Topical Anomaly Detection From Twitter Stream

Pramod Anantharam
Kno.e.sis Center,
Wright State University,
Dayton, OH.
pramod@knoesis.org

Krishnaprasad Thirunarayan
Kno.e.sis Center,
Wright State University,
Dayton, OH.
tkprasad@knoesis.org

Amit Sheth
Kno.e.sis Center,
Wright State University,
Dayton, OH.
amit@knoesis.org

ABSTRACT
In this paper, we spot topically anomalous tweets in twitter streams by analyzing the content of the document pointed to by the URLs in the tweets in preference to their textual content. Existing approaches to anomaly detection ignore such URLs thereby missing opportunities to detect off-topic tweets. Specifically, we determine the divergence of claimed topic of a tweet as reflected by the hashtags and the actual topic as reflected by the referenced document content. Our approach avoids the need for labeled samples by selecting documents from reliable sources gleaned from the URLs present in the tweets. These documents are used for comparison against documents associated with unknown URLs in incoming tweets improving reliability, scalability and adaptability to rapidly changing topics. We evaluate our approach on three events and show that it can find topical inconsistencies not detectable by existing approaches.

Author Keywords
Anomaly detection; spam and off-topic content detection; binary classification; twitter stream analysis

ACM Classification Keywords
H.4.0 Information Systems Applications: General

General Terms
Experimentation, Verification

INTRODUCTION
Twitter is a dynamic microblogging platform with posts on diverse topics ranging from mundane every day activities to critical information for coordination during disasters, political crisis, corruption protests, etc. As more people rely on microblogs to form opinions and make decisions, adversaries are trying to manipulate microblogs and propagate off-topic links for their selfish motives. So, it is becoming increasingly important to assess trustworthiness of microblogs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.
Copyright 2012 ACM 978-1-4503-1228-8...$10.00.

Figure 1 is a controversial tweet by Kenneth Cole which has hashtag related to egypt political crisis but has a link pointing to a fashion website. Since the tweet in Figure 1 is from a well known author, approaches that use author's reputation or just the tweet content are ineffective. Figure 2 shows a tweet that is stuffed with hashtags of trending events, though it has a link that corresponds to none of these events.

The existing approaches to microblog spam detection focus on the explicit structure of a tweet, its author or the account generating the tweet. The "trending topic exploit" uses on-topic trending keywords in a tweet to include off-topic URLs (which are typically opaque about the referred content), that when visited by the tweet consumer serves the selfish motives of the tweet abuser. We propose to exploit inconsistency between the topic of a tweet and the topic of the document referred by URLs in the tweet, so that such "abusive" tweets can be discovered and flagged for further action.

Our work is similar to that on spam detection and quality content selection from twitter since it separates tweets that can potentially harm or mislead users. It can also detect novelty. Our approach is innovative in that it analyzes referenced documents in addition to analyzing tweet content for assessing its relevance to an event/topic. This is an improvement over existing approaches that rely only on tweet content for two reasons: (a) Tweets are short and often do not have sufficient terms in common among them since they may describe different aspects of an event. (b) Tweets can be stuffed with relevant keywords while the URLs in the tweet may point to off-topic or other misleading/malicious content. This is important considering that 90% of spam tweets contain URLs [2, 4] to content unrelated to the claimed subject (as evident from the hastags or keywords used in the tweet). So, we expect topical consistency check to be effective in detecting and removing off-topic content. To address the challenges for detecting off-topic content, we propose an approach that does not require labelled tweets to learn from or determine classification threshold. Instead, we rely on trusted sources (such as news outlets) to provide reliable documents to compare against. This has two important benefits: (i) It is scalable because the need for a large labelled training set is replaced by the use of relatively smaller number of reliable news sources. We expect that the rate of tweets on each topic is bounded and topic-specific tweet streams can be processed independently and in parallel. (ii) It is adaptive in that the document content can evolve in re-

"Millions are in uproar in #Cairo. Rumor is they heard our new spring collection is now available online at http://bit.ly/KCairo -KC"

Figure 1. Controversial tweet by Kenneth Cole

"RealTimeEgypt.Com 4 sale http://3DSantaClaus.com #domains #domaining #egypt #iphone #ipad #superbowl #domainfest #egyptian"

Figure 2. Abuse of trending keywords on twitter

sponse to dynamically changing events and topic drifts. The rest of the paper is organized as follows: We first present the related work in selecting quality content and spam detection in the context of tweets. We then present our approach and evaluate it against extant approaches. We conclude with discussion of future work.

RELATED WORK

There has been work on assessing trustworthiness of wikipedia articles [3], product reviews [6], and self description on dating sites [9]. Spam detection can use the structural features of a tweet and the nature of the account generating these tweets [2, 7, 10]. Truthy[1] project exploits meme diffusion patterns for detecting astroturfing, smear campaigning, and misinformation in the context of U.S political elections [8]. Selecting quality tweets from twitter for a specific event is discussed in [1]. They advocate centroid similarity approach that computes the cosine similarity of the tf-idf representation of each tweet to its associated tweet cluster centroid (where each cluster centroid term is associated with its average weight across all cluster tweets) and selecting tweets with high similarity scores. Below we analyze the shortcomings of this approach and show how the topical consistency feature we use can overcome these shortcomings.

Recall that existing techniques for off-topic content detection from microblogs ignore the contents of the URLs and consider only tweet content to train a classifier [2, 11]. These approaches are inadequate since: (a) tweets can be stuffed with relevant keywords but may have off-topic links; (b) it may not scale due to the need for event-specific training samples and (c) classifier threshold selection is not straightforward [5]. We address these challenges (i) by exploiting the content of the documents referred to by the URLs in a tweet (ii) by using the source of the document gleaned from the URL prefix in a tweet to assess its reliability, and (iii) experimentally and dynamically determining suitable threshold.

APPROACH

We present the details of how to detect topical anomaly and describe the implementation of a scalable system.

Baseline

We use the centroid-based similarity approach described in [1] as the baseline, which considers only the tweet content for similarity computation. Since [1] does not specify threshold selection, we empirically determine the threshold for which the baseline performs the best. If the maximum cosine similarity is greater than the threshold, we regard the tweet as on-topic else off-topic.

Our Approach

The key idea is to exploit the match between claimed topic of a tweet (inferred from the hashtags) and the actual topic of the document pointed to by the URLs in a tweet.

For tweets with URL, we use "background knowledge" in the form of list of reliable sources of information (e.g. CNN, BBC) that can be updated over time. For now we have used a static list of reliable sources, but a production system should provide more flexible user control over this list. As we collect the trusted documents pointed to by the reliable sources (gleaned from the embedded URLs), we compute the average cosine similarity among them (Sim_{avg}). For every incoming tweet, we compare the document pointed to by the URL in a tweet with all the trusted documents (assumed to characterize the current topics) to get the maximum cosine similarity (Sim_{max}). We flag an incoming tweet as anomalous if $\text{Sim}_{max} < \text{Sim}_{avg}$. Since a tweet may cover only some aspect of an event, the centroid of a cluster of tweets for an event may not be a good representative if the tweets do not overlap significantly. Our approach that determines similarity pairwise to obtain Sim_{avg} is more robust with respect to tweets that have independent sub-clusters related to the same event. We compare a test tweet against all the tweets (on-topic) to get the maximum cosine similarity.

For tweets without URLs, we rely on the baseline but with the threshold Sim_{avg} which is event-specific.

We address the aforementioned challenges as follows: (a) Since we do not rely just on the textual content of the tweets, but consider documents pointed to by the URLs in the tweet that has potential to be a spam among other things, tweet stuffed with trending keywords do not interfere with off-topic content detection. (b) Since we are collecting documents from trusted sources and the fact that new documents obtained as time progresses can track an evolving event, our approach remedies the need for continuous manual labeling of training examples. (c) Since we compute thresholds by analyzing trusted event-specific documents dynamically, it can deal with different events and topic drift in a uniform way with some initial delay for bootstrapping.

In our prototype, we employ LongURL API[2] to resolve the shortened URLs. We use heavyMetall API[3] to extract text from HTML pages. We use Apache Lucene[4] for finding cosine similarity among documents. The system currently expects a single stream of tweets but can be parallelized. The tweets that are off-topic are flagged.

Threshold Selection

We investigated several approaches to select cosine similarity thresholds to detect documents associated with an even-

[1] http://truthy.indiana.edu/

[2] http://longurl.org

[3] http://peweproxy.fiit.stuba.sk/metall/

[4] http://lucene.apache.org/java/docs/index.html

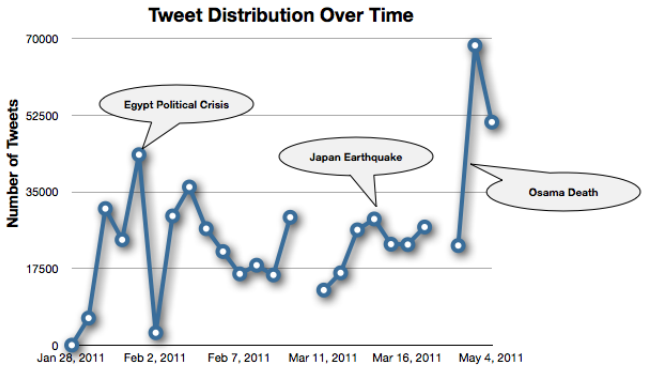

Figure 3. Tweet distribution for all the three events.

Events	Number of tweets	Time Period
Egypt Crisis	301,082	2011-01-29 to 2011-02-10
Japan Earthquake	157,193	2011-03-11 to 2011-03-17
Osama Bin Laden	142,021	2011-05-02 to 2011-05-04

Table 1. Tweets collected for evaluation

t/topic. 20 News dataset, which has a number of categories, was used for deriving thresholds based on average similarity for all documents in a category. We had three major concerns in using 20 News dataset for threshold computation: (a) There are a fixed and limited number of categories (b) Thresholds derived for each general category did not help us in analyzing specific categories (e.g. technology vs. Apple). (c) Deciding what category to use for newly found event is unclear. Overall, this approach did not yield suitable thresholds for classification.

As event-independent threshold may not exist in practice [5], we use dynamically computed thresholds based on the average cosine similarity between trusted documents for each event. Our approach (a) eliminates specifying thresholds by users, (b) avoids selecting thresholds in an ad-hoc manner, and (c) dynamically updates the threshold as the event evolves over time.

EVALUATION

The evaluation was done using the three events shown in Table 1, since these events were prone to abuse[5]. The approach is generic enough to be applicable to other such events. Figure 3 represents the distribution of tweets over the duration of the data collection. For each event, we ensured that a labelled dataset of around thousand tweets with at least twenty off-topic tweets were created. In each case, we use the approach in [1] as the baseline for comparison. Our results are presented as confusion matrix in Table 4 along with the baseline in Table 3. Table 2 has some examples of off-topic tweets detected.

Discussion

In this section, we discuss the results of our evaluation and challenges encountered.

[5]http://bit.ly/g3oWAf, http://bit.ly/yIHQXC

Event	sample off-topic tweets
Egypt Crisis	"Great Download: #1: Microsoft Exchange Server Enterprise 5.0 50 CAL: Microsoft Exchange Server Enterpris... http://amzn.to/giA3dm #Egypt" "Millions are in uproar in #Cairo. Rumor is they heard our new spring collection is now available online at http://bit.ly/KCairo -KC" "RT @dangerroom: Iran, China Block Outside Sites to Muzzle Mideast News http://wrd.tw/dGazcY #jan25 #egypt"
Japan Earthquake	"RT @amazongames: Help the victims of the #Japan earthquake and pacific #tsunami by donating to the American Red Cross http://amzn.to/fzGPCI" "Japan Relief Fund — Buy a T Shirt (Hope) & Donate — Earthquake & Tsunami Relief http://t.co/FIqLPaY" "New section in my shop: 'Charities' 50% of every sale goes to #tsunami and #earthquake relief. http://etsy.me/dPaZsx #etsy #etsybot"
Osama Death	"Digging deeper into Windows 8 App Store - http://bit.ly/jKEseE #windows8 #Win8 #leak #obl" "adebayo gbadebo commented on NewMusic's blog post 'New Hot Rod - Osama Bin Laden Is Dead!!!' http://bit.ly/jr9dvW #50CENT" "Bin Laden #Coffee #Mugs benefit families of fallen heroes (education fund) by #LTCartoons #osama #obl #UBL http://bit.ly/ieyhHi"

Table 2. Sample of topically anomalous tweets detected by our approach.

Result Description

Given that there are only 1% spam tweets on twitter[6], any classifier that assigns only on-topic labels to all the tweets would have a 99% accuracy. Thus, we do not rely just on precision/recall measures, but show the confusion matrix for better insight. For Egypt Crisis, we found 18 out of 23 off-topic tweets while misclassifying only 13 on-topic tweets as off-topic. This is significantly better than the baseline in Table 3, which detected 13 off-topic tweets but misclassified 232 on-topic tweets. Baseline continued to perform poorly by misclassifying on-topic tweets as off-topic tweets for all three events while our approach not only found more off-topic tweets in case of two events, but also had low misclassification rate. The tweets that were on-topic but classified as off-topic were not directly related to the event (e.g. stock market reaction). Sparse text content on the page also led to this confusion. In case of Japan Earthquake, our approach could detect tweets related to seeking donations as off-topic and distinguish it from tweets describing the event. Some of the documents that had event descriptions along with donation information resulting in off-topic tweets getting misclassified as on-topic. Similarly, in case of Osama Death, our approach could detect tweets promoting gifts such as photo mugs and t-shirts as off-topic and separated from the actual on-topic tweets related to the event. Our approach consistently did better than the baseline for all the three events.

Implementation Challenges

While processing twitter stream using our algorithm, we faced a number of practical challenges: (a) Dead URLs - Though majority of the tweets have URLs, many of them were broken when followed. (b) Language Diversity - Since twitter is widely used around the world, people can use their regional languages. Recognizing and analyzing multilingual content is non-trivial. (c) Multimedia content - Event-related tweets contain images and videos providing new opportunities for further improvement. (d) Dynamic content - Some linked pages that have on-topic but out-of-date content require further analysis. Tweets tend to have shortened URLs that mask the actual domain names and hence impede user authentication of an URL. This can trap users into inadvertently navigating suspicious links. Our approach expands and checks such URLs for its content and can warn users against attacks. The primary application and benefit of this work on topical

[6]http://blog.twitter.com/2010/03/state-of-twitter-spam.html

Confusion Matrix			
Event	on/off-topic	classified on-topic	classified off-topic
Egypt Crisis	on-topic	746	**232**
	off-topic	**10**	**13**
Japan Earthquake	on-topic	603	**372**
	off-topic	**5**	**20**
Osama Death	on-topic	742	**232**
	off-topic	**19**	**9**

Table 3. Confusion Matrix for the baseline

Confusion Matrix			
Event	on/off-topic	classified on-topic	classified off-topic
Egypt Crisis	on-topic	965	**13**
	off-topic	**5**	**18**
Japan Earthquake	on-topic	968	**7**
	off-topic	**10**	**15**
Osama Death	on-topic	966	**8**
	off-topic	**15**	**13**

Table 4. Confusion Matrix for our approach

anomaly is separation of on-topic and off-topic content, with further scrutiny required on the latter to distinguish "irritating" off-topic content from "dangerous" off-topic content. For example, our prototype identified web pages about stock market conditions due to Egypt crisis, wound treatment, fire bomb making manuals, and web sites that hosted rebellious content[7]. All these off-topic documents can be analyzed to derive new intelligence insights and actionable information buried in the data deluge.

CONCLUSIONS AND FUTURE WORK

We proposed a method and prototyped a system for topical anomaly detection for assessing quality of tweets. We compared our approach with the state-of-the-art on confusion-matrix metric to show how our system can improve detection of divergence in "claimed topic" (e.g. hashtags) and the actual topic of the document pointed to by the URLs in the tweet. Our approach uses documents from reliable sources to detect off-topic content, to improve reliability, scalability and adaptability in the face of large and dynamic tweet stream. The evaluation results look promising but there are several unresolved challenges.

As a future work, we propose to find topical anomalies using event specific background knowledge such as by using named entities for specific events, and incorporating more anomaly detection capabilities such as change in network connections and interactions to detect miscreants who exploit social networks for personal gains. We can also build reputation for each source. For scaling this anomaly detection approach to massive data streams, we need to exploit the inherent parallelism in themes due to multiple events, and in data and computation due to their independence. The computation of S_{max} can be speeded-up by determining similarity of a test document with a subset of reliable documents obtained by picking centroids of closely related subclusters or documents with high-centrality degree [1]. Real-time anomaly detection can be done on the server-side and flagged on the client side.

ACKNOWLEDGMENTS

We thank the twitris team (http://twitris.knoesis.org) for providing data for our evaluations. We thank Wenbo Wang, Lu Chen, and Pavan Kapanipathi for their valuable suggestions.

[7]http://www.crimethinc.com/blog/2011/02/02/egypt-today-tomorrow-the-world

REFERENCES

1. Becker, H., Naaman, M., and Gravano, L. Selecting quality twitter content for events. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011).

2. Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. Detecting spammers on twitter. In *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)* (2010).

3. Dondio, P., Barrett, S., Weber, S., and Seigneur, J. Extracting trust from domain analysis: A case study on the wikipedia project. *Autonomic and Trusted Computing* (2006), 362–373.

4. Gayo-Avello, D., and Brenes, D. Overcoming spammers in twitter-a tale of five algorithms. In *1st Spanish Conference on Information Retrieval, Madrid, Spain* (2010).

5. Kumaran, G., and Allan, J. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM (2004), 297–304.

6. Liu, H., Lim, E., Lauw, H., Le, M., Sun, A., Srivastava, J., and Kim, Y. Predicting trusts among users of online communities: an epinions case study. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, ACM (2008), 310–319.

7. Mustafaraj, E., and Metaxas, P. From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line* (Apr. 2010).

8. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. Detecting and tracking the spread of astroturf memes in microblog streams. *Arxiv preprint arXiv:1011.3768* (2010).

9. Toma, C. L., and Hancock, J. T. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, ACM (New York, NY, USA, 2010), 5–8.

10. Wang, A. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, IEEE (2010), 1–10.

11. Yardi, S., Romero, D. M., Schoenebeck, G., and Boyd, D. Detecting spam in a twitter network. *First Monday 15*, 1 (2010).