2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing

Monitoring Trends on Facebook

Irena Pletikosa Cvijikj, Florian Michahelles
Information Management
ETH Zurich
Zurich, Switzerland
{ipletikosa, fmichahelles}@ethz.ch

***Abstract*—The public nature of user generated content on social media platforms offers the possibility for trend monitoring as an insight into the topics that attract the attention of a large fraction of users. While Google Trends and Twitter have already been recognized as a valuable source of trend information by the practitioners and scholars, at the moment there are no practical implementations or research efforts in the field of trend detection over Facebook public posts. In this paper we present two contributions towards trend monitoring over Facebook public posts. We propose and evaluate a system for trend detection based on the characteristics of the posts shared on Facebook. Based on our results we propose three categories of trending topics: 'disruptive events', 'popular topics' and 'daily routines'. We analyze and compare the characteristics of the proposed categories in terms of distribution and information diffusion in order to increase the understanding of emerging trends on Facebook. Finally we draw conclusions from our findings in terms of challenges and opportunities for future work in this direction.**

Keywords-trend detection; Facebook; social networks; social media mining; text mining; information retrieval

I. INTRODUCTION

The emergence of the Web 2.0 has changed the way content is generated on the web. Rather than being just passive consumers, users became active participants by sharing information, experiences and opinions with each other. Social networks (SNs), as a part of Web 2.0 technology, provide the technological platform for individuals to connect, produce and share content online [6]. At the moment, Facebook[1] is the largest SN with more than 800 million active users[2], followed by Twitter[3] with an estimated 280 million registered users[4].

The value of the content generated on social networks as a source of information was soon recognized, resulting in individuals turning to social networks as sources of real-time news and opinions [24], [15]. This form of usage has further been supported by the platform providers, by offering the possibility for searching through the vast amount of public status updates to monitor content or find temporally relevant information [29]. In addition, they have offered the possibility to access the public status updates through their search APIs, resulting in a burst of commercial and research efforts to gather knowledge through analysis of the shared content.

Detection and analysis of trends offer valuable insights into the topics that attract the attention of a large fraction of SN users. Public opinions in form of trends are interesting not only for individuals, but also for (1) news reporters [8], pointing to the fast-evolving news stories, (2) sociologists, revealing the 'spirit of the times' [18], (3) marketing professionals, for brand image monitoring and benchmarking [11], (4) opinion tracking companies, e.g. for prediction of elections outcome [30] and (5) scholars, for explaining the underlying phenomena of SN.

As the number of available sources and the amount of online information increase, individuals and companies interested in fast discovery of trends through monitoring of the conversation on social media platforms need to rely on the tools capable of automatic topic detection and monitoring. This has motivated research into text analysis and application of the existing information retrieval and trend detection techniques to social media in order to benefit from the knowledge enclosed within the user generated content (UGC).

Trend monitoring over Twitter stream has already been the subject of attention of scholars and professionals, resulting in numerous modified and new algorithms for information retrieval and commercial online tools. However, to the best of our knowledge, at the moment there are no commercial or research efforts related to trend detection on Facebook. This situation is due to the fact that Twitter has provided the possibility to collect the public data significantly earlier compared to Facebook. In addition, privacy policies on Facebook introduce limitations into the amount and type of available data. Still, we believe that UGC on Facebook could bring valuable insights, since Facebook is the largest social network with more than 1 billion pieces of content created on a daily basis[2].

In this paper we present two contributions towards trend monitoring over Facebook public posts. We propose and evaluate a system for trend monitoring based on the characteristics of the posts shared on Facebook. Based on our results we distinguish between three categories of trending topics: (1) 'disruptive events', (2) 'popular topics' and (3) 'daily routines'. We analyze and compare the characteristics

[1] http://www.facebook.com/
[2] http://www.facebook.com/press/info.php?statistics
[3] http://twitter.com/
[4] http://harpsocial.com/2011/04/social-medias-shocking-statistics/

978-0-7695-4612-4/11 $26.00 © 2011 IEEE
DOI 10.1109/DASC.2011.150

of the proposed categories in terms of distribution and information diffusion in order to increase the understanding on emerging trends on Facebook. Finally we draw conclusions from our findings in terms of challenges and opportunities for the future work in this direction.

II. Related Work

Trend Detection (TD) is a research field that has been of interest for information retrieval and text mining applications for a long time. A recent variation of the concept appeared under the notion of Emerging Trend Detection (ETD), which according to Kontostathis et al. [14] is characterized by a "topic area that is growing in interest and utility over time". A system performing detection of emerging trends uses a document corpus as input and identifies topics that are new or show significant growth in importance within the corpus. In turn, ETD builds up on the results from the work of the Topic Detection and Tracking (TDT) initiative [1].

TDT explores the possibility and required mechanisms for topic extraction from a time-stamped corpus of documents, such as news channels. The task of topic extraction is divided into five separate research problems: (1) story segmentation, (2) first story detection, (3) cluster detection, (5) tracking, and (5) story link detection [1]. Each of these elements has caused a separate thread of research within the field of information retrieval. The results of the TDT project have further supported the development of various automated systems for detection and tracking of emerging topics through time, i.e. trend detection and monitoring [14].

Social media as a source of information has recently attracted the attention of these research communities. However, most of the efforts related to trend detection over social media focus on analysis of the long structured text discussions from blog posts, such as [3]. Recently, an additional source for trend detection has been recognized in the Google search queries, i.e. Google trends [31]. By comparison, detection of trends in UGC on social networks is still in its infancy.

The change brought about by social media towards short commentary, as introduced by social networks such as Twitter and Facebook, resulted in a significant difference in the comment structure and language, imposing additional challenges to the existing text mining techniques [28]. For that reason, the majority of the previous research over the content shared on social networks focus on understanding the users, activities, opportunities and challenges of these platforms. Furthermore, these studies mostly apply the text mining techniques on comments from Twitter. They investigate the value of tweets as online word-of-mouth [11], possibilities for movie revenue prediction [2] and opportunities for television broadcasters [9], sentiment analysis [5], avoiding traffic jams [32], web-based intelligence retrieval and decision-making [7], etc. However, the number of studies regarding Facebook is still relatively limited [23].

The efforts in the direction of trend detection on Twitter lead also to the creation of several commercial tools for monitoring trends. Apart from the official tool offered by Twitter, i.e. Twitter Search[5], there are many other platforms which provide similar features, such as Trendistic [6], Trendsmap[7], etc. In addition, research oriented platforms are being created to provide scholars with the tools that would enable investigation of the trend generation phenomena [16], [10].

From the research perspective, there have been two major streams: (1) adapting the existing and providing new algorithms for text mining, and (2) understanding the phenomena of trend occurrence and spreading. The work belonging to the first stream includes examples such as improved algorithms for first story detection based on locality-sensitive hashing [19], overcoming the problems of document summarization through definition of a notion of hybrid documents in the traditional term frequency approach [27], usage of LDA for topic identification [22], trend detection via keyword clustering [16], news recommendation [20], etc.

In the direction of the second stream, Asur et al. [2] have provided a theoretical basis for the formation, persistence and decay of trends. Becker et al. [4] have recognized the value of Twitter as a source of real-time event content. Naaman et al. [17] reveal the value of tweets for gathering information for, and about, a local community. Kwak et al. [15] investigate different characteristics of Twitter trends in terms of the number of replies, mentions, retweets, and "regular" tweets that appear in the set of tweets for each trending term. Sakaki et al. [25] study social, spatial, and temporal characteristics of earthquake-related tweets and Diakopoulos et al. [8] analyze tweets corresponding to the large-scale media events to improve reasoning, visualization and analytics.

To the best of our knowledge, at the moment there is no existing research focusing on the topic of trend detection on Facebook. In addition, there is a single commercial implementation, i.e. Facebook Trends [8], limited to the discovery of the trending topics over the public posts from German speaking Facebook users, based on term frequency weighting.

In this paper we describe and evaluate the system for fully-automatic trend detection over the full scope of public user posts shared on Facebook. Based on the obtained results we distinguish between three categories of trending topics: (1) 'disruptive events', (2) 'popular topics' and (3) 'daily routines'. We analyze and compare the characteristics of the proposed categories in terms of distribution and information diffusion in order to increase the understanding on emerging trends on Facebook. Finally we draw conclusions from our findings in terms of challenges and opportunities for the future work in this direction.

III. System Description

Trend monitoring over Facebook's public comments could be divided in two major steps: (1) data collection, and

[5] http://search.twitter.com/

[6] http://trendistic.com/

[7] http://trendsmap.com/

[8] http://www.facebook-trends.de/

(2) trend detection. Providing near real-time trend monitoring over the full set of public posts shared on Facebook assumes data collection that is (1) continuous, (2) real-time, and (3) provides access to the full scope of public posts. Trend detection is commonly based on (1) topic identification, and (2) cluster detection.

In the continuation of this paper we explain the details of both steps of the process and provide evaluation of the presented approach.

IV. Data Collection

Data collection presents one of the challenges of trend detection over the UGC on Facebook. While Twitter released the Streaming API [9] in April 2009 [10], allowing high-throughput, near real-time access to various subsets of public and protected Twitter data, Facebook provided a similar, but limited functionality through its Graph API[11] a full year later, in April 2010[12].

The Facebook Graph API provides access to the Facebook social graph via a uniform representation of objects in the graph (e.g., people, posts, pages, etc.). For our study the Post objects were of interest. Each Post object contains the following information: (1) content details for the post (message, name, caption, description), (2) Facebook user who posted the message, (3) type of the message as defined by Facebook, i.e. status, photo, link, etc., (4) time of creation, (5) application used for sharing the post, etc. All of these elements were stored in a relational database for further investigation.

Since the API does not provide the possibility to receive posts in the form of a real-time stream, we used the search feature of the Graph API, returning a list of public posts for a given keyword. In order to collect all the public posts we applied a simple algorithm which performs search for each ASCII character, thus providing a loop of 26 search queries. In addition, the Graph API offers the "limit" parameter (N_L) for the search query which defines the number of returned post objects (default is 25, maximum 500). In case there are more than N_L available posts for the given keyword, the JSON response contains the URL for the next query. This results in the possibility of having more than one sub-query for a given keyword, depending on the required time interval. Algorithm 1 shows the pseudo-code listing for collecting public posts on Facebook.

Algorithm 1: Collection of public posts from Facebook

```
1  until = getLastCollectionTime();
2  for each asciiChar in asciiList do
3     nextURL = collectPosts (asciiChar, until);
4  until (nextURL != null)
```

The selection of the data collection interval duration was based on the following two premises: (1) the interval should be long enough to be able to capture trending topics, and (2) short enough to provide possibility for near-real time monitoring while overcoming the processing challenges over the large datasets. Based on this reasoning, and the results of the parameter tuning of our system, we propose the data collection interval of 10 minutes.

V. Dataset Characteristics

Using the previously described algorithm, we collected posts from July 22, 2011 until July 26, 2011. This resulted in 2,273,665 posts in total. The average number of posts fetched with the proposed algorithm on a daily basis was 568,416 out of the estimated 1 billion pieces of content created daily on Facebook[2]. We selected these particular dates for the following reasons (1) two big events that captured the attention of the people on other forms of media (newspapers, TV) happened during these days, and (2) people access social networks and interact more frequently on weekends. The two different events, both causing great emotional reaction of the public occurred during this period are: (1) the terrorist attack in Norway[13], at 13:26 GMT on July 22nd, when 77 people died and 96 more were injured, and (2) the death of the English singer and songwriter Amy Winehouse[14] who had been attracting attention with her great talent, but also a very controversial life style, at the age of 27, on July 23rd, at 14:54 GMT.

To understand the characteristics of the Facebook posts relevant for the trend detection we performed linguistic analysis over three sets of posts. From Table 1, one could see that there are no significant differences between the three sets. The average number of sentences in a post is approximately 1.4. At the same time, the average number of words is approximately 18, a bit higher compared to the 16 words in a tweet [11]. However, looking at the full dataset, while the average post length in character didn't significantly differ from our results (108), the maximum length was found to be 754 characters, which on average corresponds to approximately 10 sentences and 122 words.

In addition, tweets allow only for textual input, while Facebook supports five different post types: (1) status, (2) video, (3) link, (4) photo and (5) music. Distribution of each of these types over the full dataset is presented in Fig. 1. It can be seen that 84% of the posts belong to the type of status post, followed by video (10%), links (4%) and photos (2%). Music posts were present with only 350 occurrences (0%) in the dataset.

In terms of the language used in the posts, based on the classification performed with the LingPipe API[15] over the three datasets, we can conclude that English was the dominant language, present in 78% of the posts, as shown in Table 1.

Finally, we assume that the cumulative distribution of posts over time of day, as presented on Fig. 2, might also have an effect over the distribution of the trending topics.

[9] https://dev.twitter.com/docs/streaming-api

[10] http://alturl.com/ekzqr

[11] http://developers.facebook.com/docs/reference/api/

[12] http://www.facebook.com/f8

[13] http://en.wikipedia.org/wiki/Norway_attacks

[14] http://en.wikipedia.org/wiki/Amy_Winehouse

[15] http://alias-i.com/lingpipe/

TABLE I. LINGUISTIC STATISTICS FOR POSTS.

	Average Length		
	5K posts	***50K posts***	***100K posts***
Sentences	1.43	1.44	1.43
Words	17.43	17.53	17.43
Characters	103.89	103.39	102.80
	Language		
English	78%	78%	77%

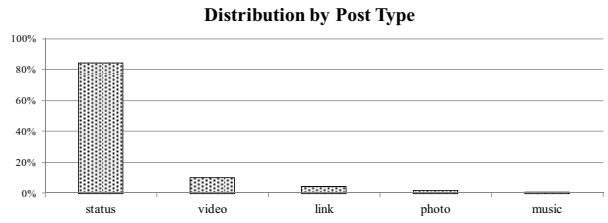

Figure 1. Distribution of posts by post type

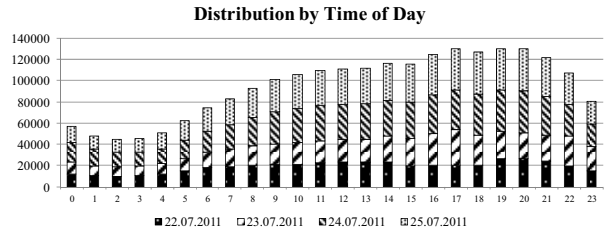

Figure 2. Distribution of posts by time of the day.

It can be seen that the lowest level of activity in terms of number of shared posts over the selected time interval occurred at 2:00 GMT, while the highest number of posts were shared between 17:00 GMT until 21:00 GMT. We will use the presented results as a basis for our further reasoning.

VI. TREND DETECTION

A. Post Topic Identification

Based on the results of the analysis of the post type distribution we base the trend detection only on the content shared in form of 'status' posts. Following the common approach [13] we begin by discovery of the most significant terms within the collection of Facebook public posts. In the context of this paper, a term is an n-gram with a length from 2 up to 5 words belonging to the same sentence within the post. Usage of bigrams as a lower limit was based on the results of our experiments with different lengths where unigrams introduced too much noise by having common words appearing high on the weighted list while carrying no topic information (e.g. "love"). Before the creation of the lists of n-grams we performed a preprocessing of the posts by applying (1) stop-words filtering, based on the predefined list only for English, and (2) removing the URLs from the posts.

For the weighting of the terms we decided to use the "Term Frequency – Inverse Document Frequency" (TF-IDF) [12] approach due to its simplicity. The method assigns a weight to a term based on two measures: (1) the frequency of occurrence of a term within a single document, and (2) the number of documents in the corpus which contain the given term. Therefore, the basic form of the formula would be:

$$w(t_i) = tf(t_i, d_j) * \log_2 \frac{N}{df(t_i)}, \quad (1)$$

where N is the total number of documents in the corpus, $tf(t_i, d_j)$ is the frequency of term t_i within the document d_j, and $df(t_i)$ is the number of documents in the corpus containing the term t_i [26]. However, this formula is not applicable for the content shared on Facebook because of the limited length of the Facebook posts, which would reduce the value of the term frequency component in the equation. Furthermore, if we calculate the frequency over the full corpus, we end up with a single document, thus losing the inverse document frequency component.

In order to overcome this problem, following the example of the previous work over Twitter [27], we use the concept of a hybrid document. In this paper the notion of a hybrid document represents a collection of posts $P = \{p_1, p_2, ..., p_K\}$, obtained within a timeframe T, which corresponds to the interval for fetching posts from Facebook in a near real-time system. Each time frame T represents a separate dataset described by a separate weighted list. In addition, based on the results of the linguistic analysis, we assume that there might be more than one occurrence of the same term within a post. For that reason, in case of large datasets, such as the one we are exploring, and due to the smoothing shape of the logarithm function, a term that appears more than once in a single post might have higher weight compared to a term which occurs in several posts. To avoid this, when calculating the term frequency, we do not use the sum of all occurrences over all posts $p_j \in P$. Instead we count only one occurrence of a term per post. Based on this discussion, the modified version of the formula we propose is:

$$w(t_i) = tf(t_i) * \log_2 idf(t_i), \quad (2)$$

$$tf(t_i) = \frac{\#\text{PostsContainingTerm}}{\#\text{AllTermsOverP}}, \text{ and} \quad (3)$$

$$idf(w_i) = \frac{|P|}{\#\text{PostsContainingTerm}}, \quad (4)$$

B. Post Clustering

Post topic identification results in an ordered list of the most significant terms in the corpus. The next step is to cluster together terms that belong to the same topic. We perform post clustering in two steps (1) clustering by distribution, and (2) clustering by co-occurrence.

Clustering by distribution is a combination of the comparison of the term weight and the intersection of the related documents. The goal is to eliminate the multiple occurrences of the similar n-grams with different lengths belonging to the same posts (e.g. terms "amy wihehouse" and "amy winehouse dead" extracted from the same post will appear as separate terms, having the same weight and containing the same information which introduces redundancy). Once the grouping is done, we replace the groups with the n-gram with the maximal length since it contains maximum information regarding the topic.

Algorithm 2: Clustering by distribution

```
1  for each term in sortedWeightList do
2    if (termWeight != previousTermWeight) then
3      createNewGroup(term);
4    else
5      for each group in topicGroups do
6        if (getPosts(group) = getPosts(term)) then
7          addTermToGroup(term, group)
8        else
9          createNewGroup(term);
10       end
11     end
12   end
13   weight = termWeight
14 end
```

Clustering by co-occurrence is based on the assumption that terms that appear frequently in same posts belong to the same topic. This step is used to further group the terms that are not semantically similar and belong to different posts, but still refer to the same topic, such as "amy winehouse" and "drug addict". The resulting list of topic groups is ordered in accordance with the highest term weight in the group.

Algorithm 3: Clustering by co-occurrence

```
1  for each term in sortedWeightList do
2    for each group in topicGroups do
3      if (getPosts(group) ∩ getPosts(term) != 0) then
4        addTermToGroup(term, group)
5      else
6        createNewGroup(term);
7      end
8    end
9  end
```

VII. Preliminary Evaluation

In order to perform a preliminary evaluation of the proposed algorithm we used the common approach of measuring the precision and recall of our system [21]. For that purpose we conducted a review of the results obtained from 10 experiments, each collecting and processing 1000 posts from different time intervals. For each experiment, evaluation was conducted over the same three topic groups that commonly occurred over the observed time interval. For each topic group, a list of corresponding posts was examined. In addition, in case of the Norway incident, two evaluations were conducted: one assuming that the group containing the majority of the related posts is the representative for the 'true positive' categorization (denoted as 'max posts'), and the second one based on the selection of the topic group that most accurately describes the actual event as a representative for the 'true positive' score ('best fit'). The results of the evaluation are presented in Table 2.

The obtained values for precision and recall show that our approach generates relatively good results for the topics of 'Amy Winehouse' and 'Harry Potter'. The commonality between these two topics lies in the fact that they are both described with a personal name containing two words, which corresponds to our minimum n-gram length.

In case of the Norway incident, the clustering algorithm didn't perform as well as for the other two topics. Instead of having a single topic group related to the events in Norway, the algorithm placed a majority of the posts within the Amy Winehouse group as a result of the co-occurrence clustering, while the remaining of the related posts were scattered over multiple topic groups, mostly consisted of a single post. This resulted in a very low value of the F-measures for both approaches. The main difference between Norway and the previous two topics is that it has occurred in a very small number of posts and with a great diversity in terms of used words within the posts resulting in clustering problems.

The average values presented in the Table 2 indicate that on overall level our algorithm performs relatively well. Still, further improvement through usage of more advanced text mining methods is needed to overcome the previously described difficulties.

VIII. Trend Categories

Applying the previously described algorithms revealed that there are differences between the topics that appear as trends. In order to analyze and understand these differences we propose the following three categories of trending topics: (1) 'disruptive events', (2) 'popular topics', and (3) 'daily routines'. Disruptive events correspond to the events that occur at a particular point in time and cause reaction of Facebook users on a global level, such as the earthquake in Japan, Wimbledon finals, etc. Popular topics might be related to some past event, celebrities or products/brands that remain popular over a longer period of time, such as Coca Cola, Michael Jackson, etc. Finally, daily routines correspond to some common phrases such as "good night", "birthday wishes", etc. In the following chapter we present the characteristics and differences between these categories.

IX. Trend Characteristics

For our further analysis we chose representatives for each of the previously described categories: (1) the death of Amy Winehouse and the Norway attacks, as examples of disruptive events, (2) Harry Potter, as a representative for the popular topic, and (3) "Happy Birthday" as a typical daily routine on Facebook. We tried to identify differences in terms of distribution through the shape and volume of the shared information. In addition, we were interested in measuring the speed and scale of information distribution on Facebook as an indicator of the possibility to use Facebook as a news media.

TABLE II. RESULTS OF THE PRELIMINARY EVALUATION

Topics	Measures		
	Precision	***Recall***	***F-measure***
Amy Winehouse	0.9475	0.7748	0.8510
Norway (max posts)	0.0736	0.6124	0.1303
Norway (best fit)	1.0000	0.0621	0.1164
Harry Potter	0.8344	0.8589	0.8115
Average Values	0.7139	0.5771	0.4773

A. Distribution

The distribution in terms of the volume of the posts shared on Facebook regarding a certain topic is a clear indicator of a level of interest of users for the related topic. In addition, the shape of the distribution is an indicator of a topic belonging to the category of 'daily routines' that is always present in the conversation at some relatively equalized level, or if it relates to an event occurring at a particular point in time.

Understanding the differences between distributions that relate to the 'daily routines' and 'popular topics' on one side, and the distributions related to 'disruptive events' on the other, gives us the possibility to train the systems for automatic trend detection in order to distinguish between these different types of trends. Fig. 3 illustrates the time series for the selected four topics in the observed time interval.

It can be seen that the topic of Amy Winehouse has a burst of posts immediately after the time of her death. The same effect, although not with such intensity, can also be seen for the Norway attack. An interesting observation is that at the day of the event, the number of posts related to Norway is significantly lower compared to the next day. Furthermore, the big peak on the Norway graph corresponds with the initial peak for Amy Winehouse. Analyzing the post clusters obtained through our algorithm showed that these two topics indeed appeared in same posts.

Regarding the "Harry Potter" and "Happy Birthday" topics, the curve shows almost regular peaks throughout the interval as could be expected, however, these variations in the volume are not as big as those in the previous case. Comparison to the cumulative distribution of posts over time of day shows that the peaks on the 'popular topics' correspond with the peaks on the daily post distribution graph, while peaks for 'daily routine' are the opposite. We explain this as a result of people wanting to congratulate as early as possible. On a more general level, daily routines are usually related to a certain period of time in a day, for example, "good night" appears as a trending topic only in the evenings. Furthermore, these two topics are present and trending during the whole time interval, indicating a popular topic, but not something new.

Descriptive statistics for the selected topics are presented in Table 3. The obtained values indicate big differences in the distributions. Differences in standard deviation can be used as an indication of the 'disruptive event', while sum and mean do not provide such a clear distinction. In addition, kurtosis corresponds to the variations between peaks and has higher value for the 'disruptive events' compared to the popular and common topics. Finally, skewness illustrates that the majority of the posts have been grouped at one segment of the time interval, again as an indication of a significant peak in the distribution.

B. Speed and Flow of Information Diffusion

Time of occurrence of the first post regarding a certain topic is interesting from the perspective of evaluating the possibility to use Facebook as a news media. In addition, the speed of the information diffusion can be measured by the time interval between the event and the time the topic became a trending topic. Fig. 4 illustrates the distribution of posts for both big events during the first two hours. It can be seen that the first post for Amy Winehouse occurred at 16:16 GMT, approximately one hour after the announcement. In addition, the topic became a trend with 78 occurrences in the second interval of data collection after the first post, i.e. at 16:30 GMT.

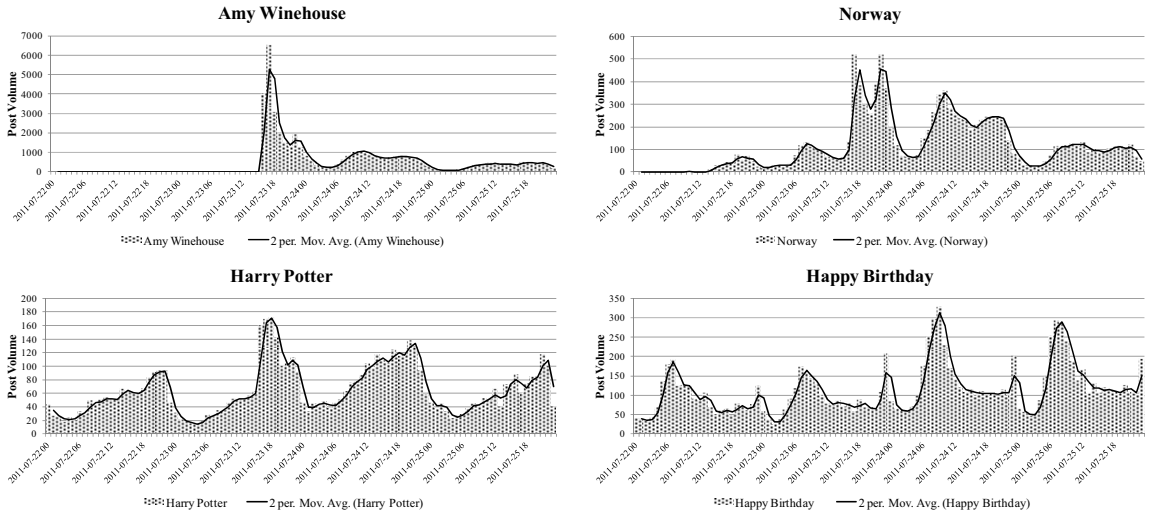

Figure 3. Time series of posts related to disruptive events (Amy Winehouse, Norway), popular topics (Harry Potter) and daily routines (Happy Birthday).

TABLE III. DESCRIPTIOVE STATISTICS FOR TRENDS

Measures	Trending topics			
	Amy Winehouse	***Norway***	***Harry Potter***	***Happy Birthday***
Mean	464.66	114.78	64.44	113.66
Standard Deviation	888.69	115.31	36.27	63.86
Kurtosis	25.02	2.16	0.46	1.70
Skewness	4.44	1.50	0.96	1.35
Min/Max	0 / 6516	0 / 520	15 / 172	25 / 328
Sum	44607	11019	6186	10911

When it comes to the events in Norway, the situation is different. Regarding the time difference between the event and the first post it can be seen that the first post was shared a bit faster, approximately 50 minutes after the event. However, the spread of this topic significantly differs from the previous one. Posts are shared on irregular intervals and the number of posts is relatively low with an average below 1 post per 10 minutes. As such this topic positioned itself very low on the weighted topic group list.

Based on these two examples we might find similarity in terms of the fact that both topics appeared on Facebook relatively fast, however, the example with Norway clearly illustrates that 'disruptive events' not necessarily become trending topics, even in case of such events that are attracting a lot of attention on other, more traditional sources of media.

X. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a system for trend detection over the Facebook public focusing on two problems: (1) data collection and (2) trend detection. We have shown that the proposed data collection algorithm provides the possibility to collect large datasets despite the existing Facebook privacy policies. We did not evaluate whether the amount of collected data is a representative sample for the content shared on Facebook. Instead, our goal was to confirm that Facebook can be used as an additional valuable source of information regarding the topics that attract the attention of a large fraction of people.

In addition, we have performed analysis of the obtained dataset to determine the structure of the posts in terms of the (1) length, (2) language, (3) post type and (4) posting distribution over the time of day. Based on the obtained results from the dataset analysis we proposed an algorithm for trend detection over Facebook public posts which (1) takes in consideration only 'status' posts present within 84% of the dataset, (2) does not consider multilingualism apart from English stop words filtering due to the fact that English language is used in 78% of the posts, and (3) limits the minimum length of the n-gram based terms to 2 in order to avoid noise and improve performance.

Based on the preliminary evaluation and the previously presented discussion over the results obtained through the proposed system we can conclude that our simple approach performs well only on certain topic groups. In particular, the applied clustering algorithm is very greedy, resulting in problems with topics such as the Norway incident, where there is a little overlap between separate terms belonging to the same topic group and an existing overlap with the more dominant topic group of Amy Winehouse. Therefore, this algorithm needs to be further improved to achieve optimal results. We propose the combination of named entity detection over the unigrams and latent semantic indexing (LSI) as possible approach towards overcoming the observed problems.

Through analysis of the results of the proposed trend detection algorithm, we have identified three different categories of trending topics: (1) 'disruptive events', (2) 'popular topics' and (3) 'daily routines'. Based on the comparison of their characteristics we have shown that statistical measures, such as standard deviation, kurtosis and skewness can be used for distinction of 'disruptive events' among the trending topics and information travels and spreads relatively fast. However 'disruptive events' do not necessarily become trending topics even in cases when they attract a lot of attention on more traditional sources of media.

We plan to continue our research in the direction of improving the proposed algorithm, in particular in the segment of clustering by applying more advanced methods such as latent semantic analysis (LSA), LDA models or network community detection. In addition, we would like to perform the analysis over a larger period of time to be able to confirm our results.

REFERENCES

[1] J. Allan, "Topic Detection and Tracking," in Event-based Information Organization, Kluwer Academic Publishers, 2002.

[2] S. Asur, and B. A. Huberman, "Predicting the Future with Social Media," Proc. Int. Conf. on Web Intelligence and Intelligent Agent Technology (IEEE/WIC/ACM 10), IEEE Press, Sep. 2010, pp. 492–499.

[3] N. Bansal and N. Koudas, "Blogscope: A System for Online Analysis of High Volume Text Streams," Proc. 33rd Int. Conf. on Very Large Data Bases (VLDB 07), pp. 1410–1413.

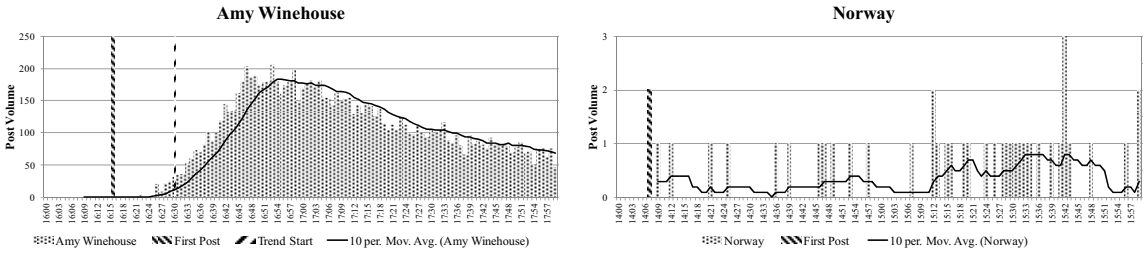

Figure 4. The speed and scale of information diffusion for disruptive events

[4] H. Becker, M. Naaman, and L. Gravano, "Beyond Trending Topics: Real-world Event Identification on Twitter," Proc. 5th Int. AAAI Conf. on Weblogs and Social Media (ICWSM 11), AAAI Press, Jul. 2011.

[5] A. Bermingham, and A. F. Smeaton, "Classifying Sentiment in Microblogs: Is brevity an Advantage?" Proc. 19th ACM Int. Conf. on Information and Knowledge Management (CIKM 10), ACM, Oct. 2010, pp. 1833-1836, doi:10.1145/1871437.1871741.

[6] D. M. Boyd, and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," Comput.-Mediat. Comm., Vol. 13 (1), Oct. 2007, pp. 210–230.

[7] M. Cheong, and V. Lee, "Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base," Proc. 2nd ACM Workshop on Social Web Search and Mining (SWSM 09), ACM, Nov. 2009, pp. 1-8, doi:10.1145/1651437.1651439

[8] N. A. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry," IEEE Symposium on Visual Analytics Science Technology (IEEE VAST 09), Oct. 2009.

[9] N. A. Diakopoulos, and D. A. Shamma, "Characterizing Debate Performance via Aggregated Twitter Sentiment," Proc. 28th Int. Conf. on Human Factors in Computing Systems (CHI 10), ACM, Apr. 2010, pp. 1195-1198, doi: 10.1145/1753326.1753504.

[10] S. Goorha, and L. Ungar, "Discovery of Significant Emerging Trends," Proc. 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 10), ACM, Jul. 2010, doi:10.1145/1835804.1835815.

[11] B. J. Jansen, and M. Zhang, "Twitter Power: Tweets as Electronic Word of Mouth," American Society for Information Science, vol. 60 (11), Nov. 2009, pp. 2169-2188.

[12] K. S. Jones, "A Statistical Interpretation of Term Specificity and its Applications in Retrieval," Documentation, vol. 28 (1), 1972, pp. 11–21.

[13] J. Karlgren, "Information Retrieval Systems: Statistics and Linguistics," in Legal Management of Information Systems: Incorporating Law in E-solutions, 2nd ed., C. M. Sjoberg, Eds. Studentlitteratur AB, 2010, pp. 295-336.

[14] A. Kontostathis, L. Galitsky, W. M. Pottenger, S. Roy, and D. J. Phelps, "A Survey of Emerging Trend Detection in Textual Data Mining," in Survey of Text Mining: Cluster Classification and Retrieval, W. M. Berry, Eds. Springer-Verlag, 2003, pp. 185-224.

[15] H. Kwak, C. Lee, H. Park, and S. M. Kwak, "What is Twitter, a Social Network or a News Media?" Proc. 19th Int. Conf. on World Wide Web (WWW 10), ACM, Apr. 2010, pp. 591-600, doi:10.1145/1772690.1772751.

[16] M. Mathioudakis, and N. Koudas, "TwitterMonitor: Trend Detection over the Twitter Stream," Proc. ACM Int. Conf. on ACM Special Interest Group on Management of Data (SIGMOD 10), ACM, Jun. 2010, pp. 1155-1158, doi:10.1145/1807167.1807306.

[17] M. Naaman, H. Becker, and L. Gravano, "Hip and Trendy: Characterizing Emerging Trends on Twitter." American Society for Information Science and Technology, vol. 62 (5), May 2011, pp. 902-918.

[18] E. Noelle-Neumann, The Spiral of Silence: Public Opinion-Our Social Skin, Chicago: Univ. Chicago Press, 1980.

[19] S. Petrovic, M. Osborne, and V. Lavrenko, "Streaming First Story Detection with Application to Twitter," North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 10), Association for Computational Linguistics, Stroudsburg, PA, USA, Apr. 2010, pp. 181-189.

[20] O. Phelan, K. McCarthy, and B. Smyth, "Using Twitter to Recommend Real-Time Topical News," Proc. 3rd ACM Conf. on Recommender Systems (RECSYS 09), ACM, Oct. 2009, pp. 385-388. doi:10.1145/1639714.1639794.

[21] V. Raghavan, P. Bollmann, and G. S. Jung, "A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance," ACM Trans. Inf. Syst., vol. 7, pp. 205-229, 1989.

[22] D. Ramage, S. Dumais, and D. Liebling, "Characterizing Microblogs with Topic Models," Proc. 4th Int. AAAI Conf. on Weblogs and Social Media (ICWSM 10), AAAI Press, May 2010, pp. 130-137.

[23] D. Richter, K. Riemer, and J. vom Brocke, "Internet Social Networking: Research State of the Art and Implications for Enterprise 2.0 (State of the Art)," Wirtschaftsinformatik, vol. 53 (2), Apr. 2011, pp. 89-103.

[24] M. Ringel Morris, J. Teevan, and K. Panovich, "What do People Ask their Social Networks, and Why?: A Survey Study of Status Message Q&A Behavior," Proc. 28th Int. Conf. on Human Factors in Computing Systems (CHI 10), ACM, Apr. 2011, pp. 1739-1748, doi:10.1145/1753326.1753587.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," Proc. 19th Int. Conf. on World Wide Web (WWW 10), ACM, Apr. 010, pp. 851–860, doi:10.1145/1772690.1772777.

[26] G. Salton, and M. J. Mcgill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.

[27] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Experiments in Microblog Summarization," Proc. 2nd IEEE Int. Conf. on Social Computing (SOCIALCOM 10), IEEE Press, Aug. 2010, pp. 49-56, doi:10.1109/SocialCom.2010.17

[28] W. Simm, M.-A. Ferrario, S. Piao, J. Whittle, and P. Rayson, "Classification of Short Text Comments by Sentiment and Actionability for VoiceYourView," Proc. 2nd IEEE Int. Conf. on Social Computing (SOCIALCOM 10), IEEE Press, Aug. 2010, pp. 552–557, doi:10.1109/SocialCom.2010.87.

[29] J. Teevan, D. Ramage, and M. R. Ringel, "#TwitterSearch: A Comparison of Microblog Search and Web Search," Proc. 4th ACM Int. Conf. on Web Search and Data Mining (WSDM 11), ACM, Feb. 2011, pp. 35-44, doi: 10.1145/1935826.1935842.

[30] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," Proc. 4th Int. AAAI Conf. on Weblogs and Social Media (ICWSM 10), May 2010.

[31] H. R. Varian, and H. Choi, "Predicting the Present with Google Trends," Google Research Blog, Apr. 2009, http://googleresearch.blogspot.com/2009/04/predicting-present-with-google-trends.html

[32] C. Zeidler, "Avoiding Traffic Jams with Twitter & iPhone," SAP.info, Feb. 2010, http://en.sap.info/avoiding-traffic-jams-with-twitter-iphone/23754