Semantic Expansion of Hashtags for Enhanced Event Detection in Twitter

Ozer Ozdikis, Pinar Senkul, Halit Oguztuzun
Middle East Technical University
Ankara, Turkey
ozer.ozdikis, senkul, oguztuzn@ceng.metu.edu.tr

ABSTRACT

In this work, we present an event detection method in Twitter based on clustering of hashtags and introduce an enhancement technique by using the semantic similarities between the hashtags. To this aim, we devised two methods for tweet vector generation and evaluated their effect on clustering and event detection performance in comparison to word-based vector generation methods. By analyzing the contexts of hashtags and their co-occurrence statistics with other words, we identify their paradigmatic relationships and similarities. We make use of this information while applying a lexico-semantic expansion on tweet contents before clustering the tweets based on their similarities. Our aim is to tolerate spelling errors and capture statements which actually refer to the same concepts. We evaluate our enhancement solution on a three-day dataset of tweets with Turkish content. In our evaluations, we observe clearer clusters, improvements in accuracy, and earlier event detection times.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering*

General Terms

Algorithms, Performance

Keywords

Event Detection, Clustering, Twitter, Hashtags, Tweets in Turkish, Semantics, Word Co-occurrences

1. INTRODUCTION

With the introduction of micro-blogging platforms like Twitter, people can write about anything they want as their status messages and share these messages with their friends and followers [1]. Important events like elections, disasters, concerts, and football games can have immediate and direct impact on the density of status updates. As a result, Twitter provides a great medium for situation awareness using the messages written by millions of real world users. In Twitter environment, these messages are called "tweets". A tweet can be at most 140 characters long. Moreover, additional information can be given in a tweet by using special characters. For example, the hashtag symbol "#" can be used to attach descriptive key phrases, or a Twitter user can be referenced by typing a "@" symbol before his name.

There are studies that utilize Twitter platform for situation awareness [2]. The aim is to detect events happening in the real world by observing the tweets written by users and shared with the Twitter community. Event detection can be considered as a clustering process, where tweets with similar contents are grouped in order to identify popular concepts. Hashtags provide valuable information in that respect. By clustering tweets with the same (or similar) hashtags, it is possible to have an opinion about the popular events in Twitter. However, the same event can be annotated with different hashtags. In addition to that, people can make spelling errors while writing a hashtag. We believe that by tolerating such spelling mistakes and identifying semantically related hashtags, clustering results can be improved. We aim to detect such related hashtags by making word co-occurrence analysis on tweet contents and improve the performance of event detection by applying a lexico-semantic expansion on hashtags. In our previous work, we studied a similar idea using the whole tweet contents, without focusing on the hashtags [3]. The results for the use of hashtags indicated improvements in event detection accuracy and event detection time.

The novelty of this work is that we use tweet contents just in order to measure semantic similarities among the hashtags. Then these hashtags and their pairwise similarity scores are utilized while expanding the tweet vectors before the clustering process. In this paper, we first present previous studies related to event detection, semantic expansion techniques, and statistics of word co-occurrences in Section 2. Then we introduce our methods for enhanced event detection in Section 3, which is followed by the evaluation and analysis in Section 4. We finally present our concluding remarks.

2. RELATED WORK

Event detection is a task of Topic Detection and Tracking (TDT) studies [4]. TDT techniques are applied earlier for event detection using newspaper texts [5][6]. After the introduction of social networks, and especially Twitter, TDT approaches have started to be adapted for and applied on social networks. In one of the earliest studies about Twitter, features of tweets and users are analyzed in order to determine for which purposes Twitter is used [7]. In [8], characteristics of tweets and Twitter users are studied to discover trending topics, and to discover which features are important to detect trends. A recent study, TwitterStand [9], aims to detect events using tweets and to assign a geographic location to these events. An early earthquake

VLDB 2012 WOSS, August 31, 2012, Istanbul, Turkey

detection system in Japan using Twitter is introduced in [10]. There are other studies that aim to detect events from social networks other than Twitter [11].

Considering that tweets are maximum 140 characters long and they may have spelling differences for the same concepts, the event detection algorithms can be enhanced if we can identify similar words and use them interchangeably. Such a lexico-semantic expansion method has been applied for different purposes as in [12] and [13]. However, both of these studies require well-defined and mature dictionaries. Even if such comprehensive dictionaries were available in all languages, due to idiosyncratic way of spelling and very different writing conventions in Twitter, they would not be as useful as expected. Therefore, in order to infer semantic relationships between words, their co-occurrences statistics can be utilized. Co-occurrence relations are classified as syntagmatic (first-order) and paradigmatic (second-order) relations [14][15]. Syntagmatic relations are observed if two words appear together very frequently in texts (e.g. "blue" and "sky"). On the other hand, a paradigmatic relationship exists between two words if these words can be used interchangeably without affecting the structure and grammar of the sentences. A paradigmatic relation between two words is identified if they co-occur very frequently with the same set of other words (hence, second-order). For example, "blue", "dark", "bright" can be used interchangeably in texts together with the word "sky".

In our previous work, we applied a lexico-semantic expansion technique on tweet contents [3]. We analyzed first-order and second-order relationships between words, identified pairs whose similarities are above a threshold, and identified word pairs that are semantically related. During the clustering process, we expanded tweet vectors with constant values in accordance with these semantic relationships among words. Finally clusters whose tweet counts are above a threshold are identified as representatives of events. We briefly explain a part of our previous work in Section 3.3.1 and 3.3.2. On the other hand, in this work, we mainly focus on hashtags. Our semantic relationship analysis aims to identify similar hashtags by using other words in tweets as context descriptors. Moreover, we no longer use constant values while expanding a tweet vector or while identifying clusters for specific events. Instead, we utilize the similarity scores of relevant hashtags during semantic expansion, and make outlier analysis while deciding on important clusters. To the best of our knowledge, identifying relevant hashtags in tweets, and using them for event detection by applying a lexico-semantic expansion to tweet contents, has not been studied before. Moreover, since we use co-occurrence based statistical methods for the identification of semantic relationships, the methods that we present do not depend on any dictionary, and thus they are language-independent.

3. PROPOSED METHODS

In this work, for event detection, we implemented a clustering algorithm and applied it on tweet vectors generated by using four different methods. First two methods consider the whole tweet contents, without discriminating annotations from the tweet texts. These first and second methods differ in whether a lexico-semantic expansion on tweet vectors is applied or not. Third and fourth methods, on the other hand, use only the hashtags in tweets while generating tweet vectors. In the fourth method, the effect of applying a lexico-semantic expansion on hashtags is experimented.

3.1 Data Collection

Twitter provides an API to access Twitter data through its REST and Streaming API services. REST API is used to send requests about some specific users, tweets, locations or other objects in Twitter and get the response in JSON or XML format. Streaming API on the other hand, provides a stream of filtered tweets, i.e. tweets satisfying some desired criteria. In order to gather tweets posted by Turkish users, we defined geographic bounding boxes that cover almost all of Turkey, and added them as our filters for the streaming service. Whenever a Twitter user whose current location is set to be in these bounding boxes, posts a tweet, the streaming service notifies our client application. This results in tweets from random users, and thus tweets related to any topic, which may be useful for us to detect bursts on any kind of event. In addition to the textual content, tweets contain further attributes as their meta-data. Together with the tweet content, we save the tweet id, creation time, user id and hashtags in our database. Using this approach, we collected approximately 150K tweets per day. In our experiments, we used 388K tweets gathered between March 16, 2012 and March 19, 2012. For accessing Twitter services, we used [1]Twitter4J, an open source Java library that provides an abstraction and facilitates the use of Twitter services.

In the preprocessing step, we tokenize tweet contents into words by using space and punctuations as separators. The extracted words are stemmed using a freely available morphological analyzer for Turkish, named TRMorph [16]. After stop word elimination, we remove tweets with non-alpha numeric characters and the ones that indicate location check-ins. As a result, we obtained a set of 150K tweets to be used as an input for the event detection process.

3.2 Clustering Algorithm

Agglomerative text clustering is a widely used technique for event detection [17]. In agglomerative techniques, items (i.e. tweets in this context) are clustered according to their similarity in vector space model. In our agglomerative clustering implementation, values in tweet vectors, i.e. weights of the corresponding terms for each tweet, are set as TF-IDF values [4]. The tweet vectors generated using TF-IDF values are normalized in order to obtain unit-length vectors. Tweet creation times are also attached as attributes of these vectors. Clusters are represented with vectors as well. A cluster vector is calculated by taking the arithmetic mean of values in tweet vectors in each dimension. For a tweet to be added to a cluster, the similarity of their vectors must be above a threshold, calculated by applying the cosine similarity [6] on the corresponding vectors. After our empirical tests, we have chosen a threshold value of 0.6 in all clustering executions in this work. While processing tweets chronologically, if the highest similarity value between the current tweet and the cluster vectors is above the threshold, the tweet is added to the most similar cluster, resulting in updates on the cluster vector. Otherwise, if a similar cluster cannot be found for a tweet, the tweet initiates a new cluster on its own.

[1] http://twitter4j.org

As the result of clustering process, there may appear clusters with only a few tweets, which probably do not correspond to any specific event. On the other hand, it is possible to observe clusters grouping thousands of similar tweets and such clusters have potential to represent events. In order to distinguish ”event clusters“ (i.e. clusters with significantly higher number of tweets), we applied an outlier detection method using the empirical rule (also known as the three-sigma or 68-95-99.7 rule) [18]. The mean and standard deviation (σ) of the number of tweets in clusters are calculated, and the clusters with more than mean+3σ tweets are identified as ”event clusters”.

3.3 Tweet Vector Generation

We developed four different tweet vector generation methods and compared their performances in our clustering algorithm. First of two methods have been proposed in [3]. In the present work, the emphasis is on the last two methods, namely tweet vector generation by using hashtags with and without semantic expansion. We aim to evaluate the event detection performance by using only hashtags instead of the whole tweet contents.

3.3.1 Method-1: Using Words in Tweets without Semantic Expansion

In the first method, in order to define tweet vectors, we first extract distinct words from tweets. In our dataset of 150K tweets, we identified 32766 distinct words that appear in at least two different tweets. These words can be numbers, hashtags or even meaningless terms. Based on the vector space model, each tweet is represented as a vector of length 32766 whose values are TF-IDF values of the corresponding words. Applying our clustering process on these vectors produced 3288 clusters, 50 of which are identified as event clusters (i.e. outliers).

3.3.2 Method-2: Using Words with Semantic Expansion

The second method applies a lexico-semantic expansion to tweet contents before the application of clustering. Our semantic expansion process inserts a positive value to the tweet vector for a word w_i if 1) w_i does not already appear in the tweet, and 2) the tweet contains a word w_j which is semantically related to w_i. Since second-order relations give us words that can be used interchangeably, we mainly focus on second-order relationships for semantic expansion. In order to identify second-order relations, each word is represented by vectors of co-occurrence values as given in equation 1. In this vector, c_{ij} is the co-occurrence count of words w_i and w_j in tweets, and W represents the set of words in our corpus.

$$\vec{v}_i = (c_{i,1}, c_{i,2}, \ldots c_{i,i-1}, 0, c_{i,i+1}, \ldots, c_{i,|W|-1}, c_{i,|W|}) \qquad (1)$$

At this step, by using the similarity between two vectors, we find the degree of commonality in their contexts and identify word pairs that are semantically related. We measured the pairwise distances among word co-occurrence vectors using city-block distance, and marked closest pairs as semantically related. City-block distance is simply the sum of the differences of vector values in each dimension, and it is stated in [14] that it yields good results. Since the co-occurrence vectors can be very sparse, we consider the words with a document frequency of higher than 100. Without this condition, apparently unrelated words can have high similarity values, since they have zeroes in most of their dimensions. Using this threshold, we obtained 442 relationships with city-block distance [3]. Several examples of paradigmatically related words are given in Table 1.

Table 1. Examples for Paradigmatic Relations

Word-1	Word-2	Reason for relation
alexsandra	alexandra	A frequent spelling mistake for the name “alexandra”
galatasaray	fenerbahce	These football clubs had a derby match in that time period

Before the clustering process, given the original content of a tweet, we look for the semantically related words that do not already appear in the tweet. Since these words do not actually appear in the original tweet content, they should be added to the tweet vector with a TF value that is smaller than 1. In our experiments, we used a TF value of 0.5 during this semantic expansion process. As a result, clustering by using second-order relationships resulted in 3178 clusters. Among them, 44 clusters are detected as outliers (i.e. group significantly higher number of tweets) and thus identified as event clusters.

3.3.3 Method-3: Using Hashtags without Semantic Expansion

Hashtags are important annotations for tweets that provide hints about their topics. In our tweet corpus of three days, we observed that more than 17K of tweets contain at least one hashtag. In this vector generation method, we constructed tweet vectors by using only the hashtags, ignoring the non-hashtag words. Moreover, it is also observed that some users write several hashtags in a tweet, even if they are completely irrelevant with each other or with the tweet content. Therefore, in clustering, we used only tweets with a single hashtag. Since we work with a much lower volume of data (only tweets with single hashtags), in comparison to vector generation by using the whole tweet content, the clustering process resulted in fewer clusters. 868 clusters are generated with this method, 13 of which are identified as event clusters.

3.3.4 Method-4: Using Hashtags with Semantic Expansion

Similar to the lexico-semantic expansion we applied on words in Section 3.3.2, we identify the semantic relationships between hashtags and apply a semantic expansion to tweet vectors generated by using the hashtags. However, here we apply a different co-occurrence analysis and similarity metric than in Method-2. Our basic intuition is that a hashtag should be a summary of or contextually very relevant to other non-hashtag words in a tweet. In other words, non-hashtag words can be used as a kind of context descriptor for the hashtags. For example, consider two tweets like “galatasaray plays against fenerbahce #derby” and “fenerbahce must win against galatasaray #match”, and assume there are many other tweets with similar content. Since the hashtags “#derby” and “#match” are used together with the same words in tweets, their co-occurrences with words like “galatasaray”, “fenerbahce” or “against” should indicate us that these hashtags are semantically relevant. Therefore, we count the co-occurrences of hashtags with non-hashtag words in tweets and generate co-occurrence vectors for them. The similarity metric we used here is cosine distance. Cosine distance gives us a numerical similarity value in the range of [0, 1], which will be used as TF values in the semantic expansion of tweet vectors. Moreover, these vectors are even sparser than the word co-occurrence vectors, and we do not want to restrict this similarity analysis based on document frequency as we did for Method-2. As a result, pairwise cosine similarities are found and for each hashtag the most similar three hashtags are stored in our database together with their cosine distance. Some examples of most similar hashtags and their cosine distances (i.e. similarity scores) are presented in Table 2. The meanings of these hashtags in English can be seen in Figure 1 and Figure 2.

Table 2. Examples for Relations between Hashtags

Word	Related Words	Similarity Score
#canakkalegecilmez	#çanakkalegeçilmez	0.77
	#18martcanakkalezaferi	0.59
	#canakkalagecilmez	0.56
#bugungunlerdenfenerbahce	#yilinderbisi	0.71
	#galatasaray	0.64
	#fenerbahcederlerbenimadima	0.59
#yilinderbisi	#bugungunlerdenfenerbahce	0.71
	#bukorkusizeyeter	0.66
	#bugungunlerdengalatasaray	0.63

While applying semantic expansion to tweet vectors generated from hashtags, instead of giving a constant TF value for the semantically related hashtags (such as 0.5 as we did in Method-2), we use their cosine similarity scores as TF values. If two hashtags have completely the same meaning with the same co-occurrence vectors, then their similarity score will be 1. Otherwise, the score will definitely be less than 1. Using these similarity values, a tweet vector with a single hashtag is expanded with three similar hashtags and then used in the clustering process. This method produced 761 clusters, 10 of which are marked as event clusters.

Cluster counts and event cluster counts obtained under each of the described tweet vector generation methods are presented in Table 3. It is worth pointing out the differences in the number of clusters, their sizes and the number of event clusters. Using hashtags, there are fewer clusters but with larger number of tweets.

Table 3. Cluster Statistics

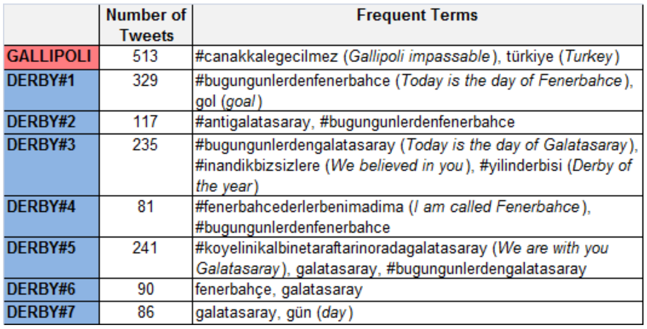
	M#1	M#2	M#3	M#4
Cluster Count	3288	3178	868	761
Event Cluster Count	50	44	13	10
Mean	10.328	10.231	16.154	18.425
Std. Dev.	23.174	23.715	80.401	99.710

The choice of coefficients and threshold values has been made by manually observing the outputs, and they can be improved by applying a more detailed analysis. However, we would like to remind that the focus of this work is to show that by keeping all constants and conditions intact, event detection is still feasible by using hashtag information only.

4. EVALUATION

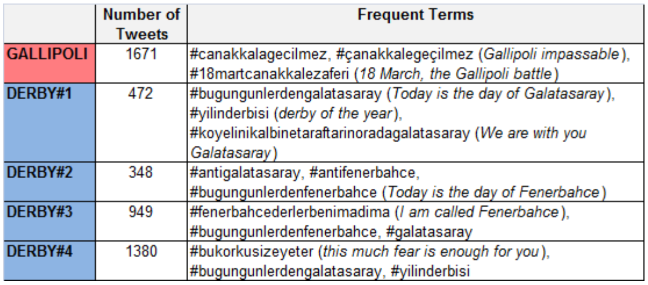
During the 3-day tweet collection period between March 16 and March 19, 2012, we know that there were two important events in Turkey. The first one is the anniversary of a historic battle in Canakkale, known as the "Battle of Gallipoli", an episode of the "Dardanelles Campaign" (March 18, 1915). We named this event "GALLIPOLI" in this study. The second event was an important football match between Fenerbahce and Galatasaray, which was followed by millions of football fans in Turkey. We refer this event as "DERBY". Our goal was to detect these two events. Since we did not implement any credibility filter in order to decide whether a tweet is about an event or it is a dialog between two people, some event clusters that we found may not belong to either of these events [19]. We classify them as "CHAT" and simply ignore them in this study.

We manually annotated a subset of tweets in our corpus by using a search-guided annotation technique [20], and identified 4331 tweets for DERBY (3659 of them contain a hashtag) and 2026 tweets for GALLIPOLI (2104 of them contain a hashtag). As a result of our clustering process, we obtained several event clusters, some of which are about the DERBY or GALLIPOLI events. When we analyzed the contents of event clusters, we found that all four vector generation methods result in a single event cluster for the GALLIPOLI event. However, for the DERBY event there are more than one but different number of event clusters. The event cluster counts for the DERBY event that are generated by our four methods are 7, 7, 6, and 4 respectively. For illustrative purposes, we present the details of event clusters produced by Method-1 in Figure 1, and the ones produced by Method-4 in Figure 2.



	Number of Tweets	Frequent Terms
GALLIPOLI	513	#canakkalegecilmez (*Gallipoli impassable*), türkiye (*Turkey*)
DERBY#1	329	#bugungunlerdenfenerbahce (*Today is the day of Fenerbahce*), gol (*goal*)
DERBY#2	117	#antigalatasaray, #bugungunlerdenfenerbahce
DERBY#3	235	#bugungunlerdengalatasaray (*Today is the day of Galatasaray*), #inandikbizsizlere (*We believed in you*), #yilinderbisi (*Derby of the year*)
DERBY#4	81	#fenerbahcederlerbenimadima (*I am called Fenerbahce*), #bugungunlerdenfenerbahce
DERBY#5	241	#koyelinikalbinetaraftarinoradagalatasaray (*We are with you Galatasaray*), galatasaray, #bugungunlerdengalatasaray
DERBY#6	90	fenerbahçe, galatasaray
DERBY#7	86	galatasaray, gün (*day*)

Figure 1 - Relevant Event Clusters Produced Method-1



	Number of Tweets	Frequent Terms
GALLIPOLI	1671	#canakkalagecilmez, #çanakkalegeçilmez (*Gallipoli impassable*), #18martcanakkalezaferi (*18 March, the Gallipoli battle*)
DERBY#1	472	#bugungunlerdengalatasaray (*Today is the day of Galatasaray*), #yilinderbisi (*derby of the year*), #koyelinikalbinetaraftarinoradagalatasaray (*We are with you Galatasaray*)
DERBY#2	348	#antigalatasaray, #antifenerbahce, #bugungunlerdenfenerbahce (*Today is the day of Fenerbahce*)
DERBY#3	949	#fenerbahcederlerbenimadima (*I am called Fenerbahce*), #bugungunlerdenfenerbahce, #galatasaray
DERBY#4	1380	#bukorkusizeyeter (*this much fear is enough for you*), #bugungunlerdengalatasaray, #yilinderbisi

Figure 2 - Relevant Event Clusters Produced by Method-4

In these figures, it is obvious that the number of tweets collected in event clusters using hashtags is much higher. Moreover, since there are fewer number of event clusters, it is easier for users to distinguish popular events in Twitter.

After comparing these results with our manually annotated tweet dataset, we observed a high precision for all methods we used. On the other hand, there is a remarkable improvement in coverage when hashtags are used for event detection. Moreover, application of semantic expansion results in a slight increase in the overall accuracy, for both cases of using words and hashtags. The results of our precision/recall analysis for both events are presented in Figure 3 and Figure 4.

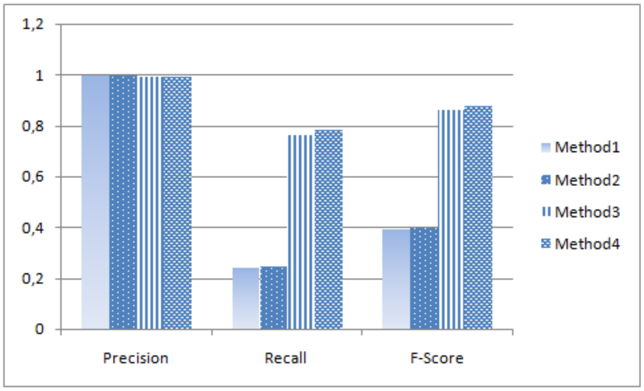


Figure 3 - Accuracy Analysis for the GALLIPOLI event

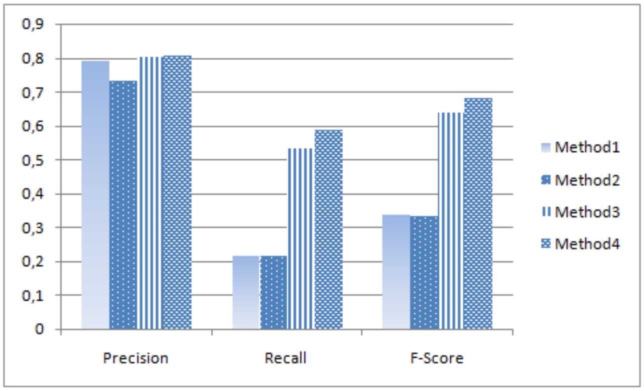


Figure 4 - Accuracy Analysis for the DERBY event

In addition to the precision/recall analysis, we also analyzed the posting times of tweets in event clusters. This information gives us the event detection times and their time span, as shown in Figure 5. The positive impact of semantic expansion is obvious in lifetimes of event clusters. A lexico-semantic expansion, no matter whether it is applied on all words or only on hashtags, tolerates the spelling mistakes and consolidates the tweets with similar meanings, resulting in larger event lifetimes. Especially for the DERBY event, the effect of expansion is several hours of earlier event cluster generation.

Other improvements that we observed by using only the hashtags in tweet vectors are about the execution times of similarity calculations and clustering processes. Since hashtag-based vectors consist of a limited number of attributes, similarity calculations between these vectors take much shorter time in comparison to those of word-based tweet vectors.

5. CONCLUSION

In this work, we analyze the effectiveness of hashtag-based methods for event detection in Twitter in comparison to that of word-based methods. We propose a co-occurrence based method for identifying the semantic relationships among the hashtags, and use these relationships in order to enhance event detection techniques. We present four different implementations by using all words in tweets (and applying a semantic expansion on them), and by using only hashtags in tweets (and applying another semantic expansion on them).

The experimental results show that by using hashtag-based tweet representation, higher accuracy is obtained for event detection. Moreover, events are detected earlier in the analysis. This effect is further increased under the lexico-semantic expansion technique that we implemented. Therefore, our findings suggest that hashtags can be representative for events in Twitter and event detection can be performed by using this compact information.

The results we obtain should be regarded as preliminary findings and the study can be extended in several directions. First of all, we plan to extend the analysis for longer tweet collection periods, including other events. In addition, studying on tweets including several hashtags for event detection is another future extension.

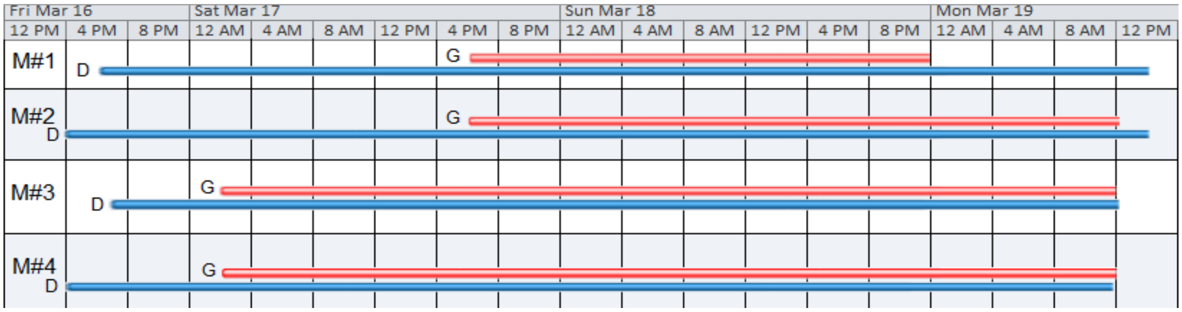


Figure 5 - Event Detection Times of four Methods for the Events GALLIPOLI (G) and DERBY (D)

6. REFERENCES

[1] Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B., and Magoulas, R. 2008. Twitter and the Micro-blogging Revolution. O'Reilly Radar Report. http://weigend.com/files/teaching/haas/2009/readings/OReillyTwitterReport200811.pdf

[2] Petrovic, S., Osborne, M., and Lavrenko, V. 2010. Streaming First Story Detection with Application to Twitter. In Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT), LA, California

[3] Ozdikis, O., Senkul, P., and Oguztuzun, H. 2012. Semantic Expansion of Tweet Contents for Enhanced Event Detection in Twitter. International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Istanbul/Turkey (to be published)

[4] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic Detection and Tracking Pilot Study: Final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop

[5] Teitler, B., Lieberman, M.D., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. 2008. NewsStand: A New View on News. In Proc. of ACM SIGSPATIAL GIS. CA, USA, pp: 144-153

[6] Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H.C., and Uyar, E. 2010. New event detection and topic tracking in Turkish. Journal of the American Society for Information Science and Technology. Vol. 61, No. 4, pp: 802-819

[7] Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why We Twitter: Understanding Microblogging Usage and Communities. In Proc. of SNA-KDD 2007 Workshop on Web mining and social network analysis. San Jose, California, pp: 56-65

[8] Naaman, M., Becker, H., and Gravano, L. 2011. Hip and Trendy: Characterizing emerging trends on twitter. Journal of the American Society for Information Science and Technology (JASIST). Vol: 62(2), pp:902–918

[9] Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., and Sperling, J. 2009. TwitterStand: News in Tweets. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 04-06, Seattle, Washington

[10] Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. Earthquake Shakes Twitter users: Real-time Event Detection by Social Sensors. Proceedings of the 19th international conference on World Wide Web. North Carolina, USA

[11] Sayyadi, H., Hurst, M., and Maykov, A. 2009, Event Detection and Tracking in Social Streams. In Proceedings of ICWSM 2009, San Jose CA, USA

[12] Pembe, F.C., and Say, A.C.C. 2004. A Linguistically Motivated Information Retrieval System for Turkish. Lecture Notes in Computer Science. 3280, pp: 741–750

[13] Frasincar, F., IJntema, W., Goossen, F., and Hogenboom, F. 2011. A Semantic Approach for News Recommendation. In Business Intelligence Applications and the Web: Models, Systems and Technologies. IGI Global

[14] Rapp, R. 2002. The computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In Proceedings of COLING, Taiwan

[15] Rapp, R. 2004. A Freely Available Automatically Generated Thesaurus of Related Words. In Proceedings of 4th International Conference on Language Resources and Evaluation (LREC), Portugal

[16] Coltekin, C. 2010. A Freely Available Morphological Analyzer for Turkish. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)

[17] Steinbach, M., Karypis, G., and Kumar, V. 2000. A comparison of document clustering techniques. Tech.Rep. 00\034, Univ. Minnesota

[18] Wang, M., Hsu, P., and Chuang, Y.C. 2011. Mining Workflow Outlier with a Frequency-Based Algorithm. Journal of Control and Automation, Vol.4, No.2

[19] Castillo, C., Mendoza, M., and Poblete, B. 2011. Information Credibility on Twitter, In Proceedings of World Wide Web Conference

[20] Fiscus, J.G., and Doddington, G.R. 2002. Topic detection and tracking evaluation overview. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers, pp:17-31