Characterizing and Predicting Bursty Events: The Buzz Case Study on Twitter

Mohamed Morchid, Georges Linarès, Richard Dufour

LIA - University of Avignon (France)
{mohamed.morchid, georges.linares, richard.dufour}@univ-avignon.fr

Abstract

The prediction of bursty events on the Internet is a challenging task. Difficulties are due to the diversity of information sources, the size of the Internet, dynamics of popularity, user behaviors... On the other hand, Twitter is a structured and limited space. In this paper, we present a new method for predicting bursty events using content-related indices. Prediction is performed by a neural network that combines three features in order to predict the number of retweets of a tweet on the Twitter platform. The indices are related to popularity, expressivity and singularity. Popularity index is based on the analysis of RSS streams. Expressivity uses a dictionary that contains words annotated in terms of expressivity load. Singularity represents outlying topic association estimated via a Latent Dirichlet Allocation (LDA) model. Experiments demonstrate the effectiveness of the proposal with a 72% F-measure prediction score for the tweets that have been forwarded at least 60 times.

Keywords: Bursty events detection; Latent Dirichlet Allocation; Neural network

1. Introduction

Information disseminated in micro-blogging platforms may become very popular and massively shared depending on the interest of the users for this information. This study deals with this particular information behavior by proposing to prior characterize and detect these bursty events (called *buzz*). The *buzz* is defined by (Yi, 2005) as a media activity explosion triggered by an information.

Predicting these bursty events is a difficult task since this phenomenon depends on various features related to the information itself, the public awareness and also the media dynamic aspects that tend to self-sustaining. These difficulties increase with the size, the dispersion and the fragmentation of the Web information.

Authors in (Bass, 2004; Goldenberg et al., 2001) study the impact of the *mouth-to-ear* and the *viral advertisement* in the context of information propagation. Others studies focused on information propagation using threshold-based models (Kempe et al., 2003). From this point-of-view, Twitter is an experimental space that is easier to study than the global Internet network. In this article, we propose to predict the retransmission peaks of the posted messages in Twitter. By this way, the number of retransmissions (*i.e. retweets*) is used as a measure of the performance of *buzz* prediction. Concretely, we consider that a message is considered as a *buzz* if its retweet number is higher than an *a priori* defined threshold.

In the context of the Twitter platform, its popularity is due to the capability to relay messages (*i.e. tweets*) posted by users. This particular mechanism, called *retweet*, allow users to massively share tweets[1] they consider as potentially interesting for others. The description of Twitter user behavior is a recent case study (Larceneux, 2007; Yang et al., 2010; Morchid et al., 2013). Various applications have then been made: prediction of natural disasters (Vieweg et al., 2010), learning support (Grosseck and Holotescu, 2008), political analysis (Tumasjan et al., 2010; Golbeck et al., 2010), or marketing (Wright, 2009). Studies also focused on the particular retweet mechanism. Usually, three types of features contribute to the retweet behavior: the content and the context in which the tweet is produced and shared (Kwak et al., 2010; Suh et al., 2010; Peng et al., 2011), the popularity of users (Cha et al., 2010; Romero et al., 2011), and the relation between users (Peng et al., 2011).

Nonetheless, few works are related to the *buzz* prediction. In (Hong et al., 2011), authors use the number of retweets not to predict the bursty events but to evaluate their popularity. Authors in (Kleinberg, 2003) propose a statistical hierarchical model of bursty text streams evaluated in the particular context of e-mails.

In this article, we focus on the *buzz* prediction using the textual content of the messages. We propose an approach based on the mapping of source documents in a reduced semantic space in which some descriptors could be determined by a Latent Dirichlet Allocation (LDA) analysis (Blei et al., 2003). Three features are particularly studied: the popularity, the thematic singularity, and the expressivity of the tweet content. Prediction is performed by a neural network that combines the three features. We demonstrate that this approach allows to detect information that may generate a *buzz* in large collection of heterogeneous data.

The article is organized as follows. Section 2. described the proposed system, while section 3. presents the experimental protocol. Results are then detailed in section 4. before concluding in section 5.

2. Buzz prediction system

The proposed system includes three main steps. The first one extracts keywords associated to a tweet. This keyword selection step lies on a thematic model estimated from the Latent Dirichlet Allocation (LDA) approach. Then the

[1] A *tweet* is a short text message composed of up to 140 characters.

popularity, singularity and expressivity descriptors are extracted from the message and its thematic representation. Finally, a neuronal network is used to determine the number of retweets for each tweet using the considered descriptors. Figure 1 presents the architecture of the *buzz* prediction system, where 6 successive steps are needed:

1. Representation of a tweet t with a feature vector W^t.
2. Estimation of a LDA model on a large corpus D of documents to produce a topic space T_{spc}.
3. Projection of W^t into T_{spc} to select a subset of topics $S^z \subset T_{spc}$ representing the tweet.
4. Extraction of a subset S^w representing the tweet keywords from S^z regarding W^t.
5. Extraction of an index vector from S^w where each coefficient represents the score of popularity, expressivity and singularity.
6. Training of a neuronal network to predict the tweet *buzzability*.

The three main steps are described below.

2.1. Extraction of keywords

The Twitter platform limits the size of messages to 140 characters. This constraint causes the use of a particular vocabulary that is often unusual, noisy, full of new words, including misspelled or even truncated words (Choudhury et al., 2007). Only using the tweet words is insufficient (Morchid et al., 2013).
In order to compensate these particularities, two methods are compared to increase the initial tweet lexicon from an additional corpus of documents: a classical word representation with the TF-IDF-RP method (Salton, 1989) and a topic space representation with the LDA approach (Blei et al., 2003). In our experiments, this additional corpus is composed of 100,000 Wikipedia and AFP documents containing around 1 billion words.

2.1.1. TF-IDF-RP

Let's D be a corpus of n_d documents d and n_w be the vocabulary size. Each tweet t can be represented as a point of $\mathbb{R}^{n_w}$ by the vector W_i^t of size n_w where the i^{th} feature $(i = 1, 2, \ldots, n_w)$ combines: the Term Frequency (TF), the Inverse Document Frequency (IDF) and the Relative Position (RP) (Salton, 1989) of a word w_i of t:

$$W_i^t = tf_i.idf_i.rp_i \ , \tag{1}$$

where

$$tf_i = \frac{|\{w_i : w_i \in t\}|}{|t|}, \ idf_i = log\frac{|C|}{|\{d : w_i \in d\}|}, \ rp_i = \frac{|t|}{fp(w_i)} \ . \tag{2}$$

Here, $|\cdot|$ is the number of elements in the corresponding set and fp_i is the position of the first occurrence of w_i in the tweet t.
This approach allows a simple extraction of the n most representative words in W^t of a tweet. The system extracts the 10 words that obtain the highest TF-IDF-RP scores (Zhang et al., 2010).

2.1.2. Combination of latent topics

Latent Dirichlet Allocation (LDA) is a generative model which considers a document model (seen as a *bag of words* (Salton, 1989)) as a mixture probability of latent topics. These latent topics are characterized by a distribution of word probabilities which are associated with them. At the end of LDA analysis, we obtain a set of topics with, for each, a set of words and their emission probabilities.
LDA is applied on a corpus D containing a vocabulary of m_w words. Firstly, a topic model is built using a feature vector V_i^z associated with each topic z of the semantic space T^{spc}. Each i^{th} feature $(i = 1, 2, \ldots, m_w)$ of V_i^z represents the probability of the word w_i knowing the topic z.
A semantic space of 5,000 topics is obtained (empirically defined). For each LDA class, the 50 words with the maximum weights are selected.
The tweet t is then mapped into T_{spc}. A similarity measure is computed between a tweet (W_i^t) and a topic (V_i^z) using the cosine metric:

$$\delta(t, z) = \frac{\sum\limits_{w_i \in t} V_i^z.W_i^t}{\sqrt{\sum\limits_{w_i \in z} {V_i^z}^2. \sum\limits_{w_i \in t} {W_i^t}^2}} \ . \tag{3}$$

Finally, the keywords associated to a tweet are obtained by searching the intersection between the main topics and the tweet. This intersection $S^w = \{s(w_1), s(w_2), \ldots, s(w_{|S^w|})\}$ is composed with the $|S^w|$ words w_i with the higher scores:

$$s(w_i) = \sum_{z \in S^z} \delta(t, z).P(w_i|z) \ . \tag{4}$$

2.2. *Buzzability* descriptors

We propose to study the contribution of three indicators to the *buzz* phenomenon. The first one is the recent "popularity" of words based on a statistical analysis of RSS feeds. The second one is based on the probability of associating dominant themes of the *tweet*; this is a measure of saliency that uses unlikely thematic associations as a factor favoring the audience. The last indicator assesses the expressivity of the *tweet* words from a *sensitivity lexicon* previously annotated.

2.2.1. Popularity

This dynamic descriptor provides the frequency of the "recent" popular words of a tweet in the media. Thus, the RSS feeds R of four French major national newspapers (*Le Monde, Libération, Le Figaro and L'Equipe*) are extracted between 2009/01/01 and 2011/07/01. The *popularity* p of each word $w \in S^w$ in the RSS feeds is calculated by:

$$p(w) = \frac{1}{\hat{p}(w)} \sum_{d \in R} p(w|d) \ , \ p(t) = \underset{w \in S^w}{\text{argmax}} \ (p(w)) \ , \tag{5}$$

where d is a page of R, $\hat{p}(w)$ is the maximum number of occurrences of w in R and $p(w|d)$ is the number of occurrences of w in the page d. Thus, a classification of the words found in R from the most "popular" $(p(w) = 1)$ and

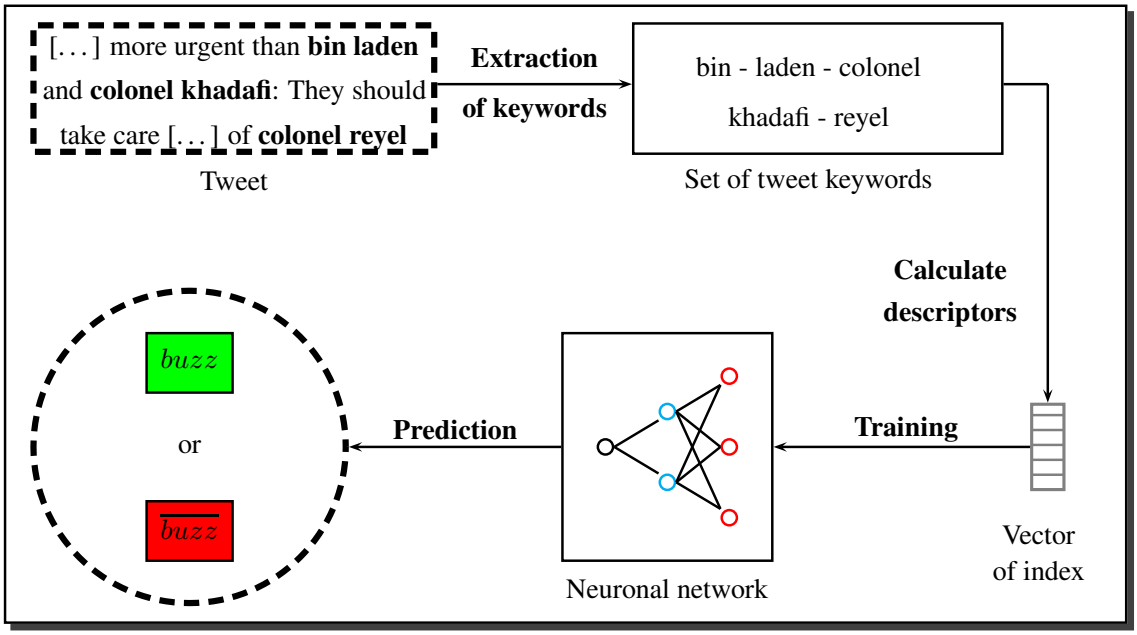

Figure 1: Architecture of the *buzz* prediction system.

the less "popular" ($p(w) \approx 0$) words is formed. The RSS feeds studied are prior to the emission of the tweets.

2.2.2. Singularity

We consider that an ordinary subject (*i.e.* frequent) is less likely to be released than a more unusual one. The idea developed behind is that the salience of a document promotes its popularity.

The most natural approach is to empirically estimate, from a large corpus, the probability of topic co-occurrences. Due to the overall complexity (semantic space of 5,000 topics) the number of representative topics of a tweet is limited to two. The aim is then to determine the probability $P(z_i \in t, z_j \in t)$ $(i \neq j)$ that these two topics are present in the same tweet t. Rather than attempting to directly estimate this probability in the semantic space $(5,000 \times 5,000)$, we use the fact that the classes themselves come from a co-occurrences analysis of words.

In this case, we consider that the probability of two topics to appear in a tweet depends of the intersection of their most representative words. We construct a graph whose vertex is a topic. Arc values are evaluated by the symmetric Kullback-Lieber divergence ψ normalized between each of the two topics:

$$\psi(z_i, z_j) = \frac{1}{2}\left(\sum_{w \in z_i} p_i log\frac{p_i}{p_j} + \sum_{w \in z_j} p_j log\frac{p_j}{p_i}\right) , \quad (6)$$

where p_n is the probability that the word w appears in the topic z_n.

Overall, the distance n between any of the two topics is then determined using the shortest path connecting them with the Dijkstra algorithm (Dijkstra and Eindhoven, 1971) ϕ normalized by the number of arcs. This yields to a descriptor d $(0 \leq d \leq 1)$:

$$n(z_i, z_j) = \frac{1}{|a(i,j)|}\phi(z_i, z_j) ,$$

and

$$d(t) = \underset{z_i, z_j \in S^z}{\text{argmax}} \left(n(z_i, z_j)\right) . \quad (7)$$

2.2.3. Expressivity

The expressivity of a tweet is measured using a sensitivity lexicon (Paroubek, 2010) containing **976 words** from ANEW (Affective Norms of English Words) (Bradley and Lang, 1999). To determine the sensitivity of a word, the authors introduce a new specific measure named *valence*. This measure is defined for each word and is calculated as the number of times that this word appears near a positive **:)** or a negative **:(** emoticon. This value is calculated considering the probability that the word is in a positive context $P(M^+|w)$. The valence v gives a sensitivity score of a word w:

$$v(w) = 8 \times P(M^+|w) + 1 . \quad (8)$$

The valence value varies from 1 (negative context) to 9 (positive context). If a word appears in a positive or negative context, this word can be considered as a sensitive word. For this reason, we introduce a measure $\delta(w)$. This measure $\delta(w)$ shows if a word w appears in a sensitive context (positive or negative) and if it is centered between 0 (any) and 1 (sensitive):

$$\delta(w) = 2 \times |P(M^+|w) - \frac{1}{2}| ,$$

and

$$\delta(t) = \underset{w \in S^w}{\text{argmax}} \left(\delta(w)\right) . \quad (9)$$

2.3. Predictive model

The prediction is realized with a neuronal network that evaluates the *buzzability* from the three descriptors (p, d, δ) described above. This is a multi-layer perceptron with one hidden layer. This neuronal network is trained with the gradient algorithm.

All networks have a 2-cells in the hidden layer and the input/output layers are limited to one cell. In all cases, the input cells receive the descriptors and the output cells product a value between 0 and 1. The training of these networks depends on the behavior expected. In particular, we adopted the rule that a tweet is considered as a *buzz* if it exceeds a defined number of retweets.

A threshold t^b (number of minimal retweets to *buzz*) is then varied from 0 to 90 retweets in our experiments. For example, a threshold of 10 means that a tweet retweeted more than 10 times generates a *buzz*.

3. Experimental protocol

A corpus of 4,500 French tweets extracted with the Twitter API[2] is used (90% for training and 10% for testing). All the processed data (corpus and tweets) are lemmatized and filtered trough a stop list. Table 1 shows examples of tweets.

Experiments consist in evaluating the variation of the F-measure according to a threshold (*i.e.* number of retweets, see section 2.3.). We can expect an easier detection of "extreme" examples that are either very few or widely retweeted. The system is evaluated with two methods: TF-IDF-RP (see section 2.1.1.) and intersection of topics (see section 2.1.2.). Each descriptor is evaluated separately and then combined with the others.

4. Results

The figures 2 and 3 present the evolution of the F-score depending on the number of retweets[3], for the intersection of topics (figure 2) and TF-IDF-RP (figure 3) methods. We can firstly note that keyword extraction performance using a LDA approach (figure 2) is better than the one obtained with a classical TF-IDF-RP method. These results confirm the idea that have motivated this approach: passing trough an intermediate representation improves the robustness of the system for the "noisy" Twitter language.

The LDA results are more consistent than those obtained with the TF-IDF-RP approach which tend to produce unstable results for all individual descriptors. The results obtained with the combination of the three descriptors (figure 2) also show their complementarity, although individual results are quantitatively similar.

Moreover, the "sensitivity" descriptor obtains the best individual results and, more surprisingly, outperforms the combination of low retweet thresholds . This phenomenon indicates the ability of this descriptor to detect the retweets (even low) rather than the *buzz*. The table 1 shows that most of the tweets contain at least a "sensitive" word (see figure 4).

[2] dev.twitter.com

[3] A threshold beyond considers the tweet as a *buzz*

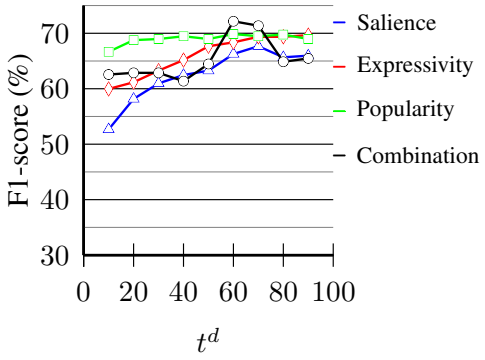

Figure 2: F-score performance when varying the number of retweets threshold using the LDA-based methods.

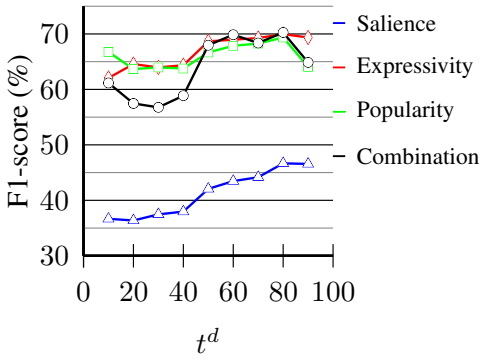

Figure 3: F-score performance when varying the number of retweets threshold using the TF-IDF-RP method.

5. Conclusion and perspectives

In this paper, we proposed a method for predicting the *buzz* (*i.e.* bursty events) on the Twitter platform. Three descriptors have been evaluated alone, and then combined. Results show their complementarity: the best system achieved a 72% F-score.

The *buzz* is a dynamic phenomenon where the prediction could be based on models that include not only the content (this is the way that we explore in this work) but also the information speed spreads. Incorporate the dynamic and/or

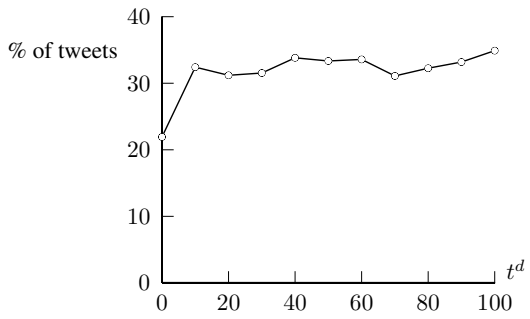

Figure 4: Proportion (%) of tweets containing at least a "sensitive" word regarding t^d.

tweet	date of emission	retweets gap
What am i supposed to do? I adore you but the circumstances keep us apart	03/03/2011	[0; 10[
Michael has put his whole self to create beauty & magic!	02/05/2011	[0; 10[
A police officer came up to me yesterday and asked: Where We You between four and six? I said: Kindergarten...	02/05/2009	[10; 20[
My yoga teacher is from Vancouver!!! I don't know why, but this makes me really happy :-) She has my accent...	09/05/2011	[10; 20[
I'll be back with an Oscar!	09/01/2010	[20; 30[
Feminism is not about women trying to be men. It's about women wanting to be respected, and wanting feminine values to ...	10/05/2011	[20; 30[
Obama... Who Needs Him? Just Turn The Teleprompter Towards The Camera and We'll Read It Ourselves	23/04/2011	[30; 40[
There's nothing scarier than getting what you want, because that's when you really have something to lose.	27/04/2011	[30; 40[
Who Needs Education When We have Google	26/06/2009	[40; 50[
The first thing in my mind when I opened my eyes in the morning, it's You	22/06/2011	[40; 50[
#Sagittarius (Woman) are very adventurous in the bedroom and are very open minded	18/08/2010	[50; 60[
Dear older married people: You know you can each have an email address of your own, right?	22/05/2011	[50; 60[
why do my friends always think im in a bad mood ? im just a GEMINI sometimes we need personal space	17/05/2011	[60; 70[
It's painful to say goodbye to someone you don't want to let go, but more painful to ask someone to stay when you ...	20/05/2011	[60; 70[
My HEART is bigger than yours because its been beated up too much by your heart	24/04/2011	[70; 80[
I saw you, I wanted you. I got you, I liked you. I loved you, I lost you, I miss you. That's the way it goes sometime ...	09/05/2011	[70; 80[
I think, the day of the very LAST Monster Ball we should trend	14/08/2009	[80; 90[
We are investigating CSS issues with mobile.twitter.com and hope to resolve the issue shortly.	23/04/2010	[80; 90[
You must have been born on a highway 'cause that's where most accidents happen	18/02/2010	[90; 100]
Lies will eventually be discovered, sooner or later. Cause no lies last forever.	09/05/2011	[90; 100]

Table 1: Examples of tweets with the date of emission and the number of retweets gap.

structural aspects of the diffusion mechanism could significantly improve the quality of the prediction.

6. References

F.M. Bass. 2004. Comments on "a new product growth for model consumer durables the bass model". *Management science*, 50(12 supplement):1833–1840.

D.M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

M.M. Bradley and P.J. Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. *University of Florida: The Center for Research in Psychophysiology*.

M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

M. Choudhury, R. Saraf, V. Jain, S. Sarkar, and A. Basu. 2007. Investigation and modeling of the structure of texting language. In *IJCAI-Workshop on Analytics for Noisy Unstructured Text Data*, pages 63–70.

E.W. Dijkstra and Technische Hogeschool Eindhoven. 1971. A short introduction to the art of programming.

J. Golbeck, J.M. Grimes, and A. Rogers. 2010. Twitter use by the us congress. *Journal of the American Society for Information Science and Technology*, 61(8):1612–1621.

J. Goldenberg, B. Libai, and E. Muller. 2001. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review*, 9:1–18.

G. Grosseck and C. Holotescu. 2008. Can we use twitter for educational activities. In *International Conf. on e-Learning and software for education, Bucharest, Romania*.

L. Hong, O. Dan, and B.D. Davison. 2011. Predicting popular messages in twitter. In *ACM international conference companion on World wide web*, pages 57–58.

D. Kempe, J. Kleinberg, and É. Tardos. 2003. Maximizing the spread of influence through a social network. In *ACM SIGKDD International Conf. on Knowledge discovery and data mining*, pages 137–146.

J. Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

H. Kwak, C. Lee, H. Park, and S. Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

F. Larceneux. 2007. Buzz et recommandations sur internet: quels effets sur le box-office? *Recherche et applications en marketing*, pages 45–64.

M. Morchid, R. Dufour, and G. Linarès. 2013. Thematic representation of short text messages with latent topics: Application in the twitter context. In *PACLING*.

A.P.P. Paroubek. 2010. Construction d'un lexique affectif pour le français à partir de twitter.

H.K. Peng, J. Zhu, D. Piao, R. Yan, and Y. Zhang. 2011. Retweet modeling using conditional random fields. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 336–343. IEEE.

D. Romero, W. Galuba, S. Asur, and B. Huberman. 2011. Influence and passivity in social media. *Machine Learning and Knowledge Discovery in Databases*, pages 18–33.

G. Salton. 1989. Automatic text processing: the transformation. *Analysis and Retrieval of Information by Computer*.

B. Suh, L. Hong, P. Pirolli, and E.H. Chi. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 177–184. IEEE.

A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conf. on Weblogs and Social Media*, pages 178–185.

S. Vieweg, A.L. Hughes, K. Starbird, and L. Palen. 2010. Microblogging during two natural hazards events: what

twitter may contribute to situational awareness. In *ACM International Conf. on Human factors in computing systems*, pages 1079–1088.

A. Wright. 2009. Mining the web for feelings, not facts. *New York Times*, 24.

Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. 2010. Understanding retweeting behaviors in social networks. In *ACM international conference on Information and knowledge management*, pages 1633–1636.

J. Yi. 2005. Detecting buzz from time-sequenced document streams. In *IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE)*, pages 347–352.

W. Zhang, T. Yoshida, and X. Tang. 2010. A comparative study of tf*idf, lsi and multi-words for text classification. *Expert Systems With Applications*.