VGI and crowdsourced data credibility analysis using spam email detection techniques

Journal:	*International Journal of Digital Earth*
Manuscript ID	TJDE-2016-0238.R1
Manuscript Type:	Original Research Paper
Keywords:	VGI, Crowdsourced Data, Credibility, Bayesian Networks, Spam emails

VGI and crowdsourced data credibility analysis using spam email detection techniques

Abstract

Volunteered Geographic Information (VGI) can be considered a subset of Crowdsourced Data (CSD) and its popularity has recently increased in a number of application areas. Disaster Management is one of its key application areas in which the benefits of VGI and CSD are potentially very high. However, quality issues such as credibility, reliability and relevance are limiting many of the advantages of utilising crowdsourced data. Credibility issues arise as CSD come from a variety of heterogeneous sources including both professionals and untrained citizens. VGI and CSD are also highly unstructured and the quality and metadata is often undocumented. In the 2011 Australian Floods, the general public and disaster management administrators used the Ushahidi Crowd-mapping platform to extensively communicate flood related information including hazards, evacuations, emergency services, road closures and property damage. This study assessed the credibility of the Australian Broadcasting Corporation's (ABC) Ushahidi Crowdmap dataset using a Naïve Bayesian network approach based on models commonly used in spam email detection systems. The results of the study reveal that the spam email detection approach is potentially useful for CSD credibility detection with an accuracy of over 90% using a forced classification methodology.

Keywords: VGI, Crowdsourced Data, Credibility, Bayesian Networks, Spam emails

1. Introduction

Volunteered Geographic Information (VGI) (Goodchild 2007), with its geographic context, is considered a subset of Crowdsourced Data (CSD) (Howe 2006; Goodchild and Glennon 2010; Heipke 2010; Koswatte, McDougall, and Liu 2016). In recent times, there has been an increased interest in the use of CSD for both research and commercial applications. VGI production and use have also become simpler than ever before with technological developments in mobile communication, positioning technologies, smart phone applications and other infrastructure developments which support easy to use mobile applications. However, data quality issues such as credibility, relevance, reliability, data structures, incomplete location information, missing metadata and validity continue to limit its usage and potential benefits (Flanagin and Metzger 2008; De Longueville, Ostlander, and Keskitalo 2010; Koswatte, McDougall, and Liu 2016). Therefore, researchers are now seeking new approaches for improving and managing the quality of VGI and CSD in order to increase the utilisation of this data.

VGI quality can be described in terms of quality measures and quality indicators (Antoniou and Skopeliti 2015). The quality measures of spatial data have largely focused on quantitative measures such as completeness, logical consistency, positional accuracy, temporal accuracy and thematic accuracy whilst the quality indicators are often more difficult to measure and refer to areas such as purpose, usage, trustworthiness, content quality, credibility and relevance (Senaratne et al. 2016). However, in CSD it may not always be appropriate to trust the information provided by the volunteers as their experience and expertise varies dramatically and assessing the credibility of the provider may be impractical. In particular, the volunteers in a disaster situation are often extremely heterogeneous and their input only occurs during a short period. Hence, it is difficult to profile these contributors, unlike many users of Twitter which may have a long history of activity. Therefore, a key challenge is to

assess the credibility of the provided data in order to utilise it for future decision making.

A popular approach to assess credibility in spam email detection is to numerically estimate the "degree on belief" (Robinson 2003) by analysing the email content using natural language processing and machine learning techniques. Natural language processing is a commonly used term to describe the use of computing techniques to analyse and understand natural language and speech. These approaches have been successfully applied to the detection of spam in Twitter messages (Wang 2010). The objective of this research is to investigate and test the use of spam email detection processes for credibility detection of crowdsourced disaster data.

The data for this research was collected through the Uhsahidi[1] CrowdMap platform which has been successfully used in a range of disasters including the 2011 Australian floods, the Christchurch earthquake and the 2011 tsunami in Japan. The Ushahidi platform was initially developed to easily capture crowd input via cell phones or emails (Bahree 2008; Longueville et al. 2010) and was utilised to report the election violence in Kenya. Over time, its popularity has increased and the platform has been successfully deployed in a number of disasters around the world.

This paper discusses the use of a Naïve Bayesian network based model to detect the credibility of CSD using a similar approach to spam email detection. The paper is structured as follows: Section two discusses the background of CSD credibility detection and the use of Naïve Bayesian networks for spam email detection. Section three explores the methods used in the study. Section four details the results of the study and discusses their implications. Finally, section five provides some concluding remarks and some future suggestions for research.

[1] https://www.ushahidi.com

2. Crowdsourced data credibility

Hovland, Janis, and Kelley (1953) defined credibility as "the believability of a source or message" which comprises primarily of two dimensions, trustworthiness and expertise. However, as identified by Flanagin and Metzger (2008), the dimensions of trust and expertise can also be considered as being subjectively perceived, as the study of credibility is highly interdisciplinary and the definition of credibility varies according to the field of study. While the scientific community view credibility as an objective property of information quality, the communication and social psychology researchers treat credibility more as a perceptual variable (Fogg and Tseng 1999; Flanagin and Metzger 2008). According to Fogg and Tseng (1999) credibility is defined as "a perceived quality made up of multiple dimensions such as trustworthiness and expertise" or simply as believability.

Credibility analysis approaches and the methods will vary depending on the context. Studies conducted by Bishr and Kuhn (2007), Noy, Griffith, and Musen (2008), Janowicz et al. (2010), Sadeghi-Niaraki et al. (2010), and Shvaiko and Euzenat (2013) have identified the importance and usefulness of spatial semantics and ontologies in assessing the quality of CSD. Most approaches tackle CSD quality by qualifying contributors and contributions (Brando and Bucher 2010). Various authors have investigated the classification of users based on their purpose (Coleman, Georgiadou, and Labonte 2009), their geographic location (Goodchild 2009) and trust as a reputational model (Bishr and Kuhn 2007). Quality based on contributions has mostly been validated using rating systems (Brando and Bucher 2010; Elwood 2008) or using a reference data set (Haklay 2010; Goodchild and Li 2012). Longueville et al. (2010) proposed an approach which consisted of a workflow that used prior information about the phenomenon. The key to their approach was to extract valid information from CSD using cross validation, cluster processing and ranking. A similar but extended approach for the automated assessment of the quality of CSD was proposed by

Ostermann and Spinsanti (2011).

Given the variability of contributors of CSD during a disaster event, and the complexities in qualifying the expertise or experience of contributors, it was decided that a content analysis approach would provide the greatest likelihood of success for this research.

2.1. Statistical approaches for CSD credibility detection in disaster management

Disaster related CSD is quite different in the sense of its lifetime and contributors. Data are often collected over a very short period of time with many different contributors during the event. Recent research conducted by Hung, Kalantari and Rajabifard (2016) identified the possibility of using statistical methods to assess the credibility of VGI. They used the 2011 Australian flood VGI data set as the training data and the 2013 Brisbane floods data as the testing data set. Their approach was to use binary logistic regression modelling to achieve an overall accuracy 90.5% for a training model and 80.4% accuracy for the testing data set. They highlighted the potential of using statistical approaches for efficiently analysing the CSD credibility and for rapid decision making in the disaster management sector even without real-time or near real-time information.

Kim (2013) developed a framework to assess the credibility of a VGI dataset from the 2010 Haiti earthquake based on a Bayesian Network model. The outcomes of this earthquake damage assessment study were compared with the results from official sources. The author reported that 'the experiments have not only demonstrated microscopic effects on the individual data, but also showed the macroscopic variations of the overall damage patterns by the credibility model'. Both of these models were identified as being more suitable for post disaster management purposes.

In filter based classification processes, it is important to simplify the message content using

transformations including tokenizing, stemming and lemmatizing (Figure 1) which may improve the classification accuracy and performance (Guzella and Caminhas 2009). This research followed a similar approach by incorporating natural language processing techniques and enhancing a 'bag of words' model with tokenizing (extracting words), stemming (removing derivational affixes), lemmatizing (remove inflectional endings and returning the base or dictionary form of the word) and removing stop-words (Common words in English).

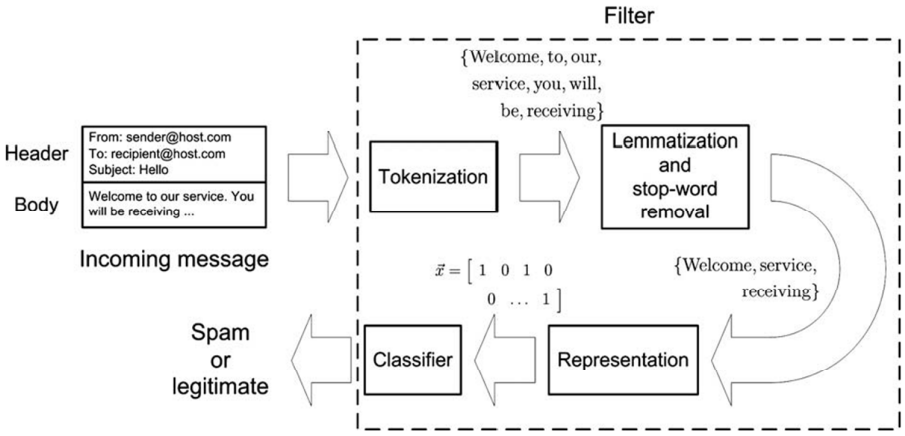


Figure 1: Main steps involved in filter based email classification (Guzella and Caminhas 2009)

Credibility can be calculated and rated into different levels which may be useful for disaster management staff. However, in critical events such as disaster management, a binary form of credibility representation would be simpler and less confusing for the general public (Ostermann and Spinsanti 2011). This research has adopted a similar binary approach by classifying the credibility using a “credible/credibility unknown” labelling. The term “credibility unknown” is used to describe those messages or reports that were not classified as “credible”.

2.2. Why use spam email detection as an approach for CSD credibility detection?

Spam email is considered as 'unsolicited bulk email' in its shortest definition (Blanzieri and Bryl 2008). Spam emails cost industries billions of dollars annually through the misuse of computing resources and the additional time required by users to sort emails. Spam emails can often carry computer viruses and also violate users' privacy (Blanzieri and Bryl 2008). Compared to the spam emails, CSD has some similarities and differences. Firstly, CSD also has a mixture of content that varies in credibility and the CSD events often generate large volumes of data. Emails, including spam emails, often have a specified structure (sender, body text and header), however, CSD often lacks structure. Finally, the aim of the filtering data to identify legitimate or credible content is similar in both cases.

Spam email detection (Pantel and Lin 1998; Cranor and LaMacchia 1998; Metsis, Androutsopoulos, and Paliouras 2006; Robinson 2003; Lopes et al. 2011), junk-email detection (Sahami et al. 1998) or anti-spam filtering (Androutsopoulos et al. 2000; Schneider 2003) research has a long history which grew from the commercialization of the internet in mid 1990s (Cranor and LaMacchia 1998). Researchers have explored various approaches with Content Based Filters (CBF) or Bayesian filters being the most popular anti-spam systems (Lopes et al. 2011). Wang (2010) tested a Bayesian classifier for spam detection in Twitter and confirmed that Bayesian classifiers performed highly in terms of weighted recall and precision, and outperformed the decision tree, neural network, support vector machines, and k-nearest neighbour's classifications.

Castillo, Mendoza, and Poblete (2011) analysed the news worthiness of tweets using a supervised classifier whilst Kang, O'Donovan, and Höllerer (2012) analysed the "credible individual tweets or users" based on three models (social model, content model and hybrid model) using Bayesian and other classifiers. These studies support the use of a modified

Bayesian approach for assessing the credibility of crowd sourced data.

2.3. A Naïve Bayesian network based model for CSD credibility detection

The Bayesian Networks (BN) were initially identified as powerful tools for knowledge representations and inference. With the advent of Naïve Bayesian networks, which are simple BNs that assume all attributes are independent, the classification power of BNs were expanded (Cheng and Greiner 1999). The credibility CSD detection engine proposed in this research was developed using a Naïve Bayesian based spam detection model.

A credibility detection function can be defined as,

$$f(m,\theta) = \begin{cases} t_{credible} & \text{if } f(m,\theta) > T \text{ message is credible} \\ t_{crediblility\ unknown} & \text{Otherwise message classified as credibility unknown} \end{cases}$$

where m is a message to be classified, θ is a vector of parameters, and $t_{credible}$ and $t_{crediblility\ unknown}$ are tags to be assigned based on the threshold T to the messages.

The vector of parameters θ is the result of training the classifier on a pre-collected dataset:

$$\theta = \Theta(M)$$

$$M = \{(m_1, l_1), (m_2, l_2), \ldots (m_n, l_n)\},\ \ l_i \in \{t_{credible}\ , t_{crediblility\ unknown}\}$$

where $m_1, m_2 \ldots m_n$ are previously collected messages, $l_1, l_2 \ldots l_n$ are the corresponding labels, and Θ is the training function.

As Guzella and Caminhas (2009) defined; if a given message is represented by $\vec{x} = [x_1, x_2, \ldots\ x_n]$ which belongs to class $c \in (s: spam, l: legitimate)$, the probability $\Pr(c|\vec{x})$ that a message is classified as c and represented by $\vec{x}$ can be written as,

$$Pr(c|\vec{x}) = \frac{Pr(\vec{x}|c)\cdot Pr(c)}{Pr(\vec{x})} = \frac{Pr(\vec{x}|c)\cdot Pr(\mathrm{c})}{Pr(\vec{x}|s)\cdot Pr(s)+Pr(\vec{x}|l)\cdot Pr(l)} \quad (1)$$

Where;

$\Pr(c)$ is overall probability that any given message is classified as c

$\Pr(\vec{x})$ is the a priori probability of a random message represented by $\vec{x}$

$\Pr(\vec{x}|s)$ and $\Pr(\vec{x}|l)$ are the probabilities that a message is classified as spam or legitimate respectively

$\Pr(s)$ and $\Pr(l)$ are overall probabilities that any given message is classified as spam or legitimate respectively.

The naïve classifier assumes that all feature in $\vec{x}$ are conditionally independent to every other feature and the probability $Pr(\vec{x}|c)$ can be defined considering *N* number of messages as,

$$Pr(\vec{x}|c) = \prod_{i=1}^{N} Pr(x_i|c) \quad (2)$$

So, the equation (1) becomes,

$$Pr(\vec{x}|c) = \frac{\prod_{i=1}^{N} Pr(x_i|\mathrm{c})\,.\,Pr(c)}{\prod_{i=1}^{N} Pr(x_i|s)\,.\,Pr(s) + \prod_{i=1}^{N} Pr(x_i|l)\,.\,Pr(l)} \quad (3)$$

with $Pr(x_i|c), c \in [s,\ l]$ given by,

$$Pr(x\_i\,|\,c) = Pr(X\_i = x\_i\,|\,c) = f(Pr(t_i|c, \mathbb{D}_{tr})\,, x_i)$$

Where function f depends on the representation of the message. The probability $Pr(t_i|c, \mathbb{D}_{tr})$ is determined based on the occurrence of term t_i in the training dataset $\mathbb{D}_{tr}$.

3. Methods

During the 2011 Australian Floods, the Australian Broadcasting Corporation[2] (ABC) developed a customised version of the Ushahidi Crowdmap to report/map disaster communications (Koswatte, McDougall, and Liu 2016). This data comprised primarily of text based content that was submitted by volunteers during the flood event. The data included input from a heterogeneous range of volunteers who submitted reports during a relatively short period of time (approximately 7 days) via various channels including a mobile app, a website, SMS messages, emails, phone calls and Twitter.

3.1. CSD credibility detecting algorithm based on spam email detection approach

An algorithm for the CSD credibility detection based on the Naïve Bayesian network was developed for the analysis. The Java[3] programming language was used for coding the system within the NetBeans[4] Integrated Development Environment (IDE). The pseudo code of the algorithm consisted of two phases including training and testing, and is listed below.

Phase 1: Start training
Select Classifier and Training Data set
    for each Message m_i in Training Dataset D_{tr} **do**
        for each Word in the Corpus **do**
        Calculate the *Credible* and *Credibility unknown* Probabilities and store in Hash Table
        end for
    end for
End training
Phase 2: Start classification
Select Classifier, Testing Dataset and Hash Table

[2] www.abc.net.au

[3] https://java.com

[4] https://netbeans.org/

for each Message m_i in the Training Dataset D_{tr} **do**

for each Word in the Corpus **do**

Calculate the Word Probability for being *Credible and Credibility unknown*

Update Hash Table

end for

Calculate combined Probability for the Message

if combined Probability > Threshold

Label Message as *Credible*

else

Label Message as *Credibility unknown*

end if

end for

End classification

The probability threshold was determined after the initial testing and was set at the 0.9 probability level.

Figure 2 illustrates the key steps in CSD credibility detection approach based on the Naïve Bayesian network and the classical "bag of words" model popular in email spam detection.

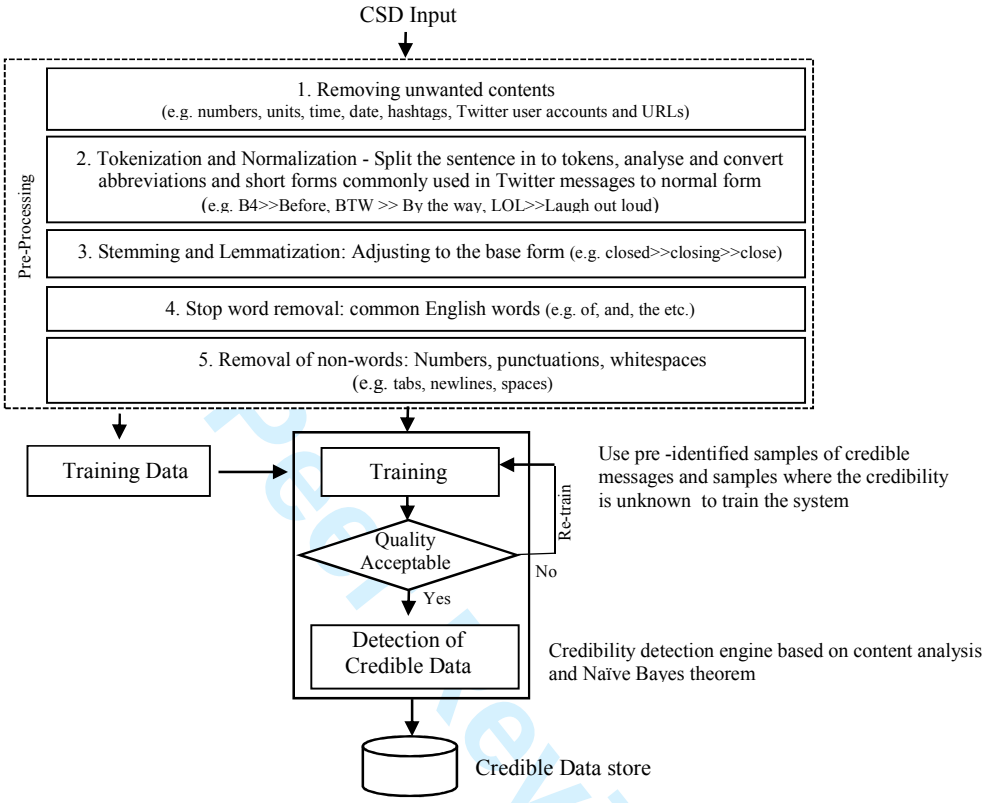


Figure 2: CSD credibility detection workflow

The ABC's 2011 Australian Flood Crisis Map dataset (Ushahidi Crowdmap) was used as the input CSD. The dataset was initially pre-processed using the steps explained in Figure 2. After the data pre-processing, the system was trained using a training sample dataset.

Within the ABC's Ushahidi Crowdmap, there were approximately 700 reports during the period of 9th -15th of January 2011 which often included information about the location where the report had originated. After the initial duplicates were removed, there were 663 unique Ushahidi Crowdmap reports remaining. The duplicates of the dataset were removed using the

'Remove duplicates' tool of the MS Excel[5] software.

For training and testing purposes, approximately 20% of the total reports (143 reports) were randomly selected from this Ushahidi Crowdmap dataset. Eighty percent of these reports (110 reports) were then selected as training data and remaining 20% selected as the testing data (33 reports). The remainder of the full dataset (520 reports) was then used for the credibility detection analysis.

The whole dataset was initially pre-processed to prepare for the training, testing and credibility detection. The training data set was classified through a manual decision process which identified messages that where either credible or where the credibility was unknown based on the credibility of terms within the message. The classification was undertaken by a reviewer who had local and expert knowledge of the disaster area.

Some examples of the manually classified credible messages and messages where the credibility was classified as unknown are shown below:

Credible Message: *Queensland Police Service: The D'Aguilar Highway at Kilcoy is now closed in both directions. Police remind motorists not to attempt to cross flooded roads or causeways.*

Message where credibility unknown: *thanks local baker keep spirit keep bake provide bread otherside town picture nothing*

The system was then trained and tested using the testing data set under two different environments namely, unforced and forced conditions, to test the accuracy and performance improvements.

In the unforced training, the data processing of the test data followed the normal pre-

[5] https://products.office.com/en-au/excel

processing steps and was then used directly for refining the training of the system. The results of this unforced training provided a report on the level of possible false positives in the classification. A high level of false positives is indicative of a possible bias in the classification process and is often referred to as *Bayesian poisoning* (Graham-Cumming 2006). The purpose of the forced training was then to review the false positives and other classified data to improve the quality of the classification process and hence re-train the system. In some instances, a number of terms which had artificially increased the credibility of the messages were identified and removed. This enabled the training of the system to be further refined and to more effectively distinguish the credible messages. The forced training process consisted of the following stages:

- The location terms were removed/disabled from both the credible and credibility unknown messages
- Highly credible terms such as *flooding*, *evacuation centre*, *road close, police, hospital* etc. were removed from messages where the credibility was unknown to give more weight to similar terms in the credible messages and to avoid Bayesian poisoning
- Removing remaining messages which could cause a high False Positive rate and therefore avoid Bayesian poisoning

When location terms appeared frequently in messages, these terms tended to increase the probability of the message being credible when in reality this was not the case. This impacted both the credible and credibility unknown messages. This impact was reduced by removing all the location terms in both credible and credibility unknown training sample messages. The Queensland Place Names Gazetteer was used as the basis for removing location terms as it provided a list of registered geographic locations and places. All incoming message terms

were cross checked against the gazetteer list and discarded if found.

The full message structure from the Ushahidi reports included information on *message number*, *incident title*, *incident date*, *location*, *description*, *category*, *latitude* and *longitude*. For example:

"101, Road closure due to flooding, , 9/01/2011 20:00, Esk-kilcoy Rd, Fast running water over the road at the bottom of the decent below lookout, Roads Affected, -27.060215, 152.553593".

Some of the message descriptions were very brief in the Ushahidi Crowdmap data. The content of these messages were further reduced when some of the pre-processing activities were undertaken including the removal of numbers, units, time, dates, hashtags, Twitter user accounts and URLs. If the number of characters of these messages were less than 30 characters, the data columns "*Incident Title*" and "*Description*" were combined (see Table 1) to make the descriptions more comprehensive and meaningful.

Table 1: Example of the combination results of the *Incident title* and *Description* of the Ushahidi Crowdmap message fields

Incident title	Description	Combined message
Road Closed-Manly Rd between new Cleveland Rd and Castlerea St, Manly	Road closed due to flooding	Road Closed-Manly Rd between new Cleveland Rd and Castlerea St, Manly road closed due to flooding

In some cases, this combination did not provide a meaningful result and did not satisfy the above condition. Therefore, the "*Location*" column was also combined in these situations (see Table 2) to improve the message meaning. However, a small number of messages had to be discarded as they did not succeed in any of the above operations.

Table 2: Example of the combination result of the *Incident title*, *Description* and *Location* of the Ushahidi Crowdmap message fields

Incident title	Description	Location	Combined message
Roads Affected	Not passable	Gailey Rd, St Lucia	Roads Affected Not passable Gailey Rd, St Lucia

The following example shows how the original Ushahidi Crowdmap message was processed after tokenisation, stemming, lemmatisation and stop-word removal before being used for training, testing and credibility detection.

Original Ushahidi Crowdmap message:

'Access to Stanthorpe town is severely restricted and all residents along Quart Pot Creek have been ordered to evacuate'.

Tokenized, stemmed and lemmatized message:

'access to Stanthorpe town be severely restrict and all resident along Quart Pot Creek have be order to evacuate'.

Stop word removed message:

'access stanthorpe town severely restrict resident along quart pot creek order evacuate'.

4. Results and discussion

4.1. Results of initial training and testing using different sized training data

The system was initially trained using two different sized training data sets to assess any variations in the outcomes based on the size of the training data set. The first training data set consisted of 35 messages of which there were 25 credible messages and 10 messages where the credibility was unknown. The second training set was a larger training sample and consisted of 77 messages with 53 credible messages and 24 messages where the credibility

was unknown.

A dataset of 33 messages was then tested using both the smaller and larger training data sets to training the system under both forced and unforced conditions. The test dataset was also manually pre-classified to identify credible messages and messages where the credibility was unknown in order to confirm the accuracy and performance during the testing. Tables 3 to 6 show the classification results for the four test environments. Test 1 utilised the smaller training data set (35 messages) with the 33 test messages under unforced training conditions. Test 2 utilised the smaller training data set (35 messages) with the 33 test messages under forced training conditions. Test 3 utilised the larger training data set (77 messages) with the 33 test messages under unforced training conditions. Finally, Test 4 utilised the larger training data set (77 messages) with the 33 test messages under forced training conditions.

The terms True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) were used to compare the results of the classification. A True Positive result correctly predicts a "Credible" outcome when it is "Credible", a True Negative result correctly predicts a "Credibility unknown" outcome when the "Credibility is unknown", a False Positive result falsely predicts a "Credible" outcome when the "Credibility is unknown", and finally, a False Negative result falsely predicts a "Credibility unknown" outcome when it should be "Credible".

Table 3: Test 1 – Unforced training using the small training sample (35 messages) and 33 test messages.

	Classified credible	**Classified as credibility unknown**	**Total**
Actually credible	24 (TP)	1 (FN)	25
Actual credibility unknown	7 (FP)	1 (TN)	8
Total	31	2	33

Table 3 results indicates that the system correctly classified 24 out of 25 credible messages during unforced training, but only one out of the eight messages where the credibility was unknown was correctly classified. This outcome resulted in a high number of False Positives for the unforced training which indicated that further training was required.

When the system utilised the same training data set but ran under forced training conditions the results as expected varied (Table 4). Of the 25 credible messages 23 messages were correctly classified and only two messages incorrectly classified. These results only varied slightly from the unforced training outcomes in regard to detecting credible messages correctly. However, there was a significant improvement in the correct detection of messages where the credibility was unknown with all messages being correctly classified during this test. Overall, the results were considered acceptable with a high classification accuracy for both the credible messages classification and the classification where the credibility of the messages was unknown and hence validated the forced training conditions.

Table 4: Test 2 - Forced training using small training sample (35 messages) and 33 test messages.

	Classified credible	**Classified as credibility unknown**	**Total**
Actually credible	23 (TP)	2 (FN)	25
Actual credibility unknown	0 (FP)	8 (TN)	8
Total	23	10	33

Next, the size of the training sample was increased from 35 messages to 77 messages and then the unforced and forced training was repeated on the same test data set. The results of unforced training are shown in Table 5 and identify that for the credible message classification, 21 out of 25 messages were correctly classified which was a small decrease in

accuracy compared to the previous result (Table 3). However, the classification accuracy where the credibility of the message was unknown, improved from one correctly classified message to five correctly classified messages out of the eight to be classified.

Table 5: Test 3 – Unforced training using the larger training sample (77 messages) and 33 test messages.

	Classified credible	Classified as credibility unknown	Total
Actually credible	21 (TP)	4 (FN)	25
Actual credibility unknown	3 (FP)	5 (TN)	8
Total	24	9	33

Finally, Table 6 shows the results of the classification using the larger training data set under forced training conditions. The results of the testing are identical to the forced training using the smaller training data set with 23 out of 25 credible messages correctly classified and all eight messages where the credibility was unknown were also correctly classified. This indicated that the forced training conditions were consistent and were not impacted by the changed training sample size.

Table 6: Test 4 - Forced training using the larger training sample (77 messages) and 33 test messages.

	Classified credible	Classified as credibility unknown	Total
Actually credible	23 (TP)	2 (FN)	25
Actual credibility unknown	0 (FP)	8 (TN)	8
Total	23	10	33

A number of measures such as accuracy, precision, sensitivity and the F1 score provided an indication of each classification's effectiveness. The accuracy, which is the ratio of correctly predicted observations, was calculated by the formula (TP+TN)/(TP+TN+FP+FN). The precision or Positive Predictive Value (PPV) is the ratio of correct positive observations. The PPV was calculated by TP/(TP + FP). The F1 score (F1) is used to measure classification performance using the weighted recall and precision, where the recall is the percentage of relevant instances that are retrieved and was calculated by 2*TP / (2*TP + FP + FN). The sensitivity or True Positive Rate (TPR) was calculated by TP / (TP + FN).

The classification quality for the four tests are summarised in Table 7. The accuracy and precision was higher for the forced training outcomes for both training sample sizes and indicates the importance of the forced training. It can also be seen that the classification accuracy and precision increased slightly for the unforced training outcomes when the larger training sample size was utilised. However, the precision and accuracy outcomes for the forced training were similar and indicate that there may be a lesser dependency on the size of the training data set when force training is utilised. The F1-Score did not change with the sample size but the measures indicate that the forced training again performed better than the unforced training scenarios. Finally, the classification sensitivity remained constant for the forced training for both training sample sizes but dropped slightly with the larger training sample size for the unforced training test outcomes.

Table 7: Quality of the CSD classification

Test Scenario	Accuracy	Precision	F1-Score	Sensitivity
Test – 1 Unforced **Using the small training sample (35 messages) and 33 test messages**	76	77	86	96
Test -2 Forced **Using the larger training sample (77 messages) and 33 test messages**	94	100	96	92
Test – 3 Unforced **Using the small training sample (35 messages) and 33 test messages**	79	88	86	84
Test – 4 Forced **Using the larger training sample (77 messages) and 33 test messages**	94	100	96	92

4.2. Results of the full Ushahidi Crowdmap data CSD analysis

After the training testing of the system was completed to an acceptable classification quality, the full Ushahidi Crowdmap sample of remaining 433 messages was analysed for credibility. As the Figure 3 (a) indicates, 54% (234 out of 433) of the messages were identified as credible using an unforced training classification. However, when the system was run under forced conditions, 77% (334 out of 433) of the messages were identified as credible (Figure 3 (b)). This was a more confident value than the previous result as the accuracy and precision of the credibility detection was higher.

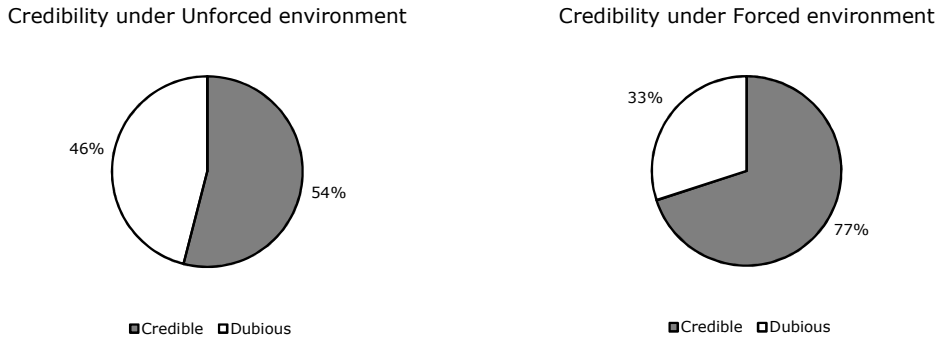


Figure 3: Credibility of 2011 Australian flood's Ushahidi Crowdmap data

5. Conclusion

The CSD message credibility detection is a challenging task due to the high degree of variability of the data, the lack of a consistent data structure, the variability of the data providers and the limited metadata available. This study identified that Bayesian spam email detection approaches can be applied successfully to the challenge of classifying the credibility of CSD. However, the training approaches and the size of the training data set can influence the quality and performance of the training outcomes.

Due to the variability of the data, it is recommended that forced training is undertaken to achieve the highest accuracy and performance. In particular, the forced training provided a higher level of confidence in eliminating the number of False Positive (FP) outcomes which were the incorrect classification of messages. The size of the training data set was found to be less critical when a forced training approach was utilised with the results of the classification outcomes being similar for both the smaller and larger training data sets.

However, if the system training was unforced, a larger training data set is recommended.

Although this study focussed on the issue of credibility, it should be recognised that the *relevance* of that dataset is another critical dimension in the quality assessment of the crowd sourced datasets. It is often not enough to just have a credible source of information as it is also important that the information is relevant to the purpose of the operational activity. For example, in the case of a flood disaster, the relevant information should relate to useful and relevant data regarding the support of the flood operations or emergency services. It is therefore important that future studies analyse both the *credibility* and the *relevance* of the crowd sourced datasets.

Acknowledgments

Authors wishes to acknowledge the Australian Government for providing support for the research work through the Research Training Program (RTP) and Monique Potts, ABC – Australia for providing the 2011 Australian Flood's Ushahidi Crowdmap data.

References

Androutsopoulos, Ion , John Koutsias, Konstantinos V Chandrinos, and Constantine D Spyropoulos. 2000. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. Paper presented at the Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.

Antoniou, V, and A Skopeliti. 2015. "Measures and Indicators of Vgi Quality: AN Overview." *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5:345-51. doi: 10.5194/isprsannals-II-3-W5-345-2015.

Bahree, Megha. 2008. "Citizen Voices." *Forbes Magazine* 182 (12):83.

Bishr, Mohamed, and Werner Kuhn. 2007. "Geospatial information bottom-up: A matter of trust and semantics." In *TThe European information society: Leading the way with geo-information*, edited by S Fabrikant and M Wachowicz, 365-87. Berlin: Springer.

Blanzieri, Enrico, and Anton Bryl. 2008. "A survey of learning-based techniques of email spam filtering." *Artificial intelligence review* 29 (1):63-92.

Brando, Carmen, and Bénédicte Bucher. 2010. Quality in user generated spatial content: A matter of specifications. Paper presented at the Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. Paper presented at the Proceedings of the 20th international conference on World wide web.

Cheng, Jie, and Russell Greiner. 1999. "Comparing Bayesian network classifiers." In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 101-8. Stockholm, Sweden: Morgan Kaufmann Publishers Inc.

Coleman, David J, Yola Georgiadou, and Jeff Labonte. 2009. "Volunteered geographic information: The nature and motivation of produsers." *International Journal of Spatial Data Infrastructures Research* 4 (1):332-58.

Cranor, Lorrie Faith, and Brian A. LaMacchia. 1998. "Spam!" In *Communications of the ACM*, 74-83.

De Longueville, Bertrand, Nicole Ostlander, and Carina Keskitalo. 2010. "Addressing vagueness in Volunteered Geographic Information (VGI)–A case study." *International Journal of Spatial Data Infrastructures Research* 5:1725-0463.

Elwood, Sarah. 2008. "Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS." *GeoJournal* 72 (3-4):173-83.

Flanagin, Andrew J, and Miriam J Metzger. 2008. "The credibility of volunteered geographic information." *GeoJournal* 72 (3-4):137-48.

Fogg, BJ, and Hsiang Tseng. 1999. The elements of computer credibility. Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems.

Goodchild, Michael F. 2007. "Citizens as sensors: the world of volunteered geography." *GeoJournal* 69 (4):211-21. doi: 10.1007/s10708-007-9111-y.

Goodchild, Michael F. 2009. "NeoGeography and the nature of geographic expertise." *Journal of Location Based Services* 3 (2):82-96. doi: 10.1080/17489720902950374.

Goodchild, Michael F., and J. Alan Glennon. 2010. "Crowdsourcing geographic information for disaster response: a research frontier." *International Journal of Digital Earth* 3 (3):231-41. doi: 10.1080/17538941003759255.

Goodchild, Michael F., and Linna Li. 2012. "Assuring the quality of volunteered geographic information." *Spatial Statistics* 1:110-20. doi: 110-120. doi: 10.1016/j.spasta.2012.03.002.

Graham-Cumming, John. 2006. "Does Bayesian poisoning exist." In *Virus Bulletin*, 69.

Guzella, Thiago S., and Walmir M. Caminhas. 2009. "A review of machine learning approaches to Spam filtering." *Expert Systems with Applications* 36 (7):10206-22. doi: 10.1016/j.eswa.2009.02.037.

Haklay, Mordechai. 2010. "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets." *Environment and planning. B, Planning & design* 37 (4):682-703. doi: 10.1068/b35097.

Heipke, Christian. 2010. "Crowdsourcing geospatial data." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (6):550-7. doi: 10.1016/j.isprsjprs.2010.06.005.

Hovland, Carl I, Irving L Janis, and Harold H Kelley. 1953. "Communication and persuasion; psychological studies of opinion change."

Howe, Jeff. 2006. "The rise of crowdsourcing." In *Wired magazine*, 1-4.

Hung, Kuo-Chih, Mohsen Kalantari, and Abbas Rajabifard. 2016. "Methods for assessing the credibility of volunteered geographic information in flood response: A case study in Brisbane, Australia." *Applied Geography* 68:37-47. doi: 10.1016/j.apgeog.2016.01.005.

Janowicz, Krzysztof, Sven Schade, Arne Broring, Carsten Kebler, Patrick Maue, and Christoph Stasch. 2010. "Semantic enablement for spatial data infrastructures." *Transactions in GIS* 14 (2):111-29. doi: 10.1111/j.1467-9671.2010.01186.x.

Kang, Byungkyu, John O'Donovan, and Tobias Höllerer. 2012. Modeling topic specific credibility on twitter. Paper presented at the Proceedings of the 2012 ACM international conference on Intelligent User Interfaces.

Kim, Heejun. 2013. "Credibility assessment of volunteered geographic information for emergency management: a Bayesian network modeling approach." University of Illinois.

Koswatte, Saman, Kevin McDougall, and Xiaoye Liu. 2016. "Semantic Location Extraction from Crowdsourced Data." *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*:543-7.

Longueville, Bertrand De, Gianluca Luraschi, Paul Smits, Stephen Peedell, and Tom De Groeve. 2010. "Citizens as sensors for natural hazards: A VGI integration workflow." *Geomatica* 64 (1):41-59.

Lopes, Clotilde, Paulo Cortez, Pedro Sousa, Miguel Rocha, and Miguel Rio. 2011. "Symbiotic filtering for spam email detection." *Expert Systems with Applications* 38 (8):9365-72. doi: 10.1016/j.eswa.2011.01.174.

Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. 2006. Spam filtering with naive bayes-which naive bayes? Paper presented at the CEAS.

Noy, Natalya F, Nicholas Griffith, and Mark A Musen. 2008. "Collecting community-based mappings in an ontology repository." In *The Semantic Web-ISWC 2008*, edited by A. Sheth, S Steffen, M Dean, M Paolucci, D Maynard, T Finin and K Thirunarayan, 371-86. Springer.

Ostermann, Frank O, and Laura Spinsanti. 2011. A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management. Paper presented at the Proceedings of AGILE, University of Utrecht, Utrecht

Pantel, Patrick, and Dekang Lin. 1998. Spamcop: A spam classification & organization program. Paper presented at the Proceedings of AAAI-98 Workshop on Learning for Text Categorization.

Robinson, Gary. 2003. "A statistical approach to the spam problem." *Linux journal* 2003 (107):3.

Sadeghi-Niaraki, Abolghasem, Abbas Rajabifard, Kyehyun Kim, and Jungtaek Seo. 2010. Ontology Based SDI to Facilitate Spatially Enabled Society. Paper presented at the Proceedings of GSDI 12 World Conference.

Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz. 1998. A Bayesian approach to filtering junk e-mail. Paper presented at the Learning for Text Categorization: Papers from the 1998 workshop.

Schneider, Karl-Michael. 2003. A comparison of event models for Naive Bayes anti-spam e-mail filtering. Paper presented at the Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1.

Senaratne, Hansi, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechai Haklay. 2016. "A review of volunteered geographic information quality assessment methods." *International Journal of Geographical Information Science*:1-29. doi: 10.1080/13658816.2016.1189556.

Shvaiko, Pavel, and Jérôme Euzenat. 2013. "Ontology Matching: State of the Art and Future Challenges." *IEEE Transactions on Knowledge & Data Engineering* 25 (1):158-76. doi: 10.1109/TKDE.2011.253.

Wang, Alex Hai. 2010. Don't follow me: Spam detection in Twitter. Paper presented at the Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, 26-28 July 2010.

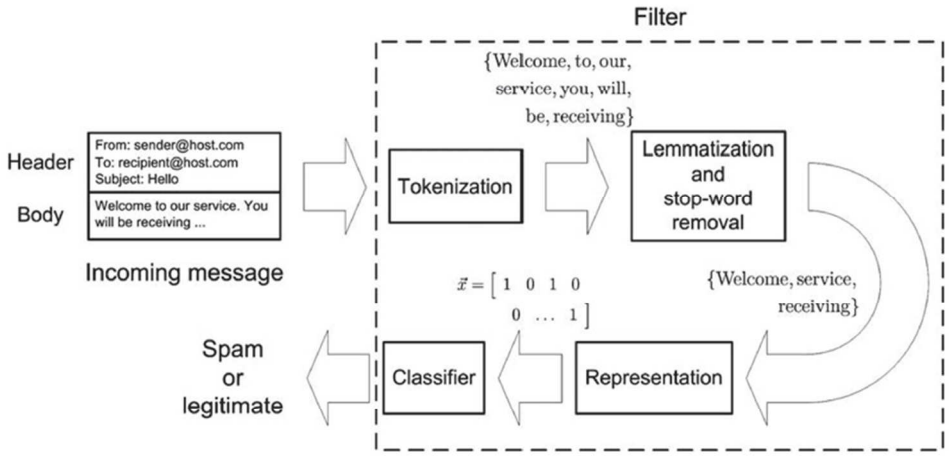


Figure 1: Main steps involved in filter based email classification (Guzella and Caminhas 2009)

71x37mm (300 x 300 DPI)

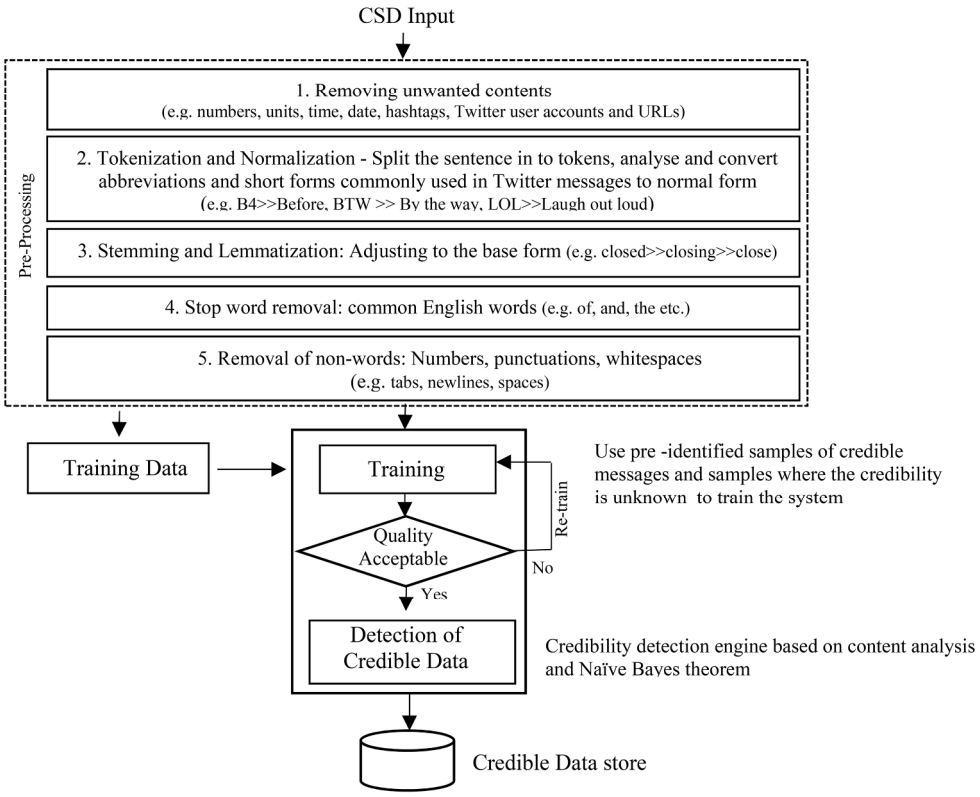


Figure 2: CSD credibility detection workflow

111x89mm (600 x 600 DPI)

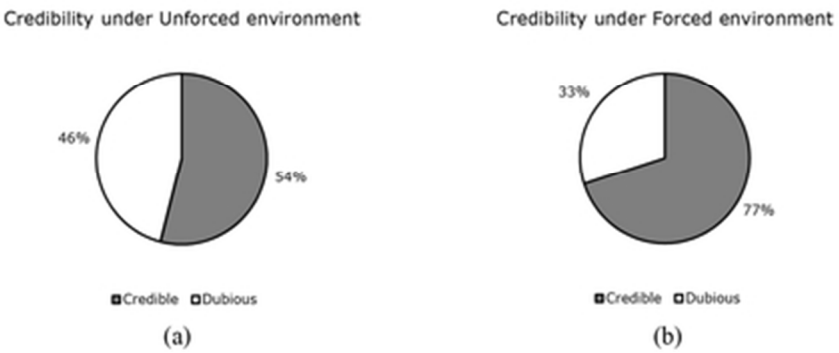


Figure 3: Credibility of 2011 Australian flood's Ushahidi Crowdmap data

20x7mm (600 x 600 DPI)

VGI and crowdsourced data credibility analysis using spam email detection techniques

Abstract

Volunteered Geographic Information (VGI) can be considered a subset of Crowdsourced Data (CSD) and ~~has recently become~~its popularity has recently increased in ~~many fields~~a number of application areas. Disaster Management is one of its key application areas in which the benefits of VGI and CSD are potentially very high. However, quality issues ~~like~~ such as credibility, reliability and relevance ~~may be reducing~~are limiting many of ~~real~~ the advantages of utilising crowd-sourced~~ing of~~ data. Credibility issues arise as CSD come from a variety of heterogeneous sources ~~captured~~ including~~by~~ both ~~of~~ professionals and ~~untrained amateurs~~untrained citizens. ~~Moreover,~~ VGI and CSD are also highly unstructured and the quality and metadata is often undocumented. In the 2011 Australian Floods, the general public and disaster management administrators used the Ushahidi Crowd-mapping platform to extensively communicate flood related information including hazards, evacuations, ~~help~~ emergency services, road closures and property damage. This study ~~has~~ assessed the credibility of the Australian Broadcasting ~~Commission's~~ Corporation's (ABC) Ushahidi Crowdmap dataset using ~~a~~a Naïve Bayesian ~~Network~~ network ~~based on a model~~ approach ~~which~~ based on ~~is~~models ~~commonly~~ used in spam email detection systems. The results of the study reveal that the spam email detection ~~a~~approach~~es~~ is~~are~~ potentially ~~feasible~~ useful for CSD credibility detection with an accuracy of ~~approximately~~ over ~~80% of the reports identified as credible and a detection accuracy close to~~ 90~~5~~% using a forced classification methodology.

Keywords: VGI, Crowdsourced Data, Credibility, Bayesian Networks, Spam emails

Formatted: Font: 11 pt, Pattern: Clear

Formatted: Font: 11 pt, Pattern: Clear

Formatted: Font: 11 pt, Pattern: Clear

1. Introduction

Volunteered Geographic Information (VGI) (Goodchild 2007), with its geographic context, is considered a subset of Crowdsourced Data (CSD) (Howe 2006; Goodchild and Glennon 2010; Heipke 2010; Koswatte, McDougall, and Liu 2016). as it comes with a geographic reference and In recent times, has there has been angained increased interest in popularity in the use of CSD for both research and commercial utilisation and researchapplications. The VGI production and use have also become simpler than ever before with technological developments in the areas of mobile communication, computingpositioning technologies, software appssmart phone applications and other infrastructure developments which supporting easy to use mobile applications. However, data quality issues like such as credibility, relevance, reliability, data structuresal limitations, incomplete location information, documentation missing metadata and validity continue to limit its usage and its potential benefits (Flanagin and Metzger 2008; De Longueville, Ostlander, and Keskitalo 2010; Koswatte, McDougall, and Liu 2016). Research in the field of VGI is now very active and Therefore, researchers are now seeking to find the meansnew approaches of for improving and managing its the quality of VGI and CSD in order to open upincrease the utilisation of application avenues. this data.

CSD and VGI quality can be described improvement research has identified two themes. One theme is to assess the spatial quality measures (accuracy) and the other is assessing the quality of the information (credibility) (Antoniou and Skopeliti 2015)in terms of quality measures and quality indicators (Antoniou and Skopeliti 2015). The spatial quality measures can be limited as the VGI and CSD spatial quality and metadata are largely undocumented. Hence, the general of spatial data have accuracy assessment parameters likelargely focused on quantitative measures such as completeness, logical consistency, positional accuracy, temporal accuracy and, thematic accuracy whilst the quality indicators are often more

Formatted: Pattern: Clear

difficult to measure and refer to areas such as purpose, usage, trustworthiness, content quality, credibility and relevance~~, purpose, usage, lineage~~ (Senaratne et al. 2016)~~ (Antoniou and Skopeliti 2015) (Haklay 2010; Girres and Touya 2010; Goodchild and Li 2012) , attribute accuracy (Girres and Touya 2010), semantic accuracy (Goodchild and Li 2012), definition, coverage, legitimacy and accessibility (Kim 2013) are still questionable in VGI/CSD quality assessment based on accuracy~~. ~~Another commonly tested approach for VGI and CSD quality assessment is based on information quality in terms of credibility. However, the existing methods and processes in this area of research are still relatively immature.

~~Credibility detection can be defined as filtering of irrelevant and dubious information to identify useful and credible information. In general, if the source can be trusted the information can also be trusted, so from a statistical perspective, there is a higher probability of the information being credible if the source is credible.~~ However, in CSD it may not always be appropriate to trust the information ~~source as the information providers vary with the situation being considered.~~provided by the volunteers as their- experience and expertise varies dramatically and assessing the credibility of the provider may be impractical. In particular, the volunteers in a disaster situation are often extremely heterogeneous and their input only occurs during a short period. Hence, it is difficult to profile these contributors, unlike many users of Twitter which may have a long history of activity. Therefore, a key challenge is to assess the credibility of the provided data in order to utilise it for future decision making.~~It is also the case during a particular event that the information provider's input may be limited to a specific time span. Therefore, assessing the source credibility may be impractical for CSD for an event like a flood. In such events, a possible approach for identifying credibility is to assess the content of the message.~~

A popular approach to assess credibility in spam email detection is to numerically estimate

Formatted: Pattern: Clear

Formatted: Font: 12 pt

Formatted: Font: 12 pt

Formatted: Font: 12 pt

the "degree on belief" (Robinson 2003) by analysing the email content using natural language ~~p~~processing and machine learning techniques. Natural language processing is a commonly used term to describe the use of computing techniques to analyse and understand natural language and speech. These ~~same~~ approaches ~~have~~ been successfully applied ~~for~~ to ~~spam~~ the detection of spam in Twitter messages ~~by~~ (Wang 2010). The ~~purposes~~ objective of this research is ~~are t~~ to ~~: (1) identify the similarities~~investigate and test the use of spam email detection processes for ~~and CSD~~ credibility detection of crowdsourced disaster data ~~, and (2) examine the possibility of using spam email detection techniques to assess the credibility of CSD.~~

The data for this research was collected through the Uhsahidi[1] CrowdMap platform which has been successfully used in a range of disasters including the 2011 Australian floods, the Christchurch earthquake and the 2011 tsunami in Japan. The Ushahidi platform was initially developed to easily capture crowd input via cell phones or emails (Bahree 2008; Longueville et al. 2010) and was utilised to report the election violence in Kenya. Over time, its popularity has increased and the platform has been successfully deployed in a number of disasters around the world.

This paper discusses the use of a Naïve- Bayesian n~~N~~etwork based model to detect the credibility of CSD using a similar approach to spam email detection. The ~~remainder of the~~ paper is structure~~d is~~ as follows: Section two discusses the background of CSD credibility detection and the use of Naïve Bayesian n~~N~~etworks for spam email detection. Section three explores the methods used in the study. Section four ~~describes~~ details the results of the study and discusses their implications~~ion of the study~~. Finally, section five provides ~~the~~ some concluding remarks ~~along with~~and ~~the~~ some future suggestions for research~~directions~~.

[1] https://www.ushahidi.com

Formatted: Pattern: Clear

Formatted: Pattern: Clear

Formatted: Pattern: Clear

Formatted: Pattern: Clear

Formatted: Pattern: Clear

2. ~~The C~~crowdsourced data credibility

~~T~~Hovland, Janis, and Kelley (1953) defined credibility as "the believability of a source or message" which comprises primarily of two dimensions, trustworthiness and expertise. However, as identified by Flanagin and Metzger (2008), the dimensions of trust and expertise can also be considered as being subjectively perceived, as the study of credibility is highly interdisciplinary and the definition of credibility varies according to the field of study ~~(Flanagin and Metzger 2008)~~. While the scientific community view credibility as an objective property of information quality, the communication and social psychology researchers treat credibility more as a perceptual variable (Fogg and Tseng 1999; Flanagin and Metzger 2008). According to Fogg and Tseng (1999) credibility is defined as "a perceived quality made up of multiple dimensions such as trustworthiness and expertise" or simply as believability.

Formatted: Pattern: Clear (Background 1)

Formatted: Pattern: Clear (Background 1)

Credibility analysis approaches and the methods will vary depending on the context. ~~Previous~~ Sstudies conducted by Bishr and Kuhn (2007), Noy, Griffith, and Musen (2008), Janowicz et al. (2010), Sadeghi-Niaraki et al. (2010), and Shvaiko and Euzenat (2013) have identified~~y~~ the importance and usefulness of spatial semantics and ontologies in assessing the quality of CSD. Most approaches tackle CSD quality by qualifying contributors and contributions (Brando and Bucher 2010). Various authors have investigated the classification of users based on their purpose (Coleman, Georgiadou, and Labonte 2009), their geographic location (Goodchild 2009) and trust as a reputational model (Bishr and Kuhn 2007). ~~The q~~Quality based on contributions has~~are~~ mostly been validated using rating systems (Brando and Bucher 2010; Elwood 2008) ~~and~~ or using a reference data set (Haklay 2010; Goodchild and Li 2012). Longueville et al. (2010)~~'s~~ proposed an approach which consisted of a workflow that~~which~~ use~~ds~~ prior information about the phenomenon. The key to their approach was to extract valid information from CSD using cross validation, cluster processing and ranking. A similar but extended approach for the automated~~ically~~ assessment of the quality of CSD was

proposed by Ostermann and Spinsanti (2011).

Given the variability of contributors of CSD during a disaster event, and the complexities in qualifying the expertise or experience of contributors, it was decided that a content analysis approach would provide the greatest likelihood of success for this research.

2.1. Statistical approaches for CSD credibility detection in disaster management

Disaster related CSD is quite different in the sense of its lifetime and contributors. Data are often collected ~~over a~~over a very short period of time ~~and the~~with many different contributors during~~may also vary with~~ the event. ~~Credibility analysis through source reputation analysis can be highly challenging in this context and often can be very problematic. A more feasible option is the analysis of information credibility.~~

Recent research conducted by Hung, Kalantari and Rajabifard (2016) identified the possibility of using statistical methods to assess the credibility of VGI. They used the 2011 Australian flood VGI data set as the training data and the 2013 Brisbane floods data as the testing data set. Their approach was to use binary logistic regression modelling ~~at a threshold of 0.917~~ to achieve an overall accuracy 90.5% for a training model ~~while~~ and 80.4% accuracy ~~was achieved~~ for the testing data set. They highlighted the potential of using statistical approaches for efficiently analysing the CSD credibility and for rapid decision making in the disaster management sector even without real-time or near real-time information.

Formatted: Pattern: Clear

Kim (2013) developed a framework to assess the credibility of a VGI dataset from the 2010 Haiti earthquake based on a Bayesian Network model. The outcomes of this earthquake damage assessment study ~~has been~~were compared with the results from official sources. The author reported that 'the experiments have not only demonstrated microscopic effects on the individual data, but also showed the macroscopic variations of the overall damage patterns by

the credibility model'. Both of these models ~~are~~ were identified as being more suitable for post disaster management purposes. ~~The proposed model is more suitable for post disaster management purposes as the model specifically focused on natural disaster damage assessments and includes a number of manual processing steps. None are capable of assessing the credibility in real-time or near-real-time context which is important in time critical applications like disaster management.~~

Formatted: Not Strikethrough

In filter based classification processes, it is important to simplify the message content using transformations ~~like~~ including tokenizing, stemming and lemmatizing (Figure 1) which may improve the classification accuracy and ~~the~~ performance (Guzella and Caminhas 2009). This research followed a similar approach by incorporating natural language processing techniques and enhancing a 'bag of words' model with tokenizing (extracting words), stemming (removing derivational affixes), lemmatizing (remove inflectional endings and ~~to~~ returning the base or dictionary ~~-~~form of the word) and removing stop-words (Common words in English).

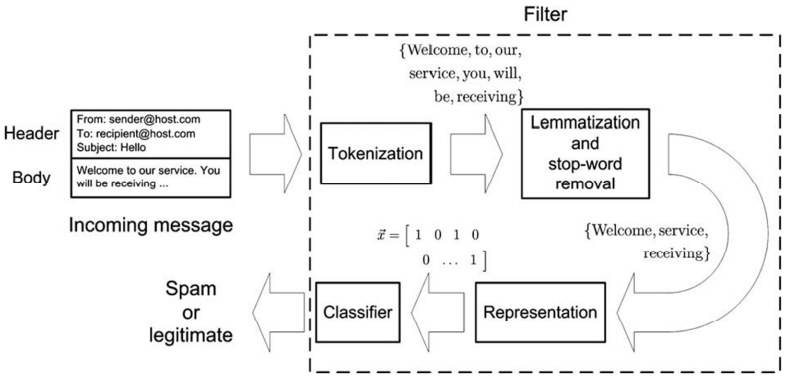


Figure 1: Main steps involved in filter based email classification (Guzella and Caminhas 2009)

Credibility can be calculated and rated into different levels which may be useful for disaster management staff. However, i~~l~~n critical events ~~like~~ such as disaster management, a binary form of credibility representation would be ~~more~~ simpler and less confusing for the general public (Ostermann and Spinsanti 2011). ~~To avoid confusion~~ T~~t~~his research has adopted a similar binary approach by~~looks to~~ classifying the credibility ~~in binary format~~ using a "credible/credibility ~~dubious~~unknown" labelling. The term "credibility unknown" is used to describe those messages or reports that were not classified as "credible".

2.2. Why use spam email detection as an approach for CSD credibility detection?

S~~A~~ spam email is considered as 'unsolicited bulk email' in its shortest definition (Blanzieri and Bryl 2008). Spam emails cost industries billions of dollars annually through the misuse of computing resources and the additional time required by users to sort emails. Spam emails can often carry computer viruses and also violate users' privacy ~~Direct marketers send spam emails to thousands of recipients without any cost, advertising anything from vacations to get-rich schemes (Androutsopoulos et al. 2000; Sahami et al. 1998) uses numerous issues such as direct financial losses, misuse of computer resources, wasting manpower to sort additional mails and violating privacy rights etc.~~ (Blanzieri and Bryl 2008). Compared to the spam emails, CSD has some similarities and differences. ~~The~~ Firstly, CSD ~~is~~ also has a mixture of content ~~of~~ that varies in credibility and the CSD events often generate ~~credible and dubious messages and it comes~~ ~~in~~ large volumes of data. ~~The s~~Epam emails, including spam emails, ~~are highly targeted and~~often have a specified structure (sender, body text and header), h~~business oriented, h~~owever, CSD ~~in general,~~ often lacks structure. ~~is created by the general public for different purposes.~~ Finally, the aim of the filtering data ~~is similar in both cases, which is~~ to identify ~~the~~ legitimate or credible content is similar in both cases. ~~messages by filtering the spam content.~~

~~The~~ ~~S~~spam email detection (Pantel and Lin 1998; Cranor and LaMacchia 1998; Metsis, Androutsopoulos, and Paliouras 2006; Robinson 2003; Lopes et al. 2011), junk-email detection (Sahami et al. 1998) ~~detection~~ or anti-spam filtering (Androutsopoulos et al. 2000; Schneider 2003) research has a long history ~~as the issue~~which ~~began growing~~grew ~~with~~ from the commercialization of the internet in mid 1990s (Cranor and LaMacchia 1998). Researchers have explored~~working on~~ various approaches with ~~with the~~ Content Based Filters (CBF) or ~~the~~ Bayesian filters being the most popular anti-spam systems (Lopes et al. 2011). Wang (2010) tested a Bayesian classifier for spam detection in Twitter and confirmed that ~~the~~ Bayesian classifiers ~~as having the best overall performance~~performed highly in terms of ~~F-measure (also called F1-Score) which is an indicator used to measure the test's accuracy calculated by~~ weighted recall and precision, and ~~(where recall is the fraction of relevant instances that are retrieved while precision is the fraction of retrieved instances that are relevant as defined in Wikipedia[2]). In this comparison, the Bayesian classifier~~ outperformed the decision tree, neural network, support vector machines, and k-nearest neighbour's classifications. ~~This finding provides support for the use of same approach for assessing the credibility of CSD.~~

Castillo, Mendoza, and Poblete (2011) analysed the news worthiness of tweets using a supervised classifier ~~and~~whilst Kang, O'Donovan, and Höllerer (2012) analysed the "credible individual tweets or users" based on three models (~~i.e.~~ social model~~: positive credibility indicators from social networks,~~ content model~~: probabilistically identifying positive retweets and user ratings,~~ and hybrid model~~: a combination of the above~~) using Bayesian and other classifiers~~, which are also identified as significant for the scope of this study.~~ These studies support the use of a modified Bayesian approach for assessing the credibility of crowd

[2] ~~https://en.wikipedia.org/wiki/Precision_and_recall~~

Formatted: Pattern: Clear

Formatted: Not Strikethrough, Pattern: Clear

Formatted: Not Strikethrough

sourced data.

2.3. A Nnaïve Bayesian nNetwork based model for CSD credibility detection

The Bayesian Networks (BN) were initially identified as powerful tools for knowledge representations and inference. With the advent of Naïve- Bayesian networks, which are simple BNs ~~which~~ that assume all attributes are independent, the classification power of BNs were ~~revealed~~ expanded (Cheng and Greiner 1999). The credibilitye CSD detection engine proposed in this research wasis developed using a Naïve -Bayesian ~~theorem~~ based ~~email~~ spam detection ~~system~~model. ~~There are number of Bayesian network probabilistic event models based on the first Naïve Bayes network based anti-spam classifier proposed by Sahammi et al. (1998).~~

~~With reference to the machine learning a~~A credibility detection function can be defined as,

$$f(m,\theta) = \begin{cases} t_{credible} & \text{if } f(m,\theta) > T \text{ message is credible} \\ t_{\sout{dubious}credibility\ unknown} & \text{Otherwise message classified as } \sout{dubious}\text{credibility unknown} \end{cases}$$

where m is a message to be classified, θ is a vector of parameters, and $t_{credible}$ and $t_{\sout{dubious}credibility\ unknown}$ are tags to be assigned based on the threshold T to the messages.

The vector of parameters θ is the result of training the classifier on a pre-collected dataset:

$$\theta = \Theta(M)$$

$$M = \{(m_1, l_1), (m_2, l_2), \ldots (m_n, l_n)\},\ l_i \in \{t_{credible}\ , t_{\sout{dubious}\ credibility\ unknown}\}$$

where $m_1, m_2 \ldots m_n$ are previously collected messages, $l_1, l_2 \ldots l_n$ are the corresponding

Formatted: Line spacing: single

labels, and Θ is the training function.

As Guzella and Caminhas (2009) defined; if a given message is represented by $\vec{x} = [x_1, x_2, \ldots x_n]$ which belongs to class $c \in (s: spam, l: legitimate)$, the probability $\Pr(c|\vec{x})$ that a message is classified as c and represented by $\vec{x}$ can be written as,

$$Pr(c|\vec{x}) = \frac{Pr(\vec{x}|c)\cdot Pr(c)}{Pr(\vec{x})} = \frac{Pr(\vec{x}|c)\cdot Pr(\mathrm{c})}{Pr(\vec{x}|s)\cdot Pr(s)+Pr(\vec{x}|l)\cdot Pr(l)} \quad (1)$$

Where;

$\Pr(c)$ is overall probability that any given message is classified as c

$\Pr(\vec{x})$ is the a-priori probability of a random message represented by $\vec{x}$

$\Pr(\vec{x}|s)$ and $\Pr(\vec{x}|l)$ are the probabilities that a message is classified as spam or legitimate respectively

$\Pr(s)$ and $\Pr(l)$ are overall probabilities that any given message is classified as spam or legitimate respectively.

~~In here, T~~the naïve classifier assumes that all feature in $\vec{x}$ are conditionally independent to every other feature and the probability $Pr(\vec{x}|c)$ can be defined considering *N* number of messages as,

$$Pr(\vec{x}|c) = \prod_{i=1}^{N} Pr(x_i|c) \quad (2)$$

So, the equation (1) becomes,

$$Pr(\vec{x}|c) = \frac{\prod_{i=1}^{N} Pr(x_i|\mathrm{c})\,.\,Pr(c)}{\prod_{i=1}^{N} Pr(x_i|s)\,.\,Pr(s) + \prod_{i=1}^{N} Pr(x_i|l)\,.\,Pr(l)} \quad (3)$$

with $Pr(x_i|c)$, $c \in [s,\ l]$ given by,

$$Pr(x\_i \mid c) = Pr(X\_i = x\_i \mid c) = f(Pr(t_i|c, \mathbb{D}_{tr}), x_i)$$

Where function f depends on the representation of the message. The probability $Pr(t_i|c, \mathbb{D}_{tr})$ is determined based on the occurrence of term t_i in the training dataset $\mathbb{D}_{tr}$.

3. Methods

~~The proposed CSD credibility detection approach consisted of two distinct phases including a system training phase and a detection phase. An algorithm was developed in the research design stage and later programmed using the Java[3] language.~~ During the 2011 Australian Floods, the Australian Broadcasting Corporation[4] (ABC) developed a customised version of the Ushahidi Crowdmap to report/map disaster communications~~. The Ushahidi crowd-mapping platform's initial development focus was on reporting and mapping post-election violence of the 2008 election in Kenya (Okolloh 2009). However, over time its applications were diversified and now the application is much popular in natural crisis mapping (Gao et al. 2011;~~ (Koswatte, McDougall, and Liu 2016). ~~This research used part of that dataset to train the CSD detection system and tested credibility of remainder of the dataset.~~This data comprised primarily of text based content that was submitted by volunteers during the flood event. The data included input from a heterogeneous range of volunteers who submitted reports during a relatively short period of time (approximately 7 days) via various channels including a mobile app, a website, SMS messages, emails, phone calls and Twitter.

3.1. CSD credibility detecting algorithm based on spam email detection approach

An algorithm for the CSD credibility detection based on the Naïve Bayesian network was

[3] ~~https://www.java.com~~

[4] www.abc.net.au

developed for the ~~programing work~~analysis. The Java[5] programming language was used for coding the system ~~using~~ within the NetBeans[6] Integrated Development Environment (IDE). ~~The algorithm consisted of two phases including training and testing.~~ The pseudo code of the algorithm ~~is~~ consisted of two phases including training and testing, and is listed below.

Phase 1: Start training

Select Classifier and Training Data set

for each Message m_i in Training Dataset D_{tr} **do**

for each Word in the Corpus **do**

Calculate the *Credible* and ~~*Dubious*~~*Credibility unknown* Probabilities and store in Hash Table

end for

end for

End training

Phase 2: Start classification

Select Classifier, ~~and~~ Testing Dataset and Hash Table

for each Message m_i in the Training Dataset D_{tr} **do**

for each Word in the Corpus **do**

Calculate the Word Probability for being *Credible and* ~~*Dubious*~~*Credibility unknown*

Update Hash Table

end for

Calculate combined Probability for the Message

if combined Probability > Threshold

Label Message as *Credible*

else

Label Message as ~~*Dubious*~~*Credibility unknown*

end if

end for

End classification

[5] https://java.com

[6] https://netbeans.org/

Formatted: Font: Italic

Formatted: Font: Not Bold

Formatted: Font: Italic

Formatted: Font: Italic

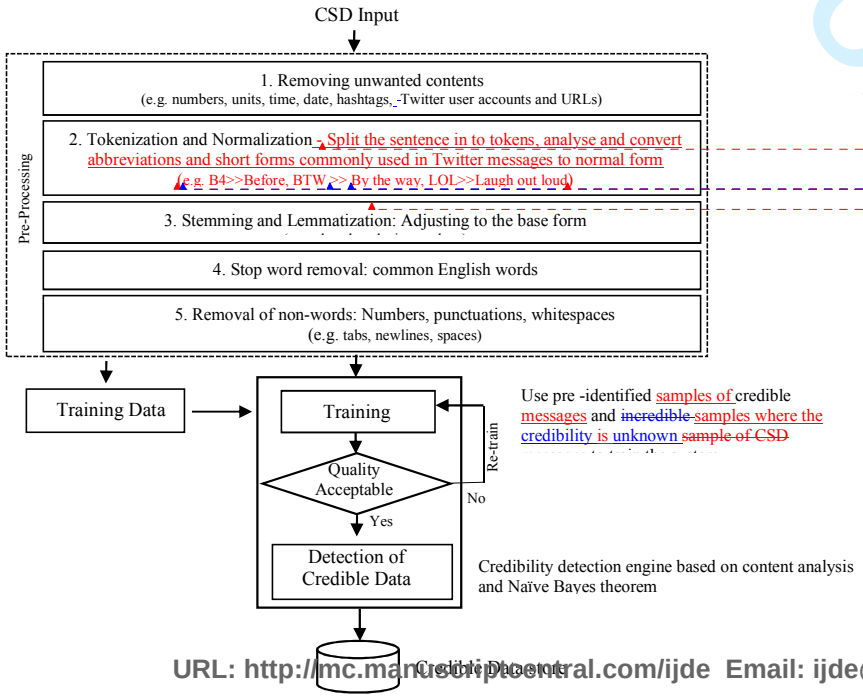
Formatted: New paragraph

Formatted: Line spacing: single

Formatted: English (Australia)

The probability threshold was determined after the initial testing and was set at the 0.9 probability level.

Figure 2 illustrates the key steps in CSD credibility detection approach based on the Naïve Bayesian network and the classical "bag of words" model popular in email spam detection.



Formatted: New paragraph

Formatted: Font: 8 pt

Formatted: Font: 8 pt

Formatted: Font: 7 pt

Formatted: Font: 7 pt

Formatted: Font: 7 pt

Formatted: Font: 8 pt

Formatted: Font: 8 pt

Figure 2: CSD credibility detection workflow

The ABC's 2011 Australian Flood Crisis Map dataset (Ushahidi Crowdmap) was used as the input CSD. The dataset was initially pre-processed using the steps explained in Figure 2 ~~and in the section 3.2~~. After the data pre-processing, the system was trained using a training sample dataset.

Within the ABC's Ushahidi Crowdmap, there were approximately 700 reports during the period of 9th -15th of January 2011 which often included information about the location where the report had originated. After the initial duplicates were removed, there were 663 unique Ushahidi Crowdmap reports remaining. The duplicates of the dataset were removed using the *'Remove duplicates'* tool of the MS Excel[7] software.

For training and testing purposes, approximately 20% of the total reports (143 reports) were randomly selected from this Ushahidi Crowdmap dataset. Eighty percent of these reports (110 reports) were then selected as training data and remaining 20% selected as the testing data (33 reports). The remainder of the full dataset (520 reports) was then used for the credibility detection analysis.

The whole dataset was initially pre-processed to prepare for the training, testing and credibility detection. ~~Part of pre-processed dataset was used for training and the other part~~

[7] https://products.office.com/en-au/excel

Formatted: Paragraph, No bullets or numbering

Formatted: Font: Not Bold

Formatted: Paragraph, No bullets or numbering

Formatted: English (Australia)

~~was used for CSD credibility detection.~~ The training data set was classified through a manual decision process which identified messages that where either credible or where the credibility was unknown based on the credibility of terms within the message. The classification was undertaken by a reviewer who had local and expert knowledge of the disaster area.

Some examples of the manually classified credible messages and messages where the credibility was classified as unknown are shown below:

Credible Message: *Queensland Police Service: The D'Aguilar Highway at Kilcoy is now closed in both directions. Police remind motorists not to attempt to cross flooded roads or causeways.*

Message where credibility unknown: *thanks local baker keep spirit keep bake provide bread otherside town picture nothing*

~~The training sample was split into two samples as being credible and dubious messages. This was done manually based on the pre-defined credible and dubious terms which were identified within the messages. Moreover,~~ T~~he~~ the system was then trained and tested using the testing data set under two different environments namely, ~~i.e.~~ unforced and forced conditions, ~~-~~ to test the accuracy and performance improvements.

In the unforced training, the data processing of the test data followed the normal pre-processing steps and was then used directly for refining the training of the system. The results of this unforced training provided a report on the level of possible false positives in the classification. A high level of false positives is indicative of a possible bias in the classification process and is often referred to as *Bayesian poisoning* (Graham-Cumming 2006). The purpose of the forced training was then to review the false positives and other classified data to improve the quality of the classification process and hence re-train the system. In some instances, a number of terms which had artificially increased the credibility

Formatted: Line spacing: Double, Pattern: Clear
Formatted: Font: 12 pt
Formatted: Font: 12 pt, Bold
Formatted: Font: 12 pt
Formatted: Font: 12 pt, Italic
Formatted: Font: 12 pt
Formatted: Font: 12 pt, Bold
Formatted: Font: 12 pt
Formatted: Font: 12 pt, Italic
Formatted: Paragraph, No bullets or numbering

of the messages were identified and removed. This enabled the training of the system to be further refined and to more effectively distinguish the credible messages. The forced training process consisted of the following stages:

- The location terms were removed/disabled from both the credible and credibility unknown messages
- Highly credible terms ~~like~~such as *flooding*, *evacuation centre*, *road close, police, hospital* etc. were removed from messages where the credibility was unknown to give more weight to similar terms in the credible messages and to avoid Bayesian poisoning
- Removing remaining messages which could cause a high False Positive rate and therefore avoid Bayesian poisoning

When location terms appeared frequently in messages, these terms tended to increase the probability of the message being credible when in reality this was not the case. This impacted both the credible and credibility unknown messages. This impact was reduced by removing all the location terms in both credible and credibility unknown training sample messages. The Queensland Place Names Gazetteer was used as the basis for removing location terms as it provided a list of registered geographic locations and places. All incoming message terms were cross checked against the gazetteer list and discarded if found.

Formatted: New paragraph

~~In the force training, credible and dubious messages in the training sample were modified as explained in the section 3.2.2. During the training phase the system's classification quality was assessed using different parameters such as accuracy, precision etc. When the classification quality was satisfactory, the CSD credibility detection was carried out using the remaining pre-processed CSD.~~

3.2. Ushahidi Crowdmap data for training, testing and credibility detection

Within the ABC's Ushahidi Crowdmap, there were approximately 700 reports during the period of 9th –15th of January, 2011, which included the originated location information. There were 663 unique Ushahidi Crowdmap messages after the initial duplicates were removed. For training and testing purposes, 150 random reports were selected from this Ushahidi Crowdmap dataset. The remainder of the dataset were then used for the credibility detection analysis. The whole dataset was initially pre-processed to prepare for the training, testing and credibility detection. The pre-processing steps included:

1. Removing unwanted contents (e.g. numbers, units, time, date, hashtags, Twitter user accounts and URLs)
2. Tokenization and Normalization: Split the sentence in to tokens, analyse and convert abbreviations and short forms commonly used in Twitter messages to normal form (e.g. B4>>Before, BTW>>By the way, LOL>>Laugh out loud)
3. Stemming and Lemmatization: Adjusting to the base form (e.g. Closed>>closing>>close)
4. Stop word removal: common English words (e.g. of, and, the etc.)
5. Removal of non-words: Numbers, punctuations, whitespaces (tabs, newlines, spaces)

From the training sample, 80% of the total messages were selected as training data and other 20% selected as the testing data. However, there were only 110 messages out of 150 messages remaining for training and testing and 433 messages out of 513 messages for the credibility detection were remaining when the pre-processing and duplicates removal were performed. The training and testing sample of 110 messages consisted of 53 credible, 24 dubious and 33 testing messages.

The full message structure from the Ushahidi reports included information on *message number*, *incident title*, *incident date*, *location*, *description*, *category*, *latitude* and *longitude*. For example:

"101, Road closure due to flooding,, 9/01/2011 20:00, Esk-kilcoy Rd, Fast running water over the road at the bottom of the decent below lookout, Roads Affected, -27.060215, 152.553593".

Some of the message descriptions were very brief in the Ushahidi Crowdmap data. The content of these~~ose~~ messages ~~contents~~ were further reduced when some of the pre-processing activities were ~~carried out~~undertaken ~~like~~ including the removal of ~~the~~ numbers, units, time, dates, hashtags, Twitter user accounts and URLs. ~~were removed.~~ If the number of characters of ~~such~~ these messages were ~~<~~ less than 30 characters, the data columns "*Incident Title*" and "*Description*" were combined (see Table 1) to make the descriptions more comprehensive and ~~more~~ meaningful.

Table 1: Example of the combination results of the *Incident title* and *Description* of the Ushahidi Crowdmap message fields

Incident title	Description	Combined message
Road Closed-Manly Rd between new Cleveland Rd and Castlerea St, Manly	Road closed due to flooding	Road Closed-Manly Rd between new Cleveland Rd and Castlerea St, Manly road closed due to flooding

In some cases, this combination did not provide a meaningful result and did not satisfy the above condition. Therefore, the "*Location*" column was also combined in ~~such~~ these situations (see Table 2)~~.~~ to improve the message ~~The end result of those operation were mostly meaningful.~~meaning. However, a small number~~few~~ of ~~the~~ messages had to be

Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Pattern: Clear
Formatted: Pattern: Clear
Formatted: Pattern: Clear
Formatted: Paragraph, Line spacing: 1.5 lines
Formatted: Paragraph
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Pattern: Clear

discarded as they did not succeed in any of the above operations.

Table 2: Example of the combination result of the *Incident title*, *Description* and *Location* of the Ushahidi Crowdmap message fields

Incident title	Description	Location	Combined message
Roads Affected	Not passable	Gailey Rd, St Lucia	Roads Affected Not passable Gailey Rd, St Lucia

The following example shows how the original Ushahidi Crowdmap message was ~~transformed~~ processed after tokenisation, stemming, lemmatisation and stop-word removal before being~~in to the final form~~ used for training, testing and credibility detection.

Original Ushahidi Crowdmap message:

'Access to Stanthorpe town is severely restricted and all residents along Quart Pot Creek have been ordered to evacuate'.

Tokenized, stemmed and lemmatized message:

'access to Stanthorpe town be severely restrict and all resident along Quart Pot Creek have be order to evacuate'.

Stop word removed message:

'access stanthorpe town severely restrict resident along quart pot creek order evacuate'.

~~The training of the CSD credibility detection system was conducted in unforced and forced environments to test the accuracy and performance improvements. Following two sections describe the forced and unforced training.~~

~~*3.2.1. Unforced training*~~

~~In the unforced training environment, the credible and dubious training sample messages~~

Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Font: Italic
Formatted: Pattern: Clear
Formatted: Indent: First line: 0"

which were pre-processed and used directly for the training. The training sample messages were processed with the following changes.

- The location terms were removed/disabled from both of the credible and dubious messages
- Removing highly credible terms like *flooding*, *evacuation centre*, *road close*, *police*, *hospital* etc. from dubious messages to give more weight to similar terms in the credible messages and to avoid Bayesian poisoning
- Removing doubtful dubious messages which could cause a high False Positive rate and to avoid Bayesian poisoning

The impact of location terms were high if they were appearing in the messages. This can happen in both credible and dubious messages. Therefore, this impact was reduced by removing all the location terms in both credible and dubious training sample messages. As the system tends to learn from new incoming messages other than the training sample, this issue will not be completely resolved by only removing location terms from the training sample. Thus, the impact of location terms should be avoided by disabling all possible locations. The Queensland place name gazetteer was used as the basis for removing location terms as it provided a list registered geographic locations and places. All incoming message terms were cross checked against the gazetteer list and discarded if found.

3.2.2. Forceful training

It is often very hard to distinguish credible and dubious data from Ushahidi Crowdmap reports in their raw forms. Generally, in spam email filtering it can be easy to identify unique terms which commonly occur in spam type messages as opposed to legitimate email messages. It is not same in Crowdmap type reports and the credible terms appear both in

credible and dubious messages. When the system was trained with a similar sample, it can cause more false positives which is identified as *Bayesian poisoning* (Graham-Cumming 2006). In this research it was decided to forcefully train the system by removing more credible types of terms from the dubious messages. However, removing of credible type of terms from the dubious messages did not solve the issues in all cases. So, after careful examination, some of the dubious messages were totally removed from the training sample.

4. Results and discussion

4.1. Results of initial tTraining and testing using different sized training data results

The CSD credibility analysis using Naïve-Bayesian Network processing provided some promising initial results. Initially, Tthe system was initially trained and tested using the detection accuracy under different situations using two different sized training data sets to assess any variations in the outcomes based on the size of the training data set. The first training data set consisted of 35 messages of which there were 25 credible messages and 10 messages where the credibility was unknown. The second training set was a larger training sample and consisted of 77 messages with 53 credible messages and 24 messages where the credibility was unknown.

A dataset of 33 messages was then tested using both the smaller and larger training data sets to training the system under both forced and unforced conditions. Theis testinget dataset was also a manually pre-classified sample asto identify credible messages and messages where the credibility wasis unknown in order to use as the ground truth data for classificationconfirm the accuracy check and performance during the testingtests. (i.e. with variable sample sizes, and under forced and unforced environments as explained in Section 3.2.1 and Section 3.2.2). Tables 3 to 6 show the classification results using different training sample sizes under two different environments which were unforced and forcedfor the four test environments. Test 1

Formatted: Pattern: Clear
Formatted: Pattern: Clear
Formatted: Pattern: Clear
Formatted: Pattern: Clear
Formatted: Pattern: Clear
Formatted: Pattern: Clear

utilised the smaller training data set (35 messages) with the 33 test messages under unforced training conditions. Test 2 utilised the smaller training data set (35 messages) with the 33 test messages under forced training conditions. Test 3 utilised the larger training data set (77 messages) with the 33 test messages under unforced training conditions. Finally, Test 4 utilised the larger training data set (77 messages) with the 33 test messages under forced training conditions.

The terms True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) were used to compare the results of the classification. ~~The~~ A True Positive result ~~is~~ correctly predict~~sing a label i.e. predicted~~ a "Credible" outcome when it ~~, and~~ is "Credible", a True Negative result ~~is~~ correctly predict~~sing the other label i.e. predicted~~a "~~Dubious~~Credibility unknown" outcome~~,~~ when the ~~and is~~ "~~Dubious~~Credibility is unknown", a False Positive result~~is~~ falsely predict~~sing a label i.e. predicted~~ a "Credible" outcome when the~~, but is~~ "~~Dubious~~Credibility is unknown", and finally, a False Negative result~~is~~ falsely predict~~sing the other label i.e. predicted~~ a "~~Dubious~~Credibility unknown" outcome when it should be~~, but is~~ "Credible".

Table 3: Test 1~~ing – 1 (~~Unforced training using the small training sample (35 messages) and ~~) results with 25 credible, 10 dubious and~~ 33 test~~ing~~ messages.

	Classified credible	Classified as ~~dubious~~credibility unknown	Total
Actually credible	24 (TP)	1 (FN)	25
Actual~~ly dubious~~credibility unknown	7 (FP)	1 (TN)	8
Total	31	2	33

Formatted: Line spacing: single

Formatted: Line spacing: single

Formatted: Line spacing: single

~~The~~ Table 3 results indicates that the system ~~under unforced training could~~ correctly classified~~y~~ 24 out of 25 credible messages during unforced training, but only one o~~ out of 25~~

~~messages and only one message was incorrectly classified. O~~ut of the eight messages where the ~~dubious~~credibility was unknown was correctly classified.~~messages, the system correctly identified only one message.~~ This outcome resulted in a high number of ~~It was clear that the system required further training as the number of~~ False Positives for the unforced training ~~were very high~~which indicated that further training was required.

When the system utilised the same training data set but ran under forced training conditions the results as expected varied (Table 4). Of the 25 credible messages 23 messages were correctly classified and only two messages incorrectly classified. These results only varied slightly from the unforced training outcomes in regard to detecting credible messages correctly. However, there was a significant improvement in the correct detection of messages where the credibility was unknown with all messages being correctly classified during this test. Overall, the results were considered acceptable with a high classification accuracy for both the credible messages classification and the classification where the credibility of the messages was unknown and hence validated the forced training conditions.

Table 4: Test~~ing~~ 2 - Forced training using small training sample (35 messages) and 33 test messages. ~~(Forced) results with 25 credible, 10 dubious and 33 testing messages.~~

	Classified credible	Classified as ~~dubious~~credibility unknown	Total
Actually credible	23 (TP)	2 (FN)	25
Actual~~ly dubious~~credibility unknown	0 (FP)	8 (TN)	8
Total	23	10	33

~~When the system ran under forced conditions, 23 credible messages out of 25 were correctly~~

Formatted: Paragraph

Formatted: Line spacing: single

Formatted: Line spacing: single

~~classified and only two messages incorrectly classified (Table 4). All the dubious messages were correctly classified in this test. Overall, the results were acceptable as the classification accuracy was high in both credible and dubious messages identification. Moreover, the results of the test encouraged the running of the system under forced conditions.~~

Next, the size of the training sample was increased from 35 messages to 77 messages and ~~the system was run under normal conditions.~~ then the unforced and forced training was repeated on the same test data set. The results of ~~this test~~unforced training are shown in ~~(~~Table 5 and ~~)~~ ~~show~~identify ~~that there is an impact of sample size increment for the classification accuracy. In this instance~~that for the credible message classification, 21 ~~credible messages~~ out of 25 messages were correctly classified which was a small decrease in accuracy~~slight drop~~ compared to the previous result (Table 3). However, the classification accuracy ~~of~~ where the ~~dubious~~credibility of the message was unknown, ~~messages~~ improved from one correctly classified message to ~~as~~ five correctly classified messages out of the eight to be classified.

~~were correctly classified.~~

Table 5: Test 3 – Unforced training using the larger training sample (77 messages) and 33 test messages.

~~Testing 3 (Unforced) results with 53 credible, 24 dubious and 33 testing messages.~~

	Classified credible	Classified as ~~dubious~~credibility unknown	Total
Actually credible	21 (TP)	4 (FN)	25
Actual~~ly dubious~~credibility unknown	3 (FP)	5 (TN)	8
Total	24	9	33

~~The~~ Finally, Table 6 shows the results of the classification using the larger training data set

Formatted: Line spacing: single

Formatted: Line spacing: single

under forced training conditions ~~and with increased sample size~~. ~~However, according to the results it can be seen that there is no impact of change of sample size when the system run under forced conditions.~~The results of the testing are identical to the forced training using the smaller training data set with 23 out of 25 credible messages correctly classified and all eight messages where the credibility was unknown were also correctly classified. ~~The testing 2 results (Table 4) and testing 4 results (Table 6) were similar and the classification results were identical.~~ This indicated that the forced training conditions were consistent and were not impacted by the changed training sample size.

Table 6: Test 4 - Forced training using the larger training sample (77 messages) and 33 test messages.~~Testing 4 (Forced) results with 35 credible, 20 dubious and 33 testing messages.~~

	Classified credible	Classified as ~~dubious~~credibility unknown	Total
Actually credible	23 (TP)	2 (FN)	25
Actual~~ly dubious~~credibility unknown	0 (FP)	8 (TN)	8
Total	23	10	33

Formatted: Line spacing: single

Formatted: Line spacing: single

A number of measures such as accuracy, precision, sensitivity and the F1 s~~S~~core provided ~~some~~ an indication~~s~~ of ~~the~~each classification's ~~outcomes~~effectiveness. The accuracy, which is the ratio of correctly predicted observations, ~~can be~~was calculated by the formula (TP+TN)~~-~~/ (TP+TN+FP+FN). The precision or Positive Predictive Value (PPV) is the ratio of correct positive observations. The PPV ~~can b~~was~~e~~ calculated by TP~~-~~/~~-~~(TP + FP). The F1 s~~S~~core (F1) is used to measure classification performance using the~~the~~ weighted recall and precision, where the recall is the percentage of relevant instances that are retrieved and ~~as explained in section two~~ was ~~and can be~~ calculated by 2*TP / (2*TP + FP + FN). ~~and~~ T ~~t~~he sensitivity or True Positive Rate (TPR) ~~is~~ was calculated by TP / (TP + FN).

The classification quality ~~was tested in each training and test phase~~for the four tests are summarised in ~~.~~ Table 7. ~~shows the classification quality using different indicators.~~ The accuracy and precision was higher for the forced training outcomes for both training sample sizes and indicates the importance of the forced training. It can ~~be~~also be ~~clearly~~ seen that the classification accuracy and precision increased slightly for the unforced training outcomes when the larger training sample size ~~is~~ was utilised~~increased~~. ~~both in the forced~~However, the precision and accuracy outcomes for ~~and~~ the ~~un~~forced trainings~~teps~~ were similar and indicate that there may be a lesser dependency on the size of the training data set when force training is utilised. The F1-Score did not change with the sample size but the measures indicate that the ~~-~~forced training again performed better than the unforced training scenarios. Finally, t~~T~~he classification sensitivity remained constant for the forced training for both training sample sizes but dropped slightly with the larger training sample size for the unforced training test outcomes.~~increments whilst still providing a good result.~~

Formatted: New paragraph

~~When the system was run under a forced environment, all the indicators improved except the sensitivity which was remained high. However, the change in the sample size had limited~~

~~impact for the classification results and all indicators remained largely unchanged.~~ Table 7: Quality of the CSD classification

Test Scenario	Accuracy	Precision	F1-Score	Sensitivity
Test – 1 Unforced Using the small training sample (35 messages) and 33 test messages	76	77	86	96
Test -2 Forced Using the larger training sample (77 messages) and 33 test messages	94	100	96	92
Test – 3 Unforced Using the small training sample (35 messages) and 33 test messages	79	88	86	84
Test – 4 Forced Using the larger training sample (77 messages) and 33 test messages	94	100	96	92

Formatted Table
Formatted: Line spacing: single
Formatted: Font: 10 pt
Formatted: Font: 10 pt
Formatted: Font: 10 pt
Formatted: Font: 10 pt
Formatted: Font: 10 pt
Formatted: Line spacing: single
Formatted: Font: 10 pt
Formatted: Font: 10 pt
Formatted: Line spacing: single
Formatted: Line spacing: single
Formatted: Font: 10 pt
Formatted: Font: 10 pt
Formatted: Font: 10 pt

4.2. Results of the full Ushahidi Crowdmap data CSD analysis

~~After~~ After the training testing ~~training~~ of the system was~~has~~ completed ~~with~~ to an acceptable classification quality, the full Ushahidi Crowdmap sample of remaining 433 messages ~~which was allocated for credibility testing (this was the remainder of the full dataset kept for testing which was 520 and it became this number after the pre-processing and further duplicates were removed)~~ was~~as~~ analysed for credibility. As the Figure 3 (a) indicates, 54% (234 out of 433) of the messages were identified as credible using an unforced training classification~~in the Crowdmap data~~. However, w~~W~~hen the system was run under forced conditions, 77% (334 out of 433) of the messages were identified as credible (Figure 3 (b)). This was a more confident value than the previous result as the accuracy and precision of the credibility detection was higher.

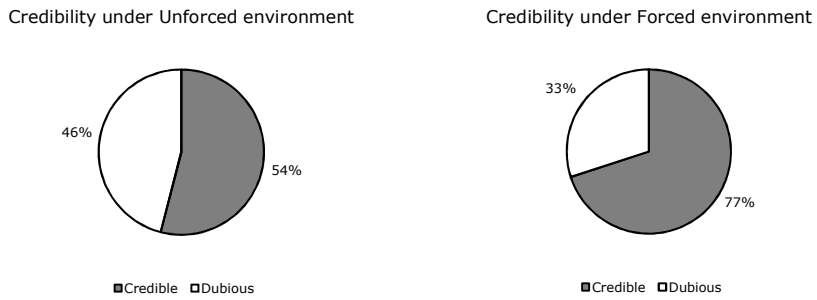


Figure 3: Credibility of 2011 Australian flood's Ushahidi Crowdmap data

5. Conclusion

The CSD message credibility detection is a challenging task ~~as identified by various researchers.~~due to the high degree of variability of the data, the lack of a consistent data structure, the variability of the data providers and the limited metadata available. This study ~~has~~ identified that ~~the~~ Bayesian spam email detection approaches can be applied successfully to the challenge of classifying ~~and~~ the ~~CSD~~ credibility of CSD.~~detection have some conceptual similarities.~~ However, the training approaches and the size of the training data set can influence the quality and performance of the training outcomes.

-Due to the variability of the data, it is recommended that forced training is undertaken to achieve the highest accuracy and performance. In particular, the forced training provided a higher level of confidence in eliminating the number of False Positive (FP) outcomes which were the incorrect classification of messages. The size of the training data set was found to be less critical when a forced training approach was utilised with the results of the

Formatted: Left, Tab stops: 5.19", Left

classification outcomes being similar for both the smaller and larger training data sets. However, if the system training was unforced, a larger training data set is recommended.

~~This study analysed the CSD credibility using an approach well accepted and commercially used in the field of email spam detection. The results of the study indicate that a modified spam email detection approach may be appropriate for CSD credibility detection. However, it is important to ensure the accuracy and performance of this approach over other available spam detection approaches such as machine learning and statistical techniques are considered.~~

~~The study concludes that in regard to CSD credibility detection models,~~

- ~~CSD credibility analysis and spam email analysis are somewhat conceptually similar, however differing approaches are required;~~
- ~~CSD credibility detection models need to be trained under very careful and highly controlled conditions; and~~
- ~~The impact of the size of the training sample can be influenced by forceful training of the system~~

Although this study focussed on the issue of credibility, i~~I~~t should be recognised that ~~the~~ the *relevance* of that dataset is another critical dimension in the quality assessment of the crowd sourced datasets~~. is incomplete until the *relevance* of that dataset is also assessed.~~ It is often not enough to just have a credible source of information as~~,~~ it is also important that the information is relevant to the purpose of the operational activity. For example, in the case of a flood disaster, the relevant information should relate to useful and relevant data regarding the support of the flood operations or emergency services. It is therefore important that f~~F~~uture ~~work of this~~ studies~~y is planned to~~ analyse both the *credibility* and the *relevance* of the

Formatted: Font: 12 pt, Italic

~~Ushahidi Crowdmap~~crowd sourced dataset~~s. in the case of the flooding operational event. The study has identified that the CSD quality can be understood using credibility and relevance parameters, however it is not certain whether the relevant CSD would always be credible or the other way round. To answer this important question, it is planned to assess the impact of CSD credibility for its relevance and vice versa after the relevance of the Ushahidi Crowdmap data set has been identified.~~

Acknowledgments

Authors wishes to acknowledge the Australian G~~g~~overnment for providing support for the research work through the Research Training Program (RTP) ~~funds~~ and Monique Potts, ABC – Australia for providing the 2011 Australian Flood's Ushahidi Crowdmap data.

Formatted: Font: 12 pt, Pattern: Clear

Formatted: Font: 12 pt, Pattern: Clear

References

Androutsopoulos, Ion , John Koutsias, Konstantinos V Chandrinos, and Constantine D Spyropoulos. 2000. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. Paper presented at the Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.

Antoniou, V, and A Skopeliti. 2015. "Measures and Indicators of Vgi Quality: AN Overview." *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5:345-51. doi: 10.5194/isprsannals-II-3-W5-345-2015.

Bahree, Megha. 2008. "Citizen Voices." *Forbes Magazine* 182 (12):83.

Bishr, Mohamed, and Werner Kuhn. 2007. "Geospatial information bottom-up: A matter of trust and semantics." In *TThe European information society: Leading the way with geo-information*, edited by S Fabrikant and M Wachowicz, 365-87. Berlin: Springer.

Blanzieri, Enrico, and Anton Bryl. 2008. "A survey of learning-based techniques of email spam filtering." *Artificial intelligence review* 29 (1):63-92.

Brando, Carmen, and Bénédicte Bucher. 2010. Quality in user generated spatial content: A matter of specifications. Paper presented at the Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimarães, Portugal.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. Paper presented at the Proceedings of the 20th international conference on World wide web.

Cheng, Jie, and Russell Greiner. 1999. "Comparing Bayesian network classifiers." In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 101-8. Stockholm, Sweden: Morgan Kaufmann Publishers Inc.

Coleman, David J, Yola Georgiadou, and Jeff Labonte. 2009. "Volunteered geographic information: The nature and motivation of produsers." *International Journal of Spatial Data Infrastructures Research* 4 (1):332-58.
Cranor, Lorrie Faith, and Brian A. LaMacchia. 1998. "Spam!" In *Communications of the ACM*, 74-83.
De Longueville, Bertrand, Nicole Ostlander, and Carina Keskitalo. 2010. "Addressing vagueness in Volunteered Geographic Information (VGI)–A case study." *International Journal of Spatial Data Infrastructures Research* 5:1725-0463.
Elwood, Sarah. 2008. "Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS." *GeoJournal* 72 (3-4):173-83.
Flanagin, Andrew J, and Miriam J Metzger. 2008. "The credibility of volunteered geographic information." *GeoJournal* 72 (3-4):137-48.
Fogg, BJ, and Hsiang Tseng. 1999. The elements of computer credibility. Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems.
Goodchild, Michael F. 2007. "Citizens as sensors: the world of volunteered geography." *GeoJournal* 69 (4):211-21. doi: 10.1007/s10708-007-9111-y.
Goodchild, Michael F. 2009. "NeoGeography and the nature of geographic expertise." *Journal of Location Based Services* 3 (2):82-96. doi: 10.1080/17489720902950374.
Goodchild, Michael F., and J. Alan Glennon. 2010. "Crowdsourcing geographic information for disaster response: a research frontier." *International Journal of Digital Earth* 3 (3):231-41. doi: 10.1080/17538941003759255.
Goodchild, Michael F., and Linna Li. 2012. "Assuring the quality of volunteered geographic information." *Spatial Statistics* 1:110-20. doi: 110-120. doi: 10.1016/j.spasta.2012.03.002.
Graham-Cumming, John. 2006. "Does Bayesian poisoning exist." In *Virus Bulletin*, 69.
Guzella, Thiago S., and Walmir M. Caminhas. 2009. "A review of machine learning approaches to Spam filtering." *Expert Systems with Applications* 36 (7):10206-22. doi: 10.1016/j.eswa.2009.02.037.
Haklay, Mordechai. 2010. "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets." *Environment and planning. B, Planning & design* 37 (4):682-703. doi: 10.1068/b35097.
Heipke, Christian. 2010. "Crowdsourcing geospatial data." *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (6):550-7. doi: 10.1016/j.isprsjprs.2010.06.005.
Hovland, Carl I, Irving L Janis, and Harold H Kelley. 1953. "Communication and persuasion; psychological studies of opinion change."
Howe, Jeff. 2006. "The rise of crowdsourcing." In *Wired magazine*, 1-4.
Hung, Kuo-Chih, Mohsen Kalantari, and Abbas Rajabifard. 2016. "Methods for assessing the credibility of volunteered geographic information in flood response: A case study in Brisbane, Australia." *Applied Geography* 68:37-47. doi: 10.1016/j.apgeog.2016.01.005.
Janowicz, Krzysztof, Sven Schade, Arne Broring, Carsten Kebler, Patrick Maue, and Christoph Stasch. 2010. "Semantic enablement for spatial data infrastructures." *Transactions in GIS* 14 (2):111-29. doi: 10.1111/j.1467-9671.2010.01186.x.
Kang, Byungkyu, John O'Donovan, and Tobias Höllerer. 2012. Modeling topic specific credibility on twitter. Paper presented at the Proceedings of the 2012 ACM international conference on Intelligent User Interfaces.
Kim, Heejun. 2013. "Credibility assessment of volunteered geographic information for emergency management: a Bayesian network modeling approach." University of Illinois.

Koswatte, Saman, Kevin McDougall, and Xiaoye Liu. 2016. "Semantic Location Extraction from Crowdsourced Data." *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*:543-7.

Longueville, Bertrand De, Gianluca Luraschi, Paul Smits, Stephen Peedell, and Tom De Groeve. 2010. "Citizens as sensors for natural hazards: A VGI integration workflow." *Geomatica* 64 (1):41-59.

Lopes, Clotilde, Paulo Cortez, Pedro Sousa, Miguel Rocha, and Miguel Rio. 2011. "Symbiotic filtering for spam email detection." *Expert Systems with Applications* 38 (8):9365-72. doi: 10.1016/j.eswa.2011.01.174.

Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. 2006. Spam filtering with naive bayes-which naive bayes? Paper presented at the CEAS.

Noy, Natalya F, Nicholas Griffith, and Mark A Musen. 2008. "Collecting community-based mappings in an ontology repository." In *The Semantic Web-ISWC 2008*, edited by A. Sheth, S Steffen, M Dean, M Paolucci, D Maynard, T Finin and K Thirunarayan, 371-86. Springer.

Ostermann, Frank O, and Laura Spinsanti. 2011. A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management. Paper presented at the Proceedings of AGILE, University of Utrecht, Utrecht

Pantel, Patrick, and Dekang Lin. 1998. Spamcop: A spam classification & organization program. Paper presented at the Proceedings of AAAI-98 Workshop on Learning for Text Categorization.

Robinson, Gary. 2003. "A statistical approach to the spam problem." *Linux journal* 2003 (107):3.

Sadeghi-Niaraki, Abolghasem, Abbas Rajabifard, Kyehyun Kim, and Jungtaek Seo. 2010. Ontology Based SDI to Facilitate Spatially Enabled Society. Paper presented at the Proceedings of GSDI 12 World Conference.

Sahami, Mehran, Susan Dumais, David Heckerman, and Eric Horvitz. 1998. A Bayesian approach to filtering junk e-mail. Paper presented at the Learning for Text Categorization: Papers from the 1998 workshop.

Schneider, Karl-Michael. 2003. A comparison of event models for Naive Bayes anti-spam e-mail filtering. Paper presented at the Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1.

Senaratne, Hansi, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri, and Mordechai Haklay. 2016. "A review of volunteered geographic information quality assessment methods." *International Journal of Geographical Information Science*:1-29. doi: 10.1080/13658816.2016.1189556.

Shvaiko, Pavel, and Jérôme Euzenat. 2013. "Ontology Matching: State of the Art and Future Challenges." *IEEE Transactions on Knowledge & Data Engineering* 25 (1):158-76. doi: 10.1109/TKDE.2011.253.

Wang, Alex Hai. 2010. Don't follow me: Spam detection in Twitter. Paper presented at the Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, 26-28 July 2010.