

2014


Predicting Events Surrounding the Egyptian Revolution of 2011 Using Learning Algorithms on Micro Blog Data

Benedikt Boecking
Carnegie Mellon University

Margeret A. Hall
University of Nebraska at Omaha, mahall@unomaha.edu

Jeff Schneider
Carnegie Mellon University

Follow this and additional works at: <https://digitalcommons.unomaha.edu/interdiscipinformaticsfacproc>

 Part of the [Communication Technology and New Media Commons](#), [Political Science Commons](#), and the [Social Media Commons](#)

Please take our feedback survey at: https://unomaha.az1.qualtrics.com/jfe/form/SV_8cchtFmpDyGfBLE

Recommended Citation

Boecking, Benedikt; Hall, Margeret A.; and Schneider, Jeff, "Predicting Events Surrounding the Egyptian Revolution of 2011 Using Learning Algorithms on Micro Blog Data" (2014). *Interdisciplinary Informatics Faculty Proceedings & Presentations*. 6.

<https://digitalcommons.unomaha.edu/interdiscipinformaticsfacproc/6>

This Conference Proceeding is brought to you for free and open access by the School of Interdisciplinary Informatics at DigitalCommons@UNO. It has been accepted for inclusion in Interdisciplinary Informatics Faculty Proceedings & Presentations by an authorized administrator of DigitalCommons@UNO. For more information, please contact unodigitalcommons@unomaha.edu.

Predicting Events Surrounding the Egyptian Revolution of 2011 Using Learning Algorithms on Micro Blog Data

Benedikt Boecking

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

C3141537@UON.EDU.AU

Margeret Hall

Karlsruhe Service Research Institute (KSRI), Karlsruhe Institute of Technology, Karlsruhe, Germany

HALL@KIT.EDU

Jeff Schneider

School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

JEFF.SCHNEIDER@CS.CMU.EDU

Abstract

We aim to predict activities of political nature in Egypt which influence or reflect societal-scale behavior and beliefs by using learning algorithms on Twitter data. We focus on capturing domestic events in Egypt from November 2009 to November 2013. To this extent we study underlying communication patterns by evaluating content-based and meta-data information in classification tasks without targeting specific keywords or users. Classification is done using Support Vector Machines (SVM) and Support Distribution Machines (SDM). Latent Dirichlet Allocation (LDA) is used to create content-based input patterns for the classifiers while bags of Twitter meta-information are used with the SDM to classify meta-data features. The experiments reveal that user centric approaches based on meta-data can outperform methods employing content-based input despite the use of well established natural language processing algorithms. The results show that distributions over users-centric meta information provides an important signal when detecting and predicting events.

1. Introduction

Machine Learning algorithms for supervised and unsupervised tasks have evolved over the years to be successfully employed on structured as well as unstructured data to enable advancements in many areas such as image and voice recognition or for the navigation of autonomous ve-

hicles (Mitchell, 1997). The methods are invaluable as they help both machines and humans to gain a high-level understanding of large collections of data. Today, many algorithms have matured to a degree that they can also be used by scientists outside the field of Machine Learning. Since these algorithms learn from experience and thereby help in finding hidden patterns in datasets of varying types, the approaches have also gathered attention amongst scholars who previously used more traditional means of data analysis. More recently, scholars have also taken to using machine learning approaches on social media data to enrich or even supplement traditional data sources. The general interest of governments, businesses and researchers in social media stems from the rich variety of information contained within as well as its often public nature and ready availability. Machine learning algorithms are popular because the techniques are designed to deal with noise, are highly scalable and able to work with almost arbitrary types of data.

We aim to predict events in a timeframe of four years surrounding the Egyptian Revolution of 2011 on the basis of Twitter data. 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. In our study we target domestic events of political nature in Egypt which can be associated with either the standing government or the opposition. We further examine which input features provide the signal for event detection and prediction. Our work focuses on methods which have limited bias in gathering input data and require a low amount of parameter selection.

Section 2 will cover literature on Twitter data in relation to events and Machine Learning. Our setup for event detec-

tion and prediction is described in Section 3. The results of our experiments will be presented in Section 4 followed by a discussion of the findings.

2. Related Work

Twitter has already been the focus of numerous studies resulting in a large body of research. A look at literature concerning Twitter data and the use of Machine Learning algorithms revealed limited research on the prediction of events. The task of detecting and predicting events in general is interesting as it could be used for the benefit of the public, e.g. in tracking natural hazards, epidemics, or outbreaks of violence. It also poses a way for policy makers to understand the environment in which decisions are being made and to understand their consequences. In this section we present publications that are related to our study and focus on classification tasks, event detection, prediction tasks, and the use of Twitter during events and emergency situations as well as during 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. (Kwak et al., 2010; Lotan et al., 2011; Khondker, 2011; Wolfsfeld et al., 2013; Anderson, 2011; Starbird and Palen, 2012)). The studies provide evidence for the coverage of the Egyptian Revolution on Twitter making event prediction in a wider timeframe surrounding the revolution an appealing task. With regard to the information content of tweets during the Arab Spring (Kwak et al., 2010) show that the majority (over 85%) of trending topics they observe during their study are headline news or persistent news in nature. Further, (Starbird and Palen, 2012) examine information diffusion activity through a subset of tweets during the Egyptian Revolution of 2011 and state that the protesters were clearly using social media services to coordinate their actions and garner support as the retweeted messages contained information like required supplies, meeting times, information on injuries, and on violence. The authors extracted the 1,000 most highly retweeted users by use of popular hashtags related to the protests. They highlight that 30% of the users appear to have been in Cairo during the event, many of which were amongst 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)).

With regard to learning algorithms, Twitter data has been used in various problem settings (e.g. (Sriram et al., 2010; Sakaki et al., 2010; Asur and Huberman, 2010; Conover et al., 2011; Puniyani et al., 2010; Li et al., 2012)). The articles include tasks such as the assignment of tweets and users to categories, the inference of network structures, and

the detection of events, i.e. 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 term frequencies, measures from reconstructed networks measures, and topic modeling (see e.g. (Conover et al., 2011)), as well as sentiment analysis (see e.g. (Asur and Huberman, 2010)). Although there are no studies that directly relate to event prediction using Twitter data, prediction tasks on the basis of Twitter data have been studied in other settings. For example, (Asur and Huberman, 2010) predict box office revenues using a linear regression model simply based on term frequencies of movie names extracted from Twitter data.

In relation to events, several studies have investigated information flow and user behavior on Twitter during common emergency and crisis situations (e.g. (Vieweg et al., 2010; Starbird and Palen, 2010; Mendoza et al., 2010; Starbird and Palen, 2011; Qu et al., 2011; Starbird et al., 2014)). 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.

Some studies have already used Twitter data for event detection. (Petrović et al., 2010), in a first story detection setting, aim to detect new Twitter events as they unfold through the use of streaming algorithms. The authors present a modified locality sensitive hashing technique, a technique for finding the approximate nearest neighbor of term frequencies in vector space based on the cosine distance with a limit on the amount of comparisons of documents. To detect significant events, threads of similar tweets are uncovered and only the fastest growing threads are deemed new and noteworthy events. (Sakaki et al., 2010) train a binary classifier on tweets using a SVM for the purpose of detecting a specific target event such as typhoons and show that twitter location information can be exploited to track and map the events.

3. Method

We use binary classifiers to predict and detect event occurrences on a daily basis. The classifiers process a numerical feature input based on Twitter data and assign a qualitative output signal called label that determines the occurrence of an event. These classifiers are trained and evaluated on labeled training data. We create content based features as well as features based on meta data. This section will cover the classifiers we used and the methods followed to create features. The decision to use Twitter data was made due to the public nature of the service and the amount of data available from a sampling stream we had access to. In addition, tweets can be geotagged by their authors which allows the extraction of tweets that can be attributed to a region and make event detection within a nation's borders possi-

ble. 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 several times. The only restriction employed on the available Twitter data in this work is with regard to location information that is used to extract tweets originating within Egypt.

3.1. Classifiers

Two different but related algorithms for classification are used during the experiments. One is the Support Vector Machine (SVM) which is a nonlinear maximum margin classifier (Cortes and Vapnik, 1995). The other classifier is the Support Distribution Machine (SDM), an extension of the SVM that was chosen because it enables classification on sets using an approach that estimates divergences of unknown distributions from data samples (Poczos et al., 2012). To establish a base level of comparison for classification accuracy, the values are compared to the performance achieved by a majority classifier. This majority classifier is static and will always predict the label that belongs to the majority class of the training set.

3.1.1. SUPPORT VECTOR MACHINE

Given pairs of patterns and labels $(x_1, y_1) \dots (x_m, y_m)$ which constitute the set of training observations, the dual form of the *nonlinear soft-margin* SVM can be written as (Schölkopf and Smola, 2001):

$$\begin{aligned} \max_{\alpha \in \mathbb{R}^m} \quad & \hat{\alpha} = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j), \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, i = 1, \dots, m, \\ \text{and} \quad & \sum_{i=1}^m \alpha_i y_i = 0. \end{aligned}$$

where C is a penalty parameter, α are the Lagrange multipliers, and K is a Mercer kernel, i.e. a function that quantifies similarity of two elements of a set by assigning a distance value and which fulfills Mercer's condition (see (Schölkopf and 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.

3.1.2. 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 general-

izes kernel machines so that classification can be done on groups of data points by treating the patterns within as 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 $\{X_1, \dots, X_T\} \in \mathcal{X}$ with labels $Y_t \in \mathcal{Y} = \{\pm 1\}$. Let t -th input $X_t = \{X_{t,1}, \dots, X_{t,m_t}\}$ be m_t independent and identically distributed samples from density p_t . The objective function of the SVM dual changes to:

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

subject to $\sum_i \alpha_i y_i = 0, 0 \leq \alpha_i \leq C$. Now kernels on i.i.d. sample sets need to be defined. Kernels can be defined on the basis of the following quantity:

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

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:

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

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

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

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 (Poczos et al., 2012; Higham, 2002)).

3.2. Input Feature Creation

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

3.2.1. FEATURES FROM TEXT

Latent Dirichlet Allocation (LDA) is a widely adopted probabilistic generative topic model (Blei et al., 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, i.e. 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 $\text{Dir}_V(\eta)$ denote a V -dimensional symmetric Dirichlet distribution with scalar parameter η . Let $\text{Dir}_K(\vec{\alpha})$ 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 and Lafferty, 2009):

- (1) For each topic k ,
 - (a) draw a distribution over words $\vec{\beta}_k \sim \text{Dir}_V(\eta)$.
- (2) For each document d ,
 - (a) draw a vector of topic proportions $\vec{\theta}_d \sim \text{Dir}_K(\vec{\alpha})$
 - (b) For each word n ,
 - (i) draw a topic assignment $Z_{d,n} \sim \text{Mult}(\vec{\theta}_d), Z_{d,n} \in \{1, \dots, K\}$
 - (ii) draw a word $W_{d,n} \sim \text{Mult}(\vec{\beta}_{Z_{d,n}}), W_{d,n} \in \{1, \dots, V\}$

The hidden variables in this model are the topics $\vec{\beta}$, the proportions of topics per document $\vec{\theta}$, and the topic as-

signments per word Z . The parameter $\vec{\alpha}$ 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 et al., 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 and 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 et al., 2014) describe. To improve LDA performance, tweets are thus aggregated to produce daily input documents.

For LDA 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 number of topics overfits the data which 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 the ground truth will provide the features for the classifier that carry the hidden signal for the prediction and detection tasks most clearly. 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 and topic modeling is then conducted separately on English and Arabic sets which make up the majority of tweets. English tweets are converted to lower case 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 et al., 2005), using implementations provided by the Natural Language Toolkit (NLTK) (Bird et al., 2009) open source library. Input documents of daily tweets are created and passed on to LDA. The topic modeling output of document topic proportions is combined ex-post.

3.2.2. FEATURES FROM HASHTAGS

We also use hashtags to create feature vectors. Twitter hashtags are created by the user 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 document by 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 et al., 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.

3.2.3. FEATURES FROM META DATA

We explore different ways of using meta-data 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 author. 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 friends. User centric feature vectors per tweet are created containing any combination of the following attributes: the number of friends, number of followers, and the number of statuses sent by the respective user. Of these features, the combination of friends and followers represents degree centrality, i.e. a measure of the importance of a node in the network based on incident edges (see e.g. (Russell, 2011)). Using the SDM in these features amounts to estimating distribution divergences over degree centrality of active regions of a network. It provides a straightforward way to identify disparities in communication behavior in 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 log-normal 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.

3.3. 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 and thus a measure of magnitude with respect to these events is necessary for binary label creation. One of the few publicly accessible event databases which covers the necessary time period is the Global Database of Events, Language, and Tone (GDELT) (Leetaru and Schrodt, 2013)¹ 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). Target categories of events are selected from the different categories that GDELT captures on the basis that the actions can be associated with being in favor of either the standing government or the opposition. The event dataset is used on a highly aggregated macro level and the recorded number of mentions across news sources is used as a measure of importance. To detect spikes that deviate from the baseline of daily event mentions, a threshold is set on this value that determines the target significance. Based on that measure, a day is labeled as positive on which one or more events occur that fit the target profile and a target significance, while all other days are labeled as negative. Figure 1 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 dataset is mitigated 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 will go unnoticed. However, it also represents the approach that uses a minimal amount of assumptions. 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 dataset. A maximum window size of three days was chosen to prevent an excessive number of positive labels. If one or more days of event data within the window

¹Recently, 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.

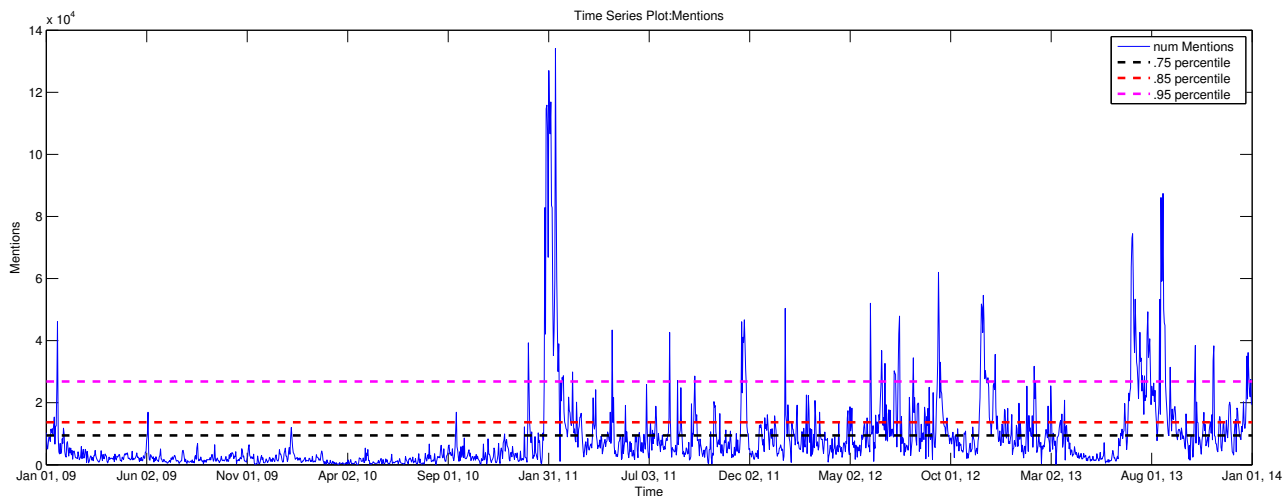


Figure 1. The smoothed number of mentions for extracted GDELT events and different thresholds for binary label creation.

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 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, i.e. 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 dataset to be positive.

3.4. Twitter Data

Sampling Twitter data through target users, keywords, or hashtags can be sensible in problem settings where a concise set of keywords or users can be anticipated. However, in the case of sampling by hashtags (González-Bailón et al., 2014) show that this strategy introduces a bias and may artificially crop a periphery of activity. Also, for the purpose of detecting and predicting events during times of social change, a predefined set of keywords or lead users could mean that relevant tweets and tweeting behavior escape the sample. 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 geo-

location to gather tweets that most likely stem from the target area while making no assumptions on their content or on user behavior. The Twitter data used for our work stems from a dataset established through sampling from the Twitter streaming API with Gardenhose streaming access at an approximate sample rate of 10% of the entire Twitter feed.² If location information for a tweet is available and determined to be in Egypt, the tweet is extracted. Unfortunately, the percentage of geotagged tweets on Twitter is very low, and it generally lies between 1 – 2%. If the tweet’s location is unavailable, the user’s location associated with their profile is used as a proxy for their current location.

1.3 Million tweets located in Egypt were 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 dataset contains messages from 57,238 unique users and the messages contain 81,253 unique hashtags.

4. Experiments

All results quoted in this section were achieved with 10-fold cross-validation. 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.



Figure 2. A word cloud showing the top 40 stemmed words of a topic from a model trained with 100 topics on documents of daily English tweets. The word size is proportionate to its probability in the topic, but not proportionate to its frequency in the corpus.

4.1. SVM Classification of Document Topic Feature Vectors

For LDA, a total of 1,042,680 tweets were used because the language guessing algorithm allocated them to be in either English or Arabic. The word cloud of Figure 2 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. As explained in the previous section, models were fitted to the corpora for different k to assess a reasonable number of topics based on classification performance. The results were evaluated across several dimensions. Features from Arabic and English tweets were classified separately as well as in combination. The performance was compared across different threshold levels for binary label creation as well as different label window sizes for detection and prediction. Figure 3 gives an example of how classification performance changes along different numbers of topics. The best results in terms of classification accuracy for both detection and prediction with a three day label window were achieved with features from a model with $k = 70$ topics. Most results from hereon will be quoted

²Brendan O'Connor at Carnegie Mellon University's School of Computer Science helped assemble the Twitter dataset.

Table 1. This table shows prediction results of SVM classification with topic modeling features. The table indicates prediction results for different window sizes in comparison to the majority classifier for a label threshold value at the 0.75 percentile.

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

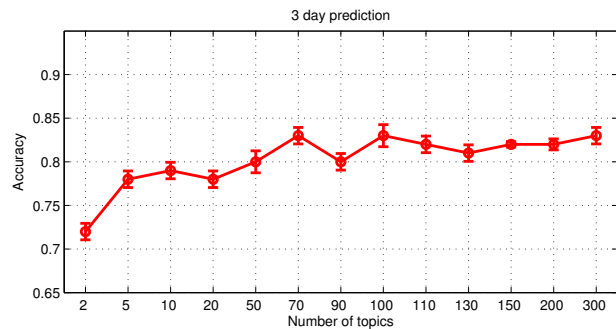
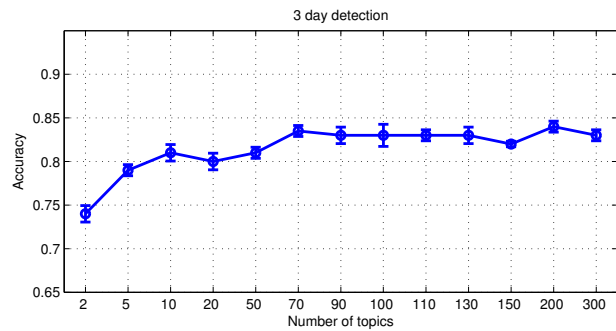


Figure 3. SVM classification accuracy of topic model features. Bars denote standard error. The top plot shows a three day detection window with the classified day in the center. The bottom plot shows a three day prediction with the classified day preceding the label window.

for features from a model with a reasonable $k = 100$ topics as the accuracy appears to stabilize around this value. 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. It is possible in this case that due to the noise both in the Twitter data as well as in the label data an increasing window to a sensible three days makes the detection and prediction task easier. The classification task was also conducted using English and Arabic topic modeling output separately as shown in Table 3. The results suggest that a combination of both signals provides significantly better classification results in terms of accuracy.

4.2. Hashtag Features

Classification results using vectors from reduced hashtag count matrices cannot match the classification performance of their topic modeling counterparts. Also, in an effort to use the SDM on similar vectors, we created count vectors using unique hashtags per hour. However, SVM classifica-

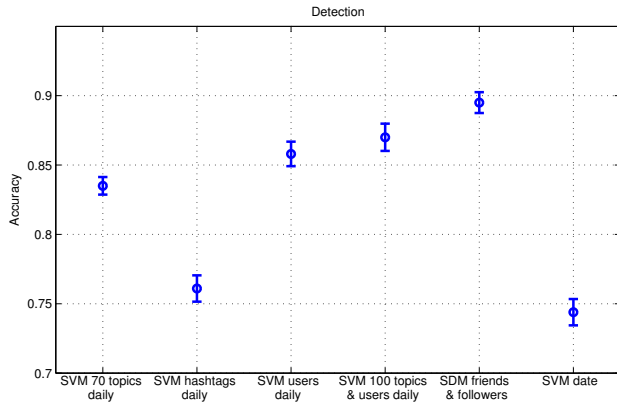


Figure 4. Summary of the best detection classification accuracies for a label window of size three, label threshold at the 0.75 percentile.

tion 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 too small a sample of feature vectors per bag and that these bags of features additionally violate the SDM assumption of having an independent and identically distributed sample input. The results can be seen in Table 2.

4.3. Meta Data 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. The feature beats the performance achieved by using topic modeling based features and hashtag-based features for a three day label window. A combination of topic features and these user features lead to another small improvement in terms of accuracy, however standard errors for these results 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 2.

4.3.1. CLASSIFICATION OF META-DATA BASED FEATURES WITH THE SUPPORT DISTRIBUTION MACHINE

The meta information that was established consistently throughout our Twitter dataset was the number of friends, the number of followers, the number of statuses sent, and message length. These features are available on a per tweet basis and 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 1-dimensional feature vector was achieved by using the number of friends (see Table 4 in

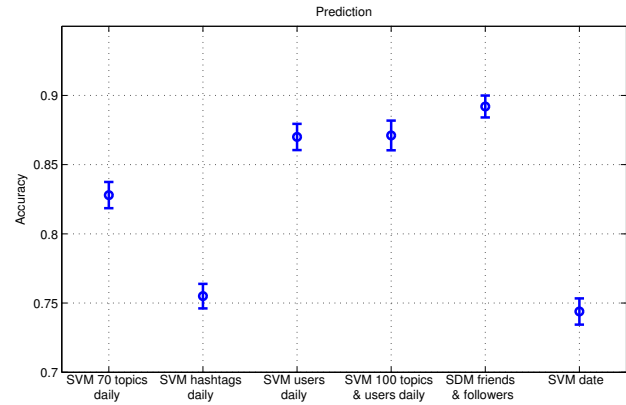


Figure 5. Summary of the best prediction classification accuracies for a label window of size three, label threshold at the 0.75 percentile.

the Appendix). 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 2-dimensional feature vector comprised of the number of friends and followers beating the classification performances of all other approaches (see Table 5 in the Appendix). A summary of the detection and prediction results in Figure 4 and Figure 5 also visualize the standard error, showing that the accuracy using this approach provides a significant improvement over previous content based efforts. Attempts of achieving similar results using the SVM on daily feature vectors of the mean and variance as well as distribution parameters of a lognormal distribution fitted to the daily meta data did not achieve competitive results.

5. 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 in so far successful as they comfortably outmatch a time series feature and a majority classifier in terms of classification accuracy. Interestingly, distribution divergences over user centric features of meta-data outperform content-based features in the prediction and detection settings. A purely content based approach of feature vectors created using topic modeling did deliver good error reduction in comparison to the majority classifier but could not match the classification performance of the best meta data based features. We have also seen that the noise in both the event dataset used for label creation and the Twitter dataset makes exact detection and prediction a difficult task. With a label window spanning three days, the error reductions

Table 2. This table shows SVM classification accuracies of hashtag features, user ID based features, topic model based features, and a combination of topic model and user features. The label window was set to three days, the label creation threshold was set to the 0.75 percentile.

TASK	50-DIM USER	100-DIM USER	100 TOPICS	100 TOPICS & 100-DIM USER	SVM #-DAILY 200 DIM	SVM #-DAILY 100 DIM	SDM #-HOURLY 200 DIM	SDM #-HOURLY 100 DIM
PREDICTION	0.843	0.870	0.826	0.871	0.738	0.755	0.701	0.711
DETECTION	0.854	0.858	0.834	0.870	0.752	0.761	0.692	0.689

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 1, 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 dataset 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 feature. Our results show that the topic modeling based approach, as well as the best result which was achieved using meta-data, outperform this date feature comfortably.

5.1. Classification Performance using Content Features

Although content based features were outperformed by user centric features in classification performance, some valuable conclusions can be drawn. The classification performance for detection and prediction tasks improved notably when the two sets of Arabic and English topic modeling features were combined. The results show that although English is the predominant language on Twitter, the inclusion of information conveyed in other languages can have a significantly positive performance impact.

Interestingly, an attempt to exploit the topical information provided by users through hashtags neither matched the accuracies achieved by the meta data 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 compared to the full texts that describe events. 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.

5.2. Meta Data 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 that were attempted. The results suggest that the approach succeeded in identifying groups of users with similar tweeting behavior which then helped the classifier to identify days where societal scale events occurred.

Event better performance was achieved using the SDM on meta information. It provided a much simpler approach compared to all attempts previously described. The best SDM classification performance was achieved through daily bags of 2-dimensional user centric features of the number of friends and followers per tweet. This corresponds to estimating distributions of degree centrality for users tweeting per day and then calculating the divergences amongst these distributions. This method thereby indirectly measures weighted activity of nodes of the 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 statuses per user combined with their number of followers or friends. 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 friends and followers does surely have its limits in the amount of information it can convey. But generated from diverse input dataset of 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 pre-processing 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 reduced. The experiments also showed that efforts in using

meta-data 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 where bags of features are present that can be used in the SDM without any preceding steps.

5.3. Limitations

The methodology followed in our study does have some drawbacks. The labels for our classification experiments were established on the basis of machine coded events from a variety of sources of international news. When working with a machine coded event dataset issues of noise, validity, and bias arise. The creation of labels to detect events relies on the sources to capture the events 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. With regard to the best features, a problem with purely relying on meta-information is that events that are not happening in Egypt could cause similar reactions in the tweeting behavior of users such that more information is needed to distinguish the reactions. These events may more easily be confused if no content or spatial information is used additionally. Unfortunately, our content based approaches did not perform well enough on a per tweet basis to combine the two features.

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 friends 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.

5.4. 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, e.g. through a combination of machine coded events and human verification. Furthermore, tweets for our dataset were only selected by their location and no keyword based extraction was performed to avoid any form of bias in data selection and to test the different methods in their ability to deal with very noisy datasets. However, in a practical application, sophisticated keyword-based retrieval may improve detection and prediction of events, especially when the target events are more closely defined. Also, (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, e.g through identifying key users ex-post.

6. Conclusion

Predicting target events through social media data can be used to provide immediate, high-level feedback on important environmental indicators in the form of event signals to policy makers. 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 geo-location 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 dataset. 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 for 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 predict and detect 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 meta data 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. 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 may otherwise be reduced when resorting to content based approaches or the reconstruction of communication networks.

7. Appendix

Table 3. SVM accuracies of daily topic features, English and Arabic tweets separately and combined. Standard error is shown in parentheses. The label window was set to three days, the binary threshold on number of mentions was set to the 0.75 percentile.

TASK	100	100	100
	TOPICS	TOPICS	TOPICS
	ARABIC	ENGLISH	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)

Table 4. This table shows SDM classification accuracies of different bags of 1-dimensional meta features. The label window was set to three days, the label creation threshold was set to the 0.75 percentile.

TASK	NO. FRIENDS	NO. FOLLOWERS	NO. STATUSES	MESSAGE LENGTH
DETECTION	0.827	0.745	0.747	0.751
PREDICTION	0.819	0.739	0.743	0.728

Table 5. This table shows SDM classification accuracies of daily bags of a user centric feature of friends and followers. The label window was set to three days, the label creation threshold was set to the 0.75 percentile.

TASK	NO. FRIENDS & FOLLOWERS	100 TOPICS	MAJORITY
DETECTION	0.884	0.834	0.509
PREDICTION	0.895	0.826	0.509

References

- Anderson, L. (2011), ‘Demystifying the arab spring: parsing the differences between tunisia, egypt, and libya’, *Foreign Aff.* **90**, 2.
- Asur, S. and Huberman, B. A. (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*, IEEE Computer Society, Washington, DC, USA, pp. 492–499.
- Bird, S., Klein, E. and Loper, E. (2009), *Natural Language Processing with Python*, 1st edn, O’Reilly Media, Inc.
- Blei, D. M. and Lafferty, J. (2009), Topic models, in A. Srivastava and M. Sahami, eds, ‘Text Mining: Classification, Clustering, and Applications’, Taylor & Francis, London, England.
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), ‘Latent dirichlet allocation’, *the Journal of machine Learning research* **3**, 993–1022.
- Conover, M., Goncalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F. (2011), Predicting the political alignment of twitter users, in ‘Privacy, security, risk and trust (passat), 2011 IEEE third international conference on social computing (SocialCom)’, pp. 192–199.
- Cortes, C. and Vapnik, V. (1995), ‘Support-vector networks’, *Mach. Learn.* **20**(3), 273–297.
- González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J. and Moreno, Y. (2014), ‘Assessing the bias in samples of large online networks’, *Social Networks* **38**(0), 16 – 27.
- Griffiths, T. L. and Steyvers, M. (2004), ‘Finding scientific topics’, *Proceedings of the National Academy of Sciences* **101**(suppl 1), 5228–5235.
- Halko, N., Martinsson, P. G. and Tropp, J. A. (2011), ‘Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions’, *SIAM Rev.* **53**(2), 217–288.
- Higham, N. J. (2002), ‘Computing the nearest correlation matrix problem from finance’, *IMA Journal of Numerical Analysis* **22**(3), 329–343.
- Khondker, H. H. (2011), ‘Role of the new media in the arab spring’, *Globalizations* **8**(5), 675–679.
- Kwak, H., Lee, C., Park, H. and Moon, S. (2010), What is twitter, a social network or a news media?, in ‘Proceedings of the 19th International Conference on World Wide Web’, WWW ’10, ACM, New York, NY, USA, pp. 591–600.
- Leetaru, K. and Schrodt, P. (2013), ‘Gdelt: Global data on events, language, and tone, 1979-2012’.
- Li, R., Lei, K. H., Khadiwala, R. and Chang, K.-C. (2012), Tedas: A twitter-based event detection and analysis system, in ‘Data Engineering (ICDE), 2012 IEEE 28th International Conference on’, pp. 1273–1276.
- Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I. and danah boyd (2011), ‘The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions’, *International Journal of Communication* **5**(0).
- McCallum, A. K. (2002), Mallet: A machine learning for language toolkit. <http://mallet.cs.umass.edu>.
- Mendoza, M., Poblete, B. and Castillo, C. (2010), Twitter under crisis: Can we trust what we rt?, in ‘Proceedings of the First Workshop on Social Media Analytics’, SOMA ’10, ACM, New York, NY, USA, pp. 71–79.
- Mitchell, T. M. (1997), *Machine Learning*, 1 edn, McGraw-Hill, Inc., New York, NY, USA.
- Petrović, S., Osborne, M. and Lavrenko, V. (2010), Streaming first story detection with application to twitter, in ‘Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics’, HLT ’10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 181–189.

- Poczos, B., Xiong, L., Sutherland, D. J. and Schneider, J. (2012), Nonparametric kernel estimators for image classification, in 'CVPR 2012'.
- Porter, M. F. (2001), 'Snowball: A language for stemming algorithms'.
URL: <http://snowball.tartarus.org/texts/introduction.html>
- Puniyani, K., Eisenstein, J., Cohen, S. and Xing, E. P. (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, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 19–20.
- Qu, Y., Huang, C., Zhang, P. and Zhang, J. (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, ACM, New York, NY, USA, pp. 25–34.
- Racette, M. P., Smith, C. T., Cunningham, M. P., Heekin, T. A., Lemley, J. P. and Mathieu, R. S. (2014), Improving situational awareness for humanitarian logistics through predictive modeling, in 'Systems and Information Engineering Design Symposium (SIEDS), 2014', pp. 334–339.
- Russell, M. A. (2011), *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*, 1 edn, O'Reilly Media.
- Sakaki, T., Okazaki, M. and Matsuo, Y. (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, ACM, New York, NY, USA, pp. 851–860.
- Schölkopf, B. and Smola, A. J. (2001), *Learning with Kernels : Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press.
- Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M. (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, ACM, New York, NY, USA, pp. 841–842.
- Starbird, K., Maddock, J., Orand, M., Achterman, P. and Mason, R. M. (2014), Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing, in 'iConference 2014 Proceedings', pp. 654 – 662.
- Starbird, K. and Palen, L. (2010), *Pass it on?: Retweeting in mass emergency*, International Community on Information Systems for Crisis Response and Management.
- Starbird, K. and Palen, L. (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, ACM, New York, NY, USA, pp. 1071–1080.
- Starbird, K. and Palen, L. (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, ACM, New York, NY, USA, pp. 7–16.
- Steyvers, M. and Griffiths, T. (2007), Probabilistic topic models, in T. Landauer, D. McNamara, S. Dennis and W. Kintsch, eds, 'Latent Semantic Analysis: A Road to Meaning', Laurence Erlbaum.
- Taghva, K., Elkhoury, R. and Coombs, J. (2005), Arabic stemming without a root dictionary, in 'Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on', Vol. 1, pp. 152–157 Vol. 1.
- Tang, J., Meng, Z., Nguyen, X., Mei, Q. and Zhang, M. (2014), Understanding the limiting factors of topic modeling via posterior contraction analysis, in 'Proceedings of The 31st International Conference on Machine Learning', pp. 190–198.
- Vieweg, S., Hughes, A. L., Starbird, K. and Palen, L. (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, ACM, New York, NY, USA, pp. 1079–1088.
- Wolfsfeld, G., Segev, E. and Sheaffer, T. (2013), 'Social media and the arab spring: Politics comes first', *The International Journal of Press/Politics* **18**(2), 115–137.
- Yao, L., Mimno, D. and McCallum, A. (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, ACM, New York, NY, USA, pp. 937–946.