SAND2017-6231C

Temporal Anomaly Detection in Social Media

Jacek Skryzalin, Richard Field, Andrew Fisher, and Travis Bauer
Sandia National Laboratories
Albuquerque, NM 87123
Email: {jskryza, rvfield, anfishe, tlbauer}@sandia.gov

***Abstract*—In this work, we approach topic tracking and meme trending in social media with a temporal focus; rather than analyzing topics, we aim to identify time periods whose content differs significantly from normal. We detail two approaches. The first is an information-theoretic analysis of the distributions of terms emitted during each time period. In the second, we cluster the 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.**

I. 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. However, taken in aggregate, they can provide insights into, for example, public health [1], political sentiment [2], and personality [3].

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.

II. Related Work

There has been a significant amount of research aimed at analyzing 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, historical speeches, news stories, and social media posts [4]–[13]. Although there are numerous such techniques, they can be broken into roughly two categories.

The techniques of the first category, generally known as topic detection and tracking (TDT) algorithms, incorporate temporal data into traditional topic modeling algorithms. Some of these algorithms use predefined categories and supervised learning techniques to classify documents into one of the predefined categories [4]. Other techniques create vectors from each document and use traditional unsupervised clustering algorithms to produce custom categories [5]. Still other TDT algorithms tackle topic detection and tracking using probabilistic Bayesian modeling techniques. Some of these Bayesian models reflect a belief that topics change slightly over time [6]–[8]; others associate each topic distribution with a temporal distribution to reflect an assumption that topics experience temporal birth and death [9].

The second category of trend-identifying algorithms consists of techniques which generate a set of memes and the periods of time when each meme is considered important. For example, Kleinberg et al. measure the importance of a meme by fitting an infinite automaton to the temporal distribution of mentions of that meme [10], [11]. 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 [12]. Swan and Allan extract important terms from temporal slices of a corpus using a χ^2 significance test [13].

Unlike the methods developed in this paper, the algorithms discussed above tend to focus more on the topics and content being tracked and less on the relative importance of different periods of time. Our focus is the identification of 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 unsupervised; instead of specifying categories or memes to be tracked, 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.

III. Methods

In this section, we discuss 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 $C_0 = C_{t_1} \cup \cdots \cup C_{t_r}$.

A. Term distribution analysis

Our first technique is an information-theoretic analysis of the distributions of terms seen across varying time periods.

For each term w, we construct a probability $p_t(w)$ associated with the term w and the corpus C_t via one of the following:

- Document frequency: $p_t(w)$ is the proportion of documents in C_t containing w.
- Term frequency: $p_t(w)$ is the proportion of all terms in C_t which are equal to w.
- Weighted term frequency: $p_t(w)$ is a document-weighted proportion of all terms in C_t constructed so that all documents are weighted equally:

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

To measure the extent to which $\{p_t(w)\}_w$ (i.e., the set of all values $p_t(w)$ for varying w while keeping t fixed) differs from the baseline probabilities $\{p_0(w)\}_w$, we use a variant of the Kullback-Leibler divergence. Because 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

$$\begin{aligned} AKL\left(p_t \| p_0\right) &= KL\left(p_t \| p_0\right) - KL\left(p_0 \| p_t\right) \\ &= \sum_w \left(p_t(w) + p_0(w)\right) \log\left(\frac{p_t(w)}{p_0(w)}\right). \end{aligned}$$

When analyzing the trends of a corpus C_t, we find it useful to analyze the term-wise contributions to $AKL\left(p_t \| p_0\right)$. We thus define the pointwise antisymmetric Kullback-Leibler ($PAKL$) score of a term w as

$$PAKL_{(p_t \| p_0)}(w) = \left(p_t(w) + p_0(w)\right) \log\left(\frac{p_t(w)}{p_0(w)}\right).$$

The value $PAKL_{(p_t \| p_0)}(w)$ satisfies the following properties:

1) $PAKL_{(p_t \| p_0)}(w)$ is positive if $p_t(w) > p_0(w)$ and is negative if $p_t(w) < p_0(w)$.
2) $PAKL_{(p_t \| p_0)}(w)$ nears 0 as $p_t(w)$ approaches $p_0(w)$.
3) $\left|PAKL_{(p_t \| p_0)}(w)\right|$ increases both when $p_t(w)$ stays constant while $p_t(w)$ approaches 0, and when $p_0(w)$ stays constant while $p_t(w)$ approaches 0.

Thus, a term's $PAKL$ scores reflect whether the relative frequency of a term is increasing, decreasing, or staying constant in time. To construct a score which measures the relative trendiness (resp. "anti-trendiness") exuded by C_t, we can sum all $PAKL$ scores, all positive (resp. negative) $PAKL$ scores, or the most n positive (resp. negative) $PAKL$ scores in each corpus C_t.

It is the third property above which 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 $PKL_{(p_t \| p_0)}(w)$ via

$$PKL_{(p_t \| p_0)}(w) = p_t(w) \log\left(\frac{p_t(w)}{p_0(w)}\right).$$

Note, however, that $PKL_{(p_t \| p_0)}(w)$ is unable to differentiate between terms w such that $p_t(w) \approx p_0(w)$ and terms w such that $p_t(w) \approx 0$, since in both cases, $PKL_{(p_t \| p_0)}(w) \approx 0$.

B. Cluster coherence

Our second approach to the temporal analysis of a set of documents is based on the idea that we can construct tighter clusters of documents when there is a heightened focus on a relatively small set of concepts. This approach consists of obtaining vectors for each word, combining these to create a vector for each document, and clustering the resulting document vectors.

We acquire word vectors by running GloVe, an algorithm which uses co-occurrence statistics of the terms in a large corpus and a weighted least-squares model in order to derive a vector for each term [14]. Since our ultimate goal for 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.

Next, we derive a set of vectors to represent the content of the target corpus C_t. 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

$$\tilde{v}(d) = \sum_{w \in d} tf_d(w)\, idf_{C_0}(w),$$

where $tf_d(w)$ denotes the number of times the term w occurs in document d, and $idf_{C_0}(w)$ denotes a smoothed version of inverse document frequency of w in C_0. We define the document vector for d as $v(d) = \tilde{v}(d)/\|\tilde{v}(d)\|$. This normalization reflects our belief that documents with similar content but differing lengths should be treated as similar. We use the set of document vectors $v(d)$ as our set of "corpus vectors."

We then cluster the corpus vectors using spherical k-means clustering, which can be interpreted as a hard von Mises-Fisher (VMF) mixture model where all mixture components are forced to have the same concentration [15]. To construct a score measuring cluster coherence and tightness, we use the concentration κ derived from reinterpreting spherical k-means clustering as a VMF mixture model. For this, we rely on the techniques and formulae presented and explained in detail in [15], [16].

C. Weighted probabilistic fusion

We describe a technique for fusing the scores discussed above. Our technique is similar to that discussed in [17].

For each corpus C_t, we assume that we have generated m different scores $z_{t,1}, \ldots, z_{t,m}$ using the techniques discussed above. We assume that the values $\{z_{t,j}\}_t$ are sampled from some distribution with cumulative distribution function (cdf) F_j. 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).

We then construct fused scores s_t via

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

where $c_j > 0$ denotes the relative weight we wish to give the jth score generating technique. In our experiments, we

choose c_j so that the scores generated from term distribution analysis (Section III-A) have combined weight equal to that of the scores generated by analyzing cluster coherence (Section III-B). Finally, we fit a gamma distribution with cdf G to the set of fused scores $\{s_1, ..., s_n\}$. The value $G(s_t)$ quantifies the significance of the events occurring during time period 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 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. We then calculate s_t using the cdfs $\{F_{t,j}\}_j$. We call the fusion technique described in this paragraph "windowed fusion" in contrast to the original "global fusion" technique presented above.

IV. Experiments

A. Data

In total, three datasets are considered. We first apply our algorithm to *TwitterParisEnglish*, which consists of 50,000 tweets per day sampled uniformly at random from all English tweets acquired from the Twitter Streaming API from October 11, 2015 to November 29, 2015. Note that the sampling period for this dataset 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.

Next, we use the Twitter Search API to acquire datasets consisting of all tweets emitted by specified users. The dataset *TwitterUSUniversities* consists of all 4.2 million tweets emitted from official Twitter accounts of 2,300 United States universities from May 2014 to December 2016, and the dataset *TwitterOlympics* contains 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 all analyses, we use 25-dimensional GloVe vectors trained on 50 million English tweets sampled from the Twitter Streaming API from March 2015 to July 2015.

B. Results

We first run a $PAKL$ analysis (cf. Section III-A) for our *TwitterParisEnglish* dataset using the "document frequency" option, segmenting our corpus by day. The terms with the highest $PAKL$ scores for select days can be seen in Table I. We include terms from both uneventful days (Oct. 26, 2015 and Nov. 4, 2015) and anomalous days (Nov. 13, 2015 and Nov. 26, 2015). We note that the top $PAKL$ scores for anomalous days tend to be higher than those for normal days.

We construct subcorpora of *TwitterParisEnglish* by varying the number of tweets per day between 10,000 and 50,000. We observe surprisingly little variation in our scores (graphs not included due to space limitations) and conclude that our score generation techniques are largely invariant to sampling rate.

For the *TwitterParisEnglish* dataset, we construct four $PAKL$ scores by summing, for each day, all $PAKL$ scores, all positive $PAKL$ scores, and the highest 50 and 200 $PAKL$ scores. We also acquire cluster scores by constructing 9 clusterings, each with $k = 50$ clusters. We weight the scores so that the $PAKL$ and clustering scores each account for 50% of the total fused scores. Fig. 1 depicts these fused scores, with dashed vertical lines denoting the date of the Paris attacks and Thanksgiving, and a dashed horizontal line indicating the 10% significance level.

TABLE I
Top words for select days and their associated $PAKL$ scores from *TwitterParisEnglish*.

Oct. 26, 2015	
forevermore	0.0286
#pushawardslizquens	0.0154
#aldubpredictions	0.0149
the	0.0149
#aldubnewbeginnings	0.0129

Nov. 4, 2015	
#aldub16thweeksary	0.0203
i	0.0110
#showtimehousemates	0.0103
that	0.0085
it	0.0083

Nov. 13, 2015	
paris	0.1448
in	0.0682
#prayforparis	0.0582
the	0.0572
#madeintheam	0.0485

Nov. 26, 2015	
thanksgiving	0.1743
thankful	0.1159
happy	0.0692
#mtvstars	0.0602
for	0.0402

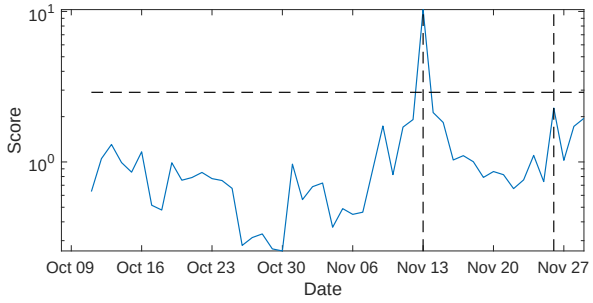

Fig. 1. Fused scores for *TwitterParisEnglish* using $F^{(\beta)}$ and global fusion.

For the *TwitterOlympics* and *TwitterUSUniversities* datasets, fusion is performed similarly with the following changes to account for the fact that these corpora are smaller in general than *TwitterParisEnglish*. 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. 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. 2) and $F^{(\text{emp})}$ (Fig. 3). Although these graphs have very similar shapes, these graphs display the general trend that fusion using $F^{(\beta)}$ tends to produce fewer significant events than fusion using $F^{(\text{emp})}$. The three periods in Fig. 2 with significant scores correspond to the various athletic events in August 2015, the 2016 Olympic trials, and the 2016 Summer

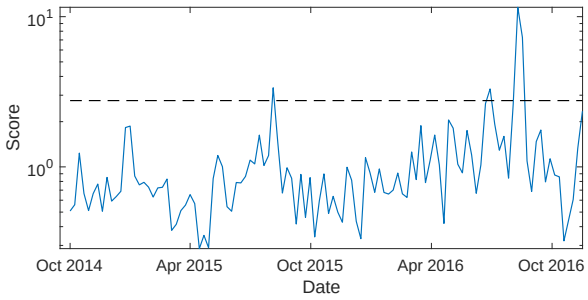

Fig. 2. Fused scores for *TwitterOlympics* using $F^{(\beta)}$ and global fusion.

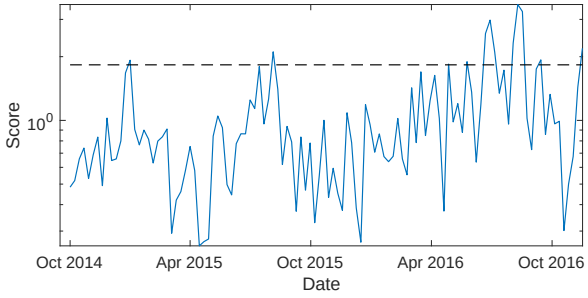

Fig. 3. Fused scores for *TwitterOlympics* using $F^{(\text{emp})}$ and global fusion.

Olympics.

For *TwitterUSUniversities*, we present a graph of the fused scores created using windowed fusion (Fig. 4). 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. 2 corresponding to the August 2015 athletic events can be attributed more to clustering scores than $PAKL$ scores. During these events, $PAKL$ scores barely rise above baseline; since each sport has its own world championship, there are no key terms with abnormally high relative frequency. Because the trending pattern of the terms associated to individual world championships cannot be differentiated from typical trending patterns, the $PAKL$ scores do not register as abnormal. However, because all terms related to sport and competition have similar vectors, the high frequency of tweets related to competition causes abnormally high cluster coherence scores.

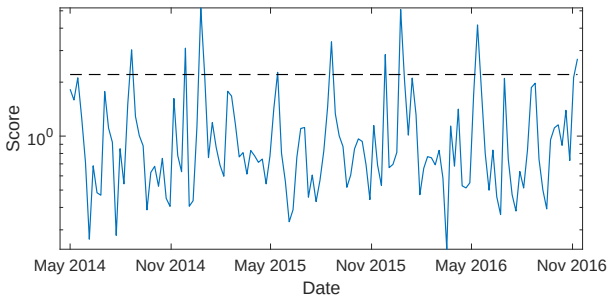

Fig. 4. Fused scores for *TwitterUSUniversities* during 2014–2016 using $F^{(\beta)}$ and windowed fusion.

REFERENCES

[1] Z. Cheng, J. Caverlee, and K. Lee, "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*, 2010, pp. 759–768.

[2] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, no. 1, pp. 178–185, 2010.

[3] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, 2011, pp. 253–262.

[4] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 251–258.

[5] S. Morinaga and K. Yamanishi, "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*, 2004, pp. 811–816.

[6] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd International Conference on Machine learning*, 2006, pp. 113–120.

[7] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Uncertainty in Artificial Intelligence (UAI)*, 2008, pp. 579–586.

[8] L. AlSumait, D. Barbará, and C. Domeniconi, "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*, 2008, pp. 3–12.

[9] X. Wang and A. McCallum, "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*, 2006, pp. 424–433.

[10] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 91–101.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 497–506.

[12] D. He and D. S. Parker, "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*, 2010, pp. 443–452.

[13] R. Swan and J. Allan, "Extracting significant time varying features from text," in *Proceedings of the 8th International Conference on Information Knowledge Management*, 1999, pp. 38–45.

[14] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation." in *EMNLP*, vol. 14, 2014, pp. 1532–1543.

[15] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, Dec. 2005.

[16] S. Sra, "A short note on parameter approximation for von Mises-Fisher distributions: And a fast implementation of $i_s(x)$," *Comput. Stat.*, vol. 27, no. 1, pp. 177–190, Mar. 2012.

[17] K. Simonson, "Probabilistic fusion of ATR results," Sandia National Laboratories (SNL-NM), Albuquerque, NM, Tech. Rep. SAND98-1699, 1998.