SAND2017-12340J

Temporal Methods to Detect Content-Based Anomalies in Social Media

Jacek Skryzalin, Richard Field, Jr., Andrew Fisher, and Travis Bauer

Sandia National Laboratories*, Albuquerque NM 87185, USA,
{jskryza, rvfield, anfishe, tlbauer}@sandia.gov

Abstract. We develop a method for time-dependent topic tracking and meme trending in social media. Our objective is to identify time periods whose content differs significantly from normal, and we utilize two techniques to do so. The first is an information-theoretic analysis of the distributions of terms emitted during different periods of time. In the second, we cluster documents from each time period and analyze the tightness of each clustering. We also discuss a method of combining the scores created by each technique, and we provide ample empirical analysis of our methodology on various Twitter datasets.

1 Introduction

Social media platforms (Twitter, Facebook, etc.) allow users to instantaneously publish small, textual utterances. Taken individually, these utterances might have little content and provide little information. Taken in aggregate, however, they can provide insights into, for example, public health [6], political sentiment [22], and personality [7].

We develop a framework which allows us to detect and understand temporal anomalies in a collection of timestamped documents, such as those produced on social media. More explicitly, we identify time periods during which the produced documents' content differs drastically from the norm or shows unusually high focus or intensity, but we do not place further restrictions or specifications on the nature of the anomaly. As such, we focus on *unsupervised* techniques which allow the detection of an anomalous state without a prior specification of the exact nature of the anomaly.

We discuss related research and its relationship to the current work in Section 2. In Section 3, we discuss two methods of detecting anomalous behavior, as well as a way to fuse the results of these two approaches. In Section 4, we present empirical results of our methods on various Twitter datasets.

* Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND-2017XXXX

2 Related Work

There has been a significant amount of research on the trends and dynamics of a corpus of timestamped documents. These methods have been used to study trends in a diverse collection of corpora, including those consisting of scientific papers [4, 8, 9], historical speeches [24], news stories [12, 23], and social media posts [11, 25]. Although there are numerous such techniques, they can be broken into roughly two categories.

The techniques of the first category are generally known as topic detection and tracking (TDT) algorithms. These algorithms attempt to incorporate temporal data into traditional topic modeling algorithms — algorithms whose primary purpose is to produce clusters of similar documents. Some of these algorithms use predefined categories (e.g., music news, sports news, political news, etc.) and supervised learning techniques to classify each new document into one of the predefined categories [11]. Other techniques create vectors from each document and use traditional unsupervised clustering algorithms to produce custom categories [13, 18].

Still other TDT algorithms tackle topic detection and tracking using probabilistic Bayesian modeling. These algorithms are usually based loosely on Latent Dirichlet Allocation (LDA) [5]. LDA represents a topic as a distribution over words and considers each document to have been generated by sampling from a mixture of topics. Some temporally sensitive variants of LDA partition a corpus into time intervals, run LDA on each time interval, and connect the topic distributions from each time interval with the topic distributions of neighboring time intervals [1, 4]. Other temporally sensitive variants of LDA associate each topic distribution with a temporal distribution to encourage each topic to occur in a relatively concentrated time period [24].

The second category of trend-identifying algorithms consists of techniques which provide the user with a set of memes, defined as (clusters of) important words or phrases, and the periods of time where each meme is considered especially important. Memes may or may not need to be specified in advance by the user, and the importance of a meme is typically related to the frequency of mentions per time. Various novel approaches have been developed to measure the importance of a meme. Kleinberg et al. measure the importance of a meme by fitting an infinite automaton to the temporal distribution of mentions of that meme [10, 12]. He and Parker construct a physical model of importance using proxies for a meme's mass and velocity derived from the temporal distribution of mentions of a meme and the context in which the meme occurs [9]. Swan and Allan extract important terms from temporal slices of a corpus using a χ^2 significance test [20]. Shasha et al. deem a meme m important in a time window w if a user-specified function $f(m, w)$ is greater than a user-specified threshold, and they have constructed efficient data structures and algorithms for identifying such memes [26, 27].

Although the algorithms discussed in this section successfully track temporal aspects of topics and/or identify trending memes, they tend to focus more on the topics and content being tracked and less on the relative importance of different

slices of time. In this work, we revisit the issue of topic tracking and meme trending with a temporal focus. Rather than analyzing topics themselves, we identify time periods with unusually high or anomalous trendiness. Moreover, our techniques satisfy two properties which allow them to function well with minimal prior configuration. First, our methods are completely unsupervised — the algorithms are able to function without specifying categories or memes to be tracked. Unlike the work of Pennebaker, et al. [14, 21], which focuses on the temporal correlation of pronoun usage and mental state, we discover both anomalous time periods and interesting textual markers which provide insights into the nature of the anomaly. Second, our methods are largely independent of the arrival rate of documents; we assume that any data we see has been sampled from a larger distribution, and we would like our methods to be able to accommodate differing sample sizes and sampling rates.

3 Methods

In this section, we present two techniques for studying term and topic trends from the perspective of identifying anomalous time periods. The first technique focuses on the variation of term distributions and highlights time periods whose term distributions differ drastically from baseline. The second technique uses clustering to construct a rough metric for topic coherence, which we expect to be higher when an unusually large percentage of documents share a topic.

We assume that we have time periods $t_1, t_2, \ldots, t_r$ and associated corpora $C_{t_1}, \ldots, C_{t_r}$ of documents, where C_{t_i} consists of all documents produced during time period t_i. We also assume that we have a corpus C_0 which serves as a "baseline" for our term distribution analysis. In our experiments, we use as the baseline corpus C_0 the union $C_0 = C_{t_1} \cup \cdots \cup C_{t_r}$.

3.1 Term Distribution Analysis

Our first technique utilizes information-theoretic analyses of the distributions of terms seen across varying time periods. Our analysis begins with Zipf's law — the observation that the nth most common word in a corpus occurs with frequency proportional to $n^{-\alpha}$ for some $\alpha > 0$ [16]. The parameter α varies based on language and corpus type (research articles, Twitter posts, etc.), yet α is surprisingly constant across different corpora of the same type. However, the distribution of terms in a corpus can vary widely, and it is this variation that we analyze.

First, for each term w, we construct a probability $p(w)$ (resp. $q(w)$) associated with the term w and some corpus C_t (resp. C_0) via one of the following:

- Document frequency: $p(w)$ is the proportion of documents in C_t containing w.
- Term frequency: $p(w)$ is the proportion of all terms in C_t which are equal to w.

– Weighted term frequency: $p(w)$ is a document-weighted proportion of all terms in C_t constructed so that all documents are weighted equally, i.e.,

$$p(w) = \frac{1}{|C_t|} \sum_{d \in C_t} \frac{\text{number of words in } d \text{ equal to } w}{\text{number of words in } d}, \tag{1}$$

where $|C_t|$ denotes the number of documents in C_t.

We note that the values $\{p(w)\}_w$ form a distribution (i.e., they are nonnegative and sum to 1) when defined using the term frequency or weighted term frequency option, but not when defined using the document frequency option. The non-weighted and weighted term frequency definitions differ in that the term frequency option assigns equal weight to each term, whereas the weighted term frequency option assigns equal weight to each document.

The Kullback-Leibler divergence $\mathrm{KL}(p\|q)$ between $\{p(w)\}_w$ and $\{q(w)\}_w$ is defined as

$$\mathrm{KL}(p\|q) = \sum_w p(w) \log\left(\frac{p(w)}{q(w)}\right). \tag{2}$$

The Kullback-Leibler divergence is an asymmetric measure of the difference between two probability distributions which measures the number of extra bits needed to encode p when using a coding scheme optimized for q rather than a coding scheme optimized for p.

Since the Kullback-Leibler divergence is asymmetric, it is common practice to use the Jensen-Shannon divergence, a symmetrized version of the Kullback-Leibler divergence, when constructing a distance metric on probability distributions. However, we define an antisymmetric version of the Kullback-Leibler divergence via

$$\mathrm{AKL}(p\|q) = \mathrm{KL}(p\|q) - \mathrm{KL}(q\|p) = \sum_w (p(w) + q(w)) \log\left(\frac{p(w)}{q(w)}\right). \tag{3}$$

When analyzing the trends of a corpus C_t, we find it most useful to analyze the term-wise contributions to $\mathrm{AKL}(p\|q)$. We thus define the pointwise antisymmetric Kullback-Leibler (PAKL) score of a term w to be

$$\mathrm{PAKL}_{(p\|q)}(w) = (p(w) + q(w)) \log\left(\frac{p(w)}{q(w)}\right). \tag{4}$$

The value $\mathrm{PAKL}_{(p\|q)}(w)$ satisfies the following properties:

1. $\mathrm{PAKL}_{(p\|q)}(w)$ is positive if $p(w) > q(w)$ and is negative if $p(w) < q(w)$.
2. $\mathrm{PAKL}_{(p\|q)}(w)$ approaches zero as $p(w)$ approaches $q(w)$.
3. $\left|\mathrm{PAKL}_{(p\|q)}(w)\right|$ increases as either (a) $p(w)$ stays constant and $q(w)$ approaches 0, or (b) $q(w)$ stays constant and $p(w)$ approaches 0.

Thus, when analyzing a set of timestamped corpora, we can monitor the time evolution of PAKL scores to determine whether the relative frequency of a term is increasing, decreasing, or staying constant in time. We can also sum all PAKL scores, all positive (resp. negative) PAKL scores, or the most n positive (resp. negative) PAKL scores in each corpus C_t in order to construct a score which measures the relative trendiness (or, in the case when relative common words experience a drop in usage, anti-trendiness) exuded by C_t.

We note that the third property listed above is key to our analysis. If we had instead chosen to analyze a "pointwise" version of the Kullback-Leibler divergence, we might have defined a score $\mathrm{PKL}_{(p\|q)}(w)$ via

$$\mathrm{PKL}_{(p\|q)}(w) = p(w) \log \left(\frac{p(w)}{q(w)} \right). \tag{5}$$

Note, however, that $\mathrm{PKL}_{(p\|q)}(w)$ is unable to differentiate between terms w such that $p(w) \approx q(w)$ and terms w such that $p(w) \approx 0$, since in both cases, $\mathrm{PKL}_{(p\|q)}(w) \approx 0$.

3.2 Cluster Coherence

Our second topic-based approach to the temporal analysis of a series of corpora is based on the idea that we can construct tighter clusters of documents during a time period when there is a heightened focus on a relatively small set of concepts. The procedure for this technique is as follows:

1. Obtain (GloVe) word vectors for the data.
2. Using the word vectors, derive a set of "corpus vectors" to represent the data in C_t.
3. Cluster the corpus vectors.
4. Obtain scores from the clustering which measure cluster coherence and tightness.

For the first step, we train GloVe vectors on a relatively large corpus consisting of data similar to the data we'll be analyzing. GloVe is an algorithm which uses co-occurrence statistics of the terms in a corpus with a weighted least-squares model in order to derive a vector for each term in a corpus such that similar terms are associated with vectors with high cosine similarity [15]. The authors of GloVe have different objectives (synonym detection and analogy completion) for their vectors and find that 300-dimensional vectors are optimal for their tasks. Such vectors are too large for our purposes. Since the ultimate goal of these vectors is to construct and cluster a set of vectors from C_t, the dimensionality of the vectors should be sufficiently small so as not to be hindered by the curse of dimensionality (i.e., the idea that as dimensionality grows, the distance between any two randomly chosen points on the unit sphere approaches $\sqrt{2}$).

In the second step, we derive a set of vectors to represent the content of the target corpus C_t. We describe the method we use here, although other methods

are possible. For each document $d \in C_t$, we construct a "document vector" $v(d)$ by taking a weighted and normalized sum of the word vectors for words occurring in d. Explicitly, we define

$$\tilde{v}(d) = \sum_{w \in d} \mathrm{tf}_d(w)\, \mathrm{idf}_{C_0}(w), \tag{6}$$

where $\mathrm{tf}_d(w)$ denotes the number of times the term w occurs in document d, and $\mathrm{idf}_{C_0}(w)$ denotes a smoothed version of inverse document frequency of w in C_0:

$$\mathrm{idf}_{C_0}(w) = \log\left(\frac{1 + |C_0|}{1 + |\{d \in C_0 \ \mid \ w \in d\}|}\right), \tag{7}$$

where $|\{d \in C_0 \ \mid \ w \in d\}|$ represents the number of documents in C_0 containing w.

We define the document vector for d as a normalized version of $\tilde{v}(d)$ defined by Eq. (6), i.e.,

$$v(d) = \frac{\tilde{v}(d)}{\|\tilde{v}(d)\|}. \tag{8}$$

This normalization reflects our belief that documents with similar content but differing lengths should be treated as similar. Finally, we use the set of document vectors $v(d)$ as our set of "corpus vectors."

In the third step, we cluster the corpus vectors. Because all of our vectors have unit length, standard Gaussian or Euclidean clusterings are not appropriate. Instead, we consider three variants of von Mises-Fisher (VMF) clustering, which are described at length in [2] and [3]. The VMF distribution is defined as the restriction to the unit sphere of a multivariate Gaussian distribution whose covariance matrix is a multiple of the identity. The probability density function of a VMF distribution with location μ (where $\|\mu\| = 1$) and concentration $\kappa \geq 0$ is given by

$$p(x; \mu, \kappa) \propto \exp\left[\kappa \mu^{\intercal} x\right]. \tag{9}$$

The vector μ is analogous to the mean of a multivariate normal distribution, and the parameter $\kappa \geq 0$ is analogous to the inverse of the variance of a normal distribution. We consider three VMF mixture models:

1. Spherical k-means clustering. Spherical k-means clustering can be reinterpreted as a hard VMF mixture model where all mixture components are forced to have the same concentration [3].
2. Hard VMF mixture model. In this model, we fit to our data a mixture of VMF components with an underlying assumption that each data point can belong to only one mixture component.
3. Soft VMF mixture model. In this model, we fit to our data a mixture of VMF components with an assumption that each datum could have been drawn (with varying probability) from any mixture component.

In the fourth step, we construct scores which measure cluster coherence and tightness. The scores that we generate are dependent on which VMF mixture model we use. There are multiple such measures; we list here only the most promising:

- The concentration κ derived from reinterpreting spherical k-means clustering as a VMF mixture model.
- The median concentration parameter from both the hard and soft VMF mixture models. After fitting to our data a mixture model consisting of k mixture components, we collect the set of concentration scores. Empirical evidence suggests that the median of the concentration scores is higher on days when relatively few topics are receiving heightened interest. We also considered the first and third quartiles as potential scores, but the signal provided by these values is comparatively weak.
- The lognormal location of the concentration parameters from the VMF mixture models. After constructing the set of concentration scores discussed above, we first discard any outlier concentration scores. Empirically, we have found that very high concentration scores result when we have a corpus with many highly similar documents. We next fit a lognormal distribution to the set of remaining concentration scores. A variable X has a lognormal distribution if $\ln(X) \sim \mathcal{N}(\mu, \sigma)$ (i.e., when $\ln(X)$ is normally distributed with mean μ and standard deviation σ). The values μ and σ are typically referred to as the location and scale of the lognormal distribution, respectively. We have found that the location parameter is typically higher on days when relatively few topics are receiving heightened interest, although this effect is more pronounced with a hard VMF mixture model than a soft VMF mixture model.

In all three methods detailed above, we rely on the techniques and formulae presented and explained in detail in [3, 19].

Remark: In future work, we would like to incorporate various successful time-sensitive Bayesian topic models into our framework [1, 4]. Bayesian topic models are typically learned using one of two techniques — Gibbs sampling and variational inference. When trained with variational inference, Bayesian topic models provide *distributions* over parameter estimates. Just as we find the concentration scores of our von Mises-Fisher mixture models helpful in identifying anomalous time periods, so too could we utilize the covariance matrices of the posterior parameter distributions in our analysis. For example, we hypothesize that the variance $var(X)$ of each parameter X in the posterior distribution is inversely proportional to the trendiness exhibited by the set of documents in the corpus.

3.3 Weighted Probabilistic Fusion

In Sections 3.1 and 3.2, we discussed numerous techniques for generating scores. In this section, we discuss a promising technique for fusing together various scores. Our technique is almost identical to that discussed in [17].

Empirical evidence suggests that our score generating techniques suffer from a lower-than-desired signal-to-noise ratio, and that the scores produced by any one technique are typically not normally distributed. As such, it would be inappropriate to use fusion techniques which return the weighted average or maximum of normalized scores as is done in other contexts. Instead, our fusion technique incorporates estimates of the various score distributions.

For each corpus C_t, we assume that we have generated m different scores $z_{t,1}, \ldots, z_{t,m}$ from one of the techniques discussed in Sections 3.1 and 3.2. We assume that the values $\{z_{t,j}\}_t$ are sampled from some distribution Z_j with cumulative distribution function (cdf) F_j. Since the true cdf F_j is not known, we approximate F_j using either the empirical cdf $F_t^{(\text{emp})}$ or by using the cdf $F_t^{(\beta)}$ of a beta distribution fit to the scores $\{z_{t,j}\}_t$ (after scaling the $z_{t,j}$ to lie strictly between 0 and 1). Empirical evidence suggests that our fusion technique produces a greater number of significant events when using $F_t^{(\text{emp})}$ than when using $F_t^{(\beta)}$.

Our fusion technique involves three steps:

1. For all scores of type j, construct a cdf F_j as described above.
2. For each time period t, construct a fused score s_t via

$$s_t = -\sum_{j=1}^{m} c_j \log\left(1 - F_j(z_{t,j})\right), \tag{10}$$

   where $c_j > 0$ denotes the relative weight we wish to give the jth score generating technique.
3. Fit a gamma distribution with cdf G to the set of fused scores $\{s_1, ..., s_n\}$. For any given time period t, the value $G(s_t)$ now quantifies the significance of the events occurring during t.

Our model assumes stationarity; that is, each cdf F_j is assumed to be time invariant. If our data spans a sufficiently large period of time, this assumption may be inappropriate. In such circumstances, we modify step (1) above and fit a separate cdf $F_{t,j}$ for each score j and time period t from the scores $\{z_{\tau,j}\}_\tau$, where τ ranges over a set of time periods which are temporally proximal to the target time period t. In step (2), we then calculate s_t using the cdfs $\{F_{t,j}\}_j$. Step (3) remains unchanged. We call the fusion technique described in this paragraph "windowed fusion" in contrast to the original "global fusion" technique presented in the enumerated list above.

We now give a rough justification of our empirically successful fusion method, recognizing that the assumptions made in our justification may be invalid in a real-world scenario. If we assume that the set of scores $\{z_{t,j}\}_{t,j}$ have been independently sampled (where $z_{t,j}$ has been sampled from a distribution with cdf F_j), then the values $\{-\log\left(1 - F_j(z_{t,j})\right)\}_{t,j}$ are iid samples from an exponential distribution. If we additionally assume that $c_j = 1$ for all j, then the values s_t are iid samples from a gamma distribution (because the sum of independent exponential random variables is a Gamma random variable).

We note that, in general, each F_j only approximates the true cdf of the corresponding score distribution, and, for any fixed time period t, the scores $\{z_{t,j}\}_j$ are far from independent. In fact, we rely on the assumption that during an anomalous time period t, all $z_{t,j}$ will be abnormally high. Furthermore, we may want to choose our score weights c_j to be nonuniform. In our experiments, we often choose c_j so that the scores generated from term distribution analysis (Section 3.1) have combined weight equal to that of the scores generated by analyzing cluster coherence (Section 3.2).

4 Experiments

Our overall motivating goal — finding content-based anomalies in temporal segments of a corpus of social media posts — is somewhat vague and underspecified. We have thus chosen to focus our analysis on the somewhat more tractable goal of finding time periods exhibiting unusually high trendiness. Yet even with this specification, we suffer not only from a lack of a clear and unambiguous definition of "trendiness" (although we have chosen to use an information-theoretic definition of "anomaly" and cluster coherence as proxies), but also from the absence of data with incontrovertible ground truth with labeled anomalous time periods. Nevertheless, we present the results of applying our methods on multiple diverse Twitter datasets to demonstrate the capabilities of the proposed algorithm.

4.1 Data

We first apply our algorithm to relatively small subsamples of the Twitter Streaming API, a free public stream consisting of social media posts containing at most 140 characters. In total, four datasets are considered. The first, referred to as *TwitterParisEnglish*, consists of 50,000 tweets per day sampled uniformly at random from all English tweets from the Twitter Streaming API from October 11, 2015 to November 29, 2015. The second dataset, *TwitterParisFrench*, consists of 53,000 tweets per day sampled uniformly at random from all French tweets from the Twitter Streaming API from October 16, 2015 to November 29, 2015. Note that the sampling period for both these datasets includes both November 13, 2015, the date of major terrorist attacks in Paris, France, and November 26, 2015, the date of the United States holiday Thanksgiving.

We next apply our algorithm to datasets consisting of all tweets emitted by specified users during a specified timeframe constructed using the Twitter Search API. In particular, we construct a dataset *TwitterUSUniversities* by collecting all 4.2 million tweets emitted from official Twitter accounts of 2,300 United States universities from May 2014 to December 2016. We further construct a dataset *TwitterOlympics* by collecting all 1.1 million tweets emitted from the accounts of 1,200 Olympians and Olympics professionals (e.g., coaches, sports journalists) from October 2014 to December 2016.

For the analysis of all our Twitter datasets except *TwitterParisFrench*, we use 25-dimensional GloVe vectors trained on roughly 50 million English tweets

sampled from the Twitter Streaming API from March, 2015 to July, 2015. For *TwitterParisFrench*, we use 25-dimensional GloVe vectors trained on roughly 5 million French tweets sampled from the Twitter Streaming API from January, 2015 to August, 2015. Note that the GloVe vectors we use are trained on tweets temporally separated from the *TwitterParisEnglish* and *TwitterParisFrench* datasets by a period of at least two months. We also feel that *TwitterUSUniversities* and *TwitterOlympics* are largely independent from the data used to train the GloVe vectors.

4.2 Results

We first run a PAKL analysis (cf. Section 3.1) for our *TwitterParisEnglish* dataset using the "document frequency" option. We segment our corpus by day, and for the analysis of day t, we consider only terms which occur at least 5 times in C_t and 20 times in the entire corpus. The terms with the highest PAKL scores for select days can be seen in Table 1. We include terms from both uneventful days (Oct. 26, 2015 and Nov. 4, 2015) and anomalous days (Nov. 13, 2015 and Nov. 26, 2015). For the anomalous days, we can successfully find terms of interest. Note also that the top PAKL scores for anomalous days tend to be higher than those for normal days.

We also wish to mention that on November 26, 2015, roughly 2% of our tweets mention "#mtvstars," "Britney Spears," and "Lana Del Rey." A post hoc analysis has revealed that the vast majority (over 98%) of these tweets were posted by accounts that are now suspended for violating the Twitter Rules. Even so, other terms associated with Thanksgiving, including "family," "turkey," and "#imthankfulfor," are included in the 20 highest scoring terms for November 26, 2015.

We also score each document $d \in C_t$ using the term PAKL scores for C_t via

$$\text{score}(d) = \frac{\ln(|d|)}{|d|} \sum_{w \in d} \text{PAKL}(w). \tag{11}$$

We report the top two documents for select days in Table 2. For anomalous days, these documents successfully capture the nature of the day's anomaly.

In order to test our methods' robustness to corpora of different sizes, we create subcorpora of *TwitterParisEnglish* containing 10,000, 20,000, 30,000, and 40,000 tweets per day. We plot the sum of all positive PAKL scores for each day in Fig. 1. We find that varying the number of tweets considered causes surprisingly little variation in the score. Similarly, we plot the concentration score for spherical k-means clustering (with $k = 50$) in Fig. 2. Although the clustering scores are less robust to the number of tweets considered each day than the PAKL scores, they still maintain a level of robustness sufficient to identify anomalous time periods with high confidence.

Fig. 3 shows the first, second, and third quartiles of the concentration scores for a hard VMF mixture model with 50 mixture components for the *TwitterParisEnglish* dataset. Similar graphs, not shown here, were produced for the

Table 1. Top words for select days and their associated PAKL scores from *Twitter-ParisEnglish*.

Oct. 26, 2015		Nov. 4, 2015	
forevermore	0.0286	#aldub16thweeksary	0.0203
#pushawardslizquens	0.0154	i	0.0110
#aldubpredictions	0.0149	#showtimehousemates	0.0103
the	0.0149	that	0.0085
#aldubnewbeginnings	0.0129	it	0.0083
everydayiloveyou	0.0142	#otwolmanilainlove	0.0082
#everydayilov...	0.0105	to	0.0078
#otwolhappytimes	0.0104	#cmaawards	0.0076
i	0.0096	#aldubnewcharacter	0.0076
you	0.0092	a	0.0075

Nov. 13, 2015		Nov. 26, 2015	
paris	0.1448	thanksgiving	0.1743
in	0.0682	thankful	0.1159
#prayforparis	0.0582	happy	0.0692
the	0.0572	#mtvstars	0.0602
#madeintheam	0.0485	for	0.0402
#aldubhappybdaylola	0.0392	britney	0.0343
is	0.0381	spears	0.0342
#paris	0.0341	rey	0.0322
and	0.0312	lana	0.0321
prayers	0.0307	del	0.0315

Table 2. Top documents for select days from *TwitterParisEnglish*.

Oct. 26, 2015	Nov. 4, 2015
EVERYDAYILOVEYOU Forevermore in the night #PushAwardsLizQuens	I'm chillin I'm good I'm straight
I LOOVE EVERYDAYILOVEYOU Forevermore #PushAwardsLizQuens	I don't know, that that's a thing that I know.

Nov. 13, 2015	Nov. 26, 2015
Sending prayers to the people in Paris #PrayForParis	thankful for everything <emoji> Happy Thanksgiving
My thoughts and prayers go out the victims in the shootings in Paris #Prayers4Paris	<emoji> Happy Thanksgiving <emoji>

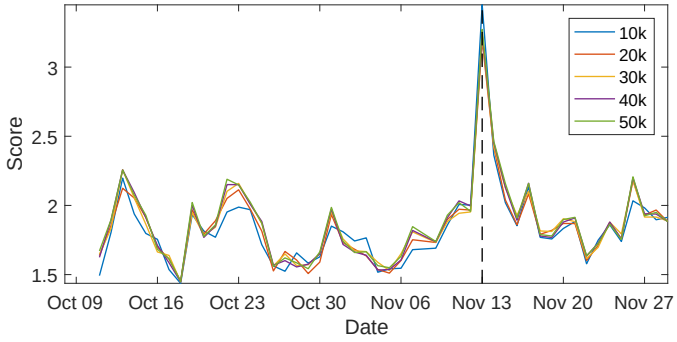

Fig. 1. PAKL scores per day for corpora of varying size taken from *TwitterParisEnglish*.

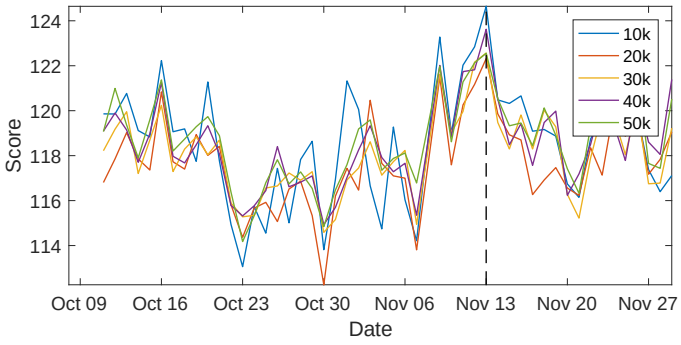

Fig. 2. Cluster scores per day for corpora of varying size taken from *TwitterParisEnglish*.

other datasets mentioned above. We normalize the scores from each quartile for a fair comparison of score quality. As mentioned previously, only the second quartile appears to be a good predictor of anomalous time periods.

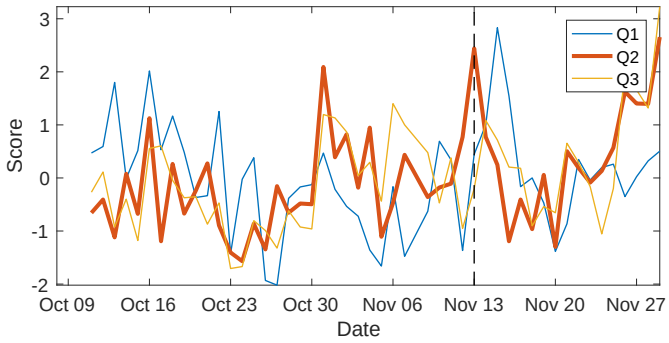

Fig. 3. VMF mixture cluster scores per day for *TwitterParisEnglish*.

We have experimented with varying the number of clusters k between 10 and 100. For k in this range, the effects of changing k are noticeable but relatively insignificant. In general, as k decreases, clustering scores become both more resilient to changing dataset size and less noisy (the randomness inherent in many clustering algorithms creates a lack of uniformity in clustering scores across different clusterings of the same data). Unfortunately, the quality of the clustering scores also tends to decrease with decreasing k.

We also present graphs produced by fusing PAKL scores with clustering scores. Unless otherwise noted, fusion for these datasets is done via $F^{(\beta)}$, and the cdfs used during fusion are calculated from the entire dataset, rather than from windows around the target time periods.

For the *TwitterParisEnglish* and *TwitterParisFrench* datasets, we construct four PAKL scores by summing, for each day, all PAKL scores, all positive PAKL scores, the highest 200 PAKL scores, and the highest 50 PAKL scores. We also construct fifteen cluster scores: we run each clustering algorithm (spherical k-means, hard VMF, soft VMF) three times with $k = 50$ clusters. From the spherical k-means clusterings, we record the concentration. From the VMF mixture models, we collect the lognormal location parameter and the median concentration. We weight the scores so that the PAKL and clustering scores each account for 50% of the total fused score. The fused scores for *TwitterParisEnglish* (resp. *TwitterParisFrench*) are shown in Fig. 4 (resp. Fig. 5). Dashed vertical lines denote the date of the Paris attacks and Thanksgiving, and a dashed horizontal line indicates the 10% significance level.

For the *TwitterOlympics* and *TwitterUSUniversities* datasets, fusion is performed similarly with the following changes to account for the fact that these

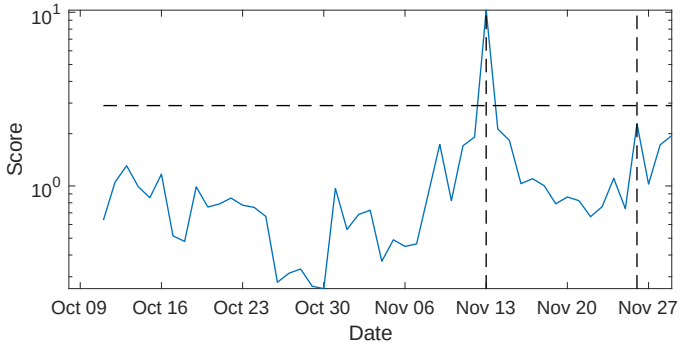

Fig. 4. Fused scores per day for *TwitterParisEnglish*.

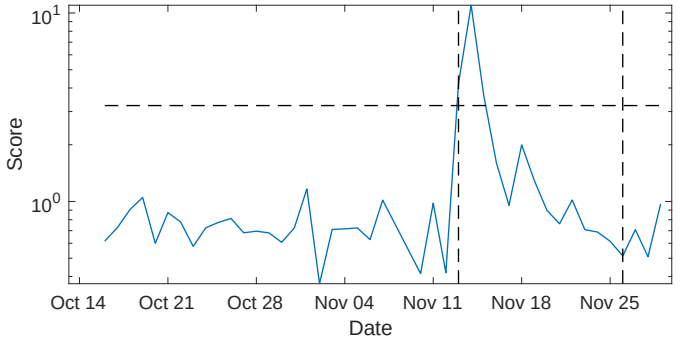

Fig. 5. Fused scores per day for *TwitterParisFrench*.

corpora are smaller in general than *TwitterParisEnglish* and *TwitterParisFrench*. First, we segment these corpora by week rather than by day. We also construct four PAKL scores, but construct scores by summing the highest 100 and 20 PAKL scores instead of the highest 200 and 50 scores as above. Finally, we run our clustering score generators with $k = 25$ instead of $k = 50$. We again use dashed horizontal lines to indicate the 10% significance level.

For *TwitterOlympics*, we produce fused scores using both $F^{(\beta)}$ (Fig. 6) and $F^{(\mathrm{emp})}$ (Fig. 7). Although these graphs have very similar shapes, fusion using $F^{(\beta)}$ tends to produce fewer significant events than fusion using $F^{(\mathrm{emp})}$. The three periods in Fig. 6 with significant scores correspond to the various athletic events in August 2015, the 2016 Olympic trials, and the 2016 Summer Olympics.

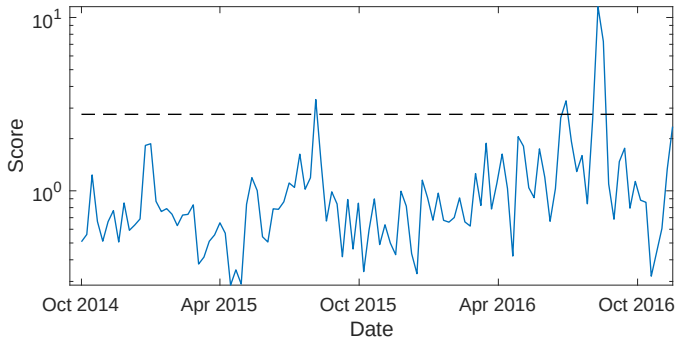

Fig. 6. Fused scores for *TwitterOlympics* using $F^{(\beta)}$.

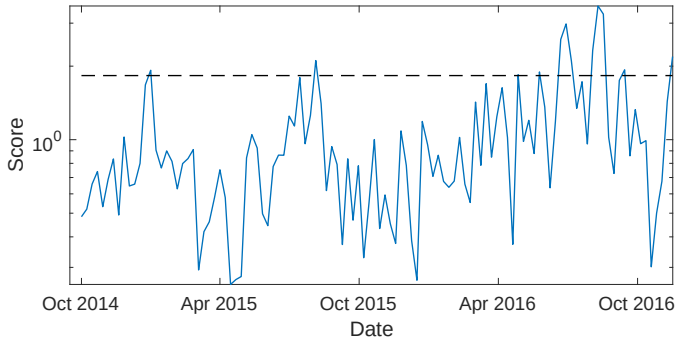

Fig. 7. Fused scores for *TwitterOlympics* using $F^{(\mathrm{emp})}$.

For *TwitterUSUniversities*, we present a graph of the fused scores based on the entire corpus (Fig. 8). We note that the Twitter feeds of many US universities changed drastically between May 2014 and November 2016. Although our algorithms are robust to changing corpus size and sampling rates, they are not robust to underlying changes in *behavior*. For example, the rate of tweet production nearly triples throughout our period of collection. Although this first appears to be a change in corpus size, we see upon further inspection that it is a change in behavior, and thus, a violation of the assumption of stationarity — later in our collection period, universities are more likely to tweet about less pressing matters, so significant events receive less attention in general. Consequently, no time periods register as significant in the latter temporal half of this corpus when using global fusion.

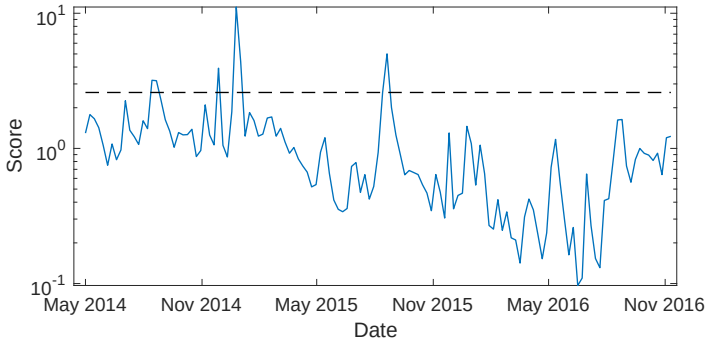

Fig. 8. Fused scores for *TwitterUSUniversities* during 2014–2016 using global fusion.

However, a much clearer pattern emerges when using windowed fusion to analyze *TwitterUSUniversities* (Fig. 9). For this analysis, we fit cdfs $F_{t,j}$ for the jth score generating technique and time period t from the scores generated by the jth score generating technique for the 15 time periods before t and the 15 time periods after t. With this modification, we see peaks for both the 2014-2015 school year and the 2015-2016 school year corresponding to the beginning of the school year, Thanksgiving break, Winter break, and the end of the school year.

We note that we have found it beneficial to fuse the clustering scores with the PAKL scores, rather than relying on either alone. For example, the first peak in Fig. 6 corresponding to the August 2015 athletic events can be attributed more to clustering scores than PAKL scores. During this event, PAKL scores barely rise above baseline; since each sport has its own world championship, the difference in term distribution from baseline is no more than expected. However, cluster coherence is particularly high during this timeframe due to the large percentage of tweets related to competition.

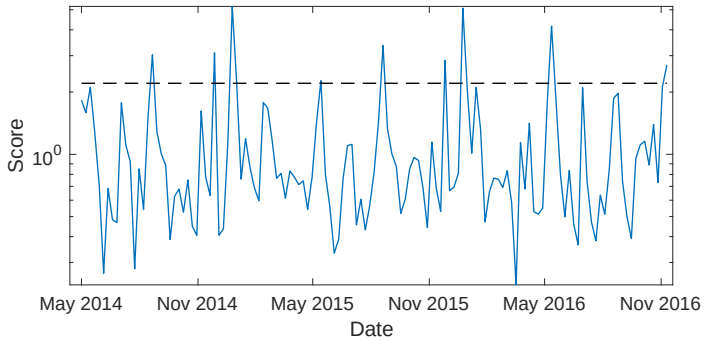

Fig. 9. Fused scores for *TwitterUSUniversities* during 2014–2016 using windowed fusion.

5 Conclusion

We have introduced two techniques which merge anomaly detection with topic detection and tracking. Our first technique relies on an information-theoretic examination of the term distributions of corpora collected over time. Our second approach produces a set of values which serve as measures for the homogeneity of the contents of the corpus. For sufficiently large corpora, both techniques are agnostic to the size of the corpus. We then explain how the scores produced from our techniques can be combined to form a single summary score. We demonstrate our algorithms on various Twitter datasets and conclude that our techniques are successful in identifying portions of a corpus with unusual and interestingly high trendiness.

References

1. AlSumait, L., Barbará, D., Domeniconi, C.: On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. pp. 3–12 (2008)
2. Banerjee, A., Dhillon, I., Ghosh, J., Sra, S.: Generative model-based clustering of directional data. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 19–28 (2003)
3. Banerjee, A., Dhillon, I.S., Ghosh, J., Sra, S.: Clustering on the unit hypersphere using von Mises-Fisher distributions. J. Mach. Learn. Res. 6, 1345–1382 (Dec 2005)
4. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine learning. pp. 113–120 (2006)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003)
6. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: A content-based approach to geo-locating Twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. pp. 759–768 (2010)

7. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: CHI '11 Extended Abstracts on Human Factors in Computing Systems. pp. 253–262 (2011)
8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101(Suppl. 1), 5228–5235 (April 2004)
9. He, D., Parker, D.S.: Topic dynamics: An alternative model of bursts in streams of topics. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 443–452 (2010)
10. Kleinberg, J.: Bursty and hierarchical structure in streams. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 91–101 (2002)
11. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.: Twitter trending topic classification. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 251–258 (2011)
12. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 497–506 (2009)
13. Morinaga, S., Yamanishi, K.: Tracking dynamics of topic trends using a finite mixture model. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 811–816 (2004)
14. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. Annual review of psychology 54(1), 547–577 (2003)
15. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
16. Piantadosi, S.T.: Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic bulletin & review 21(5), 1112–1130 (2014)
17. Simonson, K.: Probabilistic fusion of ATR results. Tech. Rep. SAND98-1699, Sandia National Laboratories (SNL-NM), Albuquerque, NM (1998)
18. Spinosa, E.J., de Leon F. de Carvalho, A.P., Gama, J.a.: Olindda: A cluster-based approach for detecting novelty and concept drift in data streams. In: Proceedings of the 2007 ACM Symposium on Applied Computing. pp. 448–452 (2007)
19. Sra, S.: A short note on parameter approximation for von Mises-Fisher distributions: And a fast implementation of $i_s(x)$. Comput. Stat. 27(1), 177–190 (Mar 2012)
20. Swan, R., Allan, J.: Extracting significant time varying features from text. In: Proceedings of the 8th International Conference on Information Knowledge Management. pp. 38–45 (1999)
21. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology 29(1), 24–54 (2010)
22. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with Twitter: What 140 characters reveal about political sentiment. ICWSM 10(1), 178–185 (2010)
23. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In: Uncertainty in Artificial Intelligence (UAI). pp. 579–586 (2008)
24. Wang, X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 424–433 (2006)

25. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. pp. 177–186 (2011)
26. Zhang, X., Shasha, D.: Better burst detection. In: Proceedings of the 22nd International Conference on Data Engineering. p. 146 (2006)
27. Zhu, Y., Shasha, D.: Efficient elastic burst detection in data streams. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 336–345 (2003)