Streaming trend detection in Twitter

James Benhardus*

Centre for Cognitive Science,
University of Minnesota,
Minneapolis, MN 55455, USA
E-mail: benha015@umn.edu
*Corresponding author

Jugal Kalita

Department of Computer Science,
University of Colorado,
Colorado Springs, CO 80918, USA
E-mail: kalita@eas.uccs.edu

Abstract: As social media continue to grow, the zeitgeist of society is increasingly found not in the headlines of traditional media institutions, but in the activity of ordinary individuals. The identification of trending topics utilises social media (such as Twitter) to provide an overview of the topics and issues that are currently popular within the online community. In this paper, we outline methodologies of detecting and identifying trending topics from streaming data. Data from Twitter's streaming API was collected and put into documents of equal duration using data collection procedures that allow for analysis over multiple timespans, including those not currently associated with Twitter-identified trending topics. Term frequency-inverse document frequency analysis and relative normalised term frequency analysis were performed on the documents to identify the trending topics. Relative normalised term frequency analysis identified unigrams, bigrams, and trigrams as trending topics, while term frequency-inverse document frequency analysis identified unigrams as trending topics. Application of these methodologies to streaming data resulted in F-measures ranging from 0.1468 to 0.7508.

Keywords: microblogs; trend detection; tf-idf; human language processing.

Reference to this paper should be made as follows: Benhardus, J. and Kalita, J. (2013) 'Streaming trend detection in Twitter', *Int. J. Web Based Communities*, Vol. 9, No. 1, pp.122–139.

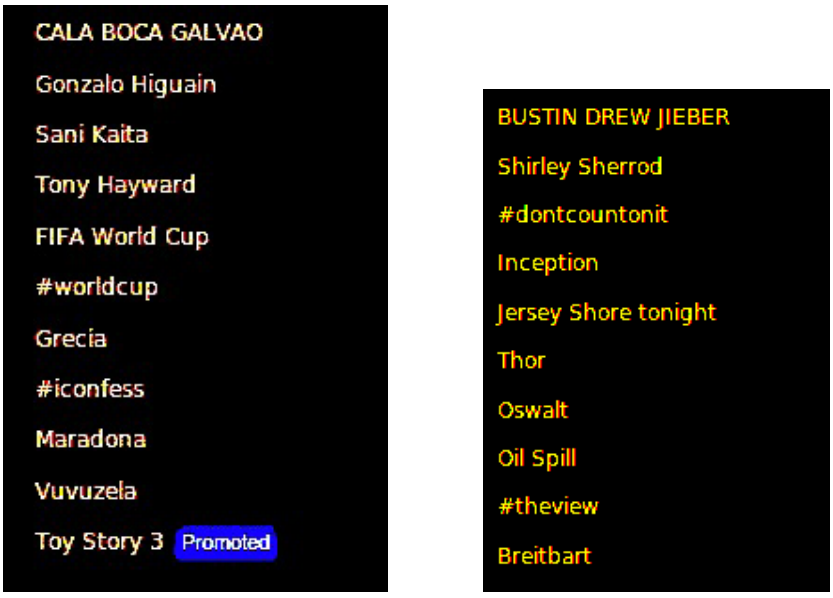
Biographical notes: James Benhardus is a graduate student at the University of Minnesota Center for Cognitive Science.

Jugal Kalita is a Professor in the Department of Computer Science at the University of Colorado, Colorado Springs.

1 Introduction

Twitter is a popular microblogging and social networking service that presents many opportunities for research in natural language processing (NLP) and machine learning. Since its inception in 2006, Twitter has grown to the point where http://Twitter.com is the 9th most visited website in the world, and the 8th most visited site in the USA[1]. Users of Twitter post short (less than or equal to 140 character) messages, called 'tweets', on a variety of topics, ranging from news events and pop culture, to mundane daily events and spam postings. As of September 2011, Twitter had over 100 million active users[2] producing over 200 million tweets per day, an average of over 2,300 tweets per second[3].

Figure 1 (a) A list of trending topics as identified by Twitter from 17 June 2010 (b) A list of trending topics as identified by Twitter from 29 July 2010 (see online version for colours)



(a) (b)

Source: http://Twitter.com

Twitter presents some intriguing opportunities for applications of NLP and machine learning. One such aspect of Twitter that provides opportunities is trending topics – words and phrases, highlighted on the main page of Twitter, that are currently popular in users' tweets. Trending topics are identified for the past hour, day and week. Examples of trending topics can be seen in Figure 1. Trending topics are supposed to represent the popular 'topics of conversation', so to speak, among the users of Twitter. Determining trending topics can be considered a type of first story detection (FSD), a subset of the larger problem known as topic detection and tracking (TDT) (Allan et al. 1998). The popularity and growth of Twitter presents some challenges for applications of NLP and machine learning, however. The length restrictions of the messages create syntactic and structural conventions that are not seen in more traditional corpora, and the size of the Twitter network produces a continuously changing, dynamic corpus. In

addition, there is quite a lot of content on Twitter that would be classified as unimportant to an outside observer, consisting of personal information or spam, which must be filtered out in order to accurately identify the elements of the corpus that are relevant to the Twitter community as a whole, and could thus be considered to be potential trending topics. The challenge of Twitter's popularity is that in order to detect and identify trending topics, one must sample and analyse a large volume of streaming data. This paper will propose methods of using NLP techniques on streaming data from Twitter to identify trending topics.

2 Related work

While there is a very large body of work pertaining to all aspects of NLP, applying NLP techniques to Twitter is a fairly recent development, due in part to the fact that Twitter has only been in existence since 2006[4]. In this relatively short span of time, however, there have been many insightful analyses of Twitter. In particular, there are many recent applications NLP techniques to Twitter.

- Twitter has been used to study the dynamics of social networks, particularly the temporal behaviour of social networks (Perera et al., 2010), or the behaviour of social networks during disasters, such as earthquakes (Kireyev et al., 2009; Sakaki et al., 2010).
- FSD has been applied to Twitter to identify the first tweets to mention a particular event (Petrovic et al., 2010a).
- Data mining from trending topics have also been applied to Twitter to summarise trending topics (Sharifi et al., 2010a, 2010b, 2010c; Inouye and Kalita, 2011) and to analyse how trending topics change over time (Cheong and Lee, 2009).
- Normalising unorthodox spelling and syntax so that tools developed for NLP can be used felicitiously has been the topic of research in Kaufmann and Kalita (2011), Lui et al. (2011) and Han and Baldwin (2011).
- Part-of-speech (POS) tagging (Gimpel et al., 2010), named entity recognition (Liu et al., 2011; Finin et al., 2010) and a host of other topics also have been lately explored.
- In addition to applications of NLP techniques to Twitter, trend and event detection techniques have also been applied to other online entities such as weblogs (Glance et al., 2004; Gruhl et al., 2004), news stories (Allan et al., 1998; Nallapati et al., 2004; Yang et al., 1998), or scientific journal collections (Shaparenko et al., 2005; Wilbur and Yang 1996).

3 Problem definition

The main goal of this project is to detect and identify trending topics from streaming Twitter data. To accurately define the problem, the first step must be to define explicitly what constitutes a trending topic. In Gruhl et al. (2004), topics are defined as consisting of a combination of *chatter*, which is characterised by persistent

discussion at a constant level and is largely user-initiated, and *spikes*, which are characterised by short-term, high intensity discussion that is often in response to a recent event. In general, trending topics consist mainly of spikes. However, trending topics can also consist of a fairly even combination of spikes and chatter, or of mainly chatter. Examples from Figure 1 of trending topics that could be considered to consist mainly of spikes are:

- CALA BOCA GALVAO
- Gonzalo Higuain
- Sani Kaita
- FIFA World Cup
- #worldcup
- Grecia
- Maradona
- Vuvuzela
- Oil spill
- Tony Hayward
- Oswalt
- Shirley Sherrod
- Breitbart.

Examples from Figure 1 of trending topics that could be considered to be a fairly even combination of spikes and chatter are:

- #theview
- Jersey Shore tonight
- Thor.

Examples from Figure 1 of trending topics that could be considered to consist mainly of chatter are:

- Inception
- #iconfess
- #dontcountonit
- BUSTIN DREW JIEBER.

The '#' symbol at the beginning of '#worldcup', '#iconfess', '#theview', and '#dontcountonit' is called a 'hashtag', and is used by Twitter users to classify their tweets as pertaining to a particular topic. In addition to spikes and chatter, a trending topic can also be the result of advertisement, as is the case for the final trending topic in Figure 1, 'Toy Story 3'. In this third possibility, the trending topic is associated with a 'promoted'

tweet – a hybrid tweet-advertisement which is displayed at the top of search results on relevant topics[5].

While the classification of a trending topic as consisting of spikes or chatter is helpful for the understanding of the nature of trending topics, it is not directly useful in the identification or classification of terms as trending topics. Our working definition of a trending topic shall be a word or phrase that is experiencing an increase in usage, both in relation to its long-term usage and in relation to the usage of other words. More technical definitions of trending topics shall be used in the actual experiments, and shall be described in the 'methodologies' section.

In addition to defining what constitutes a trending topic, we must also define what constitutes success for a particular methodology. As the goal of the project is to develop a method of identifying trending topics that is independent of the method used by Twitter, simple agreement with the Twitter-identified trending topics is both unambitious and potentially unrealistic without replicating Twitter's methodology, which happens to be proprietary. As such, we shall define a successful method as a method that produces relevant topics at a rate of at least 75% of the rate returned by Twitter's method, with an agreement of at least 50% with the terms produced by Twitter's method. These criteria, while arbitrarily assigned, represent what we considered to be reasonably attainable levels of performance given what has been achieved by similar researchers before. The details of computing relevance and agreement shall be discussed in the 'evaluation measures' section.

4 Methodologies

Multiple methodologies were implemented, making use of one or more selection criteria. Each selection criterion will be discussed in its own subsection. All methods implemented made use both of the Twitter Streaming API[6] and the Edinburgh Twitter corpus (Petrovic et al., 2010b), a collection of approximately 97 million tweets collected between November 2009 and February 2010. The Edinburgh Twitter corpus was used to provide baseline measurement against the data from the Twitter Streaming API. For each source, tweets were temporally grouped into 'bag of words' style collections, or 'documents'. These documents were be normalised by duration, meaning that each document corresponds to the tweets posted in a certain constant length of time. The Edinburgh Twitter corpus was divided into 1,212 sections, each consisting of one hour's worth of tweets. The tweets from the Twitter Streaming API were grouped into sections corresponding to either 10 minutes, 15 minutes, or one hour's worth of data collection.

4.1 Frequency

The first criterion used was simply the raw frequency of each term. This criterion was used mainly as a threshold criterion, to be used with one or more of the other criteria. Using raw frequency by itself has major drawbacks, as the most frequent terms in the stream are the terms that carry the least information, such as 'the', 'and', or 'rt' (an abbreviation for 'retweet', a term used when one Tiwtter user reposts another Twitter user's tweet). The majority of the most common words can be classified as stop words, and filtered out of the stream. Generation of a stop word list shall be discussed in the 'experiments' section.

4.2 TF-IDF

The second criterion implemented involved analysing each document using an application of tf-idf weighting. Tf-idf weighting is an information retrieval technique that weights a document's relevance to a query based on a composite of the query's *term frequency* and *inverse document frequency* (Salton and Buckley, 1988). Term frequency can be defined as either

$$tf_{i,j} = n_{i,j}$$

or

$$f_{i,j} = \frac{n_{i,j}}{N}$$

where $n_{i,j}$ is the number of times word i occurs in document j and

$$N = \sum_k n_{k,j}$$

is the total number of words in document j. The second definition of $tf_{i,j}$ is often referred to as the *normalised term frequency*. Inverse document frequency is defined as

$$idf_i = \log\left(\frac{D}{d_i}\right)$$

where d_i is the number of documents that contain word i and D is the total number of documents. Put simply, the weight of a document will be higher if the number of times a word occurs in a document is higher, or if the number of documents containing that word is lower; similarly, the weight of a document will be lower if the number of times a word occurs in a document is lower, or if the number of documents containing that word is higher (Hiemstra, 2000).

4.3 Normalised term frequency

The third criterion implemented involved utilising only the term frequency of each element, rather than both the term frequency and the inverse document frequency. For this method, a *normalised term frequency* was used, defined as

$$tf_{norm_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} * 10^6$$

where n_i is the number of times word i occurs in document j and $\sum_k n_{k,j}$ is the total number of words in document j. Due to the large number of words found in the documents, a scaling factor of 10^6 was used, meaning $tf_{norm_{i,j}}$ can be thought in terms of frequency per million words. Each word in the test document was given a *trending score*, defined as

$$ts_{i,j} = \frac{tf_{norm_{i,j}}}{atf_{norm_{i,s}}}$$

in which

$$atf_{norm_{i,s}} = \sum_{S=\{s_1,\ldots,s_p\}} \frac{tf_{norm_{i,s_k}}}{p}$$

where S is the set of p baseline documents to which the test document was compared.

4.4 Entropy

The fourth criterion implemented was entropy. To calculate the entropy of a term, all of the tweets containing that term are collected. As it is used in this project, the entropy of a term i is defined as

$$H_i = -\sum_j \frac{n_{j,i}}{N} \log\left(\frac{n_{j,i}}{N}\right)$$

where $n_{j,i}$ is the number of times word j occurs in the collection of tweets containing term i and

$$N = -\sum_j n_{j,i}$$

is the total number of words in the collection of tweets containing term i. Entropy proved to be a helpful parameter to use in filtering out terms that could be classified as spam.

5 Experiments

Two experiments were run, implementing slightly different methodologies, but following the same general format. Unless stated otherwise, the process described was used for both experiments.

5.1 Data collection

Data was collected using the Twitter streaming API, with the gardenhose tweet stream providing the input data and the trends/location stream providing the list of terms identified by Twitter as trending topics. The gardenhose streaming API is a limited stream that returns approximately 15% of all Twitter activity[7]. The trends/location stream provides a list of trending topics that is specific to a particular location. The USA was used as the location for evaluation, as both experimental methods worked almost entirely with English tweets, and most of the trending topics from the USA were in English, leading to a more accurate basis for comparison than trending topics from the entire world. The streaming data was collected automatically using the cURL data transfer tool within a shell script. The streaming data was grouped into documents of equal duration.

In order to explore the effect (if any) that different collection times would have on the performance of the methods, the collection times were varied across the experiments, but kept constant within experiments. The first experiment independently evaluated groups of documents consisting of tweets collected over ten minutes and groups of documents consisting of tweets collected over one hour of streaming. The second experiment evaluated groups of documents consisting of tweets collected over 15 minutes of streaming.

5.2 Preprocessing

The data was collected from the Twitter streaming API in JSON format and a parser was used to extract the tweets from the other information returned. Next the tweets were preprocessed to remove URLs, unicode characters, usernames, and punctuation. A dictionary containing approximately 180,000 common English words and acronyms was used to filter out tweets that did not contain at least 60% English words. It was found that setting the threshold for English tweets at 60% English words served to filter out the majority of non-English tweets while maintaining a sufficiently large sample size. Tweets were classified as spam and discarded if one word accounted for over half of the words in the tweet. After preprocessing, tweets were stored in two ways – in a collection in which each valid tweet was left intact, and in a 'bag of words' style dictionary consisting of a unigram and the frequency of the unigram in the document.

5.3 Baseline data

Baseline data was computed from the Edinburgh Twitter Corpus, a collection of over 97 million tweets collected over three months in late 2009 and early 2010. The corpus was divided into 1,212 sections corresponding to one hour's worth of tweets, consisting of two bag-of-words dictionaries for each section – one containing unigrams and one containing bigrams. For the first experiment, the resulting documents were used independently of one another. For the second experiment, the documents were compiled into a comprehensive dictionary of 805,025 words with term frequency, document frequency, and tf-idf weights computed for each word. For the first experiment, a specified number of baseline documents were used to compute average normalised term frequency. For the second experiment, the dictionary was used to provide document frequencies for terms and for the generation of a stop word list.

5.4 Stop words

For each experiment, a list of stop words was used as an additional filter after preprocessing. A stop word is defined as a word that contains no meaning or relevance in and of itself, or a word that adds to the relevance of a result to a query no more often than would a random word (Wilbur and Sirotkin, 1991). The purpose in preprocessing stop words was not to remove every single word that carried little or no meaning, but rather to determine a threshold that would remove the majority of such words from the documents.

For the first experiment, stop words were identified using a 'lossy counting' algorithm (Manku and Motwani, 2002). The lossy counting algorithm identified the most frequent words in each of the 1,212 baseline documents. All words that appeared as the

most frequent in at least 75% of the baseline documents were classified as stop words. If a word in the test data was identified as a stop word, it was immediately removed from consideration as a potential trending topic.

For the second experiment, the following criteria were found to effectively identify stop words:

- If the word appeared in over half of the 1,212 documents. Such widespread usage suggests that such a word contains little or no semantic value in relation to the tweet as a whole.
- If the word had a total frequency of at least 3,000 throughout all 1,212 documents. Gross frequency was used in addition to document frequency in order to identify words that did not appear in over half of the documents, yet still had a high frequency of occurrence across the corpus as a whole.
- If the word was classified grammatically as a preposition or a conjunction.
- If the word was a derivative of a word that that was classified as a stop word (i.e., 'can' occurs in all 1,212 documents, so it is classified as a stop word. Derivatives of 'can', such as 'can't', 'could', 'couldn't', and 'could've' carry the same amount of meaning as 'can' in relation to a sentence as a whole, and thus are also classified as stop words).

A word was considered to be a stop word if it matched one or more of the criteria given above. As with the first experiment, if a word in the test data was identified as a stop word, it was immediately removed from consideration as a potential trending topic.

5.5 Selection criteria

For the first experiment, a combination of raw frequency and relative normalised term frequency was used. The raw frequency was used as a threshold, eliminating all terms that did not occur an average of at least one time for every minute of data collection. Normalised term frequency and average normalised term frequency were calculated for each remaining term, and the terms with the highest trending scores were identified as trending topics. Analysis was performed for both unigrams and bigrams. Entropy was also calculated for both unigrams and bigrams, but was not used as a selection criterion for this experiment.

The second experiment utilised a combination of raw frequency, tf-idf weighting, and entropy to identify trending topics. Once again, the raw frequency was used as a threshold, eliminating all terms that did not occur an average of at least one time for every minute of data collection. Term frequency-inverse document frequency weights were calculated for the remaining terms. Of the remaining terms, those with a tf-idf weight below a threshold value (set at five greater than the length of data collection in minutes so as to ensure that the term was not simply novel, but popular) were removed from consideration. Terms with an entropy of less than 3.0 were removed, as such terms were likely to be found mainly in a form of spam tweets consisting of only one or a few words repeated. The remaining terms were identified as trending topics.

6 Evaluation measures

6.1 Experiment 1

The first experiment was evaluated using precision, recall, and F-measure scores in comparison to the trending topics identified by Twitter. All three measures require calculating the number of true positives – that is, the items that were identified as trending topics both by the experimental method and Twitter's method. In addition, determining precision requires calculating the number of false positives – the items identified as trending topics by the experimental method that were not identified as trending topics by Twitter, and determining recall requires calculating the number of false negatives – the items identified as trending topics by Twitter that were not identified as trending topics by the experimental method. Precision is defined as

$$P = \frac{TP}{TP + FP}$$

where TP is the number of true positives and FP is the number of false positives. Recall is defined as

$$R = \frac{TP}{TP + FN}$$

where TP is the number of true positives and FN is the number of false negatives. The F-measure is the harmonic mean of the precision and recall, defined as

$$F = 2 \cdot \frac{P \cdot R}{P + R}.$$

6.2 Experiment 2

The second experiment was evaluated using recall and relevancy scores. Recall was calculated in comparison to the trending topics identified by Twitter using two different methods of identifying true positives and false negatives. The first method only identified as true positives terms that exactly matched terms identified by Twitter as trending topics. Since the second experiment returned only bigrams, terms identified by Twitter as trending topics that were not identified by the experimental method were only considered to be false negatives if they were unigrams. The second method identified a term as a true positive if it either exactly matched a term identified by Twitter as a trending topic or if it matched one part of a multigram trending topic. Any term identified as a trending topic by Twitter that was not identified as a trending topic by the experimental method was classified as a false negative. Relevance was calculated based on the evaluations of human volunteers. Volunteers were given a list of terms identified as trending topics and marked those that they felt were valid or relevant topics. Volunteers also were shown all the tweets. The list contained both terms identified as trending topics by the experimental method and terms identified as trending topics by Twitter as a control. Relevance was calculated in the same manner as precision was calculated in the first experiment.

6.3 Human volunteers

The group of human volunteers who evaluated the relevance of the potential trending topics in the second experiment consisted of ten undergraduate students who were all currently working on similar projects in the fields of NLP and machine learning, though none of the projects were directly related to finding trending topics in Twitter. The group of volunteers consisted of eight males and two females, ranging in age from 19 to 30. All the volunteers had at least two years of undergraduate education in the fields of mathematics, computer science, electrical engineering, and/or physics, although their knowledge of said fields was not utilised in the performance of their tasks.

7 Results

For the first experiment, the hourly datasets had an average precision of 0.2167 and 0.1984 and an average recall of 0.2168 and 0.3623 for an F-measure of 0.2167 and 0.2564 for unigrams and bigrams, respectively. The ten minute datasets had an average precision of 0.3389 and 0.1212 and an average recall of 0.3167 and 0.1862 for an F-measure of 0.3274 and 0.1468 for unigrams and bigrams, respectively. The results of the first experiment can be seen in Table 1 and a graph showing the precision and recall scores for each dataset is shown in Figure 2. An initial estimate of reasonable expected performance for this experiment was a precision score of at least 0.50 and a recall score of at least 0.75, for an F-measure of at least 0.60. The initial results are well below this, but within reasonable range the results of similar work, which produced F-measures in the range of 0.30 to 0.60 (Allan et al., 1998; Yang et al., 1998).

For the second experiment, initial results gave an average precision of 0.2780 and an average recall of 0.7667 for an F-measure of 0.3988 as calculated by the first method of evaluation, and an average precision of 0.4075 and an average recall of 0.5985 for an F-measure of 0.4794 as calculated by the second method of evaluation. A table of the results of the first experiment can be seen in Table 2 and a graph showing the precision and recall scores for each dataset is shown in Figure 3. The initial results were evaluated by human volunteers as containing relevant topics 72.43% of the time, compared to 77.14% of the time for the terms identified by Twitter as trending topics. Substituting relevance scores for precision scores produces an F-measure of 0.7508 as evaluated by the first method of evaluation and F-measure of 0.66 as evaluated by the second method of evaluation. Given that the success criteria were a recall of 0.50 when evaluated with the terms identified by Twitter and a relevance of at least 75% that of the terms identified by Twitter, the data from the second experiment meets the conditions of success.

Comparing our results with Twitter's trending topics may not be the best way to evaluate our algorithms since Twitter's trending topics are produced by an unpublished algorithm. However, we feel that our results with respect to human evaluators are quite respectable. We believe that algorithms like the ones discussed in this paper can be used by those monitoring the Twitter feed to obtain a list of topics that are being highly discussed or are trendy.

Table 1 Precision, recall, and F-measure scores for both unigrams and bigrams from analysis of datasets consisting of six one-hour segments and six ten-minute segments of tweets from the Twitter streaming API

Data set	*TP*	*FP*	*FN*	*Precision*	*Recall*	*F-measure*
Hourly Unigrams 1	8	22	24	0.2667	0.2500	0.2581
Hourly Unigrams 2	9	21	20	0.3000	0.3103	0.3051
Hourly Unigrams 3	4	26	27	0.1333	0.1290	0.1311
Hourly Unigrams 4	7	23	25	0.2333	0.2188	0.2258
Hourly Unigrams 5	6	24	22	0.2000	0.2143	0.2069
Hourly Unigrams 6	5	25	23	0.1667	0.1786	0.1724
Average				*0.2167*	*0.2168*	*0.2166*
Hourly Bigrams 1	4	26	7	0.1333	0.3636	0.1951
Hourly Bigrams 2	7	7	5	0.5000	0.5833	0.5385
Hourly Bigrams 3	3	27	6	0.1000	0.3333	0.1538
Hourly Bigrams 4	4	17	9	0.1905	0.3077	0.2353
Hourly Bigrams 5	2	28	7	0.0667	0.2222	0.1026
Hourly Bigrams 6	4	16	7	0.2000	0.3636	0.2581
Average				*0.1984*	*0.3623*	*0.2472*
10 Minute Unigrams 1	13	17	19	0.4333	0.4063	0.4194
10 Minute Unigrams 2	8	22	25	0.2667	0.2424	0.2540
10 Minute Unigrams 3	12	18	19	0.4000	0.3871	0.3934
10 Minute Unigrams 4	11	19	22	0.3667	0.3333	0.3492
10 Minute Unigrams 5	8	22	24	0.2667	0.2500	0.2581
10 Minute Unigrams 6	9	21	23	0.3000	0.2813	0.2903
Average				*0.3389*	*0.3167*	*0.3274*
10 Minute Bigrams 1	2	21	7	0.0870	0.2222	0.1250
10 Minute Bigrams 2	1	23	8	0.0417	0.1111	0.0606
10 Minute Bigrams 3	2	9	8	0.1818	0.2000	0.1905
10 Minute Bigrams 4	3	9	8	0.2500	0.2727	0.2609
10 Minute Bigrams 5	1	17	8	0.0556	0.1111	0.0741
10 Minute Bigrams 6	2	16	8	0.1111	0.2000	0.1429
Average				*0.1212*	*0.1862*	*0.1423*

Figure 2 Graph of precision and recall scores for both unigrams and bigrams from analysis of datasets consisting of six one-hour segments and six ten-minute segments of tweets from the Twitter streaming API (see online version for colours)

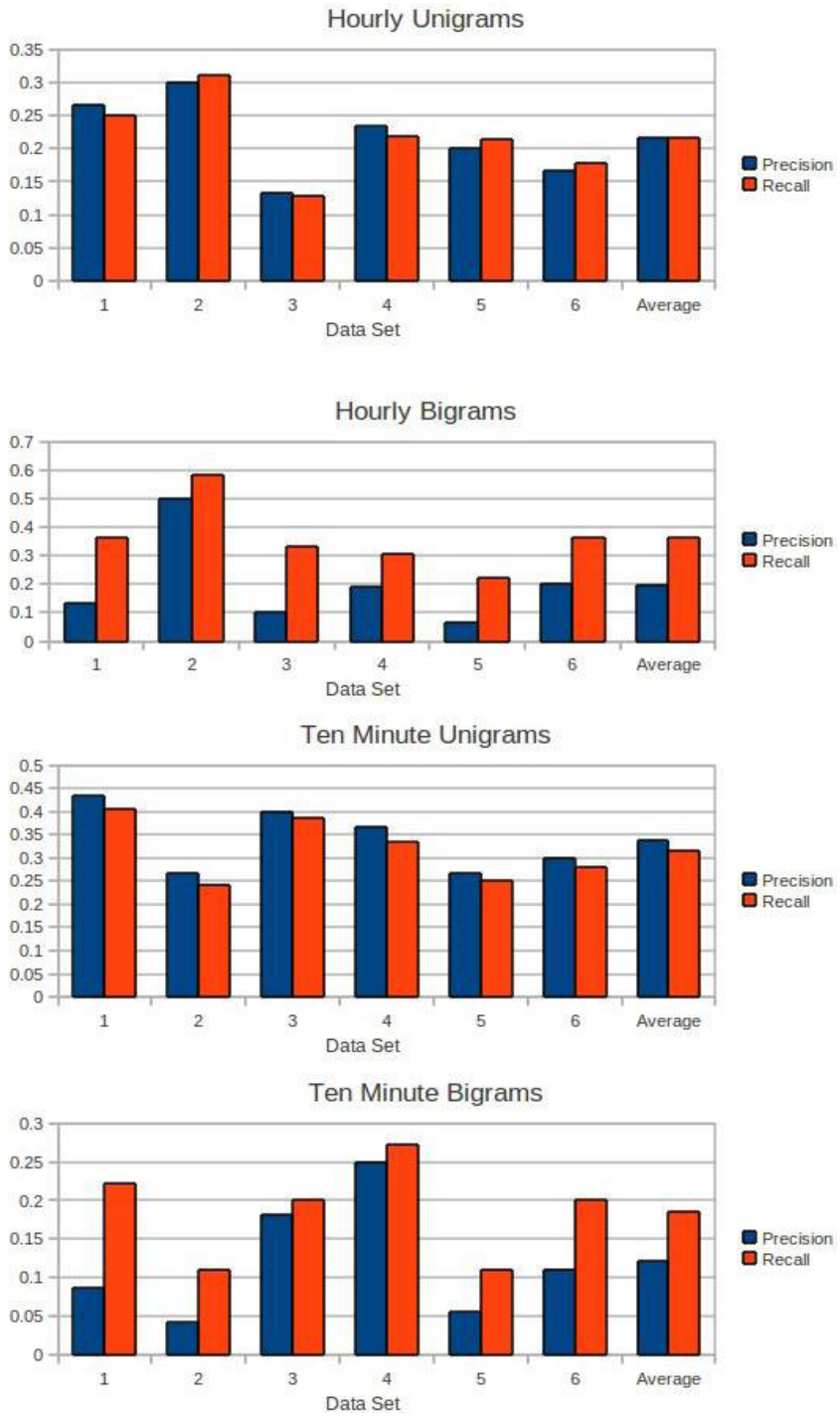
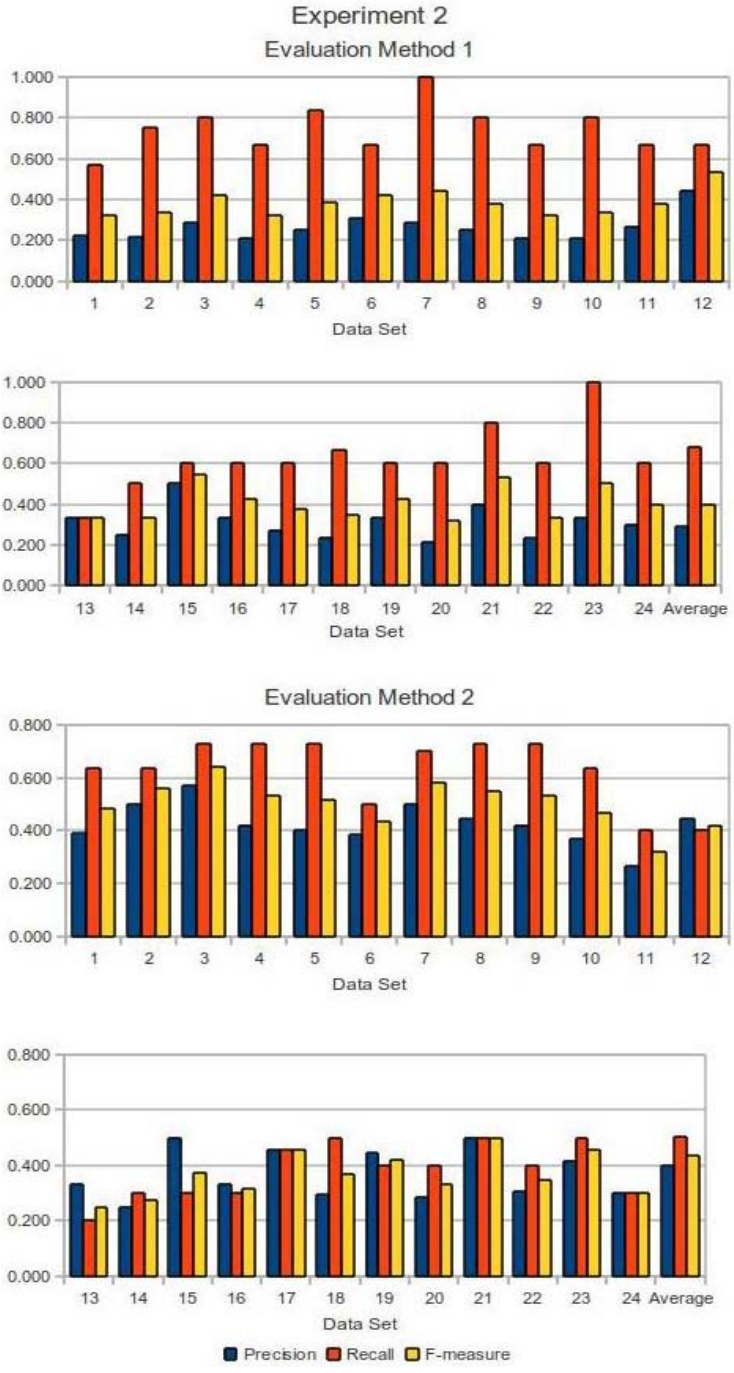


Table 2 Precision, recall, and F-measure scores for both unigrams from analysis of datasets consisting of 24 fifteen minute segments of tweets from the Twitter streaming API

Data set	*Evaluation method 1*						*Evaluation method 2*					
	TP	*FP*	*FN*	*Precision*	*Recall*	*F-measure*	*TP*	*FP*	*FN*	*Precision*	*Recall*	*F-measure*
1	4	14	3	0.222	0.571	0.320	7	11	4	0.389	0.636	0.483
2	3	11	1	0.214	0.750	0.333	7	7	4	0.500	0.636	0.560
3	4	10	1	0.286	0.800	0.421	8	6	3	0.571	0.727	0.640
4	4	15	2	0.211	0.667	0.320	8	11	3	0.421	0.727	0.533
5	5	15	1	0.250	0.833	0.385	8	12	3	0.400	0.727	0.516
6	4	9	2	0.308	0.667	0.421	5	8	5	0.385	0.500	0.435
7	4	10	0	0.286	1.000	0.444	7	7	3	0.500	0.700	0.583
8	4	12	1	0.250	0.800	0.381	8	10	3	0.444	0.727	0.552
9	4	15	2	0.211	0.667	0.320	8	11	3	0.421	0.727	0.533
10	4	15	1	0.211	0.800	0.333	7	12	4	0.368	0.636	0.467
11	4	11	2	0.267	0.667	0.381	4	11	6	0.267	0.400	0.320
12	4	5	2	0.444	0.667	0.533	4	5	6	0.444	0.400	0.421
13	2	4	4	0.333	0.333	0.333	2	4	8	0.333	0.200	0.250
14	3	9	3	0.250	0.500	0.333	3	9	7	0.250	0.300	0.273
15	3	3	2	0.500	0.600	0.545	3	3	7	0.500	0.300	0.375
16	3	6	2	0.333	0.600	0.429	3	6	7	0.333	0.300	0.316
17	3	8	2	0.273	0.600	0.375	5	6	6	0.455	0.455	0.455
18	4	13	2	0.235	0.667	0.348	5	12	5	0.294	0.500	0.370
19	3	6	2	0.333	0.600	0.429	4	5	6	0.444	0.400	0.421
20	3	11	2	0.214	0.600	0.316	4	10	6	0.286	0.400	0.333
21	4	6	1	0.400	0.800	0.533	5	5	5	0.500	0.500	0.500
22	3	10	2	0.231	0.600	0.333	4	9	6	0.308	0.400	0.348
23	4	8	0	0.333	1.000	0.500	5	7	5	0.417	0.500	0.455
24	3	7	2	0.300	0.600	0.400	3	7	7	0.300	0.300	0.300
Average				*0.287*	*0.683*	*0.394*				*0.397*	*0.504*	*0.435*

Figure 3 Graph of precision, recall, and F-measure scores for both unigrams from analysis of datasets consisting of 24 fifteen minute segments of tweets from the Twitter streaming API (see online version for colours)



8 Conclusions and future work

In this paper, we have outlined methodologies for using streaming data, tf-idf term weighting, normalised term frequency analysis, and other criteria to identify trending topics on Twitter. The methods implemented detected and identified both unigrams and bigrams as trending topics. Results for the first experiment fell significantly short of the original goals, but were reasonably close to results produced by other approaches. Results for the second experiment met the success conditions put forth in this paper. We have clearly demonstrated the ability to extract and identify pertinent information from a continuously changing corpus with an unconventional structure.

Based on the success we have achieved, there are many possible extensions and improvements we can look into. One potential extension would be to expand the functionality of the unigram and bigram algorithms to identify trigrams or higher order n-grams as trending topics, instead of single words or bigrams. Other possible extensions of this project include not only identifying but summarising trending topics (Inouye and Kalita, 2011; Sharifi et al. 2010a, 2010b, 2010c) and normalising the syntax of the summaries (Kaufmann and Kalita, 2011), or adapting the method to be used as a predictive tool. Another extension could be in the evaluation process. Terms identified as trending topics could be compared not only to topics identified by Twitter as trending, but to topics identified as trending by other sources, such as Yahoo![8] or Google Trends[9].

References

Allan, J., Papka, R. and Lavrenko, V. (1998) 'On-line New Event Detection and Tracking', in *Proceedings of ACM SIGR*, pp.37–45.

Cheong, M. and Lee, V. (2009) 'Integrating web-based intelligence retrieval and decision-making from the Twitter trends knowledge base', in *Proceedings of CIKM 2009 Co-Located Workshops: SWSM 2009*, pp.1–8.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J. and Dredze, M. (2010) 'Annotating named entities in Twitter data with crowdsourcing', in *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp.80–88.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith, N.A. (2010) 'Part-of-speech tagging for Twitter: annotation, features, and experiments', Technical Report, Carnegie Mellon University.

Glance, N., Hurst, M. and Tomokiyo, T. (2004) 'Blogpulse: automated trend discovery for weblogs', in *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. (2004) 'Information diffusion through blogspace', in *Proceedings of the 13th International Conference on the World Wide Web*, pp.491–501.

Han, B. and Baldwin, T. (2011) 'Lexical normalizations of short text messages: making sense a Twitter', in *Proc. of ACL-HLT*.

Hiemstra, D. (2000) 'A probabilistic justification for using tf×idf term weighting in information retrieval', *International Journal on Digital Libraries*, Vol. 3, No. 2, pp.131–139.

Inouye, D. and Kalita, J.K. (2011) *Comparing Twitter Summarization Algorithms,* IEEE SocialCom, October, Massachusetts Institute of Technology, Boston.

Kaufmann, J. and Kalita, J. (2011) 'Syntactic normalization of Twitter messages', *International Conference on Natural Language Processing (ICON 2011)*, Kharagpur, India, December, pp.149–158.

Kireyev, K., Palen, L. and Anderson, K. (2009) 'Applications of topics models to analysis of disaster-related Twitter data', in *NIPS Workshop on Applications for Topic Models: Text and Beyond*.

Liu, X., Wei, F., Zhang, S. and Zhou, M. (2011) 'Recognising named entities', in *Proc. of ACL-HLT*.

Lui, F., Weng, F., Wang, B. and Lui, Y. (2011) 'Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision', in the *Proc. of ACL-HLT*.

Manku, G. and Motwani, R. (2002) 'Approximate frequency counts over data streams', in *Proceedings of the 28th VLDB Conference*, Hong Kong, China.

Nallapati, R., Feng, A., Peng, F. and Allan, J. (2004) 'Event threading within news topics', in *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management*, pp.446–453.

Perera, R., Anand, S., Subbalakshmi, P. and Chandramouli, R. (2010) 'Twitter analytics: architecture, tools and analysis', *Military Communications Conference, MILCOM 2010*, pp.2186–2191.

Petrovic, S., Osborne, M. and Lavrenko, V. (2010a) 'Streaming first story detection with application to Twitter', in *Proceedings of NAACL*.

Petrovic, S., Osborne, M. and Lavrenko, V. (2010b) 'The Edinburgh Twitter Corpus', in *Proceedings of NAACL Workshop on Social Media*.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010) 'Earthquake Shakes Twitter users: real-time event detection by social sensors', in *WWW2010*.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information Processing and Management*, Vol. 24, No. 5, pp.513–523.

Shaparenko, B., Caruana, R., Gehrke, J. and Joachims, T. (2005) 'Identifying temporal patterns and key players in document collections', in *Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory, and Applications (TDM-05)*, pp.165–174.

Sharifi, B., Hutton, M. and Kalita, J. (2010a) 'Experiments in microblog summarization', in *NAACL-HLT 2010*, Los Angeles, pp.685–688.

Sharifi, B., Hutton, M-A. and Kalita, J. (2010b) 'Automatic summarization of Twitter topics, in algorithms in applications', in Utpal Sharma and Dhruba K. Bhattacharyya (Eds.), pp.121–128, Narosa, Delhi, India, ISBN 978-81-8487-082-4.

Sharifi, B., Hutton, M-A. and Kalita, J. (2010c) 'Experiments in microblog summarization', *Second IEEE International Conference on Social Computing (SocialCom 2010)*, Minneapolis, August, pp.49–56.

Wilbur, W.J. and Sirotkin, K. (1991) 'The automatic identification of stop words', *Journal of Information Science*, Vol. 18, No. 1, pp.45–55.

Wilbur, W.J. and Yang, Y. (1996) 'An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts', *Computers in Biology and Medicine*, Vol. 26, No. 3, pp.209–222.

Yang, Y., Pierce, T. and Corbonell, J. (1998) 'A study on retrospective and on-line event detection', in *Proceedings of the 21st ACM SIGR*.

Notes

1 http://www.alexa.com/siteinfo/twitter.com
2 http://blog.twitter.com/2011/09/one-hundred-million-voices.html
3 http://blog.twitter.com/2011/08/your-world-more-connected.html
4 http://en.wikipedia.org/wiki/Twitter
5 http://blog.twitter.com/2010/04/hello-world.html
6 https://stream.twitter.com/1.1/statuses/sample.json (see documentation at https://dev.twitter.com/docs/streaming-apis/streams/public)
7 https://dev.twitter.com/docs/api/1.1/get/statuses/sample
8 http://www.yahoo.com
9 http://www.google.com/trends