Does Love Change on Twitter?

The Dynamics of Topical Conversations in Microblogging

Victoria Lai and William Rand
Center for Complexity in Business
University of Maryland
College Park, Maryland 20742
{vlai, wrand}@umd.edu

Abstract—Discovering automatically what people are talking about on social media with respect to a particular topic would be useful since it would give insight into how people perceive different topics. However, identifying trending terms / words within a topical conversation is a difficult task. We take an information retrieval approach and use tf-idf to identify words that are more frequent in a focal conversation compared to other conversations on Twitter. This requires a query set of tweets on a particular topic (used for term frequency) and a control set of conversations to use for comparison (used for inverse document frequency). The terms identified as trending within a topical conversation are greatly affected by the particular control set used. There is no clear metric for whether one control set is better than another, since that is determined by the needs of the user, but we can investigate the stability properties of topics given different control sets. We propose a method for doing this, and show that some topics of conversation are more stable than other topics, and that this stability is also affected by whether only the most frequent terms are of interest (top-50), or if all words (full-vocabulary) are being examined. We end with a set of guidelines for how to build better topic analysis tools based on these results.

Keywords-social media; microblogging; trend identification; topic stability; language usage; ranking; Twitter

I. INTRODUCTION AND MOTIVATION

Users currently generate over 500 million daily status updates on Twitter and the rate of tweeting is growing. As a result, filtering through such content to identify trends becomes more contextually and computationally difficult. Trend identification provides insight into the conversations about a topic, by focusing on the most popular parts of the conversation. For example, a brand manager may want to know which keywords are commonly associated with a particular brand on social media. Using these trends, the brand manager could determine whether overall brand perception is positive or negative, or capitalize on trending subjects in future marketing campaigns. Trends are also interesting for their temporal characteristics and diffusion properties — how long is a term trending? Does a trend in one geographic area spread to another one? Being able to identify trends effectively has analytical and practical uses, whether monitoring a particular brand or studying diffusion of conversation on social media.

Using pure term frequency to identify keywords for a topic is straightforward, but may not yield useful results due to background noise on the communication medium. In other words, the terms that appear with high frequency overall need to be separated from those that have a uniquely high frequency within a specific topic's conversations. This paper lays the groundwork to investigate how background noise on Twitter changes over time, by topic, and by geography, and how it affects control set selection and trend identification within a topic.

There is no "ground truth" for true trending terms, since whether a term is trending is a somewhat subjective decision, and dependent on context. When people discuss trending terms, often what they are interested in is terms that are for some reason unique to that topic, time, and space. Of course, it is possible to simply count whether a term has increased in frequency recently, but what is driving that increase could be a seasonal pattern of behavior or simply not relevant to the query at hand. For instance, if the hashtag "#trafficjam" increases every day at 5 PM on a work day in Washington, D.C., then it may not be very informative to identify that hashtag as a trending topic. Alternatively, the term "#justinbieber" might appear frequently in a search on the topic of "school", but may not be relevant to the search at hand since "#justinbieber" appears frequently in many different topics on Twitter. As a result, taking into consideration the normal background conversations for a particular search term is necessary if "useful" trending topics are to be discovered, but since there is no clear definition of what "useful" is, this is still an ill-defined goal.

However, what can be investigated is how different methods of identifying trending terms / topics produce different results. We proceed by comparing lists of trending terms that result from varying the background noise, thereby taking into account different amounts of language change over time. By collecting tweets on a variety of topics, we find a relationship between the nature of the topic and how much the Twitter conversations surrounding it are related to background information, which affects trend identification.

To identify top keywords for a topic (the target set), we compare against control datasets (collections of pseudorandom tweets) on Twitter to filter out background noise. The control set is critical in defining the trends that will be identified, since it becomes the filter of identifying unusual terms. So, selecting the appropriate control dataset is important. Characteristics of the control set can vary by the time period, topic, and geographic location. The question is how these characteristics interrelate with the target set, and how this affects the terms

identified. We develop a framework for varying the target and control sets under these dimensions and present results in which we vary the control set by time.

This work was motivated by previous work on a visualization tool that we created known as Geo that could monitor keywords on Twitter using the streaming API and query keywords on demand using the REST API [1]. Geo used term frequency-inverse document frequency (*tf-idf*) to identify the top keywords for tweets related to the topic. The document corpus that we used for the *tf-idf* calculation was a cumulative collection including results from previous target queries and a set of background queries. These background queries were Twitter searches for common English words such as "I", "and", and "the".[1]

In developing the tool, we came across the issue of determining the frequency with which we collected the control data. How often and by how much does background noise on Twitter change over time? We also wanted to determine the best method for collecting a control set for a given target set. For instance, should the control set be collected for the same time and geographic location as the target set? And does the topic that we are interested in affect these decisions? We realized that there was a more general question here about the best way to identify trending terms, and that any developer interested in creating a social media monitoring platform needs to address these questions. However, it was necessary to first examine how the control set chosen for comparison affected the resulting trending terms. In this paper we investigate how decisions about the control set affect the stability of the resulting top terms returned from *tf-idf* and provide practical recommendations for control set selection for the purposes of keyword identification.

II. BACKGROUND AND PREVIOUS LITERATURE

There is already existing literature [2], [3] that examines the list of trending topics as identified by Twitter. Twitter uses a proprietary algorithm to identify trending topics, but its API only queries based on geographic location and not subject, and does not provide historical data. As a result, additional work is necessary to create a trending topic tool that is specific to location, subject, and time.

Other work [4], [5] detects emerging trends on Twitter in real-time. Within real-time trend identification, there is work focusing on event identification [6]. We focus on any keywords within a topic as trends, which may or may not relate to real-world events. Rather than seek out emerging trends or events from a general Twitter dataset, our paper looks for current trends within a topic. Real-time detection only concerns identifying trends at a particular timepoint; we also use historical data as control sets, which allow us to study long-term/historical monitoring of keywords within a topic.

While we use *tf-idf* to identify top keywords, Mathioudakis and Koudas (2010) identify "bursty" keywords and Cataldi,

Di Caro, and Schifanella (2010) use an aging theory to detect keywords. Benhardus and Kalita evaluate different methods of trend detection on Twitter, including *tf-idf* (2013). The other methods for topic extraction can easily replace *tf-idf* within our framework as the function f, and regardless of the method used, there is a question of the best control set to use. Benhardus and Kalita (2013) use the Edinburgh Twitter corpus [8] as baseline data when detecting trends over streaming tweets. The bursty keyword detection algorithm used by Mathioudakis and Koudas only considers real-time data and does "no optimization over older data" (2010). Our framework allows us to investigate the effects of considering older versus newer baseline data in keyword detection.

Cataldi, Di Caro, and Schifanella introduce a concept of "history worthiness", experimenting with the number of historical time slots considered (2010). They find that the parameter directly affects the trends identified. We similarly vary our control sets over time but over a much longer timeframe. Our framework also considers geographic variation in the control set — Cataldi, Di Caro, and Schifanella do not filter information by geography since it helps them identify emerging trends of global import (2010). Much of the work in trend detection and their methods aim to identify trends in all of Twitter rather than just a specific topic. Control set selection becomes even more important when the control set is used to not only identify trends in a particular topic but also filter out the irrelevant trends on Twitter.

Naaman, Becker, and Gravano use a dataset of tweets by New York City users, looking at temporal trends as specific to a geographic area (2011). Our framework allows for other kinds of questions about the importance of geography in identifying trends, particularly with respect to control set selection. Wilkinson and Thelwall examine international differences in trending topics by country, though the goal of the research is applied and not methodological (2012). They use a time series analysis method to extract the top 50 trending topics for selected English-speaking countries and find that there are differences among countries. This supports the significance of considering geography in control set selection.

III. FRAMEWORK

We represent a particular trending terms query using the following functional description:

$$f(T|C) \tag{1}$$

where f represents the method used to generate the keyword lists for a target set T given a control set C.

We use *tf-idf* as described in the Data and Processing section, but another method could be used. In general, the goal is to see how the results of f change as T is held constant and C is varied, though we will also hold C constant and vary T in some of the work below. We investigate three axes along which T and C can vary: (1) time t, (2) subject s, and (3) geography g. This allows us to refine our representation above using the following form:

---

[1]We could not use the random sampling API call of Twitter because we had to be able to specify geography for some features of Geo, which the sprinkler did not allow at the time, and as of this writing still does not allow.

$$f(T_{t^*,s^*,g^*}|C_{t,s,g}) \qquad (2)$$

where T is the target dataset, such as tweets, collected at time t^* on a subject s^* for a geographic location g^*. For example, one could collect a set of tweets on Jan. 1, 2013, containing the keyword "love" and located in the city of Washington, D.C.

The control set C is the dataset serving as a background corpus for f, collected for some time t before time t^*, subject s, which may or may not be s^*, and geographic location g, which may or may not be g^*. Along the time dimension, t could, for instance, vary as a single week ($t = 0, 1, 2, \cdots$), a cumulative dataset starting from N weeks before time 0 ($t = $ 0–N), or a sliding window period of m weeks (e.g. for $m = 2, t = $ 0–2, $t = $ 1–3, $\cdots$). Possible values for the subject s could be $s = s^*$, the same subject used for T, or the superset of all subjects queried, $s = S$. We also collect a background dataset as a control, where the queries result in a pseudorandom collection of tweets not necessarily related to any of the search terms we explore. We denote this as $s = BG$. A possible value for the geographic location g could be $g = g^*$, the same geographic location as T. Since many tweets do not have any location information attached, we could also query these tweets, $g = \emptyset$. As a result, we will define $g = G$ to be the superset of all locations queried plus any tweets without geographic information.

Some research questions that this framework can support include the following:

How does varying the control set by time affect trend identification within a target set? In the Results section, we compare $f(T_{0,s^*,G}|C_{t',BG,G})$ for different variations on t' to the baseline $f(T_{0,s^*,G}|C_{0,BG,G})$.

How do conversations about a topic change over time? We can compare $f(T_{t',s^*,G}|C_{t^*,BG,G})$ for different values of t' to the baseline $f(T_{t^*,s^*,G}|C_{t^*,BG,G})$ to investigate vocabulary change in selected topics over time.

How similar are the trends for two related topics? For example, let the first topic be "economy" and the second topic be "trade". We can examine the similarity between $f(T_{t^*,economy,G}|C_{t^*,BG,G})$ and $f(T_{t^*,trade,G}|C_{t^*,BG,G})$ to study the question at a global level. We can also ask how similar the trends for two topics are in a more specific geographic area such as a particular city. In that case, we simply use g^* instead of G for the target sets and control sets.

How similar are the trends for a topic in two geographic areas? As an example, how do Twitter conversations about the economy differ in the U.S. versus in China? Results for $f(T_{t^*,economy,U.S.}|C_{t^*,BG,U.S.})$ and $f(T_{t^*,economy,China}|C_{t^*,BG,China})$ provide a basis for comparison.

Is a particular trend in a geographic area local or global? Comparing $f(T_{t^*,s^*,g^*}|C_{t^*,BG,G})$ to $f(T_{t^*,s^*,G}|C_{t^*,BG,G})$ would allow us to see if local and global conversations about a particular topic are similar.

How similar is background noise at the local and global levels? We can compare $f(T_{t^*,s^*,g^*}|C_{t^*,BG,g^*})$ to $f(T_{t^*,s^*,g^*}|C_{t^*,BG,G})$ to see whether local and global background noise are similar enough to identify the same trends in the target set. Answering this question addresses whether one can simply use a global control set in place of a geographically specific one, when the target set is over a particular geographic area. Being able to use the global control set would allow for reusing the same control set when making queries over different geographic areas.

Though we have presented ways to collect data based on time, geography, and subject, in this paper we will primarily explore the effect of changing only time. We will always be exploring $C_{t,BG,G}$, where the subject for the control set is the background tweets and the geography is the union of all tweets with or without geographic data. As a result, for this paper we will simplify our notation, using C_0 to denote $C_{0,BG,G}$ and C_1 to denote $C_{1,BG,G}$ and so forth for all of the variations discussed above.

So now that we have described how to potentially generate different lists of trending terms, we must also discuss how to compare and contrast the results of these functions, i.e., there must be a method for comparing outputs of f and measuring similarity. In our case, we use rank correlation to measure similarity among the ranked keyword lists outputted by *tf-idf* for various sets of T and C. In order to hold the comparison constant across multiple different values of C, we establish a baseline control set, which is C_0, the underlying background noise at the same time as T_0 for all geographic locations. Establishing this control set as the standard answers the question of how results vary compared to using the most current, most general control set. In other words, we seek to compare the keyword rankings from using various control sets to the baseline rankings from using the background noise at time 0.

IV. DATA AND PROCESSING

A. Twitter Data Collection

We selected 11 broad topics in which to identify trends over time:

world economy	baseball	london riots
economy	jobs	government
fun	music	global warming
love	war	

Our background queries were "I","the","and","a", and "of", which yielded pseudorandom samples of English tweets to represent background noise on Twitter. We use these content-free stopwords rather than everyday words like "book" or "home", because the latter introduce content and biases, and as such filter out tweets that are topic-specific rather than part of the background. Ideally, we would use a completely random set of tweets, but currently there is no way through the Twitter API to collect tweets from a particular geography without specifying search terms (such as the sprinkler for global tweets), and so specifically using words that are not

| Jobs	Government	War	Love	Global Warming
job	romney	obama	much	halt
romney	jobs	romney	3	obama
create	create	president	one	ryan
hiring	obama	end	up	paul
new	people	world	know	definitely
manager	federal	barackobama	life	giving
obama	via	iraq	people	disappointed
london	more	peace	more	punish
steve	debate	women	someone	keys
plan	mitt	afghanistan	happy	uterus

Fig. 1. Rank correlation results, including the control set, over the full vocabulary when using a single-week control set varied by time. The axes are scaled to show more detail.

normally processed in language analysis, i.e., stopwords, was the best solution.

For geographic analysis, we selected 9 English-speaking cities:

Boston, MA	Los Angeles, CA	London, England
Chicago, IL	Houston, TX	New York City, NY
Washington, D.C.	Toronto, Canada	Sydney, Australia

The data consist of daily queries to Twitter's REST API at the same time each day for a time period of 47 weeks ending in October 2012, except for a server outage during one week ($t = 39$). We collected samples for the 11 topics and 5 control queries, each with a geography-unspecified query and all 9 city-specific queries.

B. Data Processing

Tf-idf is an information retrieval mechanism to identify the most common terms in a document relative to their frequency in the overall document corpus. Based on our previous work with Twitter keyword ranking [11], we group a set of tweets together as a document to represent conversations about the topic. For example, all tweets from a week's queries for "music" are grouped together as a "document" to form our target set T. We use the documents formed by control set queries as the rest of the document corpus.

The tweets for time 0, the last week of collection, were grouped by topic as individual documents. Each of these documents serves as a target set T for which we vary C by time. Essentially, the *tf* calculated from T is held constant, while the *idf* calculated from $T \cup C$ varies with C, which changes the *tf-idf* values of the trending terms. Since the topics we search for are commonly discussed on Twitter, we add 1 to the *idf* so that words with $idf = 0$, those that occur in all documents of the corpus, will not have a *tf-idf* value of 0 that puts them at the bottom of rankings.

Based on the *tf-idf* values, we get ranked keywords over the entire vocabulary. For selected topics, top keywords as identified by *tf-idf* using the baseline control set C_0 are listed in Table I.[2]

---

[2]Stopwords such as "http", "rt", "don't", and "didn't" were filtered out.
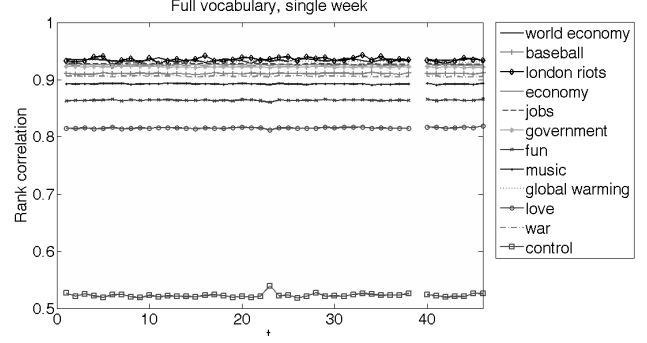
To generate results for control set variation, we first calculate the *idf* values for different C, using tweets from the control queries. We then calculate the *tf-idf* values for T given each C for each topic by determining term frequencies from the target queries and multiplying them by $1 + idf$ values for each C. The *tf-idf* values determine the keyword rankings within a topic. Finally we use the Kendall tau rank correlation coefficient to compare the ranked lists from each C to those of the standard C. We calculated the rank correlations for the entire vocabulary of T (full-vocabulary) and for the top 50 keywords (top-50). This allows us to look at how background noise variations affect rankings over the complete vocabulary as well as those of just the top keywords.

V. RESULTS

For the single week control, we compare the rankings from each of $f(T|C_1), f(T|C_2), \cdots, f(T|C_{46})$ to the baseline rankings from $f(T|C_0)$. Fig. 1 shows the full-vocabulary rank correlation results from using a single week control set. Overall the correlations for each topic are high and remain steady across weeks. The time of the control set does not significantly affect keyword rankings, but the topic determines how much using an old control set matters. For instance, the correlations for "love" are lower than those for "world economy", so using a control set other than C_0 affects rankings over the entire vocabulary for "love" more than those for "world economy".

The correlations for more subjective topics, specifically "love", "fun", and "music", are lower than for the other topics. This observation suggests that the overall vocabulary about more subjective topics is affected more by temporal changes in the control data (e.g. from week 0 to week 1) than the vocabulary about other topics. This seems to indicate that the words associated with more specific topics are fairly distinct from the background noise, while words associated with more subjective topics are more related to the background noise, so altering the background noise affects these topics. It is important to note that we are not actually changing the target set of tweets for each topic, which remains constant.

For all topics, the correlation values stay relatively constant across weeks, which reflect the lack of vocabulary variation in the underlying control sets. The differences from using differently aged control sets are not significant — e.g. the correlation between using week 0 and week 1 is about the same as the correlation between using week 0 and week 46.

However, we are more interested in how the top keyword rankings change with control set variations, as our original goal was to identify top trends within a topic. When we look only at the top 50 keywords, we discover that there is substantial variation in the week-to-week patterns of the keywords related to their variance. To identify topics with rankings that change more or less, we split the topics into high and low standard deviation groups using the median standard deviation of the rank correlation values. Fig. 2 presents the top-50 rank correlation results for each topic, split into high and low standard deviation groups.

The high correlations for all topics indicate that the top keywords are fairly consistent across using differently aged control sets. The degree of consistency seems related to the nature of the topic. Subjective topics like "love", "fun", and "music" are in the low standard deviation group. Though their overall vocabularies are more sensitive to changes in time in the control set, their top 50 keywords are more consistent. Topics with overall vocabularies changing the least with control set changes ("global warming", "london riots", and "world economy") have higher standard deviations for top-50.

As the background noise varies, the correlation changes reflect the relationship between top keywords and background noise. We interpret that less variation in the top 50 keywords means the topic's top keywords are sometimes related to background discussions and sometimes not. Thus, conversations about "global warming" are generally distinct from the background noise in the full-vocabulary results, while their top keywords are sometimes related to the background. The full vocabulary of "fun" is more similar to the background noise, and its top keywords are more stable.

VI. DISCUSSION, CONCLUSIONS, AND FUTURE WORK

We established a framework to investigate how control set selection affects trend identification in high-volume, multi-subject, geographically dispersed social media platforms, and explored this framework on Twitter. Using this framework, we analyzed results for varying control sets by time using control sets composed of a single week of tweets. Based on our research so far, we can draw a few conclusions about trend identification and background noise on Twitter.

First, in the control set variation, the top keywords for a topic as identified by *tf-idf* are fairly consistent, regardless of the age of the control set used. The term frequencies for the top keywords are so high that fluctuations in the background noise (corresponding to small changes in the *idf*) do not significantly affect keyword rankings.

Second, when using a single week control set, the differences between using an old control and an older control set are insignificant. In other words, the correlation between C_t and C_{t+h} remains relatively constant regardless of the value of h (time-independent) and only seems to depend on the topic. These first two findings indicate that if you are using a single week of data as control information for a social media monitoring tool, then it is not necessary to update this extremely frequently, i.e., more than once a year or so.

A remaining question for future work is whether the background noise on Twitter on a week-to-week basis might resemble a stationary process — whether the correlation between C_t and C_{t+h} is constant regardless of the value of t, in addition to the value of h. We hope to test this by using different starting time points within our framework and seeing if, for instance, the correlation between C_0 and C_1 is roughly the same as the correlation between C_1 and C_2.

Third, we identified a relationship between the nature of the topic of interest and how its vocabulary and keyword rankings change as background noise changes over time. For more subjective topics, the rankings over the entire vocabulary are more affected by time variations in the control set, which seems to suggest that the vocabulary related to these topics is more similar to the background noise. However, the top keyword rankings are more stable for these topics. The topics with less change in overall vocabulary rankings when the control set varies then exhibit more change in their top keyword rankings. Thus, the abstractness of a topic seems related to how much its vocabulary and top keywords are related to background noise variations. We believe that this represents how similar or distinct the topic's conversations are to background conversations on Twitter. As a result, social media modeling tools should take into account the subjective nature of a term before determining an appropriate trending topics identifier. Since we have shown that the subjective nature is readily apparent when observing the dynamics of the rank correlations of the keywords, this could be done on a dynamic basis as these correlations change over time.

These observations affect how topics queried should be selected in addition to control set selection. For example, there is a difference between tracking the "economy" and "world economy" topics because the latter is more specific than the former. The more specific topic will likely have an overall vocabulary that is more distinct from the background vocabulary, and its top keywords will vary more as the background noise changes, reflecting trends that enter and exit background discussions. Another scenario would be a brand manager selecting keywords related to a product — how one expects the vocabulary to change or relate to background conversations for "music" versus "rock music" versus a particular band name. These last two findings imply that it is important to take subject matter into account when designing a trend identification algorithm.

In future work we plan to explore varying the control set within our framework by examining subject and geography, in addition to time. We will also try other methods of comparing keyword lists besides rank correlation, such as mean average precision. We are also interested in whether our results are spe-
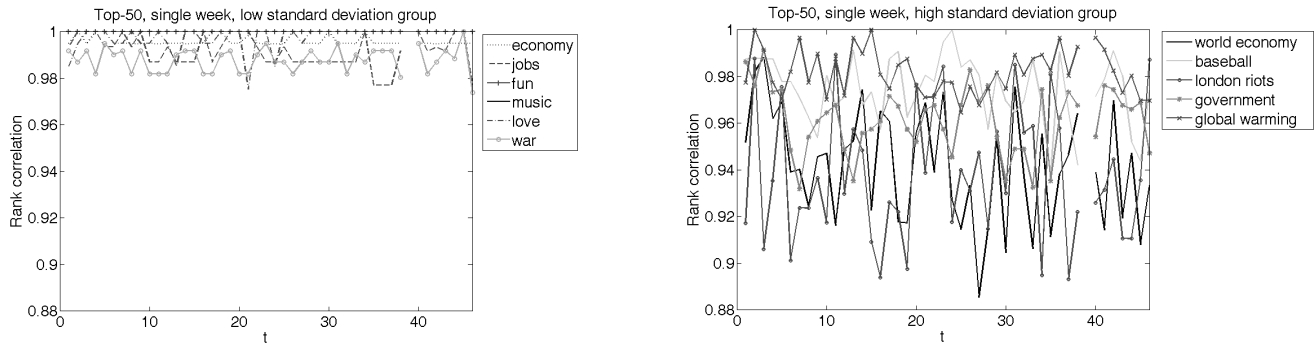
Fig. 2. Top-50 rank correlation results with single week control sets for (a) low standard deviation group and (b) high standard deviation group. The axes are scaled to show more detail.

cific to Twitter or also hold for other social media datasets or natural language datasets in general. To investigate this further, we plan to apply our methodology to blogging data previously obtained from the Spinn3r web service (http://spinn3r.com).

REFERENCES

[1] V. Lai, K. Prasad, and W. Rand, "Geographic social media visualization," University of Maryland, College Park, Tech. Rep., 2011.

[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772751

[3] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in social media: Persistence and decay," in *5th International AAAI Conference on Weblogs and Social Media*, February 2011.

[4] M. Mathioudakis and N. Koudas, "TwitterMonitor: trend detection over the Twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, ser. SIGMOD '10. New York, NY, USA: ACM, 2010, pp. 1155–1158. [Online]. Available: http://doi.acm.org/10.1145/1807167.1807306

[5] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, ser. MDMKDD '10. New York, NY, USA: ACM, 2010, pp. 4:1–4:10. [Online]. Available: http://doi.acm.org/10.1145/1814245.1814249

[6] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on Twitter," in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011. [Online]. Available: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html#BeckerNG11

[7] J. Benhardus and J. Kalita, "Streaming trend detection in Twitter," *International Journal of Web Based Communities*, vol. 9, no. 1, pp. 122–139, January 2013.

[8] S. Petrović, M. Osborne, and V. Lavrenko, "The Edinburgh Twitter corpus," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, ser. WSA '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 25–26. [Online]. Available: http://dl.acm.org/citation.cfm?id=1860667.1860680

[9] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on Twitter," *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 5, pp. 902–918, May 2011. [Online]. Available: http://dx.doi.org/10.1002/asi.21489

[10] D. Wilkinson and M. Thelwall, "Trending Twitter topics in English: An international comparison," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 8, pp. 1631–1646, Aug. 2012. [Online]. Available: http://dx.doi.org/10.1002/asi.22713

[11] V. Lai, C. Rajashekar, and W. Rand, "Comparing social tags to microblogs," in *Second International Workshop on Modeling Social Media*, July 2011.