Towards Intelligent Mining of Public Social Networks' Influence in Society

Joao Paulo Carvalho
Instituto Superior Técnico – Technical University of Lisbon
INESC-ID
R. Alves Redol 9, 1000-029 Lisboa
joao.carvalho@inesc-id.pt

Vasco Calais Pedro
INESC-ID
R. Alves Redol 9, 1000-029 Lisboa
vasco@bueda.com

Fernando Batista
ISCTE – Instituto Universitário de Lisboa
INESC-ID
R. Alves Redol 9, 1000-029 Lisboa
fernando.batista@inesc-id.pt

***Abstract* — This paper presents an overview of a proposed framework for the intelligent mining of the influence of public social networks in society. The framework consists of a data collection platform and a set of intelligent social-data processing modules. These modules identify relevant trending social topics and relevant actors within those topics, and trace back in time the evolution of those topics in the public social networks. The framework creates quantitative and qualitative indicators that can be used to analyze the role and the influence of the social networks and of their main actors in society.**

Keywords: *Public Social Networks; Twitter; Intelligent Data Mining; Natural Language Processing, Computational Intelligence.*

I. INTRODUCTION[1]

"Social networks and social networking are here to stay." This is not a controversial or novel statement: independently from how one feels towards adopting the use of social networks, no one can deny their importance in current modern world society. From event advertising or idea dissemination, to commenting and analysis, social networks have become the *de facto* means for individual opinion making and, consequently, one of the main shapers of an individual's perception of society and the world that surrounds him. The Arab Spring [24], the Indignant movement protest [1], or presidents tweeting and posting messages on Facebook instead of using official public addressing are just a few examples of how influential social networks have become. Nowadays important events are often commented in social networks even before they become "public news", and even news agencies and networks had to adapt and start using social networks as sources of information.

Despite their undeniable importance, too many questions concerning the effect of social networks in society are yet to be properly addressed. What makes events become important in social networks? Why and how they become important? How long does it take for an event to make an impact in social networks and society? Can social networks give more importance to an event than it really deserves, i.e., are social networks becoming a factor by themselves? What is the role of social networks' major actors (important journalists, bloggers, commentators, politicians, etc.) in the propagation of such events? Are such actors in the origin of the events or mere catalysts to the observations of minor role players?

This paper introduces an analysis framework that can help answering such questions. The framework enables the identification and traction of important events (topics) and key actors within those topics, as well as their origin and propagation timeline. The framework uses public social networks, such as Twitter, public blogs and webpages as the main source of data.

A succinct example of the framework *modus of operandi* is described as follows: Given an event or a set of keywords, search Twitter for relevant tweets and create statistics regarding the evolution of the number of relevant tweets through time (propagation velocity, number of posters, number of citations and retweets, etc.); Find who are the main actors (posters) for that event by checking who has most tweets, most followers, what messages have more retweets, etc.; Follow tweet links to blogs and web pages; Perform sentiment analysis; Follow main actors recent tweets/blogs back through time to find out when they became relevant concerning this event and if they are in their origin, or what/who was in the origin of the topic in the social network. These outputs (the actors' clout, the collected statistics concerning the propagation path and velocity, the sentiment analysis results, etc.) are used as data to create objective quantitative and/or qualitative measures that can be used to study and answer the above questions.

In order to be useful, the framework must be able to deal with the sheer amount of available data currently streaming in the Internet and in social networks. Presently, Facebook has more than 900 million users, Twitter has over 140 million active users, generating over 340 millions tweets daily [36], and according to BlogPulse [2], there are 152 million registered blogs. Such numbers are way beyond human processing capabilities and cannot be automatically processed without a massive (and too expensive) computer power, unless computational intelligent heuristics are employed to eliminate noise and retrieve only relevant information.

Another major obstacle addressed by the proposed framework lies in the fact that a large part of the data consists of short, unedited natural language sentences originating from different actors, which can have different ways of referring to the same event/topic, and that usually assume that the

[1] This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under projects PTDC/IVC-ESCT/4919/2012 and PEst-OE/EEI/LA0021/2011, and partially supported by DCTI - ISCTE-IUL.

information recipients are aware and familiar with the data context. Humans can naturally deal with these kinds of uncertainties and information gaps, but traditional automatic data mining methods cannot. Therefore Natural Language Processing (NLP) associated with Computational Intelligence techniques are proposed to address these problems.

This paper is organized as follows: Section II presents the literature review in what concerns social networks analysis, social data acquisition and storage infrastructures, and related processing methods. Section III presents a general overview of the proposed framework, describing each of its major components. Finally, Section IV presents our final remarks concerning the current development status.

II. Literature Review

A. Sociology and the Social Webs

The new information technologies and particularly the Internet have gained a remarkable preponderance in social sciences production in the latest years, as sociology seeks to discuss the new societal paradigm that is emerging as a result of the technological revolution, using concepts as network society, digital society, or information society [12][11][37].

The social impact of these changes has also been discussed in works that focus on: the individual and social appropriation of technologies and address usage patterns and the users' profile [8]; the digital divide and the importance of technological literacy [16]; the dynamic (sub)cultural factors that lead to the creation of an 'informatics tribe', the impact of p2p systems and the emergence of Web 2.0 and the use of technologies for the new generations [14] or the formation of the so-called virtual communities [32].

More recently, the huge impact of social networks, as Facebook or Twitter has also been under debate through several points of view [30]. Sociologists have been discussing the cultural impact of social media and particularly its results on cultural industries as music or books [28], the political impact of a new social arena in the web, a particularly important subject after the happenings in Iran in 2009 [10], or lately, the so-called Arab Spring; the presentation of self in the social networks and the impacts in sociability and friendship [5], and finally the economic dynamics linked to professional social networks [7].

B. Large Scale Social Data Acquisition and Storage

All the above studies suffer from the difficulty of dealing with the massive amount of available information in the web, and from the lack of specific computational tools to deal with it. This problem led to the development of a new research area that focuses on the use and development of computational methods to study social networks. This area has joined researchers from both sociology and computer areas. Researchers from the latter area mostly come from data mining, machine learning and natural language processing. As a result, several dedicated scientific journals and dedicated conferences have been created in the last couple of years, like for example Social Network Analysis and Mining, the International Journal of Social Network Mining, or the International Conference on Advances in Social Networks Analysis and Mining. Despite a large increase in the number of publications in recent years and of several isolated works in this area addressing/studying Twitter, Facebook, etc., several problems are unanswered since the main problem to address when trying to deal when attempting to perform research in this area is the infrastructure needed to manage the huge amount of data currently streaming in the web.

From the first successful web crawlers [6] to modern, socially aware, retrieval engines [21], data models and feature taxonomy have been established by industry practitioners with specific markets and use cases in mind. The gap between sociology and computer science is particularly evident in the strategies taken when developing social data models, since they fail to account for sociologically influenced feature taxonomy, such as in [25], where the authors fail to account for the features that might explain data sparsity, or [13], where authors take into account only simple topological features.

The direct influence of sociology within social network analysis has the potential to generate models that can account for the fundamental interactions within different relationships in a large-scale approach. New storage engines with real time map-reduce engines for analysis [29], coupled with a sound feature taxonomy would represent a tremendous step forward in modelling real time user interaction.

C. Fast and compact methods for diversity analysis

When dealing with large volumes of information and huge populations, the use of fast and compact approximated methods is mandatory to mine the data and compute statistics and frequencies of keywords, topics, actors, etc., that will be needed in the proposed framework. The most efficient method complying with those requirements is the Filtered Space Save Algorithm (FSS), that outperforms other methods [19][20][17]. Another requisite for these methods is the ability to deal with the temporal aspects that must addressed in the framework. The proposed version of FSS in a sliding window also makes it the best candidate for the problem [18].

D. Natural Language Processing Methods

The use of Natural Language Processing (NLP) techniques in this new field is obvious and has been recurrent. In the proposed framework the authors are particularly interested in the problems of Topic Detection and Tracking, Sentiment Analysis, Named Entities and Keyword Spotting. All can usually be treated as classification problems. However, the public social networks addressed in the proposed framework present particular characteristics that make standard NLP ineffective, especially in what concerns microblogs, like Twitter, where the noisy and informal nature of messages (users frequently abbreviate their posts to fit within 140 characters; the text is unedited and contains all sort of typographical errors; etc.), and the fact that one deals with a stream of data, present new challenges to NLP related tasks. Several works refer to the poor performance of standard NLP techniques trained with news data, on Twitter data, especially in what concerns named entities recognition [33]. However, Twitter includes specific phenomena, like retweets, @usernames, #hashtags, and urls, that provide important information for classification tasks, and also emoticons, which can be used together with text to improve sentiment

analysis [4][15]. Such features can be fused in the standard NLP techniques using fuzzy based systems.

III. Overview of the Proposed Framework

In order to accomplish the goal of developing a large-scale public social network data analysis framework that can help addressing the difficulties involved in obtaining treatable data for sociological social network studies, several stages are needed and must involve several distinct scientific areas. The framework is therefore divided into several blocks, of which some are mostly related with computer engineering and others involve a mix of social-minded Intelligent Data Mining, Computational Intelligence and Natural Language Processing. Figure 1 shows a diagram of the proposed framework. The following subsections provide an overview of the framework requisites, and also the major operations and goals of each of its blocks.

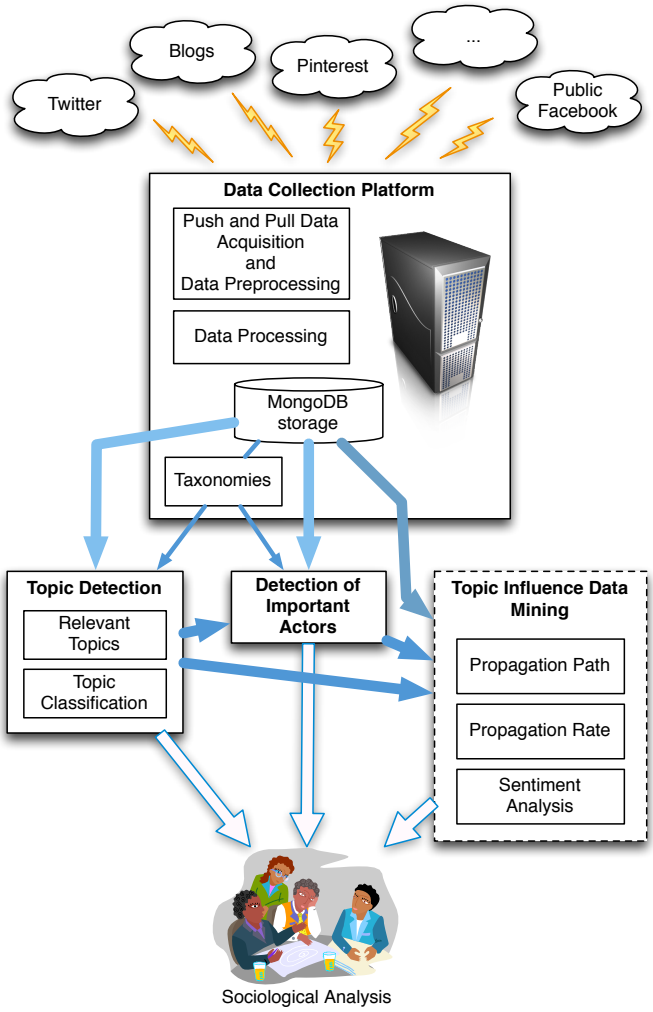

Figure 1 - Framework overview.

A. Available Data and Data representation

The proposed framework uses public social networks such as Twitter, public blogs and webpages as the main source of data. Facebook, despite its popularity, does not usually provide public data access; i.e., only friends' data can be searched for. This fact might hint that Facebook is not an attractive source of data to the problems the authors are trying to address. However, the proposed framework is also prepared for Facebook data mining since most entities that want to be heard usually let their accounts public. Another recent public social network that can possibly be mined for information is Pinterest.

The efficient mining of social streaming data requires the creation of a specific taxonomy for data representation [27]. The taxonomy essentially provides an answer for the following questions:

- What are the core characteristics of social data points?
- What kinds of characteristics exist and how can they be modeled?

In order to obtain the taxonomy specification, a sample set of data is created from the various sources and then an empirical model of the data suitable for a generic representation.

B. Public Social Networks Data Collection Platform

With 340 Million daily tweets, most of which are public, and over 150 Million registered blogs, collecting even a fraction of that data requires a combination of efficient algorithms and powerful hardware. Here one describes the procedures and the infrastructure that are necessary to collect, store and analyze the available social network data. This infrastructure creates the foundation for the remaining blocks.

1) Data Acquisition and Preprocessing

The system is composed of a main module that coordinates the acquisition and filtering process, and a set of sub-modules, each focused on a particular data source and tailored for source specific data acquisition. Each data source has specific methods of interface and a different degree of flexibility according to the type of public data. The main module is responsible for managing data sources querying, and ensures source appropriate query and uniqueness of data.

The proposed collection engine uses a combination of pull and push driven acquisition methods. In public social networks that use streaming data, such as twitter, there is the need to regulate both inflow of information as well as manage the keywords that will determine what gets streamed to the system. In other cases, such as Pinterest one needs to establish specific fuzzy dynamic query collections to acquire data and scrape the necessary information.

Data preprocessing requires deduplication, typographical and other word errors correction, normalization of dates, places and acronyms. All these tasks employ NLP techniques especially adapted to deal with the characteristics of the used corpora, and a semi-automatic process using a fuzzy-based word similarity function was devised in order to decrease the process human dependence [9]. The proposed system also normalizes locations to geo-location coordinates in order to allow for easy geographic correlation.

The platform uses Celery (a distributed task manager) for managing acquisition tasks, and rabbit MQ (an open-source message broker software) for message queues.

2) Data Processing and Storage

The proposed processing engine uses apache Mahout for batch based data processing over MapReduce [26], and S4 [29] to enable real-time stream processing. Another possibility being considered is to use Pulse, the engine developed by FeedZai [22]. Future testing will determine which system should be used on a final version. The Filtered Space Saving algorithm and its sliding window version [19][18] are used to obtain all approximated top-k lists.

The information is to be stored in a MongoDB [31][23] master/slave configuration to allow for easy read-only replication in case of scale out needs. MongoDB also has a native geo-coordinates store, allowing for easy storing and retrieval of geographic information.

C. Topics detection within the public social networks ecosystem

Topics play an essential role in the analysis of public social networks data. The proposed framework must: i) automatically detect relevant topics, corresponding to popular topics of conversation appearing in the data stream, also known as trends; and ii) classify trending topics into general categories, such as politics, economy, sports, etc., referred as topic classification.

Emerging or trending topics occur when multiple posts on a unique subject matter, distinct from the previous discourse, appear. The detection of relevant topics is of high value to analysts as they capture the public's attention and might point to fast-evolving news stories. Relevant topics are typically driven by emerging events, breaking news, and/or general topics that attract the attention of a large fraction of users (e.g. sport teams).

Topic detection and tracking is a well-known NLP processing subject. Microblogs present new interesting challenges for NLP applications, machine learning and computational intelligence techniques due to the length restrictions and unedited nature of Twitter messages, and also the size and the continuously evolving nature of the corpus. Therefore, despite an extended list of current available techniques, such as TF-IDF, semantic categorization, Random Walks or Fuzzy Clustering, specific features for social networks must be taken into account. In the particular case of Twitter, users have adopted conventions for marking relevant keywords in the text, known as #hashtags, which provide excellent cues for Keyword Spotting and are one of the most important elements for identifying relevant topics.

Topic relevance depends on a varied set of features [3] that is critical for correct determination of user's interest and expertise. The use of the feature taxonomy (III.A) allows proper modeling of topic relevance.

The first step in this module's operation consists of content filtering in order to accurately identify the elements of the corpus that are relevant [27]. This is necessary since much of the content on public social networks consists of personal information and spam. After the filtering stage, the module proceeds by detecting either topic #hashtags (keyword spotting), or keywords appearing at an unusually high rate. These are then used as a basis to identify the most relevant words associated with a given trend, which in turn are used to build topic Fuzzy Fingerprints [17]. Topic Fuzzy Fingerprints can identify messages or posts that are likely related to the topic despite not containing #hashtags. Additional information to characterize a topic is extracted from entries associated to that topic in the taxonomy.

Named entities found in the data also provide important features to improve a topic description after being marked as relevant. So the next step involves social name entity recognition.

It is natural to consider named entity detection over microblogs. The large quantity of information produced everyday provides excellent conditions to work with named entities since the same entity can be referred in multiple entries, each one containing updates and additional information. However, the performance of standard NLP tools, trained on news corpora, is poor on Twitter data [33]. Classifying named entities in tweets becomes especially difficult for two main reasons. First, tweets include a large number of entity types, like Products, Brands, and Companies, which are relatively infrequent, so even a large number of annotated tweets provide only a small number of training examples. Second, the context provided in small-sized text tweets is often insufficient for determining the entity type, without any other information.

In order to improve standard NLP methods, the proposed framework uses User Context Based Co-Reference Resolution: since each tweet refers to the user that produced it, the framework takes advantage of the user history and tweets from related users to provide additional background information about entities. Other specific phenomena, like URLs are also explored for getting related background information.

D. Detection of Important Actors within a Topic

One of the most important tasks of the proposed framework is to find who are the most important actors within a given topic streaming on a public social network.

Determining user relevance is vital to help determine trendsetters [35], as well as separate important messages from spam and garbage. The determination of a user's relevance must take into account global metrics that include not only the user's level of activity within the social network, e.g. the number of tweets posted every month, but also his impact in a given topic [38], that should take into account the number of reads in his messages, the number of followers, the number of retweets, etc. Since the whole universe of users must be considered, the use of fast and compact approximated methods is mandatory when making queries, retrieving information and computing the features that can be used to determine user relevance [20][19].

It is important to note that what consists an important feature in determining user relevance is not known as of yet, and despite several attempts at identification of relevant features in the literature [3], current models fail to account for all observed behavior of message propagation within social networks [39]. As such in this module one takes the best of breed models from current literature and augment them with features empirically extracted from sample data as well as

meta-features designed to accommodate the information generated in subsequent modules.

E. Topic Influence Data Mining

The goal of this block is to, given a topic and the key actors (as described in sub-sections *C.* and *D.*), find the origin of the topic and its evolution through time. One intends to find out the rate and path of propagation and also how the users sentiments evolved regarding the topic.

1) Topic Propagation Path

The system evaluates the path taken by topic related messages spreading across the network and the processes involved in the selection of the actors involved in that path. Analysis of the propagation path is one of the fundamental aspects for understanding relevance and influence, since it enables a topological perspective on the process, which could bring light to patterns of diffusion that rely on topological features.

The proposed block creates fuzzy queries that can automatically search tweet links to blogs and web pages, and follows main actors recent tweets/blogs back through time to find out when they became relevant concerning this event. If the main actors are not found to be in the topic origin, the queries follow the topic to find what/who was in the topic origin within the social network.

2) Topic Propagation Rate

The system evaluates which factors contribute to the increase or decrease of the amount of time it takes for a message to spread. Active recommendation systems might be in play, which might bias the rate of propagation. Not much is known about the underpinnings of social interaction that stimulate the rate of message propagation. Typically these phenomena are treated using disease infection models that do not take into account expensive features such as message context, topic predisposition and external factors such as environmental conditions outside the network.

The proposed block extracts statistics (propagation velocity, number of posters, number of citations and retweets, etc.) and creates graphs that show the propagation of the topic in the public social networks through time. One focus not only on the best indicators described in current literature, but take into account the analysis of the topic content and the topic predisposition of the actors to create refined models of propagation rates. As a result the system can answer to questions such as how fast was the propagation, or how was the growth and decay of the topic, amongst others.

3) Sentiment Analysis

This sub-block is responsible to find out how users feel concerning the topic, and how these sentiments evolve through time. The system performs sentiment analysis, which consists of assigning a given content to a predefined set of categories, depending on whether it contains positive or negative feelings. Sentiment analysis can be performed at different complexity levels, where the most basic one consists just on deciding if it contains a positive or a negative sentiment, and more complex levels may involve ranking the attitude into a set of more than two classes. The analysis can be taken further, in such a way that not only different complex attitude types can be determined, but also by finding the source and the target of such attitudes.

A classical approach for creating sentiment classification models is to collect training data, manually label the data, and then apply learning algorithms in a supervised or semi-supervised fashion. However, given the huge quantity of available data, such approach is inappropriate for the present system.

Another peculiarity lies in the fact that Twitter messages are required to be small, causing people to compress the information. This turns sentiment analysis on such data a particularly difficult problem. Moreover, words may be abbreviated, spaces may also be trimmed after punctuation marks, and texts may contain spelling and typographical errors. Also, messages sometimes simultaneously carry positive and negative feelings, and sarcasm and other playful uses of language often subvert the surface meaning.

In order to address the problem the proposed system considers the fact that besides text, tweets and blogs may include other important cues, like emoticons, that explicitly reveal a sentiment. By using emoticons, the author is explicitly annotating his text with an emotional state. That information is used for automatically labeling relevant training data [4][15]. A manually annotated golden set comprising a fixed data size, provides means for evaluating different modeling approaches.

F. Sociological Analysis

The proposed framework data output (the actors' clout, the collected statistics concerning the propagation path and velocity, the sentiment analysis results, etc.) should permit the capture of generic patterns of social behavior in public social networks. The data will be analyzed according to an analytical grid, combining three main dimensions or points of view: a thematic one, focusing on the different themes that are discussed and propagated; an agency one, focusing on actors and their specific roles; and finally a situational one, focusing on events and their dissemination on public social networks. This sociological analysis should provide the answer to questions such as what is being disseminated in public social networks, by whom, how is formation flowing, or what is the role of the public social network by itself.

IV. FINAL REMARKS

This paper describes a framework for intelligent mining the influence of public social networks in society. Composed by several blocks, it involves diverse research areas, including Computer Engineering, social-minded Intelligent Data Mining, Computational Intelligence and Natural Language Processing. The framework identifies relevant trending topics, identifies relevant actors within those topics, and traces back those topics finding their origin and their evolution through time. The analysis on a topic's propagation path and rate is a key point for understanding its relevance and influence, making it possible to answer questions, such as: how fast was the propagation, or how was the growth and decay of the topic. The analysis on how the users sentiments evolved through time and how users feel concerning a given topic is a valuable resource for analysts

or companies. In summary, the described infrastructure permits to create quantitative and qualitative indicators for analyzing the role and the influence of the social networks and of their main actors in society.

The abovementioned framework is currently under development. The data collection platform is already operational, even if not complete or optimized. Some of the components of the social-data processing blocks are under the final development stages (e.g., data pre-processing, topic detection, important actors detection), but are not yet integrated in the system. Most of the remaining components are still in an early developing phase, even if the used techniques are well identified and tested.

REFERENCES

[1] "El 15 -M sacude el sistema". El País. Retrieved 26 May 2011.

[2] "BlogPulse", The Nielsen Company. February 16, 2011.

[3] Anger, I., Kittl, C., "Measuring Influence on Twitter", Proc. of the 11th International Conference on Knowledge Management and Knowledge Technologies, 2011.

[4] Bifet, A., Frank, E., "Sentiment knowledge discovery in twitter streaming data", Discovery Science, pages 1–15, 2010.

[5] Birnbaum, M., "Taking Goffman on a tour of Facebook: College students and the presentation of self in a mediated digital environment". ProQuest, UMI Dissertation Publishing, 2011.

[6] Brin, S., Page, L., "The anatomy of a large scale hypertextual web search engine," in: Proc. 7th internatinal conference on World Wide Web, 1998.

[7] Butow, E., Taylor, K., "How to Succeed in Business Using Linkedin: Making Connections and Capturing Opportunities on the World's #1 Business Networking Site", New York, Amacom, 2008.

[8] Cardoso, G. et al, "A sociedade em Rede em Portugal", Lisboa, Campo das Letras, 2005

[9] Carvalho, J.P., Coheur, L., "Introducing UWS – A Fuzzy Based Word Similarity Function with Good Discrimination Capability: Preliminary results", Proc. of the FUZZ-IEEE 2013, Hyderabad, India, 2013 .

[10] Castells, M., "Communication power". Oxford/New York: Oxford University Press, 2009

[11] Castells, M., "The Internet Galaxy, Reflections on the Internet, Business and Society". Oxford, Oxford University Press, 2001.

[12] Castells, M., "The Rise of the Network Society", The Information Age: Economy, Society and Culture Vol. I. Cambridge, MA; Oxford, UK: Blackwell, 2000.

[13] Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K., "Measuring user influence in twitter: The million follower fallacy", In ICWSM'10, 2010.

[14] Feixa, C., "Generació @, La Joventut al seglo XXI", Generalitat de CATALUNYA, Secretaria Nacional de Joventut, Barcelona, 2001

[15] Go, A., Bhayani, R., Huang, L., "Twitter Sentiment Classification using Distant Supervision," tech. rep., Stanford University, 2009.

[16] Hamelink, C., "The Ethics of Cyberspace",Sage, 2001.

[17] Homem, N. Carvalho, J.P., "Authorship Identification and Author Fuzzy Fingerprints", Proc. of the NAFIPS2011 - 30th Annual Conference of the North American Fuzzy Information Processing Society, 2011, IEEE Xplorer

[18] Homem, N. Carvalho, J.P., "Finding top-k elements in a time-sliding window", Evolving Systems, 2(1), pp. 51-71, Jan. 2011, Springer.

[19] Homem, N. Carvalho, J.P., "Finding top-k elements in data streams", Information Sciences, 180(24), pp. 4958-4974, Dec. 2010, Elsevier.

[20] Homem, N., Carvalho, J.P., "Web User Identification with Fuzzy Fingerprints", FUZZ-IEEE 2011 - 2011 IEEE International Conference on Fuzzy Systems, pp. 2622-2629, 2011

[21] Horowitz, D., Kamvar, S.D., "The anatomy of a large-scale social search engine", Proc. 19th internatinal conference on World Wide Web, Raleigh, North Carolina, USA, 2010.

[22] http://feedzai.com

[23] http://mongodb.org

[24] Huang, C. (June 6, 2011). "Facebook and Twitter key to Arab Spring uprisings: report". The National. Retrieved April 13, 2012.

[25] Huberman, B.A., Romero, D.M., Wu, F., "Social networks that matter: Twitter under the microscope", arXiv:0812.1045v1, Dec 2008.

[26] Lammel, R., "Google's MapReduce Programming Model – Revisited", Science of Computer Programming, Vol.70 (1), pp.1-30, 2008, Elsevier

[27] Mathioudakis, M., Koudas, N., "TwitterMonitor: trend detection over the twitter stream", Proc. of the 2010 international conference on Management of data, Indianapolis, Indiana, USA, pp.1155–1157, 2010.

[28] Mjos, O.J., "Music, Social Media and Global Mobility: MySpace, Facebook, YouTube", Routledge, 2012.

[29] Neumeyer, L., Robbins, B., et al. "S4: distributed stream computing platform", In KDCloud, 2010.

[30] Papacharissi, Z., "The Virtual Geographies of Social Networks: A comparative analysis of Facebook, LinkedIn and SmallWorld", New Media & Society, 11, pp.199-220, 2009.

[31] Plugge, E., Hawkins, T., Membrey, P., "The Definitive Guide to Mongodb: The Nosql Database for Cloud and Desktop Computing". Apress, 2010

[32] Rheingold, H., "The Virtual Community: Homesteading on the Electronic Frontier", MIT Press, 2000.

[33] Ritter, A., Clark, S., Mausam, Etzioni, O., "Named entity recognition in tweets: an experimental study", Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11, pp. 1524–1534, Stroudsburg, PA, USA, ACL, 2011.

[34] Schonfeld, E.. "Mining the thought stream. techcrunch weblog article". http://techcrunch.com/2009/02/15/mining-the-thought-stream/

[35] Tinati, R., Carr, L., Hall, W., and Bentwood, J. "Identifying Communicator Roles in Twitter", Proc. of the 21st international conference companion on World Wide Web, 2012.

[36] Twitter Team, "Twitter Turns Six". Twitter Blog (blog of Twitter), March, 21, 2012.

[37] Webster, F., "Theories "The Ethics of Cyberspace" London, Sage, 2000

[38] Weng, J., Lim, E., Jiang, J., "TwitterRank: Finding Topic-sensitive Influential Twitterers," Proc. of the third ACM international conference on Web search and data mining, 2010.

[39] Yogatama, D., Heilman, M., O'Connor, B., Dyer, C., Routledge, B., and Smith, N., "Predicting a scientific community's response to an article", Proc. of EMNLP, 2011.

[40] Zhang, M., Ye, X., "A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval", Proc. of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, 2008.