

Poczos, B., L. Xiong, D.J. Sutherland, and J. Schneider. 2012. "Nonparametric Kernel Estimators for Image Classification." In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2989–96. http://dx.doi.org/10.1109/CVPR.2012.6248028.

Popescu, A.-M., and M. Pennacchiotti. 2010. "Detecting Controversial Events From Twitter." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management', CIKM '10*. New York, NY: ACM, 1873–76.

Popescu, A.-M., M. Pennacchiotti, and D. Paranjpe. 2011. "Extracting Events and Event Descriptions From Twitter." In *Proceedings of the 20th International Conference Companion on World Wide Web, WWW '11*. New York, NY: ACM, 105–6.

Porter, M.F. 2001. "Snowball: A Language for Stemming Algorithms." http://snowball.tartarus.org/texts/introduction.html.

Puniyani, K., J. Eisenstein, S. Cohen, and E.P. Xing. 2010. "Social Links From Latent Topics in Microblogs." In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, WSA '10*. Stroudsburg, PA: Association for Computational Linguistics, 19–20.

Qu, Y., C. Huang, P. Zhang, and J. Zhang. 2011. "Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake." In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11*. New York, NY: ACM, 25–34.

Racette, M.P., C.T. Smith, M.P. Cunningham, T.A. Heekin, J.P. Lemley, and R.S. Mathieu. 2014. "Improving Situational Awareness for Humanitarian Logistics Through Predictive Modeling." In *Proceedings of the Systems and Information Engineering Design Symposium (SIEDS), 2014*. IEEE, 334–39.

Ramakrishnan, N., P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang et al. 2014. "'Beating the News' With Embers: Forecasting Civil Unrest Using Open Source Indicators." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*. New York, NY: ACM, 1799–808.

Russell, M.A. 2011. *Mining the Social Web: Analyzing Data From Facebook, Twitter, LinkedIn, and Other Social Media Sites*, 1 ed. Sebastopol, CA: O'Reilly Media, Inc.

Ruths, D., and J. Pfeffer. 2014. "Social Media for Large Studies of Behavior." *Science* 346 (6213): 1063–64.

Sakaki, T., M. Okazaki, and Y. Matsuo. 2010. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors." In *Proceedings of the 19th International Conference on World Wide Web', WWW '10*. New York, NY: ACM, 851–60.

Sankaranarayanan, J., H. Samet, B.E. Teitler, M.D. Lieberman, and J. Sperling. 2009. "Twitterstand: News in Tweets." In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09*. New York, NY: ACM, 42–51.

Schölkopf, B., and A.J. Smola. 2001. *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA: MIT Press.

Sriram, B., D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. 2010. "Short Text Classification in Twitter to Improve Information Filtering." In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10*. New York, NY: ACM, 841–42.

Starbird, K., J. Maddock, M. Orand, P. Achterman, and R.M. Mason, 2014. "Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter After the 2013 Boston Marathon Bombing." In *Proceedings of the iConference 2014*, 654–62. https://www.ideals.illinois.edu/handle/2142/47257.

Starbird, K., and L. Palen, 2010. "Pass It On?: Retweeting in Mass Emergency." In *Proceedings of the 7th International ISCRAM Conference*. Seattle: International Community on Information Systems for Crisis Response and Management. http://www.researchgate.net/publication/228512367_Pass_It_On_Retweeting_in_Mass_Emergency.

Starbird, K., and L. Palen 2011. "Voluntweeters: Self-Organizing by Digital Volunteers in Times of Crisis." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*. New York, NY: ACM, 1071–80.

Starbird, K., and L. Palen. 2012. "(How) Will the Revolution be Retweeted?: Information Diffusion and the 2011 Egyptian Uprising." In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12*. New York, NY: ACM, 7–16.

Steyvers, M., and T. Griffiths. 2007. "Probabilistic Topic Models." Handbook of latent semantic analysis. *Psychology Press* 427 (7): 424–40.

Taghva, K., R. Elkhoury, and J. Coombs. 2005. "Arabic Stemming Without a Root Dictionary." In *Proceedings of the International Conference on Information Technology: Coding and Computing, 2005, ITCC 2005*. IIEEE: 152–57, Vol. 1.

Tang, J., Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. 2014. "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis." In *Proceedings of the 31st International Conference on Machine Learning*. JMLR, 190–98. http://jmlr.org/proceedings/papers/v32/tang14.pdf.

Vieweg, S., A.L. Hughes, K. Starbird, and L. Palen. 2010. "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '10*. New York, NY: ACM, 1079–88.

Wolfsfeld, G., E. Segev, and T. Sheafer. 2013. "Social Media and the Arab Spring: Politics Comes First." *The International Journal of Press/Politics* 18 (2): 115–37.

Xu, J., T.C. Lu, R. Compton, and D. Allen. 2014. "Civil Unrest Prediction: A Tumblr-Based Exploration." In *Social Computing, Behavioral-Cultural Modeling and Prediction*, eds. W. Kennedy, N. Agarwal, and S. Yang, Vol. 8393 of Lecture Notes in Computer Science. Switzerland: Springer International Publishing, 403–11.

Yao, L., D. Mimno, and A. McCallum. 2009. Efficient Methods for Topic Model Inference on Streaming Document Collections." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*. New York, NY: ACM, 937–46.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

GDELT Categories: CAMEO categories used by the GDELT event database.