Developing a Sina Weibo Incident Monitor for Disasters

Bella Robinson[*] bella.robinson@csiro.au
Hua Bai[*,**] hua.bai@csiro.au
Robert Power[*] robert.power@csiro.au
Xunguo Lin[*] xunguo.lin@csiro.au

[*] CSIRO Digital Productivity Flagship
G.P.O. Box 664
Canberra, ACT 2601, Australia

[**] School of Management
Harbin Institute of Technology
92 Xidazhi Street, Harbin,
Heilongjiang, China, 150001

Abstract

This paper presents ongoing work to develop an earthquake detector based on near-real-time microblog messages from China. The system filters earthquake related keywords from Sina Weibo messages available on the public timeline and uses a classifier to determine if the messages correspond to people experiencing an earthquake. We describe how the classifier has been established and report preliminary results of successfully using it as a detector for earthquake events in China.

We also provide an overview of how we have accessed messages in Chinese from Sina Weibo, including a summary of their structure and content. We note our experience of processing this text with Natural Language Processing packages and describe a preliminary web site for users to view the processed messages.

Our long term aim is to develop a general alert and monitoring system for various disaster event types in China reported by the public on Sina Weibo. This first case study provides a working example from which an 'all hazards' system can be developed over time.

1 Introduction

Natural disasters are a major cause of loss and damage to both lives and properties around the world. Many disasters impact regions and populations with little warning which is made worse by the projected impacts of climate change and increasing urbanization of the world's population. For these reasons, it is important to enhance the ability and effectiveness of emergency management.

One of the pressing challenges while managing an emergency or crisis event is the collection and communication of reliable, relevant and up to date information. Timely, accurate and effective messages play a vital role for disaster response. For emergencies, a rapid response is needed to make effective decisions and to mitigate serious damage. Losses can be reduced by providing information to both rescue organizations and potential victims.

Microblogging services have proven to be a useful source of information to emergency managers to gain first hand information about disaster events from the community experiencing them (Abel et al., 2012; Cameron et al., 2012; Chowdhury et al., 2013; Zhou et al., 2013). This new source of information can be used by emergency coordinators and responders to provide appropriate services to the affected area. It can also be used by the wider community to seek timely information about the event and as a means of engaging with the unfolding situation from a safe distance.

Work to date has focused predominantly on Twitter as the social media microblogging source. This service is not available in China and we wanted to explore processing techniques using messages from Sina Weibo, the Chinese equivalent to Twitter. Our initial task is to identify earthquake events and this investigation is the first step in developing a general alert and monitoring system for disaster events in China.

The rest of the paper is organised as follows. First (§2) background material is provided to motivate this investigation with a focus on China. This includes how emergency events are currently managed in China, the current use of social media services there and Sina Weibo in particular. We then provide an overview of the issues of processing Chinese text (§3) and present related work (§4). This is followed by a description of the problem (§5) and details of building the classifier (§6). We conclude with a discussion of building a prototype

Bella Robinson, Hua Bai, Robert Power and Xunguo Lin. 2014. Developing a Sina Weibo Incident Monitor for Disasters. In *Proceedings of Australasian Language Technology Association Workshop*, pages 59−68.

earthquake detection web site and planned future work (§7) and a discussion of our findings (§8).

2 Background

2.1 Motivation

As social media, especially microblogging services, become more pervasive, information can be produced, retrieved and spread more quickly and conveniently than ever before. The importance of advanced information and communication technologies has been verified by recent disaster events. For example, during the Asian Tsunami that devastated many coastal regions of Thailand and other countries in 2004, first-hand information came from online services, including news reports, rescue efforts, victims experience and emotional responses (Leidner et al., 2009).

In October 2013, hurricane 'Fitow' was accompanied by heavy rain resulting in flash flooding in Yuyao City, China causing communication blackouts and the destruction of traffic infrastructure. This resulted in rescuers not being able to access flooded areas with several towns becoming 'lonely islands'[1]. With landline communication disconnected, online social media provided an important medium for information dissemination. According to a report from the organisation Kdnet Cloud Intelligence System (KCIS), who report on Chinese Internet usage, there were more than 300,000 queries for the phrase 'Yuyao flood' on the Sina Weibo platform during the week of the event[2]. This Chinese microblogging platform was used by many people to send messages and spread information asking for help and noting their trapped location by commenting on the government's Weibo account @YuyaoPublication. This information was helpful for the rescue workers.

More generally, research from the American Red Cross (2012) showed that 20% of American citizens obtained emergency information from a mobile application, 76% would expect help to arrive in less than three hours of posting a request on social media and 40% would inform friends and loved ones they were safe if impacted by an emergency event. Similar findings (Thelwall and Stuart, 2007) note that Web 2.0 technologies were used world wide as a source of information to determine the status of an unfolding emergency situation. For example, if loved ones are safe, what the ongoing risks are, the official response activities and to emotionally connect to the event.

2.2 Emergency Management in China

The Chinese central government established the Emergency Management Office (EMO) in December 2005 (Bai, 2008) to provide integrated emergency management. It was established, along with associated laws, after the 2003 SARS epidemic and is responsible for emergency planning, natural disasters, technological accidents, public sanitation issues, security concerns, and recovery and reconstruction activities.

The EMO has a coordination role between many government agencies such as the Ministry of Civil Affairs, State Administration of Work Safety, Ministry of Public Security, Ministry of Health and other related agencies. Under the central government, all local governments follow the same structure to establish a province or city level EMO. This local level EMO has authority to coordinate between corresponding local government agencies to coordinate the emergency response, disaster relief and recovery activities as required.

In China, emergency events are classified into four levels corresponding to the required government involvement:

- Level 4, small case, less than 3 fatalities, manage at the local level.
- Level 3, major incident, 3 to 10 fatalities, escalate to city level.
- Level 2, serious, between 10 and 30 fatalities, escalate to province level.
- Level 1, extremely serious, over 30 fatalities, escalate to the state council.

2.3 Use of Social Media in China

China is the world's most populous country and second largest by land area after Russia. Recent rapid economic and technological developments have resulted in more Chinese people having access to computers or mobile phones which are used increasingly to exchange information via microblogging services. By the end of 2013, the number of users of microblogging services in China reached 281 million, with nearly 70% of users (approximately 196 million people) accessing their accounts via mobile phone (China Internet Network Information Centre, 2014). This large

[1] http://www.guancha.cn/local/2013_10_09_177085.shtml

[2] http://www.kcis.cn/4409

user group provides an opportunity to study Chinese microblogging for the purposes of situation awareness of disaster events.

2.4 Sina Weibo

Sina Weibo (now trading as the Weibo Corporation) is the most influential Chinese microblogging service (Wang et al., 2014). It was established in August 2009 by the Sina Company, one of the largest news web portal and blog services in China, in response to the government blocking access to popular social media sites such as Twitter and Facebook in July 2009, due to riots in Ürümqi.

Sina Weibo has more than 156 million active users per month and more than 69 million active users per day[3]. Sina Weibo is similar to Twitter, providing user services to create content and manage their accounts with access available from mobile devices. An Application Programming Interface (API) exists as a number of Restful Web Service endpoints to access content by providing query parameters with results returned as JSON.

3 Processing Chinese Text

3.1 Segmentation

The main difficulty with processing Chinese text is the lack of whitespace between words. Automatic word segmentation on Chinese text has been an active research topic for many years (Sproat and Shih, 1990; Chen and Liu, 1992; Nie et al., 1996; Foo and Li, 2004; Gao et al., 2005) resulting in numerous software tools. Examples include the Stanford Word Segmenter, the IK Analyzer and Microsoft's S-MSRSeg. For our classification experiments we used the ansj_seg [4] tool which is a Java implementation of the hierarchical hidden Markov model based Chinese lexical analyzer ICTCLAS (Zhang et al., 2003).

3.2 Traditional and Simplified Chinese

In mainland China, Simplified Chinese replaced Traditional Chinese in 1964 with Traditional Chinese still used in Taiwan, Hong Kong and Macau. The difference is mostly with the characters, with, as the name suggests, the Simplified Chinese characters being simpler. The same grammar and sentence structure are used for both. The main method of dealing with Traditional Chinese text is to initially convert it to Simplified Chinese. For example, 'injured' in Simplified Chinese is 伤 while in Traditional Chinese it is 傷.

3.3 Word Difficulties

Polysemous words, synonyms and variant words are particularly difficult to handle when processing Chinese text.

Polysemous words are where one word has multiple meanings which can only be resolved by context. Word sense disambiguation should be able to address this issue however this has been difficult for Chinese text due to a lack of large scale and high quality Chinese word sense annotated corpus. Similarly, synonyms are also an issue where many Chinese words are different in terms of sound and written text, but they have the same meaning.

There are numerous Chinese words that are written differently but they have the same pronunciation and meaning. For example, 唯 and 惟 are pronounced the same and both mean 'unique'. These variant words can be considered the same as synonyms and treated similarly.

Chinese synonym lists have been compiled (Jiaju, 1986; Zhendong and Qiang, 1998) which are useful however in practice auto-identification algorithms are preferred, such as the Pattern Matching Algorithm (Yong and Yanqing, 2006), Link Structure and Co-occurrence Analysis (Fang et al., 2009), and Multiple Hybrid Strategies (Lu et al., 2009).

3.4 Stop Word Lists

Stop word removal is a common pre-processing step for text analysis where stop word lists can be predefined or learned. Zou et al. (2006) describe an automatic Chinese stop word extraction method based on statistic and information theory and Hao and Hao (2008) define a weighted Chi-squared statistic based on $2 * p$ contingency table measure in order to automatically identify potential stop words for a Chinese corpus. The downside to these automatic methods is that they are computationally expensive and reliant on a specific training corpus, and so predefined stop word lists are often used instead.

There are five popular Chinese stop words lists predominantly in use (Tingting et al., 2012): Harbin Institute of Technology (includes 263 symbols/punctuation characters and 504 Chinese words); Baidu (includes 263 symbols/punctuations, 547 English words and 842

[3] http://ir.weibo.com/phoenix.zhtml?c=253076&p=irol-newsArticle&ID=1958713

[4] https://github.com/ansjsun/ansj_seg

Chinese words); Sichuan University Machine Intelligent Lab (includes 975 Chinese words); Chinese stop word list (a combination of the previous three mentioned above, includes 73 symbols/punctuations, 1113 Chinese words and 9 numbers); Kevin Bouge Chinese (includes 125 Chinese words). These lists have different features with none considered authoritative. For our work we have used the fourth list mentioned above[5], a combination of the first three.

4 Related Work

4.1 Emergency Event Detection

With the popularity of the Internet, many countries have developed disaster event detection systems. For example, earthquake detectors such as 'Did You Feel It?'[6] and 'Toretter' (Sakaki et al., 2013; Sakaki et al., 2010), make use of Web 2.0 technologies and can detect earthquakes via user reports, media news and other official information.

'Twitcident' (Abel et al., 2012) monitors Tweets targeting large gatherings of people for purposes of crowd management by focusing on specific locations and incident types. 'Tweet4act' (Chowdhury et al., 2013) performs a similar function by using keyword search to identify relevant Tweets which are then filtered using text classification techniques to categorise them into pre-incident, during-incident and post-incident classes.

There are other systems as well, 'CrisisTracker' (Rogstadius et al., 2013), the Ushahidi platform (used by volunteers during the Haiti earthquake (Heinzelman and Waters, 2010) and Hurricane Sandy[7]) and the Emergency Situation Awareness system (Power et al., 2014) which provides all-hazard situation awareness information for emergency managers from Twitter.

Since the Wenchuan earthquake in China on 12 September 2008, also known as the 2008 Sichuan earthquake, researchers began to pay more attention to Twitter's role during disaster events. Sakaki et al have developed and improved a real-time report and early warning management system using Twitter (Sakaki et al., 2013; Sakaki et al., 2010). The U.S. Geological Survey (USGS) have designed the Twitter Earthquake Detector (Earle et al., 2012) based on the ratio of the short term average of word frequencies to their long term average, referred to as the STA/LTA algorithm. In Australia, a Twitter based earthquake detector has been developed and is used by the Joint Australian Tsunami Warning Centre (JATWC) (Robinson et al., 2013b; Robinson et al., 2013a). Similarly, the Earthquake Alert and Report System (EARS) (Avvenuti et al., 2014) also uses Twitter to detect earthquake events and determine damage assessments of earthquakes in Italy.

All of these systems focus on Twitter with little attention paid to messages originating from China. Two notable exceptions are described below.

4.2 Emergency Event Detection in China

Qu et al. (2011) examined Sina Weibo messages posted after the 2010 Yushu earthquake. They collected and analysed 94,101 microblog posts and 41,817 re-posts during the 48-day period immediately after the earthquake. Two keyword search queries were used to gather the data: 玉树+地震 (Yushu AND earthquake) and 青海+地震 (Qinghai AND earthquake). They then performed three types of analysis: content analysis where they categorised the messages into four groups of informational, action-related, opinion-related and emotion-related; trend analysis where they examined the distribution of different message categories over time; and information spread analysis where they examined the reposting paths of disaster related messages.

Zhou et al. (2013) also analysed Sina Weibo messages related to the Yushu earthquake. They used a naive Bayes classifier to partition messages into five groups: 'fire brigade', 'police force', 'ambulance services', 'government' and 'other', aiming to help emergency organisations respond more efficiently during an emergency. They do not, however, provide methods for event detection.

5 The Problem

5.1 Overview

The task is to filter messages published on Sina Weibo that include earthquake related keywords or phrases and refine them using a classifier to identify those that relate to actual earthquake events being felt. There are a number of secondary aims also, mainly around reliably accessing and processing messages published on Sina Weibo. There are differences between Chinese and English for the purposes of Natural Language Processing and

[5] http://www.datatang.com/data/19300

[6] http://earthquake.usgs.gov/earthquakes/dyfi/

[7] https://sandydc.crowdmap.com/

we want to explore these in detail. While there are studies of Chinese text classification (Luo et al., 2011; Yen et al., 2010; He et al., 2000), few of them focus on short text microblog messages, especially supporting disaster response.

Our long term aim is to assess the utility of Sina Weibo as a new and relevant source of information for emergency managers to help with disaster response for different kinds of disaster events.

5.2 Preliminary Work

A user account is needed to obtain messages from the public timeline using the Open Weibo API. We had to register using a smart phone app since web browsers on a desktop failed to render the web site correctly, making user interactions ineffective. The Open Weibo API[8] provides instructions in Chinese on using the API. The 'Translate to English' feature of the Chrome web browser was used since information available on their English pages appeared to be out of date.

The API has a number of endpoints but some are not available to 'default' users. To obtain content relating to earthquakes from Sina Weibo, the search/topics endpoint could be used to get messages containing certain keywords, but this is restricted. Similarly, the place/nearby_timeline also appeared useful, but it has a maximum search radius of 11,132 metres and only returns geotagged messages. Note that this search radius limit appears to be arbitrary.

The public timeline endpoint appeared the most useful[9]. It returns the most recent 200 public messages. With a rate limit of 150 requests per hour, this endpoint is polled every 24 seconds. Our system was implemented as a Java program and the Weibo4J[10] library was used for calling the Open Weibo API. Weibo4J has a simple interface and handles the OAuth authentication required to interact with the API.

The JSON message structure is similar to that from Twitter. Each message has a unique identifier, user information (user_id, a picture, thumbnail, name, description, URL, gender), a timestamp of when the message was created, the message text, the source (application) used to send the message, the user's location (province, city, location and coordinates when provided), the user's language setting (over 98% of our messages retrieved have the ISO-639 code zh-cn, which indicates simplified Chinese characters) and so on.

5.3 Message Summary

After initial experimentation, our system has been continuously retrieving messages from the public timeline since 29 August 2014 at a rate of around 470 messages per minute, or around 28,000 per hour as shown in Figure 1. Note the dips which were due to Internet outages. By the end of October 2014, over 42 million messages have been processed.

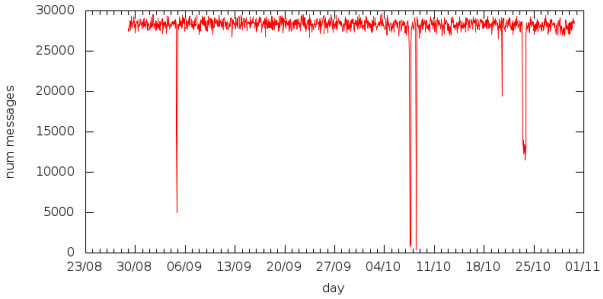

Figure 1: Hourly message counts.

Assuming there are around 90 million messages posted each day[11], this means the default rate limitations for accessing the public timeline is only providing approximately 0.8% of all messages posted. While this is only a small fraction of the content available, it is enough to detect events reported by users, as will be demonstrated below.

6 Building the Earthquake Classifier

We used the Support Vector Machine (SVM) (Joachims, 1998) method for text classification to help identify Sina Weibo messages that report feeling an earthquake. The LIBSVM (Chang and Lin, 2011) software, configured with the linear kernel function, was used to perform C-support vector classification (C-SVC) for this purpose. In this section, we describe the method used to train an SVM classifier. Relevant training data was collected and labelled (§6.1) and a range of message features to use was explored (§6.2) using ten-fold cross validation to find the most accurate feature combination.

[8] http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI

[9] http://open.weibo.com/wiki/2/statuses/public_timeline

[10] https://code.google.com/p/weibo4j/

[11] http://ir.weibo.com/phoenix.zhtml?c=253076&p=irol-products

(+)	地震地震！	Earthquake, earthquake!
(+)	原来大家都感觉到地震啦！……瞬间头好晕	Everyone felt the earthquake! Feel dizzy .....
(+)	哪里地震了？	Where is the earthquake?
(-)	2010年全球地震一览http://sinaurl.cn/hEQH1	2010 Global Earthquakes List: http://sinaurl.cn/hEQH1
(-)	那地震带会改变吗？//@薛蛮子: 科普	Will the seismic zone move? @ Xue Manzi: science
(-)	唐山地震碑林 24万人啊[泪]	Tangshan Earthquake Memorial, 240,000 people perished [tears]

Table 1: Example positive (+) and negative (-) Messages containing the phrase 'earthquake' (地震).

6.1 Training Data

The first task was to assemble a training dataset. Up to 1,000 messages from around 50,000 high scoring users were obtained. Sina Weibo attributes a user score based on their activity: the number of messages posted, comments made, friends and followers, reposted messages and so on. Not all of these users had 1,000 messages available and so a total of around 25 million were obtained. The date range for this data was February 2012 to July 2013 which was collected by a colleague with a higher level of Sina Weibo access than we do.

The messages were then filtered to those containing the word earthquake: 地震, reducing the number of messages to 21,396. To find positive examples, these messages were further filtered by only including those posted an hour after the time of known earthquake events in 2012 and 2013[12], as listed in Table 2, reducing the number to 3,549.

Location	Date/Time	Mag	Messages
Pingtung, Taiwan	26/2/12 10:34	6.0	8
Yilan, Taiwan	10/6/12 05:00	5.7	6
Xinyuan, Xinjiang	30/6/12 05:07	6.6	9
Dengta, Liaoning	23/1/13 12:18	5.1	50
Nantou, Taiwan	27/3/13 10:03	6.4	61
Lushan, Sichuan	20/4/13 08:02	7.0	228
Tongliao, Inner Mongolia	22/4/13 17:11	5.3	24
Nantou, Taiwan	02/6/13 13:43	6.7	56
Minxian, Gansu	22/7/13 07:45	6.7	25

Table 2: Earthquake events.

Next, the messages were manually examined to find positive examples of someone experiencing an earthquake. This task was performed by the two Chinese speaking authors with their individual results compared and mutual agreement reached. 467 such messages were found and the events they are associated with are indicated in Table 2. Then a collection of the same size (467) of negative messages was similarly assembled. Note that there were numerous repeated messages (reposts) such as news reports and shared prayers. For these messages only a single representative example was included with the others excluded. A sample of positive and negative messages can be seen in Table 1 where the original message in Chinese is shown, followed by a translation.

[12] http://www.csi.ac.cn/manage/eqDown/

6.2 Feature Selection

A range of message features were explored when developing the earthquake classifier: character count, word count, user mention count, hash tag count, hyperlink count, question mark count, exclamation mark count and unigrams (word n-grams of size 1).

In order to generate the SVM feature vector for each message the following steps were carried out:

1. Count the number of characters, user mentions, hash tags, hyperlinks, question marks and exclamation marks in the message.

2. Remove punctuation and replace hashtags, user mentions and hyperlinks with a constant string value (e.g. 'TAG').

3. Perform text segmentation, again using ansj_seg [4] , recording the number of tokens produced; the 'word count'.

4. Remove stop words and tokens introduced at step 2 to produce the final set of unigrams, which includes the original hashtag tokens, but not user mentions or hyperlinks.

By examining the training messages, the positive ones seemed shorter, frequently contained exclamation marks and keywords such as 'shake' (摇, 摇动) 'felt' (感, 感觉) and 'scared' (惊, 惊慌). To be certain that we didn't miss an important but less obvious feature we ran an exhaustive ten-fold cross validation process using all possible combinations of features; $2^8 - 1 = 255$ iterations in total. A selection of the results are shown in Table 3. The simple accuracy measure, which is the percentage of correct classifications, and the F_1, precision and recall scores have all been calculated. The best combination of features, indicated by the dagger† and in bold in Table 3, was

char count, link count, question mark count, exclamation mark count and unigrams. However the accuracy for this combination is only marginally better than unigrams by themselves.

Features	Accuracy	F_1 Score	Precision	Recall
unigrams	87.4%	0.876	0.862	0.893
exclamation	54.5%	0.418	0.580	0.334
question	49.6%	0.661	0.498	0.983
hyperlink	49.9%	0.445	0.498	0.573
hashtag	50.1%	0.608	0.549	0.891
user mention	50.6%	0.519	0.546	0.728
char count	64.7%	0.688	0.615	0.782
word count	63.2%	0.680	0.601	0.784
all features	88.0%	0.881	0.874	0.891
best combo†	**88.9%**	**0.890**	**0.881**	**0.902**

Table 3: Feature combination results.

6.3 Training Set Size

While the best F_1 and accuracy measures appear to be very good, we were unsure whether we had used enough training data for this process. In order to see the effect of different training set sizes we ran an experiment where the size of the training set was adjusted. For this experiment, we performed ten-fold cross validation on incrementally larger sets of training data, from 10% (84 messages) to 100% (934 messages). The results are shown in Figure 2. The features used for this experiment were the best combination identified by the feature selection process.

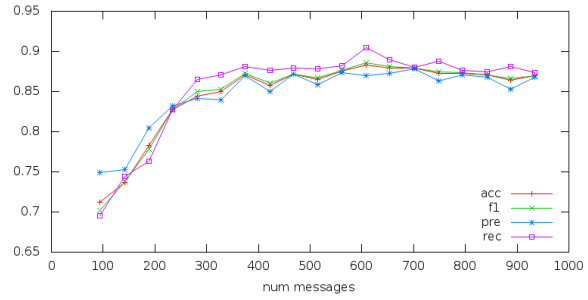

Figure 2: Adjusting the training set size.

While improvement is achieved up to the 300 message mark, only small gains are achieved after that. The maximum accuracy measure is reached at the 600 message mark, although that has a slightly larger variation between the precision and recall than the results achieved with larger training sets. It appears that the results have plateaued, which likely indicates that additional data would not improve the classifier's accuracy with the feature set we have chosen.

7 Developing an Incident Monitor

7.1 The SWIM Web Application

In order to easily view the Sina Weibo messages a web application was developed[13]. Recent messages from a selected city or province can be viewed as well as messages containing the earthquake keyword 地震. The classifier is used on the earthquake messages to highlight the positive messages, including the probability (as calculated by LIBSVM) that the message is positive. A 7-day timeline chart is also included for the earthquake messages to show if there have been any recent spikes in earthquake activity.

Figure 3 shows the web application displaying earthquake messages. The spike in the timeline chart corresponds to a 4.3M earthquake which occurred about 100km from Beijing at 18:37:41 on 6 September 2014[14]. The first positively classified message related to this event has the timestamp 18:38:24, a delay of 43 seconds. This message is followed by 55 positively (all true) and 3 negatively (1 false) classified earthquake messages in the next five minutes. Note that the messages are in reverse chronological order so the older messages are at the bottom of the page. Also, the 'Translate to English' feature of the Chrome browser has been used to translate the messages and user identifying features have been redacted.

7.2 Further Work

China is affected by a variety of natural disasters, not only earthquakes. We would like to repeat our earthquake classifier experiments to analyse messages relating to typhoon and flooding events in particular.

An earthquake detector will be developed triggered by a burst of messages containing the keyword 地震 (earthquake) which are positively classified and originate from the same geographic region. Very few messages are geotagged with exact coordinates, so we will need to approximate most message locations based on their user profile location settings (city, province and location). When an earthquake is detected, the Sina Weibo message collector can be 'zoomed' to gather messages from the affected region providing more information about the severity and impact of the earthquake.

[13] http://swim.csiro.au/

[14] http://www.csndmc.ac.cn/newweb/secondpage.jsp?id=1471

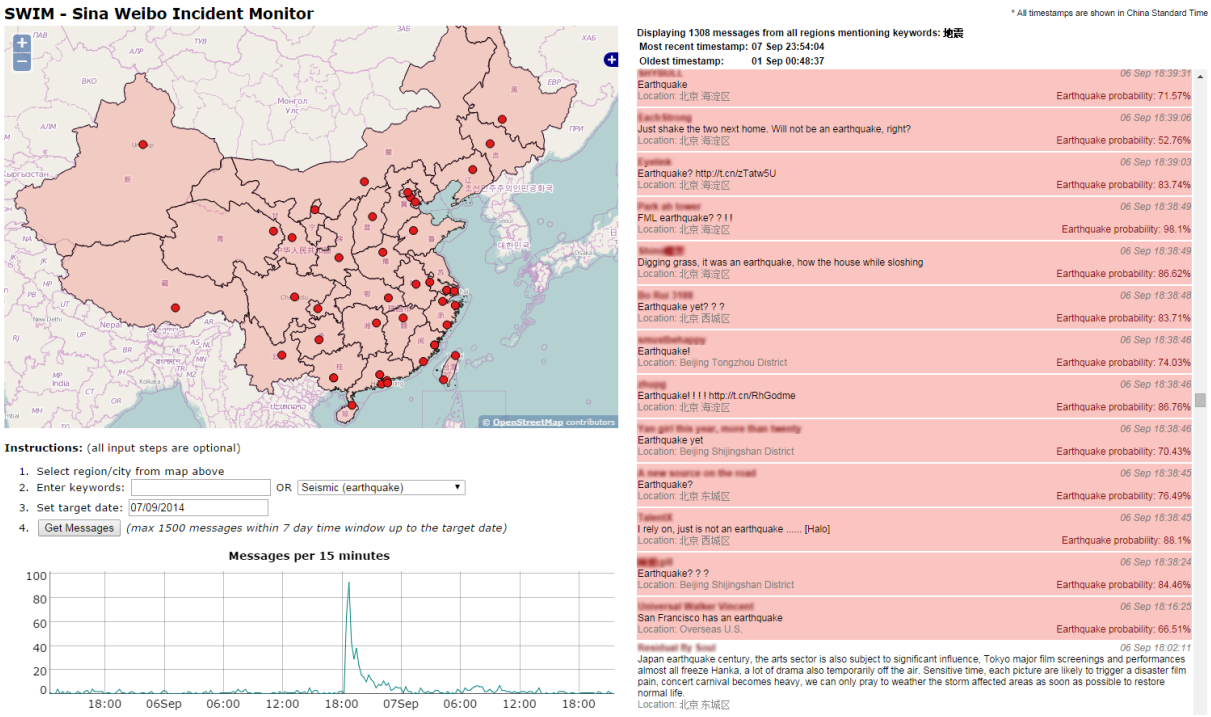

Figure 3: The Sina Weibo Incident Monitor (SWIM) Web Application

8 Conclusion

We have conducted a focused experiment to examine the feasibility of using messages from Sina Weibo to detect earthquake events. This was done by sampling messages provided on the public timeline, filtering messages that contain the keyword 'earthquake' (地震) and using a classifier to determine if messages are reports from users experiencing an earthquake.

The classifier was trained using a sample of 934 messages with an even split of positive and negative messages obtained from 'high scoring' Sina Weibo users posted soon after actual earthquake events. A comprehensive feature set was explored, with the best combination being character count, link count, question mark count, exclamation mark count and unigrams resulting in an accuracy of 88.9% and an F_1 score of 0.890.

A web site has been developed to prototype our earthquake detector. This allows users to focus on a particular province or city of interest in China or to show all messages recently posted on the public timeline. Messages containing the keyword 'earthquake' (地震) can be filtered on the display with those being positively classified highlighted in red along with an indication of the classification confidence. By correctly classifying a burst of positive messages related to the 4.3 magnitude earthquake on 6 September 2014 in Zhangjiakou City in Hebei Province, we believe that timely earthquake detection is feasible.

Future work will include setting up a notification system to report detected earthquake events; exploring the use of different training datasets with a view to improving classification accuracy; ranking messages based on a classifier's prediction of confidence to improve how notifications are interpreted; including further natural language processing techniques such as named entity recognisers, word sense disambiguation and part of speech tagging; and extending the type of events detected to include other emergency management scenarios, such as fires, typhoons and floods.

Acknowledgements

The second author, Hua Bai, thanks the China Scholarship Council for financial support and CSIRO for hosting the research project. Thanks also go to Prof. Guang Yu (School of Management Harbin Institute of Technology) and Xianyun Tian (PhD student of School of Management Harbin Institute of Technology) for providing the Sina Weibo datasets used for training the classifiers.

Hua Bai is a joint PhD student of the Harbin Institute of Technology and CSIRO.

References

Fabian Abel, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. 2012. Twitcident: Fighting fire with information from social web streams. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 305–308, Lyon, France. ACM.

American Red Cross. 2012. More Americans using mobile apps in emergencies. http://www.redcross.org/news/press-release/More-Americans-Using-Mobile-Apps-in-Emergencies, August. Accessed: 2 September 2014.

Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. 2014. Ears (Earthquake Alert and Report System): A real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1749–1758, New York, New York, USA. ACM.

Victor Bai. 2008. Emergency management in China. http://www.training.fema.gov/EMIWeb/edu/Comparative%20EM%20Book%20-%20Chapter%20-%20Emergency%20Management%20in%20China.doc. Accessed: 5 September 2014.

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, Lyon, France. ACM.

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.

Keh-Jiann Chen and Shing-Huan Liu. 1992. Word identification for Mandarin Chinese sentences. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 1*, COLING '92, pages 101–107, Nantes, France. Association for Computational Linguistics.

China Internet Network Information Centre. 2014. The 33rd statistical report on internet development in China. http://www.redcross.org/news/press-release/More-Americans-Using-Mobile-Apps-in-Emergencies, January. Accessed: 29 August 2014.

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)*, Baden-Baden, Germany, May.

Paul S. Earle, Daniel C. Bowden, and Michelle Guy. 2012. Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of GeoPhysics*, 54(6):708–715.

Huang Fang, Liu Youhua, Zhang Kezhuang, and Li Yin. 2009. Automatic recognition of Chinese synonyms using link structure and co-occurrence analysis. *Journal of Modern Information*, 29(8):125–127.

Schubert Foo and Hui Li. 2004. Chinese word segmentation and its effect on information retrieval. *Information Processing & Management*, 40(1):161 – 190.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Comput. Linguist.*, 31(4):531–574, December.

Lili Hao and Lizhu Hao. 2008. Automatic identification of stop words in Chinese text classification. In *Computer Science and Software Engineering, 2008 International Conference on*, volume 1, pages 718–722. IEEE.

Ji He, Ah-Hwee Tan, and Chew Lim Tan. 2000. A comparative study on Chinese text categorization methods. In *PRICAI Workshop on Text and Web Mining*, volume 35.

Jessica Heinzelman and Carol Waters. 2010. Crowdsourcing crisis information in disaster-affected Haiti. Technical report, United States Institute of Peace, Washington DC, USA, September.

Mei Jiaju. 1986. The function and formation of semantic systems: A new Chinese thesaurus of synonyms. *Multilingua*, 5(4):205–209, December.

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.

Dorothy E. Leidner, Gary Pan, and Shan L. Pan. 2009. The role of IT in crisis response: Lessons from the SARS and Asian tsunami disasters. *J. Strateg. Inf. Syst.*, 18(2):80–99.

Yong Lu, Chengzhi Zhang, and Hanqing Hou. 2009. Using multiple hybrid strategies to extract Chinese synonyms from encyclopedia resource. *Innovative Computing ,Information and Control, International Conference on*, 0:1089–1093.

Xi Luo, Wataru Ohyama, Tetsushi Wakabayashi, and Fumitaka Kimura. 2011. A study on automatic Chinese text classification. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 920–924. IEEE.

Jian-Yun Nie, Martin Brisebois, and Xiaobo Ren. 1996. On Chinese text retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 225–233, Zurich, Switzerland. ACM.

Robert Power, Bella Robinson, John Colton, and Mark Cameron. 2014. Emergency situation awareness: Twitter case studies. In *Proceedings of the 1st International Conference, ISCRAM-med*, volume 196 of *LNBIP*, pages 218–231, Toulouse, France, October. Springer International Publishing.

Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. 2011. Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 25–34. ACM.

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 Companion on World Wide Web*, WWW '13 Companion, pages 999–1002, Rio de Janeiro, Brazil. International World Wide Web Conferences Steering Committee.

Jakob Rogstadius, Maja Vukovic, Claudio Teixeira, Vassilis Kostakos, Evangelos Karapanos, and Jim Laredo. 2013. Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, 57(5):4:1–4:13, Sept.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference Companion on World Wide Web*, WWW '10 Companion, pages 851–860, Raleigh, North Carolina, USA. ACM.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2013. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931.

Richard Sproat and Chilin Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4(4):336–351.

Mike Thelwall and David Stuart. 2007. RUOK? Blogging communication technologies during crises. *J. Computer-Mediated Communication*, 12(2):523–548.

Zhuang Tingting, Wang Ping, and Cheng Qikai. 2012. Temporal related topic detection approach on microblog. *Journal of Information Resources Management*, 3:40–46.

Ning Wang, James She, and Junting Chen. 2014. How "Big vs" dominate Chinese microblog: A comparison of verified and unverified users on Sina Weibo. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 182–186, Bloomington, Indiana, USA. ACM.

Show-Jane Yen, Yue-Shi Lee, Yu-Chieh Wu, Jia-Ching Ying, and Vincent S Tseng. 2010. Automatic Chinese text classification using n-gram model. In *Computational Science and Its Applications–ICCSA 2010*, pages 458–471. Springer.

Lu Yong and Hou Yanqing. 2006. Automatic recognition of Chinese synonyms based on pattern matching algorithm. *Journal of the China Society for Scientific and Technical Information*, 25(6):720–724.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17*, SIGHAN '03, pages 184–187, Sapporo, Japan. Association for Computational Linguistics.

Dong Zhendong and Dong Qiang. 1998. HowNet Knowledge Database. `http://www.keenage.com`, August. Accessed: 22 August 2014.

Yanquan Zhou, Lili Yang, Bartel Van de Walle, and Chong Han. 2013. Classification of microblogs for support emergency responses: Case study Yushu earthquake in China. *2013 46th Hawaii International Conference on System Sciences*, pages 1553–1562.

Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, and Lu Sheng Wang. 2006. Automatic construction of Chinese stop word list. In *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, ACOS'06, pages 1009–1014, Hangzhou, China. World Scientific and Engineering Academy and Society (WSEAS).