Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 00 (2015) 000–000

Procedia Engineering

www.elsevier.com/locate/procedia

Humanitarian Technology: Science, Systems and Global Impact 2015, HumTech2015

# Tackling the challenges of situational awareness extraction in Twitter with an adaptive approach

Haji Mohammad Saleem*, Faiyaz Al Zamal, Derek Ruths

*School of Computer Science, McGill University, Montreal, QC H3A 0G4, Canada*

**Abstract**

Twitter is widely perceived as a potential source of valuable information for responders to mass emergencies. Despite interest in the development of extraction systems for such information, little effort has been put towards systemic methods for obtaining all posts pertaining to a disaster from the live Twitter stream. Researchers rely on keyword-based filters to extract information in spite of evidence that such markers are absent in many informational tweets, and also neglect the topic and traffic dynamics of the relevant tweets as crises progress. Previous work has shown that such practices can often lead to the loss of critical information in the context of a disaster. We introduce an adaptive filter, tailored to the idiosyncrasies of the real-time Twitter feed, intended to extract disaster-related content. Furthermore, we introduce a novel data model based on a three-label classification scheme to describe the composition of the data-stream. We use this model to simulate Twitter streams, modelling various post-disaster scenarios, for the purpose of filter performance evaluation. The filter is able to remove over 85% of the non-crisis content, and achieves a three-fold reduction in the loss of relevant contents compared to the existing approaches. In combination, the method and the model are useful tools for extracting situational awareness and highlight important directions for future work in this area.


Selection and peer-review under responsibility of the Organizing Committee of HumTech2015.

*Keywords:* Adaptive Filter, Information Retrieval, Real-time filtering, Microblogs.

## 1. Introduction

During and in the immediate aftermath of a major calamity, responders need to quickly assess conditions in the affected region. Such *situational awareness* (SA) enables the effective delivery of services and resources (e.g., humanitarian aid, and search and rescue operations) to the appropriate regions and populations [1]. However, disaster conditions impede assessment of affected locations and populations, making situational information difficult to obtain through traditional mean.

Social media has yielded a promising alternative source of such situational information: individuals in the affected regions, in many cases, readily post information about their conditions to platforms such as Twitter and Facebook. Such information may not be readily available and can contribute to SA. In fact, the analyses of recent disasters has shown that users of social media (notably Twitter) posted information would be valuable for first responders. [2,3]. As a recent example, immediately after the shooting incident in Moncton, New Brunswick, a Twitter user posted a

* Corresponding author. haji.saleem@mail.mcgill.ca.

Selection and peer-review under responsibility of the Organizing Committee of HumTech2015.

photo of the shooter, with a clear view of his armaments [1]. Such information, were it extracted rapidly and properly, could have been crucial in evaluating the threat posed by the shooter.

While Twitter data extraction is an active area of research (e.g., [4,5]), no approaches, to our knowledge, are tailored to the idiosyncrasies of disaster-related tweets in an active Twitter stream [7]. Recent work has established that disaster-related content in Twitter has several consistent characteristics:

*Small volume relative to the entire Twitter stream* - Since Twitter is used globally, an individual event generates only a small percentage of the overall traffic. This is true for disaster and non-disaster tweets alike [8].

*Absence of keyword labels* - A careful study of several disasters revealed that many early, information-laden tweets do not carry hashtags and other keywords which have been typically used for disaster-related tweet identification [9].

*Abundance of disaster tweets changes over time* - Though public response varies with disaster, the overall volume of resulting tweets can generally be split in to three phases: the *rise* demonstrating increasing frequency, the *plateau* demonstrating high frequency, and the *fall* demonstrating diminishing frequency of disaster related tweets, (see Figure 1) [8].

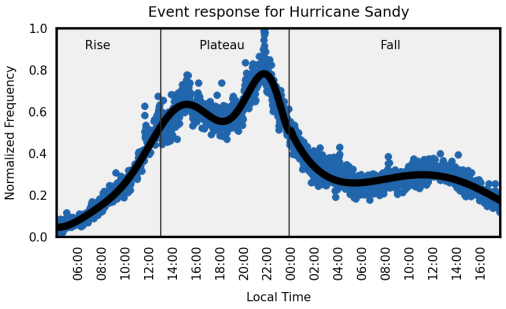


Fig. 1: *The three phases of event response on Twitter.* - The event response to Hurricane sandy on Twitter on Oct 29 and 30, 2012. The General shape can be divided into three sections, the initial rise, the middle plateau and the final fall, which is consistent with how the general population reacts to an event.

*Dramatic topical shifts* - Since an event is dynamic, the nature of public engagement changes as the event unfolds. For eg., the Moncton incident consisted of the initial shooting followed by an extended manhunt and finally the suspect's capture — all of which was reflected by discussion in Twitter. Moreover, many disasters, by their nature and coverage in the media, will engage larger Twitter populations well-outside the affected geographical area and generate additional content with diverse topics including emotional reactions, prayers for victims, and offers of support.

The need to account for such issues motivates the two major contributions of our study: (1) an adaptive filter that can accommodate for the overwhelming amount of unrelated content in the stream along with the variable volume and dynamic topical shifts of relevant content over time, and (2) a formal data model describing the composition of an active Twitter stream in terms of the relative frequencies of different types of disaster-related content.

***Adaptive filter***

Twitter generates almost 500 million tweets everyday. Therefore, any extraction system that operates on the entire data-stream has to face excessive noise in the form of non-relevant content. If not addressed carefully, this creates a learning bias towards the more dominant label, degrading the overall performance [10,11]. We propose a noise filter as an essential part of any information extraction machinery to eliminate data imbalance by effectively reducing the non-relevant content in the data-stream.

To account for the dynamic nature of the event stream, both in volume and content, we develop the filter with an adaptive framework that periodically updates its model based on recently labelled data. Actual noise filtering is carried out by a supervised learning algorithm embedded in the adaptive framework. The choice of algorithm depends on the following factors: (1) we require a low complexity algorithm to reduce computational overheads due to the real-time nature of data processing pipeline to which the filter belongs; and (2) we need an algorithm that performs reasonably well with sparse datasets due to the short document size of a single tweet. We therefore embed a Naive Bayes classifier in our filter frame work, appropriately tailored to deal with the adaptive requirements and adjusted to handle the data imbalance present in the Twitter stream.

We require pre-labelled data to test our supervised learning setup and properly measure its performance. However, labelling even a small timeframe of an entire Twitter stream is not logistically possible. To overcome this problem, we generate synthetic datasets that represent simulated Twitter streams that stem from the data model we soon explain. The adaptive Naive Bayes filter successfully removes the majority of the noise in these simulated streams, while losing only a small fraction of relevant data. Our approach outperforms keyword-based filters, non-adaptive filters or

[1] `https://Twitter.com/PatHemsworth/status/474356870686461953`

adaptive filters embedded with other learning algorithms. Notably, the performance is consistent over the 81 artificially generated datasets representing varying patterns of Twitter stream composition and therefore, affirms the applicability of the adaptive filter to characteristically diverse events.

***Data model***

The purpose of the formal data model is to define the composition of an active Twitter stream in terms of the relative abundance of tweets relevant to a disaster of interest. The model first introduces a novel three-label classification scheme to categorize the tweets relevant to a disaster, differentiating between evident and inferred relevance: (1) *obvious positive* (OP)- tweets that contain obvious textual markers (such as hashtags and keywords) of relevance to the event, (2) *inferred positive* (IP) - tweets that, though event related, lack self-evident textual markers, and therefore their relevance to the event has to be indirectly inferred and, (3) *negative* (N) - tweets that are not at all related to the event. Our model then defines a Twitter stream at a given time as a mixture of these three different types of tweets present in different quantities, which are represented by corresponding frequency distributions. To the best of our knowledge, a relevance based three-label model has not been previously introduced. Such data categorization would be applicable in any textual-data extraction problem, as it provides the ability to perform retrospective search for inferred data after a primary extraction of obvious positive content.

## 2. Background work

Social media platforms have expanded into major modes of communication [12,13], evolving from connecting people online [14], to supporting high volume information generation and flow during rapidly evolving situations [15]. They are free forums that promote the exchange of ideas, opinion and information, while providing an opportunity for global outreach in crisis communication [16]. These platforms, including Twitter, are widely regarded as sources of valuable situational awareness generated by users posting local knowledge that only they have access to [17].

***Extracting situational awareness***

In the recent years, a significant amount of work has been done on the general topic of extracting situational awareness from a post-disaster Twitter stream. However, the problem of situational awareness extraction remains without a robust solution due to the absence of a systematic approach that has been tailored to account for the multiple idiosyncrasies of Twitter data, which include, the variation in content and volume in the data-stream over time, missing hashtags in tweets with *novel situational awareness* and the excessive presence of non relevant content [8,9].

Many studies have focused on building a single stage machinery that identifies informational posts [2,4**?** ,5], and on characterizing the general kind of information such posts contain [7,18,19]. However, a notable limitation of most (if not all) existing studies is their dependence on datasets that were collected by sampling the Twitter feed using a small set of hashtags and keywords relevant to the event in question [4,5] or a significantly small handpicked dataset itself [2]. Since many tweets with situational information carry no hashtags and often fail to even use keywords that might be expected (e.g., “tornado” and “forest fire”, etc.) [9], an approach based solely on keywords can lead to loss of critical information. Training and testing extraction systems on such restricted data can be detrimental to the development the systems.

Furthermore, none of the solutions address the variability of the data or the fact that presence of noise in large proportions can lead to learning bias and affect the performance of the extraction mechanism. We therefore, expect the prior work to be non comprehensive solutions to the problem of extracting situational awareness.

***Adaptive filter***

The literature on real-time filtering of microblogs is very limited, with the exception of the microblog track at the Text Retrieval Conference, 2012. The track included a real-time filtering pilot task that required identifying tweets relevant to a query term on an aggregated dataset [20], and received multiple submissions [21]. It asked the contributors to use scaled utility (utility assigns a value or cost to each document, based on whether it is retrieved or not retrieved and whether it is relevant or not relevant and is then scaled over all topics) as the measure of performance, which is biased towards precision [22], a requirement that favours the algorithms with a low false-positive rate. A low false-positive rate can often accompany poor performance in the form of low recall, as a result of excessive data loss. In a time-critical setting, every piece of information can be of importance and therefore excessive data loss is counter productive and does not validate these performances. It is imperative to build systems that maintain high recall and therefore, retain the positive documents. As an example, entry by Harbin Institute of Technology registered

the highest utility score of 0.4117 but with a low recall at 17% [23]. Similarly, just high recall is irrelevant if majority of the documents are labelled positive [24]. While the different performance emphasis makes direct comparison awkward, TREC provided useful insights on how to setup performance measures.

## 3. Formal data model

We introduce a formal data model, that can describe the composition of an active Twitter stream in the context of a disaster. One of novel aspects of this model is the introduction of a three-label classification scheme that divides the disaster relevant content in two sets, based on how this relevance can be estimated.

***Three-label classification scheme***

Traditionally, keyword-based extraction methods have been heavily used in social media data mining literature. However, as a notable and important class of *novel SA* tweets do not use these keywords [9], it is important to differentiate these kind of tweets from the ones with keywords. We separately label them for additional focus, in an attempt to extract them and gain more comprehensive information in a time-critical situation.

Here we propose a classification scheme that creates a distinction among the tweets based on the evidence of relevance. The labels include: (1) *obvious positive* (OP) - tweets with obvious textual evidence in the form of keywords or hashtags. They are easy to extract, e.g., "Here it comes #sandy!" 2) *inferred positive* (IP) - tweets with obscure textual signature and absent keywords, e.g., "My house ain't get flooded haha!" and (3) *negative* (N) - irrelevant tweets. Recognizing this distinction enables us to model the pattern of their relative occurrences in the Twitter stream.

***Frequency distributions of different labels***

We define a temporal framework to account for the variable nature of the Twitter stream. The framework describes the individual distribution for the three labels: the *total volume frequency* ($v$), the *positive volume frequency* ($p$) and the *inferred positive volume frequency* ($ip$), and two normalizing factors ($v_{max}$, $p^{ratio}_{max}$), in order to quantify the composition of the stream. The normalizing factors are used for convert the distributions to absolute values.

*Total volume frequency distribution* (v) - The volume of generated content at a specific time is dependent on the number of active users at that instant. While this volume varies significantly over the course of a day, the traffic exhibits a recurring diurnal pattern [8].

In this work, we therefore use an averaged daily traffic distribution as a proxy for the overall Twitter traffic pattern. This distribution is modelled as an array of numeric frequencies, of length 24: $v = \langle v_1, v_2, \ldots v_{24} \rangle$, to represent hourly tweet rates, normalized to unit scale. The distribution can scaled back to the actual volumes through the factor $v_{max}$, which stores the maximum hourly traffic frequency, i.e. $v_{max} = \max_{1 \le i \le 24} v_i$.

*Positive volume frequency distribution* (p) - observing *disaster signature* from prior work [8], we came up three generalized event response phases: the *rise* during which the frequency of event relevant tweets increases, the *plateau* during which this

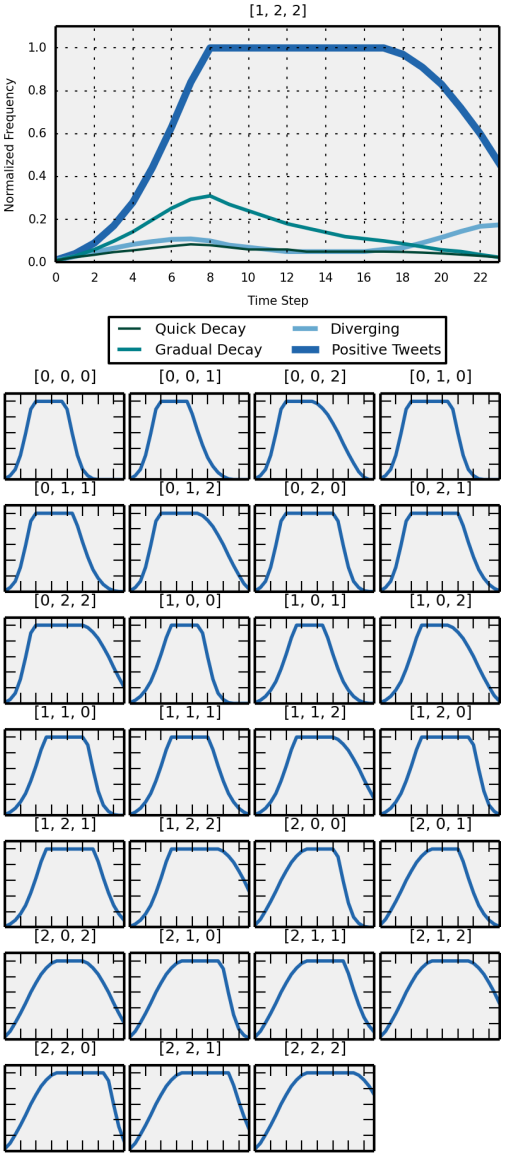


Fig. 2: *Varied frequency distributions for positive event tweets* - Event response is a combination of three phases a rise, a plateau and a fall, combined in different proportions. We use qualitative descriptors, short(0), medium(1), and long(2) that represent duration to generate 27 different signatures. The signature on the top is an example from among the generated datasets, divided by labels.

frequency attains high values and the *fall* where this frequency declines (Fig 1). Such an assessment is consistent with how general population reacts to news and the change in their interest over time. The disaster signature is therefore the set of tweets that are regarded as positive for the filter. It is represented as a numeric array: $p = \langle p_1, p_2, \ldots p_{24} \rangle$ and also normalized to unit range, to be scaled by a factor $p_{max}^{ratio}$. The factor is the ratio of highest response frequency in relation to the highest total frequency and changes according to the scale of the event.

*Inferred positive volume frequency* (ip) - The positive response combines two data labels, the obvious and the inferred positive. Defining one of the components allows us to implicitly define the other from the aggregate response. We specify the inferred positive distribution as the fraction of inferred positive tweets present in a particular timeframe of the aggregate response, mathematically represented as: $ip = \langle ip_1, ip_2, \ldots ip_{24} \rangle$, where each value is calculated as: $ip_i = \frac{num\ of\ ip\ in\ i}{num\ of\ p\ in\ i}$, $i$ being the hourly time step.

Using these three distributions ($v$, $p$, $ip$) and the two factors ($v_{max}$, $p_{max}^{ratio}$), we can comprehensively depict the multiple components of a live Twitter stream in the context of a disaster.

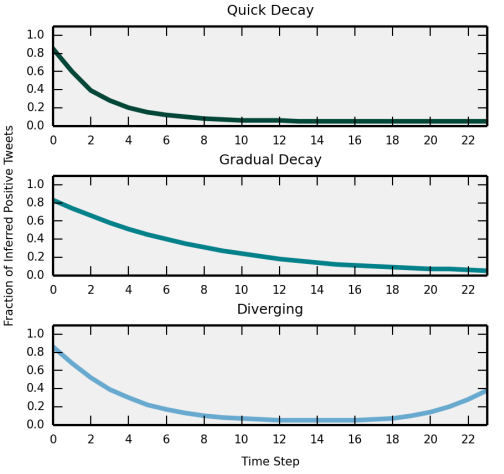


Fig. 3: *Proposed inferred positive percentage distributions* - Inferred positives are tweets that are related to an event but do not contain keywords. The three distributions represent three basic scenarios explained in text.

## 4. Adaptive filter

The filter aims to reduce the noise in an active Twitter stream to create more balanced datasets, making it applicable to any data mining pipeline that tackle high volume data feeds. The data moderation reduces the subsequent load on the more accomplished classification system that may be further deployed. We provide the basic schematics of the filter in Figure 4, along with an overview below.

**Filter components**

***Preprocessing***

*Basic text cleanup* - We performed basic text cleanup and removed punctuation, non-alphanumeric symbols, and URLs before tokenizing.

*Removing stop words* - Standard stop words were removed.

*Hashtag splitting* - We split hashtags into their component words in order to lend weight to the respective words. For example, #supersandy and #hurricanesandy both contain the token sandy, increasing its weight. Results from the hashtag splitter were added to the token list [25].

*Token stemming* - Stemming also helps in aggregating the different forms of the same word in a single token, further increasing its overall weight. We use the porter2 stemmer[2].

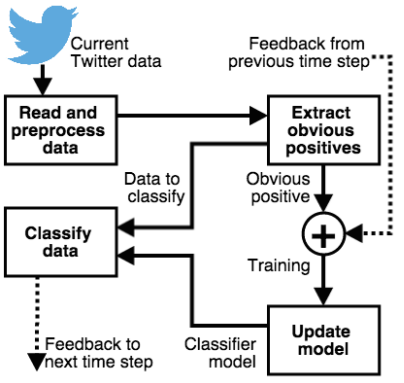


Fig. 4: *Schematics of the adaptive filter* - A simplistic overview of the different components that constitute the adaptive filter and also how they function together. This illustration represents one tilmestep.

***Obvious positive extraction***

Event descriptors are keywords that act as textual evidence of relevance to an event. Keyword-containing tweets are *obvious positives* and are initially segregated and labelled. They are added to the present training set along with recently labelled feedback. Since the initial time step lacks feedback, the training set contains obvious positive tweets along with negative tweets, sampled from before the event. The extraction is instrumental in providing the positive training samples for the learning algorithm.

[2] `https://pypi.python.org/pypi/stemming/1.0`

***Filter update***

To tackle the dynamic volume and topical shifts, the classifier updates at every time step, which in our case only requires updating individual token weights based on the additional token frequencies for the two labels. Alternate learning algorithm might require complete retraining for updating.

***Classifier***

The classifier labels tweets as disaster related/unrelated using the updated classification model. As specified, we choose to embed an adjusted version of the Naive Bayes algorithm in this work. The adjustments involved oversampling the positive tweets in the training set to increase their representation, and then, introducing a ratio-threshold-based-scheme to discriminate the labels [27]. The resulting labels are treated as feedback for the next round.

## 5. Performance evaluation of the adaptive filter

To evaluate the performance of the adaptive filter, we require (1) a dataset with temporally distributed tweets accurately labelled using the three-label classification scheme, and (2) appropriate performance measures tailored to assess the efficacy of the filter to remove as many irrelevant tweets as possible with the minimal loss of relevant tweets. Below we discuss our approaches for addressing these needs.

***Synthetic datasets***

In order to test any data analysis setup, we require labelled data. However, manually labelling the entire decahose for even a single disaster is not feasible, due to the high volume of available tweets. We therefore generate synthetic datasets that mimic various post-disaster scenarios and use it instead for testing purposes. Setting hurricane Sandy as the base event, we define a dataset as a stream of tweets (taken from real tweets posted by users) that are distributed over a 24 hour period following the corresponding tweet frequency model. This approach is closer to actual raw data than limited datasets curated through a few keywords.

*Creating data cores* - The first requirement for the creation of synthetic dataset is a large amount of labelled data. As it is not manually possible, we used disaster-specific keywords to gather a large number of positive tweets. Under our model, such tweets would be considered obvious positives. In order to construct inferred positives, we removed the specific keyword from these tweets (proxying for content that was relevant, but not tagged in an obvious way). To be concrete, consider a dataset built from hurricane Sandy content. Here, we regard any tweet that has the keywords "hurricane" or "sandy" as obvious positives. From this set of tweets, we removed keywords from a fixed percentage of them. The resulting tweets, though relevant, are devoid of the identifying keywords and meets our definition of *ip* and therefore serves our purpose of providing both obvious and inferred positive tweets. Since any tweet from before the event is definitely non relevant, we select a set of tweets at random, and remove any tweets with the keywords, which somehow might have been present. We were therefore able to create three large sets of labelled data.

*Dataset generation* - Next we create a framework to combine the labelled data, according to the data model. The overall frequency $v$ is standardized for all events. Since it is scalable, we set $v_{max} = 10,000$ (that is, highest value in the frequency distribution is 10,000), to keep a check on the size of testing datasets.

To mimic an event signature, we construct distributions for its three phases. To further represent various post event scenarios, we create response distributions as a combination of different time spans for the three phases, which were qualitatively categorized as short (0), medium (1) and long (2). We used beta distributions, that can be applied to model the behaviour of finite length random variable, to generate the distribution for response growth and decay, using the two shape parameters. The 27 generated event signatures are presented in Figure 2, with the naming convention of the datasets comes from the length of the three phases. The factor $p_{max}^{ratio}$ is set at 0.1, as a high estimate of the presence of positive tweets.

It is a common phenomenon that independently generated hashtags converge to a few that enter mainstream usage [26]. Such a convergence leads to the strongest event related keywords and can be identified as obvious positive. All the other generated content becomes a part of the inferred positive, the volume of which decreases as the convergence continues. While, at the moment we do not have the definite behaviour details for inferred positive tweets, we propose and use three frequency distributions based on the convergence of hashtags. The first distribution involves *gradual* convergence to the obvious positive, as more and more users identify with a few common hashtags. The second distribution proposes a *quick* convergence of inferred positives as the users strongly associate with the mainstream hashtags. Finally we propose a *diverging* distribution, where after the initial convergence, the event evolves and new

subtopics emerge, leading to association with new hashtags. We present these in Figure 3. We therefore, generate 81 distinct datasets (combination of 3 inferred positive models over 27 positive models), with varied event response and the degree of inferred positives present.

*Calculations* - Using the data model specifications, the framework performs calculations and combines appropriate proportions of tweets to finally generate the datasets. The mathematics behind the generation is as follows:

$$v_i^{num} = v_{max} * v_i \qquad p_i^{num} = p_{max} * p_i \qquad op_i^{num} = p_i^{num} - ip_i^{num}$$
$$p_{max} = p_{max}^{ratio} * v_{max} \qquad ip_i^{num} = p_{num} * ip_i \qquad n_i^{num} = v_i^{num} - p_i^{num}$$

where $x_i^{num}$ is the scaled volume from $x_i$ distribution and the other variables have predefined meanings.

***Performance measures*** The widely used performance measures for information retrieval systems, precision, and recall, represent the system's ability to accurately label positive documents. They however, can produce misleading results in case of data imbalance due to the overwhelming presence of negative documents. Since the aim of our filter is to remove the negative documents from the data, using recall is not suitable as the measure gauges the performance of the system on positive data. Even if the system performs with reasonably high accuracy, the number of negative documents still present can be fairly high. This can easily offset the presence of true positive labels and degrade the precision score.

We therefore, use performance measures that are noise-oriented and, therefore, more suitable for the filtering task. Their core focus lies on how well the system removes the negative documents while keeping a check on how much relevant information is lost in the process. The proposed measures capture the volume reduction aspect of filter design and provide relevant details and are, therefore, a better gauge for the performance of the system. They are as follows:

*Specificity* (Noise Removed) - Specificity relates to the algorithm's ability to exclude a non relevant content correctly and therefore a good indicator of how well the filter is able to reduce the presence of non relevant tweets. It is calculated as $TN \;/\; TN \;+\; FP$ .

*False Negative Rate*(Data Loss) - The false negative rate is a measure of algorithm's ability to identify a relevant content correctly. While trying to decrease false negative rate is the same as increasing recall, recall focusses on how well the system performed in identifying positive content fnr plays the more central role of identifying the lost tweets. It is calculated as $FN \;/\; TP \;+\; FN$.

*Negative Accuracy*(Stream Cutback) - The final measure is an indicator of the overall performance of the filter, calculating the degree of scale down the data-stream goes through. As the aim of the filter is to reduce the overall volume, negative accuracy is more relevant than positive accuracy, which would be the amount of volume that passes the filter. It is calculated as $\frac{FN + TN}{TP + TN + FP + FN}$.

## 6. Results

*Overall performance* - We view the adaptive filter as a tool to reduce the overall volume of the data-stream and the performance measures presented in Table 1 validate its development. First, the filter was designed to remove noise from the data-stream and this task is executed very well with a high average specificity of almost 0.85. More importantly, the filter achieves noteworthy results when dealing with positive data samples, as only about 11 percent are mislabelled as negatives. This result holds relevance when we compare our method to other commonly-used algorithms in the following subsection. Both these values aggregate and provide the filter with a negative accuracy of 0.77, which indicates that filter was successfully able to restrict the data-stream to less than a quarter of its original volume.

*Consistency* - Another remarkable feature of this performance, as indicated by Table 1, is the consistency of the results. The filter was tested over a set of 81 different event signatures and it was able to provide strikingly similar results for each one of them. The three performance measures have a very small standard deviation, which corroborates our observation on the consistent results.

*Quantitative analysis* - From a purely quantitative perspective, we visualize the performance of the filter based on the numeric value of the quantities involved, rather than the fractions. This representation is important for understanding the problem of data imbalance in terms of wide mismatch between positive and negative samples. We go on to present

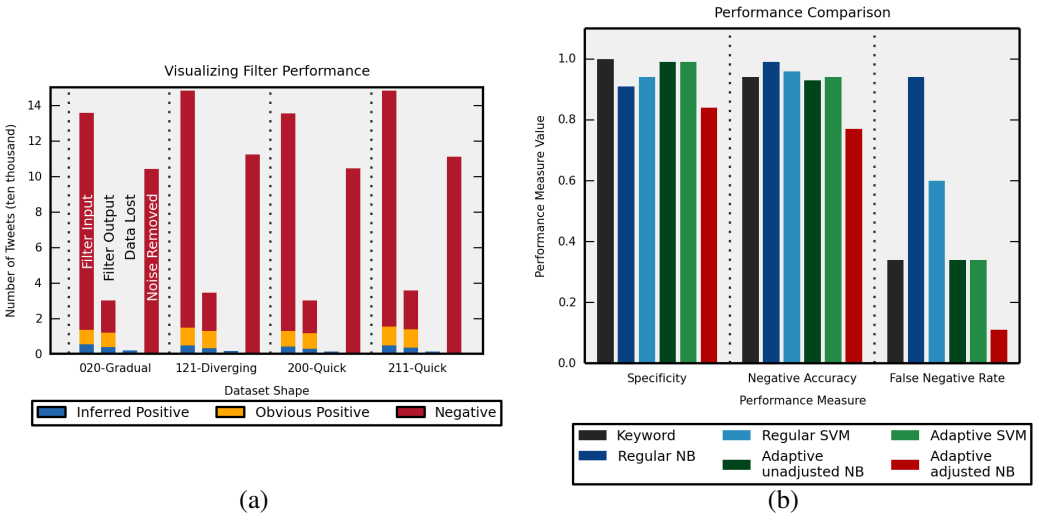


Fig. 5: (a) *Quantitative representation of the performance* - Filter performance on 4 randomly chosen. While the first two bars in a visualization represent filter input and output, the final two demonstrate the huge difference in scale of relevant data lost and noise removed. (b) *Adaptive filter vs traditional approaches* - To further investigate the performance of the adaptive filer, we analyze it against the traditional approaches that are often employed. It is clearly observable that the adaptive filter has lowest percentage of data loss while reducing a sizeable portion of the overall stream.

a before and after comparison of the composition of the data-stream with respect to the adaptive filter along with amount of data we fail to recover and the total noise we were able to successfully remove, as shown in Figure 5 a. Due to size limitations, we restricted our analysis to 4 datasets, randomly chosen from the 81 sets filter was tested on. The stark contrast between the amount of data lost and the amount noise further strengthens our performance claims.

*Comparison with different alternatives* - Through model refurbishment, the adaptive framework enables us to tackle the issues generated from both data imbalance and dynamical topical shifts. In contrast, a non-adaptive extraction method, with a prepared training set, is unable to overcome the challenges laid by these problems. We compared the performance of our adaptive adjusted Naive Bayes filter with the traditionally used filters and popular alternative algorithms, in Figure 5 b. Apart from our filter, the comparison includes: (i) keyword based filter, (ii) regular (non-adaptive) Naive Bayes filter, (iii) regular (non-adaptive) SVM filter, (iv) adaptive regular Naive Bayes filter, and (v) adaptive regular SVM filter. The keyword filter is a traditionally implemented approach which includes only the tweets that carry some predefined keywords among relevant data. The filter was considered in this comparison due to its popularity, as many existing studies collect datasets through such keyword retrieval. Since non-adaptive filters can also be employed for the purpose, albeit with a potentially significant waiting time to gather ample positive training samples, we evaluate the performance of the non-adaptive Naive Bayes and SVM classifiers (setting the waiting time till the sixth time-step) for the datasets. Finally, as our defined adaptive framework allows us to embed any classification algorithm for the actual filtering task, we include SVM and standard unadjusted Naive Bayes within an adaptive framework in our comparative analysis.

Table 1: Analysis of filter performance over all 81 datasets

| Measure | Mean | Range | Std dev |
|---|---|---|---|
| Specificity | 84.26 | 4.13 | 0.78 |
| False Negative Rate | 11.3 | 5.62 | 1.31 |
| Neg Accuracy | 77.15 | 6.37 | 1.40 |

The different filtering approaches do not yield significant variation in terms of specificity and negative accuracy. Noteworthy however is the difference in false negative rate for the six methods, varying from almost 0.9 to 0.1. Both the non-adaptive approaches have poor false negative rates, even outperformed by the keyword based filter. The non-adjusted adaptive approaches register values very close to the keyword filter, this indicates that these methods are able to detect on a tiny percentage of inferred positive tweets. Our adjusted Naive Bayes filter outperforms all other methods and is able to retain almost 90 percent of the relevant data.

The failure of the non-adaptive filters is expected, as they ignore the evolving nature of the tweet texts. Furthermore, since these supervised learning approaches require a substantial training set before the actual classification can be

initiated, the processes have to wait as they aggregate enough obvious positives to generate the classification model. This aggregation can take significant amount of time delay before the set is collected, due to the lack of obvious positive tweets in the initial phases of the disaster. To provide a perspective, the first time step in generated datasets have one positive tweet for almost 700 negative tweets, implying an extortionate imbalance. Additionally, the time critical nature of a post-disaster scenario often makes this delay unacceptable. In Figure 5b, we plot the performance of both the non-adaptive filters with a waiting time of six hours. Even this delay is unable to help the filter and they provide sub par results.

The keyword based filter, on the other hand, imposes a criteria which is too stringent, thereby intercepting many relevant tweets. While this filter yields high specificity, the resulting false negative rate is more than 30%. The use of adaptive filters with standard Naive Bayes and SVM algorithms (fourth and fifth data point in Figure 5b) produce results almost identical to the keyword based filter. In contrast, our adjusted Naive Bayes algorithm significantly improves the filter performance, reducing the false negative rate to approximately 11%. While some amount of specificity and negative accuracy is sacrificed to achieve this false negative rate reduction, the retention of a larger fraction of potentially relevant tweet, even at the cost of an increased amount of irrelevant tweets is considered more vital in this context as the multi-component data extraction pipeline can handle these irrelevant tweets at the next step using a more sophisticated classification system.

## 7. Discussion and Conclusion

This paper lays the groundwork for a framework to aid the extrication of SA from Twitter after a disaster. While previous literature has relied on keyword based searches to generate datasets, they ignore relevant documents without obvious textual markers. Besides, these works rarely take into account the dynamic contextual shifts observed and the overwhelming data imbalance present in the data-stream, which seriously undermines their ability to detect relevant tweets in real time. In this paper, we propose a filter as an essential first step to any data extraction pipeline that works on heavily imbalanced data. The objective of the filter is to reduce the level of noise present in the stream by removing the tweets that are highly unlikely to be relevant in the context of the disaster, while retaining nearly all relevant tweets. It is the task of a subsequent extraction system, next in the pipeline, to actually extract the tweets relevant to the disaster from the filtered stream that is much less unbalanced.

In order to tackle the dynamic nature of the contents, we suggest an adaptive framework for the filter, i.e. the tweets labeled as relevant from the last time steps and the tweets carrying obvious markers of relevance in the current time step are combined to form the training set for the classifier embedded in the filter. The classifier then labels the tweets in the current time step as either relevant or non-relevant that is used as the feedback for the next time step. For effective filtering in a time-critical setup, the embedded classifier must be fast and capable of handling the severe data imbalance. We chose the Naive Bayes algorithm for implementing the classifier, which can be quickly updated as new training data arrives and easily modified to handle imbalanced data sets.

Evaluation of the filter performance requires the knowledge of the true labels of the tweets from the Twitter stream. As it is not logistically possible to manually label the whole Twitter stream (or even a sampled stream like Decahose containing a small fraction of the Twitter stream) according to their relevance to a certain disaster event due to the astronomical volume of the stream, we devise a data model that describes the composition of an active Twitter stream and use this model to generate synthetic datasets simulating tweet streams following various composition patterns. The model proposes a novel three-label classification scheme for the tweets based on the presence (or absence) of the textual indicators of relevance to the disaster under consideration and uses frequency distributions delineating the relative abundance of the tweets belonging to these different labels to represent the active stream at a given time. We found that our adaptive Naive Bayes filter can perform consistently well across all these synthetic datasets. It removes 85% of the noise present in the stream and loses only about 11% of the relevant tweets; suggesting a threefold reduction of data loss compared to the keyword-based filter. The method and the model together are promising tools for SA extraction.

*Reducing data loss.* Despite the considerable improvement achieved over the other approaches, our filter is still missing about 11% of the relevant tweets. These tweets have a very low positive score, making them difficult to capture even with a considerable adjustment of the ratio threshold. This indicates that this fraction of the relevant tweets are probably textually similar to non-relevant tweets and are not probably detectable using only a text-based

classification scheme. The use of Twitter metadata, such as geolocation and user profiles may have the potential to reduce this data loss. We leave the incorporation of the metadata as a future work.

*Improving the performance of the embedded classifier.* Our adaptive filter framework requires an embedded classifier for performing the actual task of filtering. In this paper, we have used a low complexity Naive Bayes Algorithm, that was further adjusted to handle the severe data imbalance present in the Twitter stream. It is possible to update the model at every time step without the need for a complete model rebuild. Therefore, our adjusted Naive Bayes classifier is much faster than other choices where such an incremental update is not supported. We found that the runtime is reduced by almost 50% compared to the adaptive SVM classifier. However, we have not performed a comprehensive analysis of the running time covering all alternative choices of the state-of-the-art methods tailored for dealing with imbalanced datasets. Also, it might be possible to improve our Naive Bayes classifier through further adjustments and modification. We consider such analyses as important directions for future work.

*Potential application of the data model.* Our data model equips us with the ability to simulate the Twitter stream under different assumptions on the relative frequencies of different types of tweets. However, the model has potential area of application beyond Twitter stream simulation. As the Twitter response to a disaster follow distinct patterns (event signature) depending on the type of the disaster [8], it is possible to envision a system that uses the model to estimate the expected frequencies of the tweets relevant to a particular disaster at a point of time and utilize this information in building a better extractor for situational tweets. Formalizing our understanding of how the model can be used for augmenting the classifier is a direction for future work, which could deeply aid both the retrospective search of relevant tweets for a past disaster and the near-real-time extraction of ongoing disaster responses.

Moving ahead, we plan to accurately label (using crowd-sourcing) a portion of the Twitter decahose to create a gold standard for the evaluation of situational awareness extraction systems. We also plan to study the effect of the event signatures on the filter performance in more detail.

## References


[1] Endsley, M. R. (1995). *Toward a theory of situation awareness in dynamic systems.* Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 32-64.

[2] Verma, S., Vieweg, S., Corvey, W., Palen, L., Martin, J. H., Palmer, M., ... & Anderson, K. M. (2011). *NLP to the Rescue?: Extracting" Situational Awareness" Tweets During Mass Emergency.* In Fifth International AAAI Conference on Weblogs and Social Media.

[3] Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). *Microblogging during two natural hazards events: what Twitter may contribute to situational awareness.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM.

[4] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). *Earthquake shakes Twitter users: real-time event detection by social sensors.* In Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.

[5] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013, May). *Practical extraction of disaster-relevant information from social media.* In Proceedings of the 22nd international conference on WWW companion (pp. 1021-1024).

[6] Gao, H., Barbier, G., Goolsby, R., & Zeng, D. (2011). *Harnessing the crowdsourcing power of social media for disaster relief.* Intelligent Systems, IEEE 26(3):10-14.

[7] Heverin, T., & Zach, L. (2010). *Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in the Seattle-Tacoma, Washington, Area.* ISCRAM.

[8] Saleem, H. M., Xu, Y., & Ruths, D. (2014). *Effects of Disaster Characteristics on Twitter event signatures.* In Proceedings of Humanitarian Technology: Science, Systems and Global Impact 2014.

[9] Saleem, H. M., Xu, Y., & Ruths, D. (2014). *Novel Situational Information in Mass Emergencies: What does Twitter Provide?* In Proceedings of Humanitarian Technology: Science, Systems and Global Impact 2014.

[10] Kang, P., & Cho, S. (2006, January). *EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems.* In Neural Information Processing (pp. 837-846). Springer Berlin Heidelberg.

[11] Japkowicz, N., & Stephen, S. (2002). *The class imbalance problem: A systematic study.* Intelligent data analysis, 6(5), 429-449.

[12] Lenhart, A., Purcell, K., Smith, A., & Zickuhr, K. (2010). *Social Media & Mobile Internet Use among Teens and Young Adults. Millennials.* Pew Internet & American Life Project.

[13] Smith, A. (2010). *Government online.* `http://www.pewinternet.org/Reports/2010/Government-Online.aspx`, accessed on Jan, 25, 2011.

[14] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). *Social media? Get serious! Understanding the functional building blocks of social media.* Business horizons, 54(3), 241-251.

[15] Palen, L. (2008). *Online social media in crisis events.* Educause Quarterly, 31(3), 12.

[16] Wright, D. K., & Hinson, M. D. (2009). *Examining how public relations practitioners actually are using social media.* Public Relations Journal, 3(3), 1-33.

[17] Makinen, M., & Kuira, M. W. (2008). *Social media and postelection crisis in Kenya.* The International Journal of Press/Politics, 13(3), 328-335.

[18] Hughes, A. L., & Palen, L. (2009). *Twitter adoption and use in mass convergence and emergency events.* International Journal of Emergency Management, 6(3), 248-260.

[19] De Longueville, B., Smith, R. S., & Luraschi, G. (2009). *OMG, from here, I can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires.* In Proceedings of the International Workshop on Location Based Social Networks.

[20] Soboroff, I., Ounis, I., Lin, J., & Soboroff, I. (2012). *Overview of the TREC-2012 microblog track.* In Proceedings of the Twenty-First Text REtrieval Conference (TREC 2012).

[21] National Institute of Standards and Technology. (2013). *TREC 2012 Microblog Track Filtering Task Results.*, Available from World Wide Web: `http://trec.nist.gov/pubs/trec21/appendices/microblog-filtering.html`. accessed on Jun, 20, 2014.

[22] Robertson, S. E., & Soboroff, I. (2002, November). *The TREC 2002 Filtering Track Report.* In TREC (Vol. 2002, No. 3, p. 5).

[23] Han, Z., Li, X., Yang, M., Qi, H., Li, S., & Zhao, T. (2012). *Hit at trec 2012 microblog track.* In Proceedings of Text REtrieval Conference.

[24] Roegiest, A., & Cormack, G. V. (2012). *University of waterloo: Logistic regression and reciprocal rank fusion at the microblog track.* WATERLOO UNIV (ONTARIO), In Proceedings of Text REtrieval Conference.

[25] Berardi, G., Esuli, A., & Marcheggiani, D. (2012). *ISTI at TREC Microblog Track 2012: Real-Time Filtering Through Supervised Learning.* Consiglio Nazionale delle Richerche Pisa.

[26] Lin, Y. R., Margolin, D., Keegan, B., Baronchelli, A., & Lazer, D. (2013). *# Bigbirds Never Die: Understanding Social Dynamics of Emergent Hashtag*. arXiv preprint arXiv:1303.7144.

[27] Graham, P. (2003). *A plan for spam*, 2002. Available from World Wide Web: `http://www.paulgraham.com/spam.html`. accessed on Jun, 20, 2014.