Finding Fires with Twitter

Robert Power robert.power@csiro.au
Bella Robinson bella.robinson@csiro.au
David Ratcliffe david.ratcliffe@csiro.au
CSIRO Computational Informatics
G.P.O. Box 664
Canberra, ACT 2601, Australia

Abstract

This paper presents a notification system to identify in near-real-time Tweets describing fire events in Australia. The system identifies fire related 'alert words' published on Twitter which are further processed by a classifier to determine if they correspond to an actual fire event. We describe how the classifier has been established and report preliminary results.

The original notification system did not include a classifier and could not discriminate between messages unrelated to 'real' fire events. In the first three months of operation, the system generated 42 'fire' email notifications of which 20 related to actual fires and 12 of those contained Tweets that may have been of interest to fire fighting agencies. If the classifier had been used, 21 emails would have been issued: an improvement in accuracy from 48% to 78%. However, the recall score reduced from 1 to 0.8 which is not desirable for this particular task. We propose extensions to address this short coming.

1 Introduction

In Australia bushfire management is a state and territory government responsibility and each jurisdiction has its own agency which takes the lead in coordinating community preparedness and responding to bushfires when they occur. For example, the Rural Fire Services (RFS) in NSW, the Country Fire Authority (CFA) in Victoria and so on, are each responsible for firefighting activities, training to prepare communities to protect themselves, land management hazard reduction as well as situations involving search and rescue.

During the Australian disaster season, early October through to the end of March, these fire agencies continuously monitor weather conditions in preparation for responding to events when they occur. They also inform the community about known incidents, see for example the NSW RFS Current Fires and Incidents page[1].

These agencies publish incident information on social media sites such as Facebook and Twitter. This provides a new channel of communication to interact with the community to both provide information about known events and to receive crowdsourced content from the general public.

This engagement of social media is yet to be fully utilised. During crisis events, the emergency services effectively use social media to provide information to the community, but their ability to obtain information from the public is limited (Lindsay, 2011). While there are social media success stories, for example the Queensland Police Force during the Brisbane Floods in 2011 (Charlton, 2012), they are not yet widespread.

Our aim is to develop an emergency management tool that sources information from social media in near-real-time. The challenges are many: the most significant being how to reliably extract relevant information about emergency events of interest for crisis coordinators. The test case described in this paper is to extract current information published on Twitter about actual fire events.

The rest of the paper is organised as follows. First (§2) we review the use of social media for situational awareness during emergency events and describe the platform used in our study. An outline of the problem is then presented including a description of our initial notification system based on identifying 'fire' alerts from Twitter (§3). The process of incorporating a text classifier to improve our alerts is then presented (§4) and analysed (§5). We conclude with an outline of further work (§6) and a discussion of our findings (§7).

[1] http://www.rfs.nsw.gov.au/dsp_conent.cfm?cat_id=683

Robert Power, Bella Robinson and David Ratcliffe. 2013. Finding Fires with Twitter. In *Proceedings of Australasian Language Technology Association Workshop*, pages 80−89.

2 Background

2.1 Related Work

In Australia, the Victorian 2009 Black Saturday Bushfires killed 173 people and impacted 78 towns with losses estimated at A$2.9 billion (Stephenson et al., 2012). A recommendation from the subsequent Royal Commission[2] was that there needs to be improved access to information for emergency planning and response. Similarly, it has been recognised that information published by the general public on social media would be relevant to emergency managers and that social media is a useful means of providing information to communities that may be impacted by emergency events (Anderson, 2012; Lindsay, 2011).

More recently, tools are being developed that specifically focus on crowdsourced information to improve the situational awareness of events as they unfold. For example, Twitcident (Abel et al., 2012) performs real-time monitoring of Twitter messages to increase safety and security. They can target large gatherings of people for purposes of crowd management such as illegal parties, riots and organised celebrations. Their tool is adjustable to specific locations and incident types.

Tweet4act (Chowdhury et al., 2013) uses keyword methods to retrieve Tweets related to a crisis situation. Text classification techniques are then applied to automatically assign those Tweets to pre-incident, during-incident and post-incident classes. Other research (Imran et al., 2013) has used machine learning techniques to map Tweets related to a crisis situation into classes defined in a disaster-related ontology to find informative Tweets that contribute to situational awareness.

Another approach (Schulz and Ristoski, 2013; Schulz et al., 2013) for real-time identification of small-scale incidents using microblogs combines information from the social and the semantic web. They define a machine learning algorithm combining text classification and semantic enrichment of microblogs using Linked Open Data. Their approach has been applied to detect three classes of small-scale incident: car, fire and shooting.

Case studies have been reported (Stollberg and de Groeve, 2012; Beneito-Montagut et al., 2013) that demonstrate the importance of placing social media information in the correct context. Emergency managers operate under a command and control structure and while drivers exist to embrace this new technology to improve situational awareness, there are still barriers to adoption based on organisational constraints. It is our belief that these barriers will be overcome with the increasing acceptance of social media, so long as the veracity of this information is suitably characterised.

[2] http://www.royalcommission.vic.gov.au/

2.2 Social Media Platform

We started investigating the utility of information published on social media for emergency management in March 2010 (Yin et al., 2012b). When a developing emergency event was known in advance, for example Tropical Cyclone Ului (March 2010), the Twitter search API was used to gather Tweets originating from the impact area.

In late September 2011, we established eight Twitter search API captures to cover Australia and New Zealand and we have been continuously collecting Tweets from these regions since then. By this time we had developed a comprehensive toolset (Cameron et al., 2012) that includes: a statistical language model that characterises the expected discourse on Twitter; an alert detector based on the language model to identify deviations from the expected discourse; a notification system that targets specific alert keywords and generates email messages (examples can be seen in Figure 1); clustering techniques for condensing and summarising information content; and interfaces supporting forensic analysis tasks. To date, over one billion Tweets have been processed and we currently collect Tweets at a rate of approximately 1500 per minute (Robinson et al., 2013a).

3 The Problem

The task is to filter the alerts generated by our Social Media platform that match fire related keywords and refine them using a classifier to identify those that relate to actual fire events.

Fire identification provides a useful test case for our Social Media platform to extend the capabilities of the existing filtering features (by keywords) and refine the results (using classifiers). The benefit is that other use cases can be readily supported by incorporating different purpose built classifiers developed for other emergency management scenarios, for example earthquakes, cyclones, severe storms, tsunami, landslides, volcanic eruptions, floods; or for crisis management incidents, for example terrorist attacks and criminal behaviour.

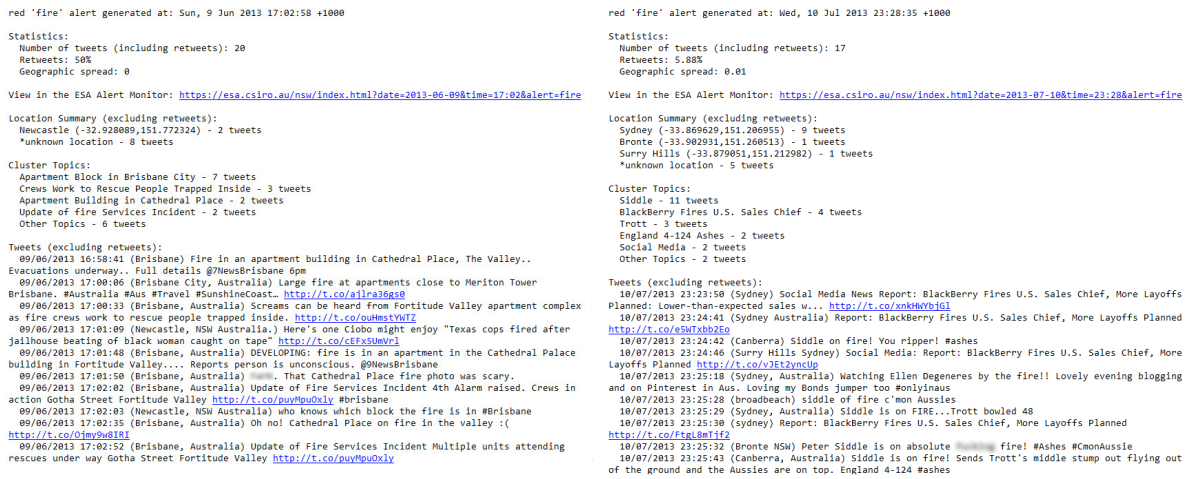

Figure 1: Examples of positive (left) and negative (right) emails.

3.1 Preliminary Work

Our Social Media platform collects Tweets from Australia and New Zealand and processes them to identify unusually frequent words that may be of interest. This processing involves extracting the individual words in the text; removing punctuation; stemming them into their common 'root' words (Porter, 1980), for example *firing*, *fires* and *fired* all have the same stem word of *fire*; calculating the observed frequency of real-time stems; and comparing this observed frequency against the historical value previously calculated and recorded in a background language model. When a stem frequency is statistically much greater than the expected value, an alert is identified.

Alerts are generated in colour from highest to lowest as: *red*, *orange*, *yellow*, *purple*, *blue* and *green*. 'Higher' alerts have a greater statistical deviation from the background language model.

In June 2013, the notification system was configured to target 17 fire related keywords, including 'fire', 'bushfire', 'grassfire', 'grass', 'bush' and 'smoke'. Each of these target keywords are associated with a different alerting colour threshold to manage the quantity of notifications generated. For example, 'smoke' and 'fire' require a high alert level (*red*) whereas 'bushfire' and 'grassfire' have a low threshold since alerts triggered from these words are considered more likely to be of interest. The notification system is currently configured to monitor alerts generated from Tweets originating from a geographic region roughly equivalent to the state of New South Wales.

The notification is delivered to registered users as an email message. Two example emails can be seen in Figure 1; both have been triggered by an alert for the keyword 'fire' and were categorised as *red* alerts. This can be seen at the top of the email which also notes the time of the alert. Currently only *red* alerts trigger an email and there must be at least two Tweets contributing to the alert; these settings are configurable. The remainder of the email is structured to help the reader decide if the alert is based on useful information sourced from Twitter describing an actual fire event. This information includes: summary statistics; a link to the web interface to explore the Tweets (the link is only accessible to authorised users) a summary of the probable locations of the Twitter users; the result of processing the Tweets into clusters; and a list of the source Tweets. Note that both examples in Figure 1 have the list of Tweets edited to save space and that expletives have been blurred.

3.2 Example Fire Alerts

The process described above, filtering 'trends' or 'bursts' from Twitter to identify words of interest, has previously been investigated for earthquake events (Robinson et al., 2013a; Robinson et al., 2013b). Specifically, they target the word 'earthquake' and its derivatives as well as the hash tag '#eqnz' and apply heuristics based on the number of retweets and tweet locations to identify first hand reports of earthquakes from Twitter. A similar process has not to our knowledge been attempted for bushfires, particularly in Australia.

For the first three months of operation the notification system described above generated 42 emails triggered by *red* 'fire' alerts, but only 20 related to real fires and of these only 12 contained Tweets that may have been of interest to fire fighting agencies. These results highlight that the word 'fire' is also used on Twitter for other purposes, as

demonstrated by the example Tweets in Table 1.

It is our expectation that using a classifier will improve the accuracy of our fire detector. Note that for this work we will attempt to use a classifier to identify Tweets related to real fires only.

4 Building the Classifiers

We have used the Support Vector Machine (SVM) (Joachims, 1998) method for text classification to identify Tweets about actual fire events. In this section, we describe the method used to develop an SVM for this purpose. We begin by identifying a test and training dataset (§4.1), and consider the features used in representing Tweets as feature vectors (§4.2). To assess whether a small labelled dataset would suffice to train an SVM with acceptable classification performance, we investigated the use of a Transductive SVM which takes a small labelled dataset and a collection of unlabelled examples, and also tested which fractions of the full dataset were required to train the standard (inductive) SVM to achieve maximum performance (§4.3). We conclude with a selection of the best available classifier to use in our goal to improve fire reporting (§4.4).

4.1 Gathering Training Data

Tweets mentioning 'fire' were identified from alerts generated by our Social Media platform during January and February 2013. This period in Australian was colloquially known as the 'Angry Summer'[3], where record high temperatures were recorded across most of the continent. Most notably, a series of devastating bushfires occurred around Coonabarabran in NSW and throughout South-Eastern Tasmania[4].

An impression of the number of candidate Tweets available for this process can be seen in Figure 2 which shows the daily count of Tweets that include the word 'fire'. Also shown are the results of processing these Tweets with the final classifier (to be described in Section 5) indicating the number of Tweets that were determined to be positive or negative. Note the gap around April. This was due to an issue with Twitter not correctly geo-locating Australian Tweets for approximately a two week period.

[3] http://climatecommission.gov.au/report/the-angry-summer/

[4] http://en.wikipedia.org/wiki/2012-13_Australian_bushfire_season

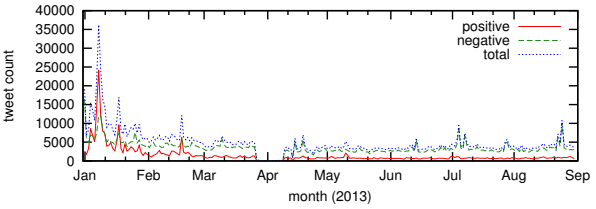

Figure 2: Daily 'fire' Tweet counts.

A selection of candidate Tweets which contributed to alerts during January and February were examined and manually labelled as positive or negative. Positive examples were first-hand witness and second-hand reports of actual fire events, as well as Tweets about fires from official sources (fire services) and news agencies. All other Tweets were considered negative examples. Our positive Tweets relate to a variety of fire events, including bushfires in Tasmania, Victoria, New South Wales and Western Australia, as well as local house and vehicle fire incidents. Negative examples were selected from 'fire' alerts that weren't to do with an actual fire. These included Tweets about fireworks, people getting fired, wood fired pizzas, fireplaces, people or sporting teams being 'on fire' and books, computer games, movies and songs with titles containing the word fire. Note that only original Tweets were labelled; retweets were excluded from this process.

A final dataset was identified consisting of 794 labelled Tweets containing the word 'fire'. This dataset consisted of an even split of positive/negative examples. Table 1 shows a sample of positive and negative Tweets from this dataset. Note that user mentions and hyperlinks that may identify user accounts have been redacted.

4.2 Feature Selection

The features selected for transforming Tweet text into feature vectors suitable for training an SVM were chosen from the following characteristics: (1) number of words; (2) user mention count; (3) hashtag count; (4) hyperlink count; (5) uni-gram occurrences; (6) bi-gram occurrences.

To determine the best combination of features to train an SVM for our problem, we performed an exhaustive search of all combinations of features ($2^6 - 1 = 63$) to train an SVM with a linear kernel. We then ranked the relative performance of each SVM by average accuracy with a 10-fold cross-validation procedure (Hastie et al., 2009), which divides the dataset into ten 90%–10% splits as training and test data (respectively). The best

(+)	Went with the other friend to the lake cause there was a HUGE fire. 2 fires actually. This photo was taken 1 km away http://t.co/...
(+)	Can see the smoke from the fire burning near craigieburn from long way. Stay safe everyone. #melbweather #hot
(+)	Fire! There's a bushfire down the road. :/
(+)	EMERGENCY WARNING issued by #TFS for uncontrolled fire at Middle Tea Tree Rd, Richmond #TAS under Extreme fire... http://t.co/...
(+)	Fires raging near ski fields in the alpine region, threatening lives and homes. Locals being told it's too late to leave #newsfeed #mthotham
(-)	the fire works are amazing this year
(-)	7 head coaches and 5 gm's got fired in the NFL and it's not 1:30 pm yet. Wow!!!
(-)	Shots fired during Auckland robberies http://t.co/...
(-)	@... @... you'll love it!! Mariah was on fire in GC.
(-)	Finally forced myself to stop reading Catching Fire. #bedtime

Table 1: Example positive (+) and negative (-) 'fire' Tweets.

result of this test is shown in Table 2, which was a combination of both (5) uni-gram occurrences and (2) user mention count as indicated by †. Subsequent rows in Table 2 show how the accuracy and F_1 scores were reduced when each of the features were excluded.

Features	Accuracy	F_1 Score
$\{2, 5\}^\dagger$	$84.54\% \pm 3.2\%$	0.831
$\{5\}$	$81.96\% \pm 4.65\%$	0.797
$\{2\}$	$54.31\% \pm 3.31\%$	0.658

Table 2: Feature combination results.

4.3 Semi-Supervised Learning

Labelling Tweets to generate training and test datasets is a labour-intensive process. To address this issue, we sought to test whether a small number of labelled positive/negative examples together with a relatively large set of unlabelled example Tweets could be used to train a Transductive SVM (TSVM) with acceptable classification performance. TSVMs have been shown to perform well for text classification problems (Joachims, 1999b), and are particularly effective over Twitter-based data (Zhang et al., 2012).

To test the performance of the TSVM relative to the standard (inductive) SVM, we used the full labelled dataset ($n = 794$) with the best determined feature combination (uni-gram occurrences and user mention count). Using this set, we aimed to test if an SVM trained on a small fraction k of the labelled examples was outperformed by a TSVM which was trained on the same set of labelled examples together with the remaining fraction $1 - k$ of the examples with their labels removed. We tested for various $k \in \{0.05, 0.10, 0.15, 0.20\}$.

For each choice of k, we created a set of experiments E where $|E| = \lceil \frac{1}{k} \rceil$. Each $e \in E$ consisted of two sets $e = \langle L, U \rangle$, where L is a randomly sampled set of $n \times k$ labelled examples (maintaining the same positive to negative example ratio as in the original dataset) which were different for each e with minimal overlap, and where U contained the remaining $n \times (1 - k)$ examples which had their labels removed, relative to each L.

For each experiment e, a two-fold cross-validation method was then used to train an SVM over set L_{train} and test over set L_{test} (where $L = L_{train} \cup L_{test}$). In each fold, a TSVM was also trained over $L_{train} \cup U$ and tested over L_{test}. The average accuracies and standard deviations for these experiments for each choice of k is shown in Table 3.

k	l/u	Type	Avg. Accuracy
0.05	40/754	SVM	61.58 ± 5.95
		TSVM	64.08 ± 8.00
0.10	80/714	SVM	69.112 ± 6.02
		TSVM	73.711 ± 4.08
0.15	120/674	SVM	68.61 ± 3.93
		TSVM	72.36 ± 3.73
0.20	159/635	SVM	69.13 ± 4.07
		TSVM	74.00 ± 5.30

Table 3: SVM versus TSVM: best features.

Experiments for both the SVM and TSVM were performed using SVM^{light} (Joachims, 1999a). As the authors of SVM^{light} have noted, aggressive feature selection has the potential to reduce the performance of a TSVM because there are often few irrelevant features in a text classification problem. For this reason, we also ran the same test for feature vectors consisting of all available features as described in §4.2, for which performance using the standard (inductive) SVM was nearly as high as the best combination of features determined by the selection process (with $n = 794$, the accuracy was $82.89\% \pm 2.84\%$ and F_1 score 80.94). The results of this test, Table 4, show that classification accuracy was not significantly different from results using the SVM and TSVM trained with feature vectors based on the best determined combination.

k	l/u	Type	Avg. Accuracy
0.05	40/754	SVM	57.89 ± 7.45
		TSVM	61.84 ± 5.21
0.10	80/714	SVM	66.60 ± 7.79
		TSVM	71.57 ± 4.99
0.15	120/674	SVM	69.72 ± 2.35
		TSVM	74.72 ± 2.37
0.20	159/635	SVM	69.00 ± 1.77
		TSVM	73.88 ± 4.42

Table 4: SVM versus TSVM: all features.

While the TSVM consistently outperformed the SVM for all cases, the improvement in accuracy was not comparable to the performance of the SVM trained on more labelled examples. It is worth noting that if we were to train a TSVM on unlabelled examples for which the proportion of positive to negative examples was unknown (which was otherwise the case in our experiment), much more experimental training and testing may be needed to determine the best assumed proportion for best performance. We did not perform such an analysis, but leave this for future work. Instead, we continued testing various proportions k in 5% increments for the SVM case to determine how the average classification accuracy changed with varying training set sizes. The results of this test are shown in Figure 3, showing that the maximum accuracy is achieved by training with around half or more ($n > 400$) of the full dataset.

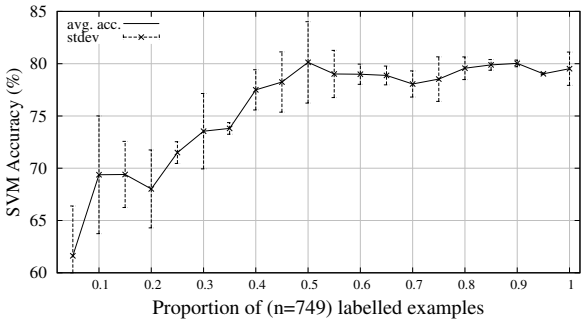

Figure 3: Learning curve for SVM^{light}.

Running the comparison experiment again for $k = 0.5$ yielded a TSVM with negligible difference in classification accuracy when compared to the SVM.

4.4 Results

As the TSVM did not outperform the standard inductive SVM in terms of classification accuracy, we opted to use the SVM trained on the best feature combination selected in §4.2 (uni-gram occurrences and user mention count) over all labelled data ($n = 794$) with a linear kernel function.

5 Improved Fire Alerting

The current email notification system has been configured to generate an email when a *new red* 'fire' alert is detected. A sequence of alerts where the gap between them is no more than 30 minutes is defined as an alert *event*. An email is generated when the first alert within an *event* that passes the notification criteria is detected; in the case of 'fire' this was configured to be a minimum alert colour of *red*. A maximum of one email is generated for each alert *event*. Figure 4 shows distribution of alerts per event for 'fire' events over an eight month period (January to August 2013). There were a lot of short events that consisted of few alerts (335 events consisted of a single alert) and on the other end of the scale there were a few long running events that consisted of a large number of alerts (the longest event had 250 alerts).

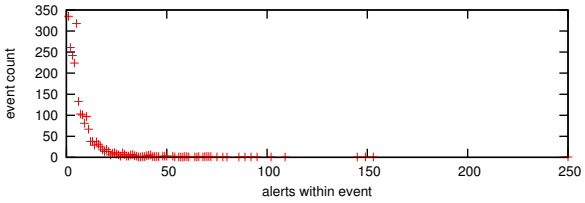

Figure 4: Alerts per event.

5.1 Analysis of Notification Emails

The 42 email notifications corresponding to *red* 'fire' alerts for June to September 2013 were examined to determine how many related to Tweets about actual fires. 20 were found to have at least one Tweet about a fire; these were labelled as true positives. The rest were labelled as false positives.

The email notification system was then configured to make use of the best performing classifier as defined in the previous section and all 'fire' alerts for the three month period were replayed to observe the effect. Various minimum positive percentage cutoff rates where also trialled to see how this affected the accuracy of the notifications.

It should be noted that as our Social Media platform is Java based, a Java implementation of LIBSVM (Chang and Lin, 2011) was used for these experiments. This was configured and trained in the same manner as the SVM^{light} software used in the classifier experiments detailed above. To verify that both SVM software packages produced equivalent results, we ran the same feature selection 10-fold cross validation experiment using LIBSVM and this produced accuracy measures that were within 5% of those achieved with

SVM^{light}. The results, Table 5, show the precision, recall, F_1 score, accuracy (percentage correct) and number of notification emails that would have been produced for each configuration.

Config	Prec	Rec	F_1	Acc	Emails
no class	0.48	1.00	0.65	47.6	42
10% pos	0.76	0.80	0.78	78.6	21
20% pos	0.81	0.65	0.72	76.2	16
30% pos	0.83	0.50	0.63	71.4	12
40% pos	0.82	0.45	0.58	69.0	11
50% pos	0.80	0.40	0.53	66.7	10

Table 5: Analysis of emails: Jun to Sep 2013.

These results show that the introduction of a classifier would have improved the overall accuracy of the email notification system, with the best result being achieved with a rule that at least 10% of Tweets contributing to the *red* 'fire' alert must be classified as positive. This improved the accuracy from 47.5% to 78.6% and the F_1 score from 0.65 to 0.78. While the number of false positives was greatly reduced (from 22 to 5) a number of false negatives were also introduced (4) which indicates that some actual fire events were missed. This is not a desirable outcome. It should also be noted that in some cases when using the classifier the generation of the notification email was delayed because the initial *red* 'fire' alerts did not pass the classification test. The notification system will keep checking follow up alerts until one passes the test and an email notification is sent.

5.2 Expected Fire Season Performance

To explore the performance of the notification system over the previous fire season, this experiment was re-run over the alerts that were generated by our Social Media platform over the months January to May 2013. During periods when there are many active bushfires it appears that the use of a classifier will not provide much benefit: for the 22 *red* 'fire' email notifications that would have been generated if our system has been running during January 2013, all of them would have been true positives without the use of a classifier. The results of this experiment are shown in Table 6 which shows the gain in accuracy by introducing a classifier is minimal and the number of emails produced is not reduced significantly.

Config	Prec	Rec	F_1	Acc	Emails
no class	0.79	1.00	0.88	79.2	48
10% pos	0.84	0.97	0.90	83.3	44
20% pos	0.91	0.84	0.88	81.3	35
30% pos	0.94	0.79	0.86	79.2	32
40% pos	0.94	0.79	0.86	79.2	32
50% pos	0.94	0.76	0.84	77.1	31

Table 6: Analysis of emails: Jan to May 2013.

6 Further Work

The notification system based on fire alerts has so far only been operating in the winter season. The results that would have been achieved if the system had been operating last summer have also been explored. The real test will be the upcoming disaster season: how well will the classifier perform? We will actively review the emails as they are generated to check they describe real fire events (true positives), while also checking the alerts that don't generate an email notification (true negatives) to verify they do not correspond to real fire events.

Our original hypothesis was that a classifier would be useful to identify real fire events from the keyword filtering of alerts generated by our Social Media platform. This was found to be true during the winter season but less so for the summer months. We will explore bypassing the alert filtering by keyword and instead focus on classification of Tweets directly. This will have performance implications, especially if there are many classifiers in operation looking for different event types.

Another avenue to explore is to analyse 'fire' alerts at a lower level than *red*. It may be possible to detect new fire events earlier, based on a smaller set of Tweets. The use of a classifier to filter out the non-fire related alerts will become more important here as the system currently generates a large number of alerts that are not at the *red* level.

There are a number of other questions to explore. The classifier developed has been trained on example Tweets from the last fire season. Will this classifier be applicable for the next fire season? Are there regional differences? For example, can the classifier trained on Australian Tweets identify fire events in New Zealand? Should Tweets from the different regions in Australia be used to train individual region specific classifiers?

To improve classification performance, we aim to try ensemble learning to combine different classifiers using a boosting strategy such as AdaBoost (Freund and Schapire, 1997; Li et al., 2008). Furthermore, the strategy we used of under-sampling the negative Tweet class to train SVMs with balanced datasets is not without drawbacks. Therefore, we aim to test learning strategies which take the underlying example imbalance

directly into account (Akbani et al., 2004). We are also interested in using the confidence or probability of individual classification determinations per Tweet to rank them in order of importance.

There are also other areas to explore with our Social Media platform. Twitter specific Natural Language Processing (NLP), Information Extraction, Word Sense Disambiguation (WSD), Part of Speech (POS) and Named Entity Recognition (NER) techniques will be investigated. For example, POS taggers (Gimpel et al., 2011; Owoputi et al., 2013; Derczynski et al., 2013) could be used to improve the identification and categorisation of fire related words. Similarly, the background language model can be extended to look at n-gram features to extend the uni-grams currently used and the existing clustering techniques can be extended to identify when alerting words are related or to make use of a WSD dictionary. NER tools can be used to better approximate the location of a Tweeter as demonstrated by Lingad et al. (2013). Also, the notification features will be extended to include other emergency use cases such as earthquakes, cyclone tracking, flood events and crisis management incidents, for example terrorist attacks and criminal behaviour.

Another area of consideration is to explore using an online incremental learning SVM similar to that described by Cauwenberghs and Poggio (2000) and Zheng et al. (2010). The aim is to dynamically refine the classifier using feedback obtained from domain specialists: incorrectly labelled Tweets can be corrected at run-time and used to re-train the classifier dynamically as an event unfolds to customize the classifier for specific events.

7 Conclusions

Our Social Media platform identifies 'alerts' based on stemmed words extracted from Tweets (Cameron et al., 2012; Yin et al., 2012a; Yin et al., 2012b). When a stem frequency is statistically much greater than the expected value, an alert is generated. These unusual events (alerts) can be filtered for keywords of interest and used as the basis for a notification system.

We have used our platform to identify occurrences of current events involving fire, such as those referring to a current *bushfire* or *grassfire*. Our system works well for words that have unambiguous and specific meanings such as these, however not for other words, such as *fire*. To improve the accuracy of the system when generating email notifications based on alerts for actual fire events, we explored the use of an SVM to discern only the relevant Tweets mentioning *fire*.

We generated a dataset of 794 Tweets with an even proportion of Tweets mentioning actual fire events to those which did not from a period during which Australia endured a particularly bad bushfire season. With this dataset, we performed an exhaustive feature selection process to train an SVM for our task. As the creation of the dataset was laborious, we also explored if a Transductive SVM (TSVM) could be used to train a model with acceptable performance with many less labelled examples in combination with more unlabelled examples, which did not prove to be the case.

Using the best trained SVM (with an accuracy of 84.54% and an F_1 score of 0.831) as a post alert filter, we found that it significantly improved the quality of the generated event notifications. In the first three months of operation, the system generated 42 'fire' email notifications where only 20 corresponded to real fire events. Filtering these alerts using the classifier resulted in 21 notifications: an improvement in accuracy from 48% to 78%, albeit with a reduction in recall from 1 to 0.8. As mentioned above however, these accuracy improvements were not obtained during the high fire danger period during the summer months.

Future work will include deploying and analysing our system in operation during the next bushfire season; exploring the use of different training datasets; improving classification accuracy using ensemble methods; ranking Tweets based on a classifier's prediction of confidence or probability to improve how notifications are interpreted; applying standard NLP techniques; and testing our system for use in other emergency management scenarios, such as earthquakes, cyclone, flood and terrorism events.

Acknowledgments

Thanks go to our colleagues Allan Yin, Sarvnaz Karimi and Jie Yin for developing some of the classification software used for our experiments. Mark Cameron was responsible for the alerting component of our Social Media platform and contributed some of the ideas for future work. Thanks also to the anonymous reviewers for their helpful comments and suggestions.

References

Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. 2012. Twitcident: fighting fire with information from social web streams. In Alain Mille, Fabien L. Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, editors, *WWW (Companion Volume)*, pages 305–308. ACM.

Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. 2004. Applying support vector machines to imbalanced datasets. In Jean-Franois Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Machine Learning: ECML 2004*, volume 3201 of *Lecture Notes in Computer Science*, pages 39–50. Springer Berlin Heidelberg.

Martin Anderson. 2012. Integrating social media into traditional management command and control structures: the square peg into the round hole. In Peter Sugg, editor, *Australian and New Zealand Disaster and Emergency Management Conference*, pages 18–34, Brisbane Exhibition and Convention Centre, Brisbane, QLD. AST Management Pty Ltd.

Roser Beneito-Montagut, Susan Anson, Duncan Shaw, and Christopher Brewster. 2013. Resilience: Two case studies on governmental social media use for emergency communication. In *Proceedings of the Information Systems for Crisis Response and Management conference (ISCRAM 2013 12-15 May, 2013)*.

Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. 2012. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 695–698, New York, NY, USA. ACM.

Gert Cauwenberghs and Tomaso Poggio. 2000. Incremental and Decremental Support Vector Machine Learning. In *NIPS*, pages 409–415.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Kym Charlton. 2012. Disaster management and social media - a case study. Technical report, Media and Public Affairs Branch, Queensland Police Service, GPO Box 4356 Melbourne VIC 3001. [Accessed: 26 April 2013].

Soudip Roy Chowdhury, Muhammad Imran, Muhammad Rizwan Asghar, Sihem Amer-Yahia, and Carlos Castillo. 2013. Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In *The 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, May.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition.

Muhammad Imran, Shady Mamoon Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Extracting information nuggets from disaster-related messages in social media. In *The 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, May.

Thorsten Joachims. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, pages 137–142, London, UK, UK. Springer-Verlag.

Thorsten Joachims. 1999a. Advances in kernel methods. chapter Making Large-Scale SVM Learning Practical, pages 169–184. MIT Press, Cambridge, MA, USA.

Thorsten Joachims. 1999b. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Xuchun Li, Lei Wang, and Eric Sung. 2008. Adaboost with svm-based component classifiers. *Engineering Applications of Artificial Intelligence*, 21(5):785 – 795.

Bruce R. Lindsay. 2011. Social media and disasters: Current uses, future options, and policy considerations. Technical report, Analyst in American National Government, GPO Box 4356 Melbourne VIC 3001, September. `http://www.fas.org/sgp/crs/homesec/R41987.pdf`.

John Lingad, Sarvnaz Karimi, and Jie Yin. 2013. Location extraction from disaster-related microblogs. In Leslie Carr, Alberto H. F. Laender, Bernadette Farias Lóscio, Irwin King, Marcus Fontoura, Denny Vrandecic, Lora Aroyo, José Palazzo M. de Oliveira, Fernanda Lima, and Erik Wilde, editors, *WWW (Companion Volume)*, pages 1017–1020. International World Wide Web Conferences Steering Committee / ACM.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Martin F. Porter. 1980. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137. `http://www.tartarus.org/~martin/PorterStemmer`.

Bella Robinson, Robert Power, and Mark Cameron. 2013a. An evidence based earthquake detector using twitter. In *Proceedings of the Workshop on Language Processing and Crisis Information 2013*, pages 1–9, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Bella Robinson, Robert Power, and Mark Cameron. 2013b. A sensitive twitter earthquake detector. In *Proceedings of the 22nd international conference on World Wide Web companion*, WWW '13 Companion, pages 999–1002, Republic and Canton of Geneva, Switzerland, May. International World Wide Web Conferences Steering Committee.

Axel Schulz and Petar Ristoski. 2013. The car that hit the burning house: Understanding small scale incident related information in microblogs. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Axel Schulz, Petar Ristoski, and Heiko Paulheim. 2013. I see a car crash: Real-time detection of small scale incidents in microblogs. In Philipp Cimiano, Miriam Fernàndez, Vanessa Lopez, Stefan Schlobach, and Johanna Völker, editors, *The Semantic Web: ESWC 2013 Satellite Events*, number 7955 in Lecture Notes in Computer Science, pages 22–33. Springer Berlin Heidelberg.

Catherine Stephenson, John Handmer, and Aimee Haywood. 2012. Estimating the net cost of the 2009 black saturday fires to the affected regions. Technical report, RMIT, Bushfire CRC, Victorian DSE, Feb.

Beate Stollberg and Tom de Groeve. 2012. The use of social media within the global disaster alert and coordination system (gdacs). In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 703–706, New York, NY, USA. ACM.

Jie Yin, Sarvnaz Karimi, Bella Robinson, and Mark A. Cameron. 2012a. Esa: emergency situation awareness via microbloggers. In Xue wen Chen, Guy Lebanon, Haixun Wang, and Mohammed J. Zaki, editors, *CIKM*, pages 2701–2703. ACM.

Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. 2012b. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6):52–59.

Renxian Zhang, Dehong Gao, and Wenjie Li. 2012. Towards scalable speech act recognition in twitter: tackling insufficient training data. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 18–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jun Zheng, Hui Yu, Furao Shen, and Jinxi Zhao. 2010. An online incremental learning support vector machine for large-scale data. In Konstantinos I. Diamantaras, Wlodek Duch, and Lazaros S. Iliadis, editors, *ICANN (2)*, volume 6353 of *Lecture Notes in Computer Science*, pages 76–81. Springer.