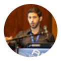
Identifying Disruptive Events from Social Media to Enhance Situational Awareness

Conference Paper · August 2015

DOI: 10.1145/2808797.2808879



Nasser Alsaedi
Cardiff University
12 PUBLICATIONS **298** CITATIONS

Pete Burnap
Cardiff University
149 PUBLICATIONS **7,449** CITATIONS

Omer F. Rana
Cardiff University
779 PUBLICATIONS **16,244** CITATIONS

2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Identifying Disruptive Events from Social Media to Enhance Situational Awareness

Nasser Alsaedi, Pete Burnap and Omer Rana
Cardiff School of Computer Science & Informatics, Cardiff University
{N.M.Alsaedi, P.Burnap, O.F.Rana}@cardiff.ac.uk

***Abstract*—Decision makers use information from a range of terrestrial and online sources to help underpin the processes through which they develop policies and react to events as they unfold. One such source of online information is social media. Twitter, as a form of social media, is a popular micro-blogging Web application serving hundreds of millions of users. User-generated content can be exploited as a rich source of information for identifying 'real-world' *disruptive events*. In this paper, we present an in-depth comparison of three types of features that could be useful for identifying disruptive events: temporal, spatial and textual. We make several interesting observations: first, disruptive events are identifiable regardless of the "influence of the user" discussing them, and over a variety of topics. Second, temporal features are the best event identifiers and hence should not be disregarded or ignored. Third, a combination of optimum textual features with temporal and spatial features achieves best performance in the event detection task. We believe that these findings provide new insights for gathering information around real-world events as well as a useful resource for improving situational awareness and decision support.**

Keywords—Data Mining, Event Detection, Feature Selection.

I. INTRODUCTION

Event identification is a concept that is crucial in event management, intelligence gathering, and decision-making. "Open source" Web-based information posted to social media sites is increasingly used as a new source of signals intelligence [2] in support of situational awareness. The Homeland Security Act of 2002 defines the term "situational awareness" as "*information gathered from a variety of sources that, when communicated to emergency managers and decision makers, can form the basis for incident management decision-making.*"[10]

Situational awareness refers to a state of understanding what is happening around you and then anticipating or predicting how it will change with time and with respect to the dynamics of the surrounding environment [3, 10]. Public safety agencies work to establish and maintain situational awareness using various sources of traditional methods [10]. The rise in popularity of social media now enables the public to request, share, and provide information in real-time using text, videos and photos. For these sources to become useful, however, various data management and analytics methods must be developed and critiqued. If integrated with traditional data, social media can help decision makers achieve and maintain situational awareness in real-time [3, 10, 12].

People tend to comment on real-world events (both local and global), when a topic suddenly captures their attention. From a situational awareness perspective, identifying events and specifically *disruptive events* – sub-events that threaten social safety and security, or could cause disruption to social order – from social media, is a key research challenge. To date, social media intelligence has helped to measure and understand hateful communications following terrorist attacks [28,29] and to observe the communications of militia leaders and terror groups such as ISIS [2, 3, 12]. However, there are several ongoing challenges. A key challenge is to distinguishing ambiguous content about everyday mundane activity, from events of public interest, in particular those that might impact on public safety. If this could be achieved then those with responsibility for managing and ensuring public safety would be able to use this information to better manage a potentially harmful situation. Understanding the features of social media content that characterise disruptive events is therefore the key motivation behind this work.

One way to optimize the identification of the patterns and signals that indicate an event is to undertake feature selection experiments. Because not all features are expected to lead to better system performance or contribute equally towards improved machine classification and/or clustering accuracy, we seek to evaluate the effectiveness of a range of features for identifying events, and, further, features that would distinguish 'normal' events from *disruptive* events. These features may be divided as follows:

- Temporal features: related to the "speed" of information diffusion over time by highlighting the "quality" of content created by users in different time frames;
- Spatial features: to approximate the *origin of posted content* or to estimate the *location of a user* or to reveal the *location of an event*; and
- Textual features: which are representative of the text published content.

In [1] we proposed a probabilistic framework that used Twitter data as a form open and programmatically accessible, real-time content to identify events and distinguish these from content not about events. We provided some initial results in identifying disruptive

ASONAM '15, August 25-28, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2808797.2808879

events. In this paper we focus specifically on optimizing feature selection to increase the performance results of event classification, and to reduce the number of features required to lower the computational overheads during calculation. Our optimized approach improved the identification accuracy from 80.85% in [1] to 83.27%, which is a significant result in event identification tasks. In addition, these results are also achieved using features selected with consideration for computational resource usage, which is important when analyzing real-time data stream surrounding ongoing events. The contributions of this paper are:

1. to present an improved model for feature selection that is suitable for microblog data such as Twitter – a real-time streamed data source
2. to explore three features; temporal, spatial and textual as well as combinations of them in order to optimise computational resource usage when analysing real-time events;
3. to identify features that improve system performance with the aim of distinguishing disruptive events from other events;
4. to perform extensive feature analysis and selection in order to demonstrate that these features contribute differently in the decision-making process regarding real-time disruptive event management.

The rest of the paper is organized as follows. In section 2 we summarise related research on feature selection. In section 3 we discuss the method used for feature selection and details of the temporal, spatial and textual features. Section 4 presents experiments and discusses the results. We conclude by highlighting some future directions for research in section 5.

II. RELATED WORK

Atefeh and Khreich [24] provide a comprehensive survey of techniques for event detection using Twitter. In this paper we focus on feature selection experimentation to determine the optimal selection of features to enhance disruptive event detection accuracy and reduce computational resource usage. In this section, we summarise the current research on temporal, spatial and textual features and the way in which it has been applied to data mining tasks.

Temporal features

Social media posts generally come with a creation time-stamp, which can be utilized for topic detection and tracking (TDT) as demonstrated in [15], which analysed the evolution of stories and topics over time. Similarly, Gabrilovich et al. [16] studied the dynamics of information novelty in some evolving news stories. There has also been much work on the community structure of the blogosphere. The authors of [13] showed that the prediction of information cascades is feasible and the relative growth of a cascade becomes more predictable as more "reshares" are observed over time – hence, these temporal features are key predictors. Rather than attempt to predict cascades, Elsas and Dumais [17] studied the dynamics of document content change to the rank documents on the basis of their temporal characteristics.

Spatial features

Several algorithms have been proposed to estimate the location of Twitter users by means of a content analysis of tweets. Eisenstein et al. [26] built geographic topic models to predict the location of Twitter users in terms of regions (reporting 58% accuracy over 4 regions) and states (predicting 48 US states with 24% accuracy). Hecht et al. [6] built Bayesian probabilistic models from words in tweets for discovering the country and state-level location of Twitter users. They were able to get approximately 89% accuracy with countries (4 countries), but only 27% accuracy for predicting states (50 states in the US). Cheng et al. [5] described a city-level location estimation algorithm, which is based on identifying local words from tweets using statistical predictive models. They achieved approximately 50% accuracy in detecting city-locations. More recently, Mahmud et al. [23] have combined time zone information and content-based classifiers in a hierarchical model at different granularities, reporting accuracy rates of 64% for cities, 66% for states, 78% for time zones and 71% for regions. Our use of spatial features relates to their predictive power when aiming to identify disruptive events – essentially, whether neighbourhood-, city-, or country-level information is a significant predictor.

Textual features

Textual features can be used as individual features (e.g. n-grams), but many studies have combined them to optimise the solution to data mining challenges, such as information diffusion [8, 12, 13], opinion mining [9, 14], spam and spammer detection [18], and identifying the most knowledgeable posts and famous users [4, 6, 7, 8, 19].

Using the topic model in [7], a set of raw features (number of original tweets, number of retweets, and number of mentions) is used for identifying the most influential Twitter users. Agarwal et al. [14] investigated two kinds of model: a feature based model and a tree kernel based model for the purpose of sentiment classification. They demonstrated that both models outperformed the unigram baseline model which was previously shown to work well for Twitter sentiment analysis. Hashtag popularity was considered by Ma, Sun, and Cong in [8]. They demonstrated that contextual features (such as the number of users, number of tweets, retweet ratio, etc.) are more effective than content features (such as tweets containing URL, the ratio of neutral, positive, and negative tweet, etc.) in predicting hashtag popularity.

III. FEATURE SELECTION

In [1], we defined a disruptive event in the context of social media data as:

"An event that interferes in the achieving of the objective of an event or interrupts ordinary event routine. It may occur over the course of one or several days, causing disorder, destabilizing securities and may result in displacement or discontinuity."

Figure 1 shows our proposed framework, which enables us to automatically identify meaningful events from Twitter. The method is based on collecting data for a given location over a predefined time frame. The five-step framework consists of data collection, pre-processing, classification, on-line clustering and summarization. (See [1] for full details of each step and the framework evaluation). In this paper, we hypothesize that a disruptive event can be characterized by three sets of features: temporal, spatial and textual. We report our work on improving feature identification for the online clustering part of the framework; this has significantly improved the accuracy of event detection.

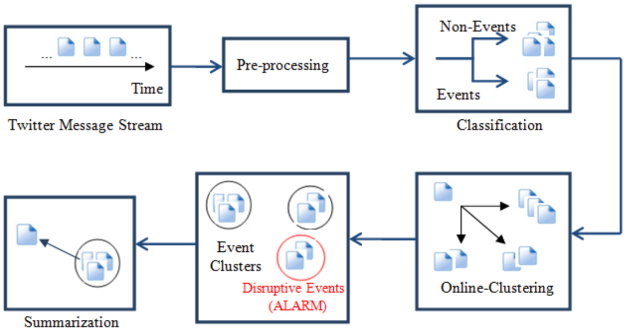


Fig. 1. Twitter Stream Event Detection Framework

Not all features are expected to improve the system's performance or lead to more accurate discrimination of the clustering algorithm. Indeed, for many reasons the inclusion of some features could result in worse behavior by the system such as introducing greater computational cost [11], may lead to overfitting [11, 25] or could result in some scalability issues [18, 21]. Thus, feature selection is a fundamental problem in mining large data sets. The problem is not limited to the total processing time but involves reducing the dimensionality to achieve better generalization. In this work, we chose for several reasons to implement an improved version of the unsupervised feature selection presented in [25]: first, it resolves the issue of the high-computational complexity involved in searching large data sets. Second, the computation time is reasonable even for large data sets where other algorithms perform well only with medium sized data sets. Third, the unsupervised feature selection results are among the best clustering performances of real-world data sets.

Let the original number of features be D and the original feature set be $O = \{F_i, i = 1, \dots, D\}$. We represent the dissimilarity between features F_i and F_j by $S(F_i, F_j)$. The higher the value of S, the more dissimilar the features. Let r_i^k represent the dissimilarity between feature F_i and its kth nearest-neighbor feature in R, where R is the reduced feature subset.

The dissimilarity between two features $\mathrm{S}(\mathrm{F_i}, \mathrm{F_j})$ is calculated by the Maximal Information Compression Index (MICI) which was proposed by Mitra et al. in [25]. The MICI is a well-known index for measuring dissimilarity between features and it has been applied in many pattern recognition and data mining tasks. The Maximal Information Compression Index is defined as:

$$\lambda(x,y) = \left[a - \sqrt{a^2 - 4b(1 - \rho(x,y)^2)}\right]/2$$

where $a = var(x) + var(y)$ and $b = var(x).var(y)$

The *correlation coefficient* is defined as $(x,y) = \frac{cov(x)}{\sqrt{b}}$, var() denotes the variance of a variable, and cov() the covariance between two variables.

Algorithm 1. Feature Selection Algorithm

```
Step 1: Choose an initial value of $k \leq D - 1$. Initialize the reduced feature subset $R$ to the original feature set $O$.
Step 2: For each feature $F_i \in R$, compute $r_i^k$
Step 3: Find feature $F_{i'}$ for which $r_{i'}^k$ is minimum.
Retain this feature in $R$ and discard $k$ nearest features of $F_{i'}$.
Let $\varepsilon = r_{i'}^k$
Step 4: If $k >$cardinality$(R) - 1$: $k =$ cardinality$(R) - 1$
Step 5: If $k = 1$: Go to Step 8.
Step 6: While $r_{i'}^k > \varepsilon$ do:
    a)  $k = k - 1,\ r_{i'}^k = \inf_{F_i \in R} r_i^k$
    b)  If $k = 1$: Go to Step 8.
        End While
Step 7: Go to step 2.
Step 8: Return feature set $R$ as the reduced feature set.
```

A. *Temporal Features*

Temporal features are important factors that have been overlooked in many event detection studies via social media. The volume of tweets and the continually updated commentary around an event suggest that informative tweets from several hours ago may not be as important as new tweets [21]. For this reason we identify the most frequent terms in the cluster across a range of time windows. In our experiments we use a range of time windows to improve the efficiency of the event clustering system in terms of accuracy and total running time.

These time windows are related to the “speed” of diffusion over time. This has been shown to be an important feature in predicting thread length on Facebook [4], a primary mechanism in predicting popularity in Twitter [8], and as the most important factor in influencing a cascade through the network [13].

B. *Spatial Features (Geospatial, Regional)*

Events are characterized by rich set of spatial and demographic features [1]. In this paper, we make use of three statistical location approaches to extract geographic content from clusters. The first one is from Twitter where the source latitude and longitude coordinates are provided by the user. The second method depends on the shared media (photos and videos) by using the GPS coordination of the capture device (if supported). Third, Open NLP (http://opennlp.sourceforge.net) and Named-Entity Recognition (NER) were implemented for geotagging the tweet content (text) to identify places, organization, street names, landmarks etc. These approaches rely purely on

Twitter with no need for user’s IP address, private login information, or external knowledge bases which give the maximum advantage [5, 23].

Once the geographic content is extracted from each tweet in a cluster, we aggregate them to determine the cluster's overall geographic focus. The higher the volume of tweets from nearby coordinates, the higher the level of confidence in the location of the event will be.

The spatial feature has been shown to be a weak event indicator due to the slow adoption of geospatial features from Twitter users as shown in [5]. We examined users' locations in our dataset which contains around 700 thousand users, we found that 12.7% (56,818 users) of the total user profiles list their locations as granular as a city name, 7.7% contain a country name and that only 15,217 (4.6%) reveal their locations as a latitude/longitude coordinate. Overall, most users tend to over generalize their location (e.g., East Region), have missing altogether, or nonsensical (e.g., Middle of the Desert) location. In addition, Twitter users often rely on shorthand and non-standard vocabulary (non-traditional gazetteer terms) for informal communication or simply users do not wish to revel their location which makes determining location-terms a non-trivial task [23]. Our results also show the differences in user behaviour across regions, languages and backgrounds across the globe (compared with [5]).

We assume that all locations provided by users are correct, although [6] found that 34% of Twitter users had entered fake locations in their profile. Some users may intentionally misrepresent their home location either to cover for their actual location, or due to privacy concerns. In addition, some users provided locations that differ from their actual location at the time because they are tweeting as they travel.

C. *Textual Features*

There are two main tasks in this paper regarding textual features: first, we analyse various textual features in order to select the best contributors to the task of event detection. Second, we rank features using performance measures and eliminate irrelevant features that introduce computational cost. First we introduce these features in detail.

- ❖ Near-Duplicate measure

The average content similarity over all pairs of tweets posted in a (1-hour time slot) cluster is calculated using:

$$\sum_{a,b \in set\ of\ pairs\ in\ tweets} \frac{similarity\ (a,b)}{|set\ of\ pairs\ in\ tweets|}$$

where the content similarity is computed using the cosine similarity over words from tweet a, b vector representation $\vec{V}(a), \vec{V}(b)$ of the tweet content:

$$similarity\ (a,b) = \frac{\vec{V}(a).\vec{V}(b)}{|\vec{V}(a)||\vec{V}(b)|}$$

- ❖ Retweet ratio

Retweeting represents the influence of a tweet beyond the one-to-one interaction domain. Popular tweets can propagate multiple hops away from the source as they are retweeted throughout the network [7]. Hence, the number of retweets can be used as an indication of popularity [19]. We calculate this attribute by normalizing the number of times a tweet appears in a timeframe to the total number of tweets in that timeframe.

- ❖ Mention ratio

A mention is a mechanism used in Twitter to reply to users, engage others or to join a conversation in a form of (@username). Regarding event reporting, users tend to mention journalists, politicians and official accounts such as news agencies or government official accounts to drive their attention about an event or to add more credibility to their event-related posts.

- ❖ Hashtag ratio

Hashtags are an important feature of social networking sites and can be inserted anywhere within a message. Some Hashtags indicate their posted messages (#bbcF1) and others are dedicated originally to events such as (#abudhabigp). In addition, topic related hashtags are used as an information seeking index on Twitter to search Twitter for more tweets belonging to the same topic [12]. The use of hashtags became a coordinating mechanism for disruptive activity on Twitter [1, 27]. The Hashtag ratio is the ratio of tweets containing hashtag over the total number of tweets in that timeframe.

- ❖ Link or Url ratio

As Twitter is limited to 140 characters per message it is common in the Twitter community to include links when tweeting to share additional information or for referencing. The co-occurrence of URLs in a cluster confirms that these tweets refer to the same event and improves the level of confidence in the event. This attribute is calculated by the fraction of tweets with URL to the total number of tweets in a timeframe.

- ❖ Tweet sentiment

Users post real-time messages in microblogging websites giving their opinions on a variety of topics (e.g. news) using positive or negative sentiment [9, 14]. Here, we first study whether sentiment polarity posts (0 indicates neutral, 1 indicates positive or negative sentiment) are significant features when reporting events. Subsequently, we investigate the influence of positive, negative and neutral sentiment on identifying disruptive events.

To calculate sentiment we use a semantic classifier based on the use of SentiStrength algorithm [9] which is suitable because it is designed for short informal text with abbreviations and slang. For a tweet, the SentiStrength algorithm computes a positive and a negative sentiment score. Then we compute the average cluster-level sentiment (set of tweets) in order to study the effect of average positive or negative sentiment with respect to events.

- **Dictionary-based feature**

One of the main objectives of this work is the ability to automatically detect messages about disruptive events such as a labor strike or a fire. To enrich such rare event identification, present tense verbs, popular event nouns and adjectives that describe events as they take place are considered typical features. This bag of words model uses a dictionary of trigger words to detect and characterize disruptive events which are manually labeled by experts from several management departments: traffic control, crisis, emergency departments, and others.

Examples of present verbs: witness, notice, observe, participate, engage, listen etc.

Examples of event nouns: breaking news, update, delay etc.

Examples of event adjectives: urgent, live, latest, severe, horrifying etc.

IV. EXPERIMENTAL EVALUATION

A. Experimental Settings

Dataset: Our dataset consists of 1,698,517 tweets, and was collected from 15 October 2013 to 05 November 2013 using Twitter's Streaming API. We have chosen to study a major sporting event - the Formula 1 Motor Racing Grand Prix - due to its global interest. The event was hosted in Abu Dhabi between 1st and 4th November 2013.

Framework Evaluation: We sampled 85,000 event-related tweets from the study dataset using the Naïve Bayes event classifier as presented in [1], which we used to train, test and evaluate our clustering algorithm. We used the first 15 days of data (from 15 Oct until 29 Oct) to train the clustering algorithm and to tune the thresholds using the validation set. Then we tested the clustering algorithm on unseen data from the 6 days between 30 Oct and 4 Nov. Threshold values were varied from 0.10 to 0.90 at graded increments of 0.05% with a total of 17 tests, in order to find the best cut-off of τ =0.45 (63 character difference).

Figure 2 illustrates the F-measure scores for different thresholds where the best performing threshold τ =0.45 seems to be reasonable because it allows some similarity between posts but does not allow them to be near-identical. In order to evaluate the clustering performance we employed three human annotators to manually label 800 clusters where the most highly retweeted post represented that cluster and the surrounding tweets were assumed also to represent the event. The task of the annotators was to choose one of the eight different categories: politics, finance, sport, entertainment, technology, culture, disruptive event and others. The agreement between annotators was calculated using Cohen's kappa (K=0.794) which indicates an acceptable level of agreement. We used only **635 clusters** which all annotators agreed o as the **gold standard**. The framework was able to achieve an average F-measure of 80.85 using all features (temporal, spatial and textual).

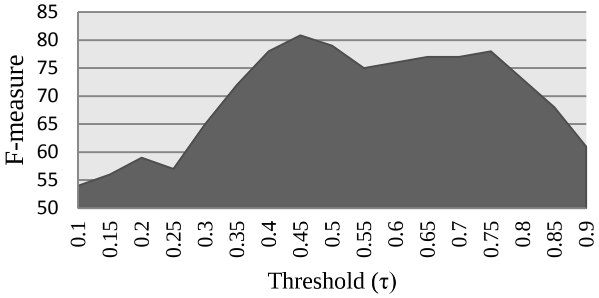


Fig. 2. F-measure of online clustering over different thresholds

B. Evaluation Matrix

To measure the effectiveness of the classifiers based on our proposed features, we used the standard classification metrics of precision, recall, accuracy, and F1 measure. Precision is a measure of false positives. Recall is a measure of false negatives. The F-measure is a harmonized mean of precision and recall. Accuracy is the proportion of correctly classified tweets to the total number of tweets.

$$\text{Precision(P)} = \frac{\text{tp}}{\text{tp+fp}} \qquad \text{Recall(R)} = \frac{\text{tp}}{\text{tp+fn}}$$

$$\text{F} - \text{measure} = \frac{2\times\text{P}\times\text{R}}{\text{P+R}} \qquad \text{Accuracy} = \frac{\text{tp+tn}}{\text{tp+fp+fn+tn}}$$

The discrimination power between different proposed features can be measured by generating a Receiver Operating Characteristics (ROC) curve [22]. ROC curves plot false positive rates on the horizontal axis and true positive rates on the vertical axis for varying thresholds. The closer the ROC curve is to the upper left corner, the higher is the overall accuracy. The coordinate (0, 1) represents 100% sensitivity (no false negatives) and 100% specificity (no false positives).

C. Experiment 1

In the first set of experiments we study each feature individually. We use accuracy and running time to select the best temporal setting. However, we use feature selection method (outlined in Algorithm 1) to eliminate spatial and textual features.

Temporal features

Here we analyse the efficiency of the proposed temporal features in terms of the *event prediction accuracy* (A) and the *total running time* (RT). We calculate A (Figure 3) and RT (Figure 4) for a range of time windows; 1 minute, 30 minutes, 1 hour, 3 hours, 6 hours, 12 hours and 24 hours. To attain the best value for temporal features, we had to look at the following optimisation problem:

$\{Clustering\ Accuracy - k \cdot Running\ Time\}$.

where *k* is the threshold which maximizes the criterion.

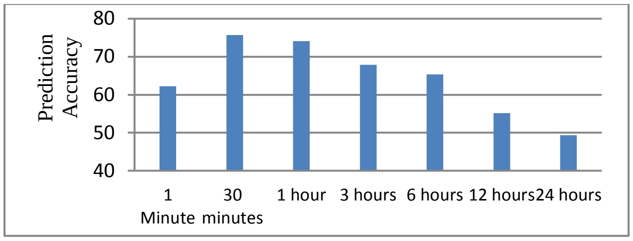


Fig. 3. Accuracy (A) obtained using various temporal settings

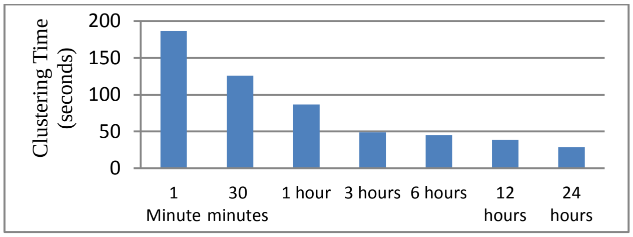


Fig. 4. Efficiency comparison with various temporal granularities (s)

The results presented in Figures 3 and 4 show that the 1–hour time window requires much less computational clustering time than for the 1 minute and 30 minute windows, while producing the second best level of accuracy after the 30 minute window. This suggests that tweets published recently are better predictors of events than older tweets, but also that a lead-in time is required (since 1 minute is too short to provide the same level of accuracy). Clustering the tweets every hour provides a small reduction in the clustering accuracy but significantly reduces the computational processing requirements; therefore for the remaining experiments we set the time window for clustering to 1 hour. Additionally, we find that reporting disruptive events are more likely in 1 hour than are for other events.

Spatial features

Here we investigate the geospatial features on our framework and evaluate them according to the three levels of the Twitter user's location: country, city and neighbourhood. Table 1 shows the results:

TABLE 1. RECALL COMPARISON USING DIFFERENT LOCATION GRANULARITIES

Location Level	**Neighborhood (Local)**	**City (Intermediate)**	**Country (General)**
Recall	49.04	53.22	17.64

As can be seen from Table 1, the city-level provides the best overall results, capturing events with around 53.2% accuracy within the city where they occurred. Comparing neighbourhood-level to city-level, we attain similar but slightly better results for the city approach, suggesting that geo-location at the level of neighbourhoods (which is more difficult to obtain via Twitter) is not necessarily required to detect events. An alternative interpretation of this result is that the location detection tools that we used could not handle the misspellings and colloquial terms used for neighbourhood level locations.

Table 1 also shows that the performance of the country-level classifier is much worse than that of other classifiers as a result of user's behaviour. Many users attempt to be more general in their tweets than mere neighborhood-level in order to get more attention. Yet they are trying to be more specific than a country or a region because of the possibility of multiple events within the same time interval. Hence users typically insert hashtags of the city (#abudhabi or #Dubai) to their tweets rather than country (#UAE). These results provide some evidence to suggest that events are inherently difficult to identify the basis of the spatial features on their own.

When it comes to disruptive events, people tend to use city names rather than country or neighbourhood names. For instance, if a user comments on a crime or a terrorist attack or other disruptive event, such as severe weather, she tends to include the city name or she might add the hashtag of the city.

Textual features

Here we investigate the discriminative power of each individual feature in classifying disruptive events in order to show the robustness of each feature individually so the least discriminative features can be removed to reduce the computational workload required to compute the results. The results are shown in Figure 5 and Table 2. Figure 5 shows the ROC curve for each feature and Table 2 presents the performance results according to the F-measure and the difference between the F-measure of each the baseline (Temporal feature is selected as the baseline).

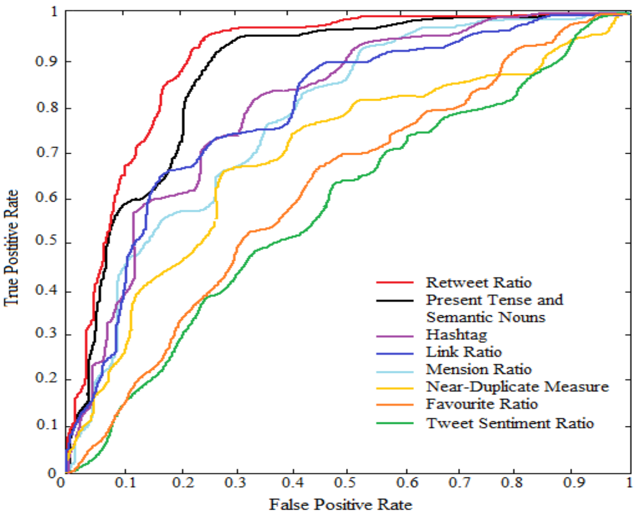


Fig. 5. ROC curves of the various proposed features

The near-Duplicate measure, Favorite ratio and Sentiment ratio are the least discriminative features, which suggest that they appear in all the different types of event, not only in disruptive ones. But the Dictionary-based model, Retweet ratio and Hashtag ratio are the most discriminative, suggesting that references to present time and references to descriptive terms (e.g. live, breaking etc.) are good discriminators. The retweet ratio suggests that other users pick up on event commentaries and propagate them further through the network.

TABLE 2. COMPARISON OF THE PERFORMANCE USING VARIOUS TEXTUAL FEATURES.

Model	F-measure	F-measure Diff
Baseline (Temporal)	74.14	-
Near-Duplicate measure	74.69	0.55
Retweet ratio	77.57	3.43
Mention ratio	75.73	1.59
Hashtag ratio	76.13	2.99
Link or Url ratio	76.81	2.67
Favorite ratio	74.16	0.02
Tweet sentiment polarity	73.63	-0.51
Dictionary-based feature	77.43	3.29

Linking content features such as Hashtags and URLs are also very predictive of events, suggesting that tweets reporting events provide evidence or further information (via URL), or are bound to an event and made more discoverable via a self-defined topic discriminator in form of a Hashtag. Overall, all proposed features have a positive improvement over the baseline except for sentiment polarity, which is further investigated in the next section.

Tweet sentiment

The goal of this experiment is to examine whether positive, neutral or negative sentiment tweets have an effect in reporting disruptive events. The main observation made from Table 3 is that tweets with negative sentiment lead to a better F-measure than the baseline (temporal) and other sentiment measures. Therefore, negative tweet sentiment has a high adoption rate regarding disruptive tweets. Due to the fact that reporting disruptive events usually involves negative terms and sentiment whereas events in general can be positive, negative or neutral. Another possible reason is that tweets with negative sentiment are more likely to be retweeted, as shown in [6, 8, 9].

TABLE 3. F-MEASURES FOR POSITIVE, NEURAL AND NEGATIVE SENTIMENT MODELS, WHICH CLEARLY SHOWS THAT THE NEGATIVE MODEL OUTPERFORMS OTHERS BY AT LEAST 1.43%

Model	**F-measure**	**F-measure Diff.**
Positive sentiment ratio	74.27	0.13
Neutral sentiment	74.40	0.26
Negative sentiment ratio	75.83	**1.69**

Ranking top textual features

After investigating each feature individually, and further probing the tweet sentiment, we discarded features with less than a 1.60 improvement over the temporal baseline, to reduce the complexity of the clustering processes and therefore the computational overhead. Only the most predictive features were used to identify the disruptive events tweets. The ranking of the most influential textual features is presented in Table 4.

TABLE 4. THE MOST EFFECTIVE FEATURES (ABOVE 1.50 DIFFERENCES)

Rank	**Feature**	**F-measure Diff**
1	Retweet ratio	3.43
2	Dictionary-based feature	3.29
3	Hashtag ratio	2.99
4	Link or Url ratio	2.67
5	Negative sentiment ratio	1.69

D. *Experiment 2*

We used a unigram model as our baseline for this experiment which is a bag-of-words textual features model (the dictionary-based feature). Figure 6 compares the performance of various models: First, we use individual feature models: temporal, spatial and textual. The temporal model uses the 1-hour setting, spatial model implements the city-level setting and the textual model uses all the features from Table 2. Second, a combination of features model: (Temporal + Spatial), (Temporal + Textual) and (Temporal + Spatial +Textual). Third, to build our optimized model we make use of the 1-hour temporal feature, city-level spatial feature and the most effective textual features from Table 4.

Overall, while each feature set is individually significantly better than the baseline, it is the temporal feature that substantially outperforms all other features, obtaining a performance score of 7.64% over textual features and 20.92% compared with the spatial features. As shown in experiment 1, using a 1-hour time window is the most effective in detecting disruptive events. This effect is less than with spatial and textual features. Using the textual feature set without temporal features, we are still able to obtain reasonable performance of 66.5%, but that is not as distinctive for disruptive events as when the temporal feature is applied. That is emphatically not the case when using only spatial features because it is a weak indicator to implement on its own.

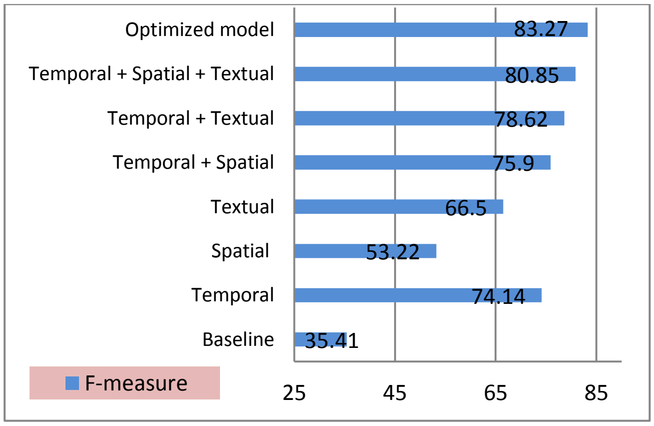


Fig. 6. Comparison of different models on the event identification task according to the F-measure. Higher is better.

Combining temporal and spatial features gives the best of both with much better performance of F-measure (75.9). More interestingly, integrating the temporal and textual features results in a better system performance than using each feature independently. It also outperforms the combination of temporal + spatial by 2.72.

A combination of all three features results in the best performance, but a further investigation which removed unnecessary textual features gave the best model performance overall. The optimized model achieved an F-score of 83.27 higher than the combination of all the features. These results support our claim that not all features are expected to improve a system's performance; instead they all contribute differently to detecting disruptive events.

V. CONCLUSION

In this paper, we presented an extensive analysis of various features related directly to Twitter data and showed how they can be used discriminatively to distinguish between disruptive events and other events. The results identify that it is not adequate to consider temporal, spatial, or content-based aspects in isolation. Rather, a combination of features covering all these aspects leads to a robust system which allows the best event detection results to operate. Our results support the claim that the use of social media for the purposes of information gathering could be adopted as complementary to traditional intelligence.

There are many directions for future work. One of the main options is to compare and validate the performance of the proposed framework against other well-known algorithms such as the state-of-the-art Labeled Dirichlet Allocation (LDA) method. Another direction is to improve the automatic summarization task of microblog posts. Finally, the detection of rumors in social media, the analysis of the distinctive characteristics of rumors and the way they propagate in the microblogging communities is to be carried out in the near future.

REFERENCES

[1] Alsaedi, N., Burnap, P. and Rana, O. 2014. A Combined Classification-Clustering Framework for Identifying Disruptive Events. SocialCom 2014, pp. 1–10.

[2] Julian E. Barnes (2014) U.S. Military Plugs Into Social Media for Intelligence Gathering, The wall street journal Politics and policy.

[3] Oh, O., Agrawal, M. and Rao, H.R. 2011. Information control and terrorism: Tracking the Mumbai terrorist attack through twitter. Information Systems Frontiers 13(1), pp. 33–43.

[4] Backstrom, L., Kleinberg, J., Lee, L. and Danescu, C. 2013. Characterizing and Curating Conversation Threads: Expansion, Focus, Volume, Re-entry, Wsdm, pp. 13–22.

[5] Cheng, Z., Caverlee, J. and Lee, K. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. CIKM '10 pp. 759–768.

[6] Hecht, B., Hong, L., Suh, B. and Chi, E. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. CHI '11, pp. 237–246.

[7] Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. ICWSM'10.

[8] Ma, Z., Sun, A. and Cong, G. 2013. On predicting the popularity of newly emerging hashtags in twitter. Journal of the American Society for Information Science and Technology 64(7), pp.1399-1410.

[9] Thelwall, M., Buckley, K. and Paltoglou, G. 2011. Sentiment in Twitter events. Journal of the American Society for Information Science and Technology 62(2), pp. 406–418.

[10] Department of Homeland Security. 2014. Using Social Media for Enhanced Situational Awareness and Decision Support. Version 5.0. June 2013, pp. 1–44.

[11] Cui, Y., Wong, W. and Cheung, D. 2009. Privacy-Preserving Clustering with High Accuracy and Low Time Complexity. DASFAA 2009, pp. 456–470.

[12] Mills, A., Chen, R., Lee, J. and Rao, H.R. 2009. Web 2.0 emergency applications: how useful can Twitter be for emergency response? Journal of Information Privacy & Security 5(3), pp. 3–26.

[13] Cheng, J., Adamic, L., Dow, P., Jon, K. and Jure, L. 2014. Can cascades be predicted? WWW '14.

[14] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. 2011. Sentiment analysis of twitter data. Proceedings of the ACL 2011 Workshop on Languages in Social Media, pp. 30–38.

[15] James, A. 2002. Introduction to topic detection and tracking. In Topic detection and tracking: event-based information, pp. 1–16.

[16] Gabrilovich, E., Dumais, S. and Horvitz, E. 2004. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. WWW '13, pp. 482–490.

[17] Elsas, J.L. and Dumais, S.T. 2010. Leveraging temporal dynamics of document content in relevance ranking. WSDM '10.

[18] Lee, K., Caverlee, J. and Webb, S. 2010. Uncovering social spammers: social honeypots+ machine learning. SIGIR '33.

[19] Petrovic, S., Osborne, M. and Lavrenko, V. 2011. RT to Win! Predicting Message Propagation in Twitter. ICWSM'11.

[20] Dou, W., Wang, X., Skau, D., Ribarsky, W. and Zhou, M.X. 2012. LeadLine: Interactive visual analysis of text data through event identification. VAST 2012, pp. 93–102.

[21] Becker, H., Naaman, M. and Gravano, L. 2011. Beyond Trending Topics: Real- Event Identification on Twitter. ICWSM'11, pp. 1–17.

[22] Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27(8), pp. 861–874.

[23] Mahmud, J., Nichols, J. and Drews, C. 2012. Where Is This Tweet From? Inferring Home Locations of Twitter Users. ICWSM'12, pp. 511–514.

[24] Atefeh, F. and Khreich, W. 2013. A Survey of techniques for event detection in twitter. Computational Intelligence 0(0).

[25] Mitra, P., Murthy, C. a and Pal, S.K. 2002. Unsupervised Feature Selection Using Feature Similarity. PAMI 24(3), pp. 301–312.

[26] Eisenstein, J., O'Connor, B., Noah, S. and Eric, X. 2010. A latent variable model for geographic lexical variation. EMNLP'10.

[27] Tsur, O. and Rappoport, A. 2012. What' s in a Hashtag ? Content based Prediction of the Spread of Ideas in Microblogging Communities. WSDM'12. pp. 16–23.

[28] Burnap, P. and Williams, M. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making, *Policy & Internet (in press)*

[29] Burnap, P., Williams, M.L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R. and Voss, A. 2014, Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack, *Social Network Analysis and Mining* 4:2