A Network-Based Model for Predicting Hashtag Breakouts in Twitter

Sultan Alzahrani[1], Saud Alashri[1], Anvesh Reddy Koppela[1], Hasan Davulcu[1(✉)], and Ismail Toroslu[2]

[1] School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287, USA
{ssalzahr,salashri,akoppela,hdavulcu}@asu.edu
[2] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
toroslu@ceng.metu.edu.tr

Abstract. Online information propagates differently on the web, some of which can be viral. In this paper, first we introduce a simple standard deviation sigma levels based Tweet volume breakout definition, then we proceed to determine patterns of re-tweet network measures to predict whether a hashtag volume will breakout or not. We also developed a visualization tool to help trace the evolution of hashtag volumes, their underlying networks and both local and global network measures. We trained a random forest tree classifier to identify effective network measures for predicting hashtag volume breakouts. Our experiments showed that "local" network features, based on a fixed-sized sliding window, have an overall predictive accuracy of 76 %, where as, when we incorporate "global" features that utilize all interactions up to the current period, then the overall predictive accuracy of a sliding window based breakout predictor jumps to 83 %.

Keywords: Information diffusion · Hashtag volumes · Prediction · Social networks · Diffusion networks

1 Introduction

Online Social Networks (OSNs) such as Twitter have emerged as popular microblogging and interactive platforms for information sharing among people. Twitter provides a suitable platform to investigate properties of information diffusion. Diffusion analysis can harness social media to investigate viral tweets and trending hashtags to create early-warning solutions that can signal if a viral hashtag started emerging in its nascent stages. In this paper, we utilize the 68-95-99.7 rule to define a simple method of hashtag volume breakouts. In statistics, the 68-95-99.7 rule, also known as the three-sigma rule or empirical rule, states that nearly all values lie within three standard deviations (σ) of the mean (μ) in a normal distribution. We utilize a fixed sized sliding window (of length 20 daily

N. Agarwal et al. (Eds.): SBP 2015, LNCS 9021, pp. 3–12, 2015.
DOI: 10.1007/978-3-319-16268-3_1

intervals), to compute a running average and standard deviation for each hashtag's volume distribution. Then, we identify non-overlapping *episodes* within a time-series of daily volumes for each hashtag whenever its daily volume exceeds ($\mu + 1\sigma$) of the previous 20 day periods. We label the 20 day periods preceeding an episode as the *accumulation period* of an episode. We categorize an episode as *breaking* if the hashtag volume goes on to exceed ($\mu + 2\sigma$) without falling below max(0, μ - 2σ), or else as a *non-breaking* episode otherwise. Next, we examine multiple network metrics associated with the accumulation period of each episode and proceed to build a classifier that aims to predict whether an episode will lead to a breakout volume or not. We employ a network based classification model and to discover latent patterns for the breakout phenomena, particularly we examine which factors contribute to make hashtag volumes breakout. We also build a visualization tool called Trending Hashtag Forecaster (THF). Our THF tool helps reveal the underlying network structures, patterns and properties that lead to breakout volumes. Our experiments showed that "local" network features during an accumulation period have an overall predictive accuracy of 76%, where as, when we incorporate "global" features that utilize measures extracted from all of the network up to the current accumulation period, then the overall predictive accuracy of the Trending Hashtag Forecaster jumps to 83%.

2 Problem Formulation

Given a set of tweets $T = t_1, t_2, t_3, ..., t_n$ where n is number of tweets in our corpus. These tweets comprise textual contents, user interactions and additional meta data. We explore and analyze both textual contents filtered by a given hashtag from hashtags set H. Then we denote tweet volume as number of tweets per day. We then compute daily means ($\mu(20)$) and standard deviation ($\sigma(20)$) for each hashtag by utilizing its volume distribution during its previous 20 days window. We experimentally determined the best window size by experimenting 10, 15, 20, 25 and 30 days windows. The 20 days window shows the best performance amongst the others.

If the hashtag frequency rises above ($\mu(20)$ $+1\sigma(20)$), then we label that period as an episode, and we mark its previous 20 days as the accumulation period of an episode. We start observing hashtag frequency for two possible outcomes:

- a breakout if hashtag volume rises above($\mu(20)$ $+2\sigma(20)$), without falling below max(0, $\mu(20)$ - $2\sigma(20)$), or
- non-breakout, if hashtag volume falls below max(0, $\mu(20)$ - $2\sigma(20)$), without rising above ($\mu(20)$ $+2\sigma(20)$)

In breakout scenario for an episode no further overlapping breakouts are allowed until its volume falls below max(0, $\mu(20)$ - $2\sigma(20)$). In both scenarios, as episode begins with its accumulation period and continues until the hashtag volume dies out (i.e. it falls below max(0, $\mu(20)$ - $2\sigma(20)$)). Figure 1, shows the histograms of all daily hashtag volumes in our corpus.

Next, in Section 3 we present related work. In Section 4 we describe out Tweet corpus. In Section 5, we describe our Trending Hashtags Forecaster visualization tool. In Section 6, we introduce our network based model, local and global network features to predict hashtag episode breakouts following accumulation periods. In Section 7, we present experimental results and findings. Section 8 concludes the paper and presents the future work.

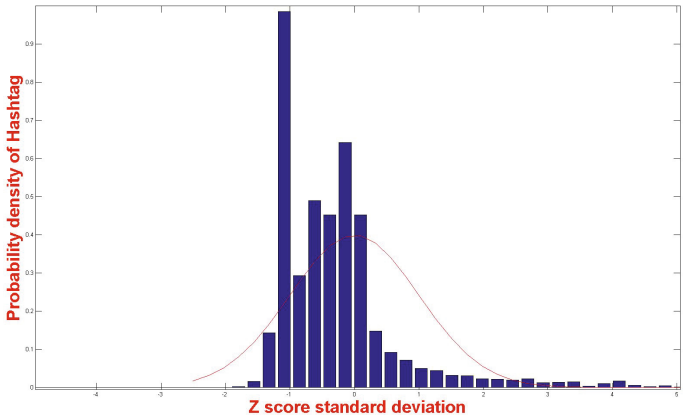

Fig. 1. Probability distribution function of all Hashtags

3 Related Work

Twitter network has more than 271 million monthly active members and 500 million tweets are generated daily [1]. The vast size and reach of Twitter enables examination of potential factors that might be correlated with breakout events and viral diffusion. We found that diffusion related studies fall into two categories. In the first category, many studies start by analyzing social networks as a graph of connected interacting nodes i.e. between users, friends or followers, and these studies investigate different factors that drive propagation and diffusion of information Arruda et al. [7] proposes that network metrics play an important role in identifying influential spreaders. They examined the role of nine centrality measures on a pair of epidemics models (i.e. disease spread on SIR model and spreading rumors on a social network). According to the authors, epidemic networks are different from social networks such that infected individuals in SIR become recovered by a probability μ while in social networks a spreader of a rumor becomes a carrier by contacts. They found centrality measures such as closeness and average neighborhood degree are strongly correlated with the outcome of spreading rumors model.

[1] https://about.twitter.com/company

The second category looked into the diffusion problem through content analysis by incorporating different natural language processing techniques. For instance, one study hypothesized that a specific group of words is more likely to be contained in viral tweets. Li et al. analyzed tweets in terms of emotional divergence aspects (or sentiment analysis) and they noted that highly interactive tweets tend to contain more negative emotions than other tweets [1], [8].

Weng et al. [5] investigated the prediction of viral hashtags by first defining a threshold for a hashtag to be viral, and then by examining metrics and patterns related to the community structure. They achieved a precision of 72% when threshold is set statically to 70. Romero et al studied the diffusion of information on Twitter and presented some sociological patterns that make some types of political hashtags spread more than others. Asur [11] presented factors that hinder and boost trends of topics on Twitter. They found content related to mainstream media sources tends to be main driver for trends. Trending topics are further spread by propagators who re-tweet central and influential individuals.

We propose a model that predicts hashtag breakouts thru adaptive dynamic thresholds, and by utilizing generic content-independent network measures that draws their information from (i) local networks corresponding to accumulation periods, as well as (2) from the global networks corresponding to the entire network history preceeding an accumulation period.Our experiments showed that local network features yield an overall predictive accuracy of 76%, and, global network features yield an overall predictive accuracy of 83%.

4 Data Source

The dataset we are using in this study is a collection of tweets from UK region. These tweets have been crawled based on a set of keywords with the aim to capture political groups, events, and trends in the UK. The dataset consists of more than 3 million tweets, 600K users, with more than 5.2 million interactions (both mentioning and retweeting) between users along with 1,334 hashtags.

5 Visualization Tool: Trending Hashtags Forecaster

In order to visualize and understand breaking hashtag phenomena, we built a visualization tool, depicted in Figure 2, that facilitate exploring temporal dynamics of hashtags and their underlying networks during accumulation period of each episode. Local and global network measures are also computed and displayed as network and node features. These network measures are utilized to train and test a predictive classifier, presented in the next section.

6 Methodology

In this study, we crawled tweets containing hashtags (case insensitive) which related to political groups in UK from June, 2013 to July, 2014. After crawling,

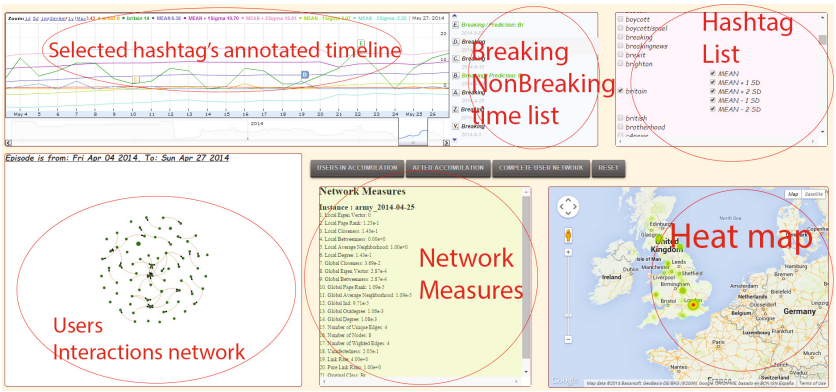

Fig. 2. THF visualization tool

we detected hashtag episodes using techniques described in Section 2. We identified the accumulation period and accumulation network of each episode, and extracted network measures corresponding to its accumulation network. Each eposide was also labeled as breaking or non-breaking based on its spread.

THF visualization tool reveals some of the discriminative patterns between breaking and non-breaking hashtags. Figure 3 shows the user interaction network for a non-breaking hashtag. User interaction network denoted by number 1 was captured during its accumulation period. Later on, this Hashtag did not breakout (i.e. did not cross its $\mu(20) + 2\sigma(20)$, but it fall back to zero volume, hence considered as a non-breaking episode. Figure 4, illustrates a breakout hashtag. Following a 20 period accumulation period, its volume exceeds $\mu(20) + 1\sigma(20)$ (denoted by network number 1), and it's volume exceeds breakout levels (by exceeding it's $\mu(20)+2\sigma(20)$) threshold (denoted by network number 2). Network 3 shows the entire reach this episode before it's demise (i.e. by falling below max(0, $\mu(20)$ - $2\sigma(20)$). An interesting observation related in the network 1 is a highly central green node, which attracts many new re-tweeters in network 2 and network 3. This observation indicates that existence of a large number of highly central nodes during the accumulation phase of an episode could be a good predictor for a following breakout. Other instances' patterns could not be cached by naked eye, yet they carry latent centrality measures correlate with our definition.

6.1 Network Based Model

In this model we investigate how users get involved in a hashtag h by mentioning, replying or retweeting. Their interactions are depicted as a directed graph G_{h_i}. We then incorporated normalized size-independent network features for directed graphs corresponding to accumulation periods of episodes. The network graph is a pair $G = (V, E)$ where V is set of vertices representing users together with a set of edges E, representing interactions between users. For instance, if a user

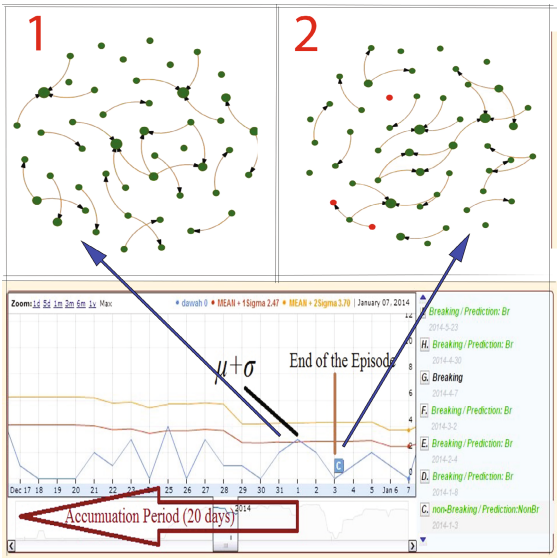

Fig. 3. Non breaking #Dawah Hashtag episode

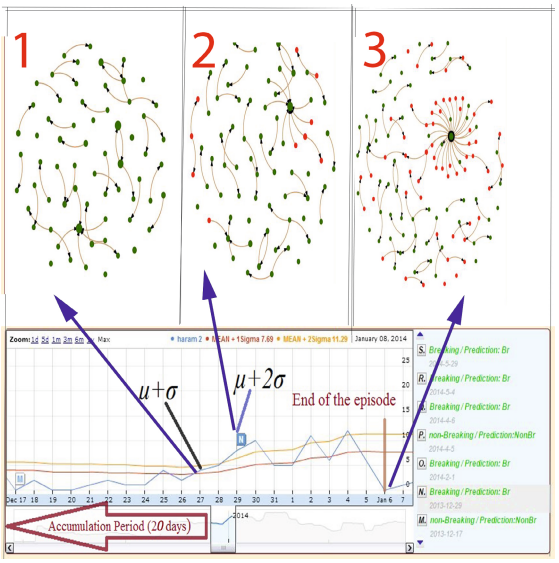

Fig. 4. Breaking #haram Hashtag episode

$u1$ mentioned, replied, or retweeted one tweet of $u2$, then a directed edge from $u1$ to $u2$ is formed.

We attempted to identify key features that contribute to the network based classification problem for breaking or non-breaking hashtags. Table 1 list all features that we used for local and global measures. Local measures are associated with user interactions during the accumulation period only, where as global measures draws their information from all interactions beginning from the start date (June 2013) until the end date of any accumulation period under consideration.

Table 1. Feature description

Feature	Description
Eigen Vector Centrality	Node's centrality depends on its neighbors centralities. If your neighbor are important you most likely are important too.
Page Rank	IVariant of Eigenvector where a node don't pass its entire centrality to its neighbors. Instead, its centrality divided into the neighbors. [3]
Closeness Centrality	A node is considered important if it is relatively close to all other nodes in the network [2].
Betweeness centrality	Measuring the importance of a node in connecting other parts of the graph [6]. This measure possesses the highest space and time complexity.
Degree centrality	It measures the number of ties a node has in undirected graph.
Indegree Centrality	It measures number of edges pointing into a node in a directed graph.
Outdegree Centrality	It is similar to the two above measure but it concerns on the number of outgoing links from a user, and it is normalized for each node.
Link Rate	Number of URLs in the tweets during the accumulation period divided by number of tweets.
Distinct Link Rate	Similar to link rate but without considering similar URLs.
Number of uninfected neighbors of early adopters	It is total number of retweets or mentioned (edges) a user has ever received globally, normalized by max-min retweets within local network in a current period being measured. [5]
Neighborhood average degree	it measures the average degree of the neighborhood of each node. [4]

7 Experiment Results and Findings

As a preprocessing step, We had 2790 for the non break out instances, while 1331 were for the break out. We sampled (without replacement) instances from both classes with oversampling for the lower represented class. We next examined the correlation between features and breaking hashtags using Principle Component Analysis (PCA). PCA is a dimensionality reduction approach that analyzes dataset to find which features give highest variance among instances and it maps the given features into lesser number of factors called components [9]. After that, in order to predict whether a given hashtag will breakout or not, we run a supervised network based learning model.

7.1 Features Correlated with Breaking Hashtags

PCA identified nine factors shown in Table 1. According to Kaiser Criterion [10], the factors to consider are the ones with eigenvalue above 1. In this study, we will focus on the first two components since they reveal interesting insights. Table 2

shows the correlation between our features and the first two components shown in Table 1. The first component is strongly correlated (negatively) with global measures, where as the second component is strongly correlated (negatively) with local measures. These two components give us a hint that global features should be grouped together and they contribute heavily (36%) to the variation in our dataset. Also, some of the local measures are also grouped together in a single factor and they somewhat contribute (21%) to the variation in our dataset.

Table 2. PCA components

Component	Eigenvalue	Variance	Cumulative Variance
1	5.79	36.16	36.17
2	3.30	20.62	56.78
3	1.669	10.43	67.21
4	1.24	7.73	74.94
5	1.01	6.31	81.24
6	0.88	5.54	86.78
7	0.663	4.14	90.92
8	0.48	3.01	93.93
9	0.42	2.65	96.57

Table 3. Correlation Between Table and Components

Feature	Component 1	Component 2	Feature	Component 1	Component 2
PageRank Local	0.14	-0.45	PageRank Global	-0.36	-0.10
Closeness Local	0.05	-0.51	Closeness Global	-0.24	0.09
Betweeness Local	0.05	-0.44	Betweeness Global	-0.35	-0.07
Avg Neighbor Degree Local	-0.11	-0.04	Avg Neighbor Degree Global	-0.3552	-0.10
Degree Cent. Local	0.11	-0.49	Degree Global	-0.3897	-0.0554
Uninfected Neighbor	0.19	0.02	In Degree Global	-0.39	-0.09
Link Rate	-0.03	0.14	Distinct Link Rate	0.0282	0.1889
Outdegree Global	-0.21	0.02	-	-	-

7.2 Network Based Model

For this model, we measured two sets of features: local and global. For local features: we have eigenvector, pagerank, closeness, betweeness, average neighborhood degree, uninfected neighbors before break out, and degree centrality. For global features we have the previous features measured globally plus in degree, out degree, and link rate. Next, we train and test a random forest classifier

with 10 fold cross-validation using three approaches: prediction using all features shown in Table 3, prediction using global features that are correlated with the first factor identified by PCA shown in Table 4, and prediction using local features that are correlated with the second factor returned by PCA shown in Table 5. We achieve the highest precision of 84%, recall of 81% and F-measure of 82% for breakout prediction with the global features. We also achieve the highest precision of 82%, recall of 85% and F-measure of 84% for non-breakout prediction with the global features. On the other hand, local features archive overall lower precision and recall of roughly 76%. These findings suggest that global measures outperform local measures in predictive accuracy.

Table 4. Break out results

Network	TP	FP	Precision	Recall	F-measure
Local	0.73	0.2	0.77	0.73	0.75
Global	**0.81**	**0.15**	**0.84**	**0.81**	**0.82**
All Features	0.8	0.16	0.83	0.8	0.81

Table 5. Non break out results

Network	TP	FP	Precision	Recall	F-measure
Local	0.79	0.27	0.75	0.79	0.77
Global	**0.85**	**0.19**	**0.82**	**0.85**	**0.84**
All Features	0.84	0.20	0.81	0.84	0.83

8 Conclusion and Future Work

In this paper, we develop a model for predicting breaking hashtags using a content independent network model comprising both local and global network features drawn from an indicative accumulation period of hashtag volumes. For the network model, we measured and experimented with the predictive accuracies of global and local features. We also examined their importance and rankings using PCA. Global features drawn for the accumulation period network showed higher predictive accuracy compared to the local features. Network based model with global centralities for the accumulation period network can be used as a general framework to predict breaking hashtags with an overall accuracy of 82%.As future work, we propose to study the utility of content based features such as sentiment analysis, and different types of sources.

Acknowledgments. This research was supported by US DoD ONR grant N00014-14-1-0477 and USAF AFOSR grant FA9550-15-1-0004.

References

1. Li, C., Sun, A., Datta, A.: Twevent: segment-based event detection from tweets. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 155–164. ACM (2012)
2. Newman, M.E.J. A measure of betweenness centrality based on random walks. Social networks 27.1, 39–54 (2005)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems **30**, 107–117 (1998). doi:10.1.16/S0169-7552(98)00110-X
4. Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. In: Proceedings of the National Academy of Sciences of the United States of America 101.11, PP. 3747–3752 (2004)
5. Weng, L., Menczer, F., Ahn, Y.-Y.: Virality prediction and community structure in social networks. Scientific reports 3 (2013)
6. Freeman, L.C.: A set of measures of centrality based on betweenness.Sociometry, 35–41 (1977)
7. Arruda, G., Barbieri, A., Rodrigues, F., Moreno, Y., Costa, L.: The role of centrality for the identification of influential spreaders in complex networks. Physical Review E **90**, 032812 (2014)
8. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted?. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 925–936. International World Wide Web Conferences Steering Committee (2014)
9. Pearson, K.: LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11), 559–572 (1901)
10. Bandalos, D.L., Boehm-Kaufman, M.R.: Four common misconceptions in exploratory factor analysis. Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences, 61–87 (2009)
11. Asur, S., et al.: Trends in social media: persistence and decay. ICWSM (2011)