

# What's Happening: A Survey of Tweets Event Detection

Amina Madani  
Blida 1 University  
LRDSI Laboratory  
Blida, Algeria  
a\_madani@esi.dz

Omar Boussaid  
Lumière Lyon 2 University  
ERIC Laboratory  
Lyon, France  
Omar.Boussaid@univ-lyon2.fr

Djamel Eddine Zegour  
High School of Computer Science  
LCSI Laboratory  
Algiers, Algeria  
d\_zegour@esi.dz

**Abstract**—Twitter is now one of the main means for spread of ideas and information throughout the Web. Tweets discuss different trends, ideas, events, and so on. This gave rise to an increasing interest in analyzing tweets by the data mining community. Twitter is, in nature, a good resource for detecting events in real-time. In this survey paper, we are going to present four challenges of tweets event detection: health epidemics identification, natural events detection, trending topics detection, and sentiment analysis. These challenges are based mainly on clustering and classification. We review these approaches by providing a description of each one.

**Keywords**—tweets mining; tweets event detection; health epidemics identification; natural events detection; trending topics detection; sentiment analysis

## I. INTRODUCTION

In recent years, hundreds of millions of users participate in online social networks and forums, subscribe to microblogging services or maintain web diaries (blogs).

Twitter, in particular, is currently the major micro-blogging service. In January 2014, the total number of active registered users on Twitter was 645,750,000 [1]. In this system, participants post short status messages that are often available publicly, or semi-publicly (e.g., restricted to the user's designated contacts). In January 2014, more than 58 millions of Twitter messages were sent every day through the world, with 9,100 the number of tweets that happen every second [1].

Recently, the exponential growth of Twitter messages has started to draw the attention of researchers from various disciplines. Numerous works in the data mining field have examined Twitter. How to automatically understand, extract and summarize useful Twitter content to detect events has therefore become an important and emergent research topic. Tweets event correspond to a lot of content generated on Twitter, which is opinions, reactions and information from users. Currently, the most promising events that exist on Twitter are: natural disasters such as earthquakes, health epidemics like influenza, trends such as world-cup and opinion about products, services or events like political election results.

The focus of this survey paper is to discuss research works done for four challenges of tweets event detection: health epidemics identification, natural events detection, trending topics detection, and sentiment analysis. Tweets event detection helps understanding what the users are really discussing about in Twitter.

This paper is organized as follows: In the next section, we explain Twitter service and its characteristics. In Section 3, we

review popular tweets event detection approaches by providing a description of each one. Section 4 discusses and highlights works studied above. Finally, we conclude the paper and further work is outlined.

## II. TWITTER, A POPULAR SAS

Twitter is a popular microblogging service, one of the main means for spread of ideas and information throughout the web. Twitter enables its users to send and read short text-based messages of up to 140 characters about “what’s happening” (either one or two sentences), known as “tweets”. Users can tweet via the Twitter website, compatible external applications (such as for smartphones), or by Short Message Service (SMS) available in certain countries.

One of the main characteristics of Twitter is that its core functions represent a *Social Awareness Stream* model. The *Social Awareness Streams* (SAS) are typified by three factors distinguishing them from other communications [2]:

- The public (or personal-public) nature of the communication and conversation;
- The brevity of posted content; and,
- A highly connected social space, where most of the information consumption is enabled and driven by articulated online contact networks.

Twitter has users of the order of hundreds of millions (Figure 1). A huge amount of content is generated every second, minute and hour of the day. Every tweet is associated with an explicit timestamp that declares the exact time it was generated. Important characteristic of Twitter is its real-time nature [3]. Tweets are data stream arriving in real-time. Data stream are data that arrive at high speed and its nature or distribution changes over time.

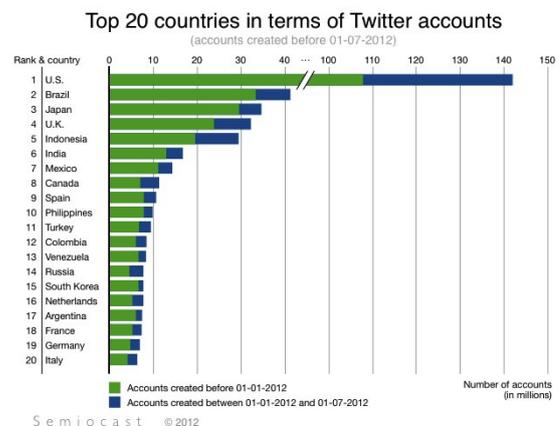


Figure 1. Top 20 countries in terms of Twitter accounts [4].

Also, another property of tweets is the rich set of fields associated with the content, which is usually presented in the form of semi-structured documents. Figure 2 presents an example of a tweet. While the contents are ostensibly 140 unstructured characters, the anatomy of a tweet reveals lots of metadata (e.g., location, location) and even the contents contain some structural information (e.g., RT indicating re-tweet or #hashtags serving as topical metadata). Moreover, every user has a well-defined profile with personal information (name, location, biographical sketch).



Figure 2. An example of a tweet.

All these characteristics gave rise to an increasing interest in analyzing tweets by the data mining community. Exploiting these characteristics can be helpful for improving tweets event detection.

Tweets event correspond to a lot of content generated on Twitter, which is opinions, reactions and information from users.

### III. TWEETS EVENT DETECTION: STATE OF THE ART

In the knowledge discovery context, there are two fundamental data mining tasks that can be considered in conjunction with Twitter data [5]:

- Graph mining based on analysis of the links amongst messages.
- Text mining based on analysis of the messages content.

In our study, we are interested in the tweets content using text mining. Tweets content have become an important channel for reporting real-world events. We describe an event by four main dimensions:

- *Event type*: what is happening?
- *Time*: when an event is happening?
- *Location*: where an event is happening?
- *Entities*: who is involved in an even?

Health epidemics identification (e.g. influenza), natural events detection (e.g. earthquakes), trending topics detection, and sentiment analysis (e.g. political events) are four considerable challenges for tweets event detection presented in Figure 3. These challenges are based mainly on clustering and classification.

In this part, we are going to explain more specifically, these challenges and present several research efforts that have focused on them for identifying events on Twitter.

### A. Health epidemics identification

In recent years, a new research area has been developed, namely “Infodemiology”. It can be defined as the “science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy” [6].

Twitter presents a promising new data source for Internet-based surveillance because of message volume, frequency, and public availability. Recent studies have begun to use Twitter data to understand their applicability in the context of Health epidemics identification. Twitter can be a low cost alternative source for tracking health epidemics. Health epidemics identification is based on classification in order to predict some illnesses while analyzing textual content of tweets.

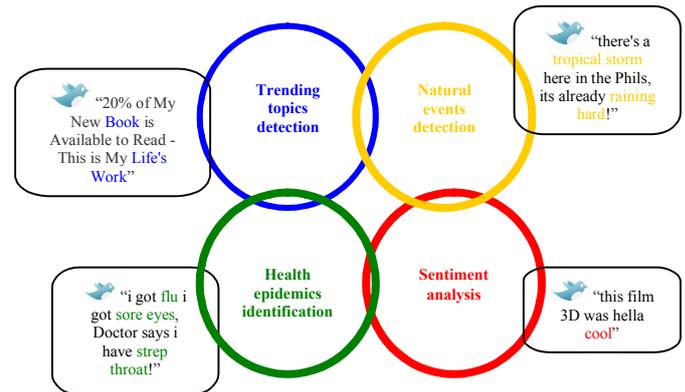


Figure 3. Challenges of tweets event detection.

Previous work in this area has focused specifically on influenza. Culotta [7] considers tweets as a valuable resource for tracking influenza for several reasons:

- The high message posting frequency enables up-to-the-minute analysis of an outbreak.
- Twitter messages are longer, more descriptive, and publicly available.
- Twitter profiles often contain semi-structured meta-data (city, state, gender, age), enabling a detailed demographic analysis.
- Despite the fact that Twitter appears targeted to a young demographic, it in fact has quite a diverse set of users.

The work of Culotta [7] explores the possibility of detecting influenza outbreaks by analyzing Twitter data. The author uses a bag-of-words classifier in order to predict *influenza-like illnesses* (ILI) rates in a population, based on the frequency of messages containing certain keywords. He compares rates with the *U.S. Centers for Disease Control and Prevention* (CDC) statistics.

A new method called *Ailment Topic Aspect Model* (ATAM) for extracting general public health information from millions of health related tweets is introduced by Paul and Dredze [8]. The approach discovers many different ailments (diseases), such as flu, allergies, or cancer and learns symptom and treatment associations. This model discovers a larger number of more coherent ailments than *Latent Dirichlet Allocation* (LDA) [9]. It produces more detailed ailment information

(symptoms/treatments) and tracks disease rates consistent with published government statistics (influenza surveillance) despite the lack of supervised influenza training data.

### B. Natural events detection (Disasters)

An important characteristic of Twitter service is their real-time nature. Users write tweets several times in a single day. Sakaki et al. [3] consider each Twitter user as a sensor and each tweet as sensory information. A sensor detects a target event and makes a report probabilistically. Each tweet is associated with a time and location, which is a set of latitude and longitude. These virtual sensors, which they call social sensors, are of a huge variety and have various characteristics. Detecting a natural event in real time is reduced to detecting an object and its location by regarding a tweet that has both spatial and temporal regions.

Some researchers have shown how these social sensors can be useful to describe the current situation during natural disasters. The objective is to mine tweets to detect natural disaster in real-time events such as earthquakes, floods or volcanic eruptions.

The automatic detection of natural events from tweets consists in developing *Business Intelligence* (BI) tools to detect information before they appear in the agencies.

A method for automatic detection of disasters in real-time is proposed by Rosoor et al. [10]. They develop a tool for journalists / librarians to detect the information before they appear in the agencies. The automatic detection method uses heterogeneous resources from the Web (blogs, tweets, RSS, etc.). This tool is based on a "salton" representation [11] of a corpus classified into different topics (flood, earthquake, and so forth). Each category is represented as a vector of words. The new text to be classified is compared to these vectors in order to identify the relevant category.

Sakaki et al. [3] investigate the real-time interaction of events, such as earthquakes in Twitter and propose an algorithm to monitor tweets and to detect a target event. To detect a target event, they devise a classifier of tweets using a support vector machine [12] based on features, such as the keywords in a tweet, the number of words, and their context. Subsequently, they produce a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location. They can detect an earthquake with high probability (96% of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more are detected) merely by monitoring tweets. The system detects earthquakes promptly and sends e-mails to registered users. Notification is delivered much faster than the announcements that are broadcast by the JMA and possibly before an earthquake actually arrives at a certain location.

The work of Cheong and Cheong [13] performs analysis on tweets during Australian floods of 2011 to identify active players and their effectiveness in disseminating critical information. They identified the most prominent users among Australian floods to be: local authorities (Queensland Police Services), political personalities (Premier, Prime Minister, Opposition Leader and Member of Parliament), social media volunteers, traditional media reporters, and people from not-for-profit, humanitarian, and community associations.

### C. Trending topics detection

Twitter messages are being posted with vast amount of new information and changes continuously. They are a live stream that contains a great wealth of information where topics of discussion shift dynamically with time. However, it is not practical for us to browse tweets manually all the time for searching about the latest most discussed issues and thus revealing the emerging topics of our interest.

Analyze tweets to detect trending topics (trends) in real time is a big challenge. Trending topics detection has primarily involved analyzing the content of tweets. We define a trending topic as an emerging keyword which links to a very recent event. This keyword is experiencing an increase in usage in Twitter messages. Trending topics are typically driven by emerging events, breaking news and general topics that attract the attention of a large fraction of Twitter users [14]. Unspecified trending topics of interest are typically driven by topics that attract the attention of a large number of users.

Twitter allows users to observe the top ten popular terms or topics of discussion at any given moment. But, we must browse related tweets manually all the time for viewing more detail about these trending topics. It is important to automatically analyse, understand, extract and summarize useful tweets content.

Twitter is a new form of newspapers. Trending topics detection in real-time is thus of high value to news reporters, analysts, and e-marketing specialists.

We present several approaches efforts that have focused on trending topics detection for identifying events on Twitter.

Cheong and Lee [15] analyze tweets to research the anatomy of trending topics. They split them into 3 categories: long-term, medium-term and short-term topics. Long-term topics occur infrequently, but over a long amount of time in the public time-line, while medium-term topics occur more frequently. However, the medium-term topics are limited to a time range of a few days. Short-term topics are heavily discussed topics, and often refer to current events. Also, they categorize the users into 3 major groups: "Personal", "Aggregator" and "Marketing". The results show that mostly users who talk about their personal life contribute to emerging trending topics.

TwitterMonitor, a system that performs trend detection is proposed by Mathioudakis and Koudas [14]. The system identifies emerging topics (trends) on Twitter in real-time. A trend is identified as a set of bursty keywords that occur frequently together in tweets. TwitterMonitor provides meaningful analytics that synthesize an accurate description of each topic. It extracts additional information from the tweets that belong to the trend, aiming to discover interesting aspects of it. Users interact with the system by ordering the identified trends using different criteria and submitting their own description for each trend.

The main objective in the work of Naaman et al. [2] is to identify different types of user activity, specifically focusing on message content and its relationship to patterns of use. To characterize the type of messages posted on Twitter they use a grounded approach to thematize and code a sample of 200 tweets. First, the three authors independently assigned categories to the downloaded messages. They then proceeded

to analyze the affinity of the emerging themes to create an initial set of coding categories. Next, they downloaded a second set of 200 posts, categorized them, then reflected on and adapted the initial categories based on the additional input.

Benhardus [16] outlines methodologies of detecting and identifying trending topics from streaming data. *Term frequency-inverse document frequency* (TF-IDF) analysis and relative normalized term frequency analysis are performed on the tweets to identify the trending topics. Relative normalized term frequency analysis identifies unigrams, bigrams, and trigrams as trending topics, while term frequency-inverse document frequency analysis identifies unigrams as trending topics.

Cataldi et al.'s approach [17] proposes a novel approach to detect in real-time emerging topics on Twitter. They extract the contents (set of terms) of the tweets and model the term life cycle according to a novel aging theory intended to mine terms that frequently occur in the specified time interval and they are relatively rare in the past. Moreover, considering that the importance of content also depends on its source, they studied the social relationships in the user network in order to determine the authority of the users. Finally, they formalized a keyword-based topic graph which connects the emerging terms with their co-occurrent ones, allowing the detection of emerging topics under user-specified time constraints.

Budak et al. [18] introduce new methods for identification of important topics that utilize the network topology. They propose two novel trend definitions called coordinated and uncoordinated trends that detect topics that are popular among highly clustered and distributed users, respectively. A novel information diffusion model called *independent trend formation model* (ITFM) has also been introduced to distinguish viral diffusion of information from diffusion through external entities, such as news media, and to capture the diffusion of an arbitrary number of topics in a social network.

#### D. Sentiment analysis

With the rapid growth of Twitter messages, users are enabled to express their opinions in terms of views, sentiments, evaluations, attitudes, appraisals and emotions towards entities, events and their properties on almost anything. Opinions can be expressed by persons or by organizations.

In recent years, sentiment analysis (also known as opinion mining, sentiment detection or sentiment classification) has emerged as a new method to study user's opinions (or feelings) in regard to some topic. Twitter sentiment analysis focuses on analyzing tweets but it is difficult to extract opinions, read them, summarize them, and organize them into usable forms. Thus, automated Twitter sentiment analysis is needed. Sentiment analysis can be cast as a classification problem where the task is to classify messages into two categories depending on whether they convey positive or negative feelings [5].

A large number of tweets include opinions about products and services. It is interesting for companies interested in knowing how users feel about their products. Another stream of research focuses on the analysis of tweets as *electronic Word Of Mouth* (eWOM) in the area of product marketing.

Word of mouth is the passing of information from person to person by oral communication.

Given its distinct communication characteristics, Twitter deserves serious attention as a form of online WOM (oWOM) or electronic WOM (eWOM). Tweets are underutilized as a source for evaluating customer sentiment. Sentiment analysis consists generally to classify an opinionated text as either positive or negative, according to the overall sentiment expressed by the author within it. A Tweet can contain polarity sentiments. For example, the word "kill" has a negative polarity, and the word "love" has a positive one.

Jansen et al. [19] consider a tweet as eWOM. They have found that 19% of a random sample of tweets contained mentions of a brand or product and that an automated classification was able to extract statistically significant differences of customer sentiment (i.e., the attitude of a writer towards a brand). Using sentiment detection, market researchers have a valuable tool to monitor how a product is accepted.

Twitter messages can also be used for political communication. It is important for political institutions to get a feel of prevalent sentiment and determine whether the public opinions are positive or negative.

The results of Tumasjan et al. [20] demonstrate that Twitter can also be considered as a valid real-time indicator of political sentiment. First, they examine Twitter messages. Their results indicate that people are finding interesting political information on Twitter which they share with their network of followers. They found that Twitter is indeed used as a platform for political deliberation. Second, they analyze the political sentiment of tweets and found that tweets reflect the current offline political sentiment in a meaningful way. To extract the sentiment of these tweets automatically, they use *Linguistic Inquiry and Word Count* (LIWC2007) [21], a text analysis software developed to assess emotional, cognitive, and structural components of text samples using a psychometrically validated internal dictionary. Third, after analyzing whether the activity on Twitter, they find that Twitter is as a predictor of the election result and even comes close to traditional election polls.

Twitter was also used to monitor the U.S. presidential debate in 2008 [22]. Tweets tended to favour Obama over McCain, and Obama really won the election afterwards. This shows that Twitter can also be used to predict political election results.

#### IV. COMPARISON AND DISCUSSION

In this section, we compare (Table 1) and discuss the approaches studied above. We begin by presenting a detailed description of the comparison criteria used:

a) *Event challenge*: is one of the fourth challenges of tweets event detection presented in section 3: Health epidemics identification, natural events detection, trending topics detection, or sentiment analysis.

b) *Event type*: describe the event type for an event challenge, specifically "what happens?". For example earthquakes, floods and volcanic eruptions are different event type for the event challenge natural events detection.

TABLE I. COMPARISON OF TWEETS EVENT DETECTION APPROACHES

Approach	Event challenge	Event type	Representation	Technique	Algorithm	Structural content	Textual content	Semantic
Culotta 2010	Health epidemics identification	Influenza	Bag-of-words	Linear regression	Supervised classifier	-	+	-
Paul and Dredze 2011		Diseases	SVM ( <i>Support Vector Machine</i> )	Machine learning approach	Supervised	-	+	-
Rosoor et al. 2010	Natural events detection	Disasters	Salton representation Bag-of-words	Statistics approach	Supervised	Time	+	-
Sakaki et al. 2010		Earthquakes	SVM ( <i>Support Vector Machine</i> )	Hybrid approach	Supervised	Time Retweets Location	+	+
Cheong and Cheong 2011		Floods	Undefined	Hybrid approach	Supervised	Username	+	-
Cataldi et al. 2010	Trending topics detection	Unspecified	Vectors of terms	Hybrid approach	Supervised Unsupervised	Time Username Retweets	+	+
Cheong and Lee 2009		Unspecified	Undefined	Statistics approach	Supervised	Username Time	+	-
Mathioudakis and Koudas 2010		Unspecified	Vectors of keyword	Statistics approach	Unsupervised	Time	+	-
Naaman et al. 2010		Unspecified	Undefined	Statistics approach	Supervised	Username	+	-
Benhardus 2010		Unspecified	Bag of words	Statistics approach	Unsupervised	Time	+	-
Budak et al. 2011		Unspecified	Undefined	Hybrid approach	Unsupervised	Username	+	+
Jansen et al. 2009		Sentiment analysis	Customer sentiment towards a brand	Undefined	Statistics approach	Unsupervised	-	+
Tumasjan et al. 2010	Political sentiment		Undefined	Linguistics approach	Unsupervised	Time Username Retweets	+	+

c) *Representation*: it is a transformation of a tweet in a format which is easier to understand. We present here the usual tweets representations.

d) *Technique*: four tweets mining techniques has been widely used for tweets:

- *Statistics Approach*: in these methods, the statistical information of the words can be used for tweets mining. Statistical methods include word frequency, TF\*IDF, word co-occurrence, etc.
- *Linguistics Approach*: it uses the linguistic features of the words including the lexical analysis, syntactic analysis, discourse analysis and so on.
- *Machine Learning Approach*: it employs the extracted keywords from training tweets to learn a model and applies the model to find keywords from new tweets. This approach includes Naïve Bayes, Support Vector Machine, etc.

- *Hybrid approach*: that combines the techniques mentioned above.

e) *Algorithm*: the major algorithms used for tweets event detection, are subdivided into supervised, unsupervised and hybrid algorithms. For each approach, we mention the direction of the algorithm used.

f) *Structural and Textual content*: textual content of a tweet is 140 unstructured characters, the structural content reveals lots of metadata (e.g., time, re-tweet). Moreover, every user has a well-defined profile with personal information that represents the structural content (name, location, biographical sketch). When dealing with tweets, according to the prior information available on the collection, it may be relevant to consider textual content alone or to consider both structure and textual content.

g) *Semantic*: Semantics is the study of meaning in language [23]. In tweets, semantic treatment (lexical not grammatical) has for goal to study the semantic relationships

between words. Hence, the problem is how to distinguish between many different senses that a word may have (polysemy) or between different words that can have the same significance (synonymy)... The objective is to exploit the semantic similarity of terms composing the textual content of tweets. The semantic treatment can use external semantic resources like ontologies, thesauruses and taxonomies. In our comparison, we study if the existing approaches take into account this aspect or not.

Since the size of tweets is small, most traditional data mining algorithms are not adapted for tweets. Twitter messages contain little informational value but the aggregation of millions of tweets can generate important knowledge.

Several supervised classification algorithms have been proposed for specified events, including for instance support vector machines [8], [3]. Most techniques are unsupervised and rely on clustering [14], [16].

The majority of approaches of tweets event detection work on tweets content. Structural information is less used. Twitter profiles often contain semi-structured metadata (city, state, gender, age), enabling a more detailed statistical analysis. For example, we can associate geographic information with each tweet in order to perform a more fine automatic detection of natural events. For health epidemics identification, we can include temporal and geospatial dynamics to track diseases across a population. Public health information can also be correlated with user location.

The majority of approaches ignore semantic of information inside tweets. Synonymy and polysemy can cause difficulties (different label that describe the same concept or a label denoting different concepts).

The major problem consists in determining, the information and the knowledge to extract from tweets to serve in different fields. Societies wish to detect some information from tweets before even their apparition in the press agencies of news. For example, automatic detection of natural events from tweets consists in developing *Business Intelligence* tools to detect information before it appears in the agencies.

Various studies have been focused for trending topics detection which is another considerable challenge for tweets event detection based mainly on clustering or classification. Analysing tweets content in real-time can help specialists (news reporters, analysts...) to understand what is happening, what emergent trends are exchanged between people.

Trend detection is also important for online marketing professionals and opinion tracking companies, as trends point to topics that capture the public's attention. The requirement for real-time trend detection is only natural for a live stream where topics of discussion shift dynamically with time. Twitter content could become key applications in the attention economy. Given the ease of monitoring any brand's sentiment, one can view tweets as a competitive intelligence source.

We think that looking at Twitter data in real-time can help people to understand what is happening, what people are thinking about brands, organisations and products, and more importantly, how they feel about them.

Other works demonstrate that Twitter can also be considered as a valid real-time indicator of political sentiment. We note that mining public opinion from freely tweets could be a faster and less expensive alternative to traditional polls.

## V. CONCLUSION AND FUTURE WORK

These last years have been marked by the emergence of microblogs. Their rates of activity reached some levels without precedent. Hundreds of millions of users are registered in these microblogs as Twitter. They exchange and tell their last thoughts, moods or activities by tweets in some words.

Tweets reveal useful event information for a variety of events. Approaches studied in this paper are interested in automatic extraction and detection of events from tweets. Although tweets are very exchanged on the web, we note that there are few works that are interested in tweets event detection, due in part to the fact that Twitter has only been in existence since 2006.

The major problem in this domain consists in determining, the event to extract from tweets to serve in different fields. For example societies wish to detect some information from tweets before even their apparition in the press agencies of news.

In terms of perspectives, we will try to take advantage of the methods studied in this paper, to propose a new approach for detecting tweets event in real-world. As a tweet is often associated with spatial and temporal information, we want to detect when and where an event happens.

## REFERENCES

- [1] Statistic brain, Twitter Statistics : Statistic Verification, <http://www.statisticbrain.com/twitter-statistics/>, 2014.
- [2] M. Naaman, J. Boase, and C. H. Lai, "Is it really about me?: message content in social awareness streams", In CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work, February 2010, pp. 189-192, Savannah, Georgia, USA.
- [3] T. Sakaki, M. Okazaki, and Y. Matsuon, "Earthquake shakes Twitter users: real-time event detection by social sensors", Proceedings of the 19<sup>th</sup> international conference on World wide web (WWW), April 2010, pp. 851-860, New York, NY, USA.
- [4] Semiocast study, [http://semiocast.com/fr/publications/2012\\_07\\_30\\_Twitter\\_reaches\\_half\\_a\\_billion\\_accounts\\_140m\\_in\\_the\\_US](http://semiocast.com/fr/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US), 2012.
- [5] A. Bifet and E. Frank, "Sentiment knowledge discovery in Twitter streaming data", In Proc 13th International Conference on Discovery Science, October 2010, pp. 1-15, Springer, Canberra, Australia.
- [6] G. Eysenbach, "Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search", Communication and Publication Behavior on the Internet . J Med Internet Res, Vol. 11(1):e11, 2009.
- [7] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages", In KDD Workshop on Social Media Analytics, July 2010, pp. 115-122.
- [8] J.M. Paul and M. Dredze, "A Model for Mining Public Health Topics from Twitter", Technical Report. Johns Hopkins University. 2011.
- [9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, Vol. 3, January 2003, pp. 993-1022.
- [10] B. Rosoor, L. Sebag, S. Bringay, P. Poncelet, and M. Roche, "When a tweet detect a natural event...", Actes du colloque Veille Stratégique Scientifique et Technologique VSST'2010, Septembre 2010, Toulouse (France).
- [11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Inf. Process. Manage., Vol. 24(5), January 1988, pp. 513-523.
- [12] T. Joachims, "Text categorization with support vector machines", In Proc. ECML'98, April 1998, pp. 137-142.
- [13] F. Cheong and C. Cheong, "Social media data mining: A social network analysis of tweets during the 2010-2011 australian floods", In PACIS, July 2011, pp. 1-16.

- [14] M. Mathioudakis and N. Koudas, "TWITTERMONITOR: trend detection over the twitter Stream", Proceedings of SIGMOD Conference, June 2010, pp.1155-1158.
- [15] M. Cheong and V. Lee, "Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base", In SWSM '09: Proceeding of the 2nd ACM workshop on Social web search and mining, November 2009, pp. 1-8, New York, NY, USA.
- [16] J. Benhardus, "Streaming Trend Detection in Twitter", 2010 UCCS Reu for Artificial Intelligence, Natural Language Processing and Information Retrieval, Final report 1, 2010.
- [17] M. Cataldi, L. D. Caro, and C. Schifanella, "Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation", MDMKDD'10, July 2010, pp. 4-13.
- [18] C. Budak, D. Agrawal, and A. El Abbadi, "Structural trend analysis for online social networks". Technical Report UCSB/CS-2011-04, UCSB, 2011.
- [19] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth", Journal of the American Society for Information Science and Technology, Vol. 60, November 2009, pp. 1-20.
- [20] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Weppe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment", In Proceedings of the 4th International Conference on Weblogs and Social Media, May 2010, pp. 178-471.
- [21] J. Pennebaker, C. Chung, and M. Ireland, "The development and psychometric properties of LIWC2007", Austin, TX, 2007.
- [22] N.A. Diakopoulos and D.A. Shamma, "Characterizing debate performance via aggregated twitter sentiment". CHI 2010 Proceedings of the SIGCHI Conference on Human Factors in Computing System, ACM, Atlanta Georgia, April 2010, pp. 1195-1198.
- [23] J. R. Hurford, "Semantics : a coursebook", Cambridge University Press, 1983.