Alternate *ACM* SIG Proceedings Paper in LaTeX Format[*]

[Extended Abstract] [†]

Ben Trovato[‡]
Institute for Clarity in Documentation
1932 Wallamaloo Lane
Wallamaloo, New Zealand
trovato@corporation.com

G.K.M. Tobin[§]
Institute for Clarity in Documentation
P.O. Box 1212
Dublin, Ohio 43017-6221
webmaster@marysville-ohio.com

Lars Thørväld[¶]
The Thørväld Group
1 Thørväld Circle
Hekla, Iceland
larst@affiliation.org

ABSTRACT

todo (200 kai kati lekseis - gia na kseroume poso xoro pianei)

CCS Concepts

•**Computer systems organization** → **Embedded systems;** *Redundancy;* Robotics; •**Networks** → Network reliability;

Keywords

ACM proceedings; LaTeX; text tagging

1. INTRODUCTION

Nowadays, the World Wide Web has transformed from a large, static library that people only browse into a vast and dynamic information resource. Relying on this, social networks is a very popular and powerful tool for expressing opinions, broadcasting news, and simply communicating with friends. People using them for commenting on significant events in real time, with several hundred micro-blogs posted each second.

The most popular micro-blogging service is Twitter. The popularity of Twitter stems from its availability on a number of different electronic devices (web and cell phones. There is a prevalence of a subculture in Twitter that encourages users to acquire a large friend pool, as well as send tweets on a wide variety of subjects, typically several times a day.

Monitoring and analysing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information, which would not have been available from traditional media outlets. Tweets can be seen as a dynamic source of information enabling individuals, corporations, and government organizations to stay informed of âĂIJwhat is happening now.âĂİ For instance, people would be interested in getting advice, opinions, facts, or updates on news or events. Companies are increasingly using Twitter to advertise and recommend products, brands, and services; to build and maintain reputations; to analyse usersâĂŹ sentiment regarding their products (or those of their competitors); to respond to customersâĂŹ complaints; and to improve decision making and business intelligence. Twitter has also emerged as a fast communication channel for gathering and spreading breaking news, for predicting election results, and for sharing political events and conversations. It has also become an important analytical tool for crime prediction and monitoring terrorist activities.

Twitter promotes an attractive style stating breaking news, as there is very little lag between the time that an event hap-

[*](Produces the permission block, and copyright information). For use with SIG-ALTERNATE.CLS. Supported by ACM.

[†]A full version of this paper is available as *Author's Guide to Preparing ACM SIG Proceedings Using LaTeX2$_\epsilon$ and BibTeX* at `www.acm.org/eaddress.htm`

[‡]Dr. Trovato insisted his name be first.

[§]The secretary disavows any knowledge of this author's actions.

[¶]This author is the one who did all the really hard work.

WOODSTOCK '97 El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06…$15.00

DOI: 10.475/123_4

pens or is first reported in the news media and the time at which it is the subject of a posting on Twitter. Twitter can be characterized as an endless database, which collects millions of real-time short text messages every second. Tweets also have a mechanism by which the user can link to other objects on the web such as articles, images or videos which is typically used to link tweets to related material on the Internet. Thereafter, the first result is that the size of information is multiplied and the variety of references is bigger, as well. These messages are not only just data, but they can be manipulated efficiently. One well-timed subject of research is to use those messages for event detection. In other words, the tweets is a source of inventing which topics are more seasonable. Event detection has also instant impact on the world, through the quick transmission of the news and necessary briefing in some cases.

With the passage of time and the effect of more and more users the topic acquire much popularity. Through this phenomenon we can form a general summarization of the event. This process is called event summarization. The massive crowd keeps close pace with the development of trending topics and provide the timely updated information. Twitter has shown its powerful ability in information delivery in many events, like the wildfires in San Diego and the earthquake in Japan. In response to searches for ongoing events, today's major search engines simply find tweets that match the query terms, and present the most recent ones. This approach has the advantage of leveraging existing query matching technologies, and for simple one-shot events such as earthquakes it works well. However, for events that have "structure" or are long-running, and where users are likely to want a summary of all occurrences so far, this approach is often unsatisfactory.

Event detection is a growing domain of research. Many different species of algorithms have been detected regarding this sector. A common approach are techniques that are based on text categorization. In addition, there are some methods that reclaim the display frequency of each term.

The computational treatment of sentiment has recently attracted a great deal of attention, in part because of its potential applications. One of the main reasons for sentiment analysis is the aforementioned increase of user-generated content on the Web which has resulted in a wealth of information that is potentially of vital importance to institutions and companies. Typically, document-based sentiment analysis processes operate at a particular level, i.e. at the word or sentence level, for extracting a document's sentiment. In machine learning, the most popular approach for sentiment analysis, the selection of appropriate features for representing a document is crucial. In sentiment identification at the word level different types of features have been introduced, which are either sentiment-based (e.g. words which express a specific sentiment), syntactic-based (e.g. part-of-speech and n-grams), or semantic-based (e.g. semantic word vector spaces which capture the meaning of each word).

Document-level polarity classification is not a special case of text categorization with sentiment -rather than topic-based categories. Hence, standard machine learning classification techniques, such as support vector machines (SVMs), can be applied to the entire documents themselves. Nevertheless, some researches presented a technique that it is easy to improve the accuracy, by integrating sentence-level subjectivity detection with document-level sentiment polarity.

As we know, in the machine learning approach, each classifier is trained using a collection of representative data. In contrast, the semantic-orientation approach does not require prior training; instead, it measures a word containing positive or negative sentiment. Each approach has its own benefits and drawbacks. For example, the machine learning approach tends to be more accurate, but the semantic-orientation approach has better generality. Recently, a new lexicon-enhanced method was accrued to generate a set of sentiment words based on a sentiment lexicon as a new feature dimension. It combines these sentiment features with content-free and content-specific features used in the existing machine-learning approach. In the evaluation stage, they showed that adding the new set of sentiment features can increase sentiment-classification performance.

The Internet and other communication technologies play a potentially disruptive role on the constraints imposed on social networks. These technologies reduce the overhead and cost for being introduced to new people regardless of geography, and help us stay in touch with those we know. Some have even gone so far as to call this âĂİthe end of geography,âĂİ where the process of relationship formation becomes disentangled from distance altogether.

However, geography still plays an important role. The reason is because of the strong relationship between event detection and geographical location each user belongs. Twitter is a social networking website, which means that users need not be viewed in isolation, but instead can be viewed as part of a large network of other users, user groups, and user cliques. Moreover, users have some meta-data information, such as description, source location, friends, which means that the social network structure in Twitter can aid in finding users that are most likely to tweet about news belonging to a particular geographic location or region.

The rise of micro-blogging services spurred various applications to mine the data coming from those services. Many such applications could benefit from information about the location of users, but unfortunately location information is currently very sparse. The main problem is that less than 1 per cent of tweets are geo-tagged and information available from the location field in usersâĂŹ profiles is unreliable at best. The benefits of mining those data promises new personalized information services, including local news summarized from tweets of nearby Twitter users, the targeting of regional advertisements, spreading business information to local customers, and novel location-based applications (e.g., Twitter-based earthquake detection, which can be faster than through traditional official channels).

There is a great number of geoinference using social networks One direction has produced approaches that claim to accurately locate the majority of posts within tens of kilometres of their true locations. Another method predicts the location of an individual from a sparse set of located users with peformance that exceeds IP-based geolocation. On the side, there is also a technique that predicts locations of Twitter users at different granularities, such as city, state, or time zone, using the content of their tweets and their tweeting behaviour.

Getting started the first section of this survey is the presentation of some techniques that aim to event detection. The survey detects both algorithms that are based on text categorization and frequency display methods. The second chapter deals with techniques of sentiment analysis. We give

more weight on techniques that use machine learning. The last chapter unfolds methods for location identification.

2. EVENT DETECTION

Given a series of twitter posts, the goal of event detection is to extract a particular event by analysing the text or hashtag of a tweet. The process of event detection is not a lenient task as tweets stream in huge volumes and the level of noise is kept high [24]. On the other hand, the huge volume of the stream allows the use of streaming algorithms, thus making event detection an accomplishable task [24].

The conventional approach for this problem is to represent the documents as term frequency vectors [24]. When a new document arrives, it is compared to all previous ones, and if its similarity to a specific document called "centroid" is below a threshold, the new document is registered as a new event [24]. Unfortunately, this simple approach doesn't perform well as the dimensionality of the data expands.

Twitter's users usually tweet about their personal life, personal opinions, taste in music etc., which makes the majority of tweets unsuitable for our purpose. Although, when a noteworthy event occurs, a lot of users tweet about it. The goal of event detection is to detect these events in an automatic fashion, keeping the noise (tweets that are not events) as low as possible [24].

2.1 Strategies for event detection

As stated before, tweets stream in high rates and capture the pulse of the moment, therefore occupying us to process them right away [25]. This is accomplished by using algorithms that are online in nature, meaning they can on inputs that are not known beforehand[25].

Tweets are inherently noisy since most of them concern small user groups. In order to improve the results of event detection we first have to improve the input by extracting only useful information from noise [25]. Even then, information that concerns us is still limited by the upper character limit set for each tweet, currently at 140 characters. Bad use of language, spelling errors and special puns also negatively affect the outcome of event detection.

Twitter is an evolving platform where users come and go, and become friends and followers with each other. When designing event detection algorithms, one should keep in mind the above aspects and make the algorithms adapt well to new knowledge, otherwise watch them fail as time passes [25].

The difference between newswire and Tweeter is that, in Tweeter, the identities of the news "reporters" in not known in advance and the news are not posted based on a schedule, but rather appear randomly among other tweets [25].

2.1.1 *Identifying users*

An approach to event detection is the identification of users that usually tweet about news [25]. A technique to do this is to manually identify these users by looking for the most common set of followers among them [25]. Another technique to do this is to assume that a user with a high amount of followers represents an influential event source into a social community [4]. As an example, we present figure 1 , which features Barack Obama as the most influential node in this group [4].

The added benefit of this method is that, not only we identify the users of interest but, we get an evaluation of a user's influence, meaning that an influential user can affect other users' tweet streams [4]. This technique cascades as the follow-like relationships spread, making it comparable to the PageRank algorithm [4]. An interesting fact to support the above claims is that studies have shown that 10% of Twitter's users are responsible for 90% of the tweets [25].

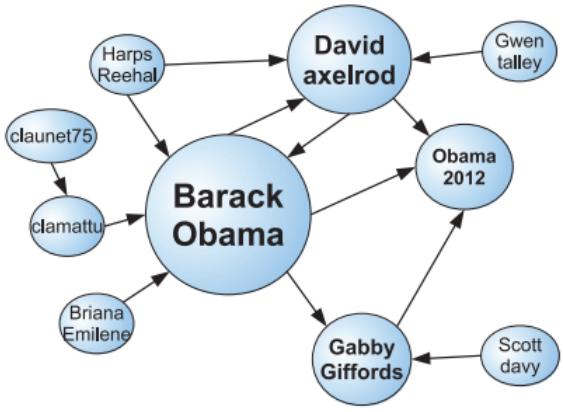

Figure 1: Graph showing user importance on Twitter, based on the number of their followers. [4]

2.2 Term and Topic Extraction

Although manual techniques exists, we will solely focus on automatic extraction techniques as they fit best on the Twitter paradigm. Filtering could be applied on all tweets, expect those posted by influential users that usually post news, leaving us with a guaranteed source of news. The goal of the extraction is to find a way to distinguish tweets that are clearly not news, and not completely eliminate the noise, which is an enormous task to take on [25]. For the above task, [25] used a naive Bayes classifier that was trained using tweets that were marked as news or not news.

2.2.1 *Setting a threshold for term extraction*

An approach to selecting emerging terms relies on a user-specified threshold parameter, as described by [4]. It is assumed that, given two keywords that are likely to be news, they can be considered as such based on user evaluation. A threshold value called critical drop is introduced, which allows the user to decide when a term is news or not.

In order for the above technique to work, a proper drop value has to be set. This a be a daunting task for a user or a machine. The difficulty lies in the fact that it's difficult to quantify a threshold given an abstract context that is, on top of other things, unstable. An automatic way of doing this, as described by [4], is shown below: The main idea is to preserve terms with high frequencies but only consider those whose frequency is higher than the average frequencies of most discussed terms. A model is introduced that works as follows:

- Terms are ranked in descending order based on their frequencies
- The maximum drop value between consecutive entries is calculated and the drop point is identified.
- The average drop between consecutive entities for all

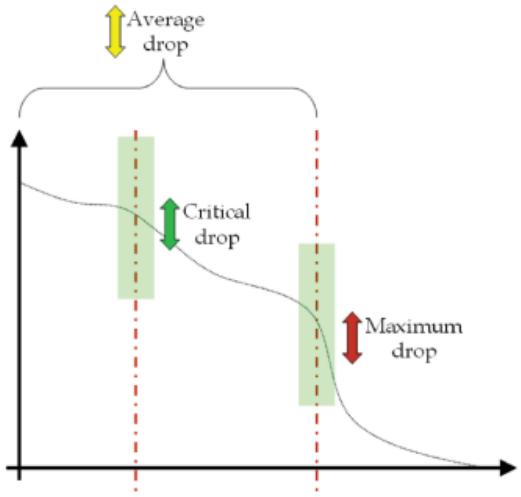

Figure 2: The critical drop cut-off: the maximum drop is the highest variation in the ordered list of weights (red mark). The average drop (between consecutive entities) is the average difference between those items that are ranked before the identified maximum drop point (yellow mark). The first drop which is higher than the computed average drop is called the critical drop (green mark). Figure also taken from [4].

those keywords that are ranked before the identified maximum drop point is calculated.

- The first drop which is higher than the computed average drop is identified as the critical drop.

At the end of this procedure, the terms ranked before the critical drop are defined as news on the specified time interval.

2.2.2 Clustering

The goal of clustering techniques is to group news tweets such that each group contains tweets about a specific news topic. This problem is similar to clustering documents with the main difference that the identity of news topics is not known beforehand [25]. This problem cannot be approached by static training sets, as news evolve over time in a random fashion. A clustering approach, presented by [25], goes as follows: Once a tweet arrives, it is added to a news cluster and remains there forever. This techniques allows for fast processing since tweets arrive at huge rates. This makes the algorithm online. Specifically, an algorithm called leader-follower clustering has been modified and used, which allowed for clustering both in content and time. Along with the clusters, a weighted list of keywords of each tweet is kept. Time is also an important aspect in this procedure: Timestamps are kept and can render a cluster inactive if its centroid is older than a specified threshold. In this case the cluster is locked and new tweets cannot be added.

Noise can greatly affect clustering techniques. Care should be taken when creating new clusters to ensure they are of good quality. A good way to ensure high quality is to only allow the creation of clusters by influential users with many followers [25]. All tweets from other users should be able to join a cluster but not create one. Unfortunately, this approach might prevent news registration for cases where simple users report an incident first. A way to combat this situation effectively is to allow cluster creation by users, but only activate a cluster after a certain amount of tweets has been added to it, thus making the topic news-worthy [25].

2.2.3 A different approach on Clustering

As [4] states, many clustering and classification strategies cannot by applies to this task due to the fact that they tend to ignore relationships among documents(tweets) related to a news event. [4] proposes a metaphor where each term is seen as a living organism: A term's vitality status increases as users use it, and slowly degrades as references drop over time. The term's vitality is weighted based on the source of the tweet. As stated before, highly influential users are more likely to post high-importance news, hence allowing us to favour terms found in their tweets.

3. CONCLUSIONS

4. ACKNOWLEDGMENTS

5. ADDITIONAL AUTHORS

6. REFERENCES

[1] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6(12):1326–1329, 2013.

[2] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[3] C. Budak, D. Agrawal, and A. El Abbadi. Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment*, 4(10):646–656, 2011.

[4] M. Cataldi, L. D. Caro, and C. Schifanella. Personalized emerging topic detection based on a term aging model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):7, 2013.

[5] D. Chatzakou, V. Koutsonikola, A. Vakali, and K. Kafetsios. Micro-blogging content analysis via emotionally-driven clustering. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 375–380. IEEE, 2013.

[6] D. Chatzakou, N. Passalis, and A. Vakali. Multispot: Spotting sentiments with semantic aware multilevel cascaded analysis. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 337–350. Springer, 2015.

[7] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[8] R. Compton, D. Jurgens, and D. Allen. Geotagging one hundred million twitter accounts with total

variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.

[9] Y. Dang, Y. Zhang, and H. Chen. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53, 2010.

[10] M. Desai and M. A. Mehta. Techniques for sentiment analysis of twitter data: A comprehensive survey. In *Computing, Communication and Automation (ICCCA), 2016 International Conference on*, pages 149–154. IEEE, 2016.

[11] C. N. Dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78, 2014.

[12] D. Gao, W. Li, and R. Zhang. Sequential summarization: A new application for timely updated twitter trending topics. In *ACL (2)*, pages 567–571. Citeseer, 2013.

[13] A. Giachanou and F. Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28, 2016.

[14] T. Hua, F. Chen, L. Zhao, C.-T. Lu, and N. Ramakrishnan. Sted: Semi-supervised targeted event detection. *KDD'13*, pages 11–14, 2013.

[15] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu, and D. Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *ICWSM*, pages 188–197, 2015.

[16] V. Kharde, P. Sonawane, et al. Sentiment analysis of twitter data: A survey of techniques. *arXiv preprint arXiv:1601.06971*, 2016.

[17] L. Kong, Z. Liu, and Y. Huang. Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment*, 7(13):1681–1684, 2014.

[18] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 721–730. ACM, 2012.

[19] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? inferring home locations of twitter users. *ICWSM*, 12:511–514, 2012.

[20] K. R. McKelvey and F. Menczer. Truthy: Enabling the study of online social networks. In *Proceedings of the 2013 conference on Computer supported cooperative work companion*, pages 23–26. ACM, 2013.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[22] G. Paltoglou and M. Thelwall. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):66, 2012.

[23] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[24] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–189. Association for Computational Linguistics, 2010.

[25] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, 2009.

[26] H. Sayyadi and L. Rschid. A graph analytical approach for fast topic detection.

[27] A. Severyn and A. Moschitti. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado*, pages 464–469, 2015.

[28] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics, 2011.

[29] C. Suen, S. Huang, C. Eksombatchai, R. Sosic, and J. Leskovec. Nifty: a system for large scale information flow tracking and clustering. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1237–1248. ACM, 2013.

[30] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL (1)*, pages 1555–1565, 2014.

[31] J. Weng and B.-S. Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.