Extracting Information Nuggets from Disaster- Related Messages in Social Media

Conference Paper · May 2013

CITATIONS
465

READS
2,872

5 authors, including:

Muhammad Imran
Qatar Computing Research Institute
154 PUBLICATIONS **6,666** CITATIONS
SEE PROFILE

Shady Elbassuoni
American University of Beirut
67 PUBLICATIONS **2,219** CITATIONS
SEE PROFILE

Carlos Castillo
University Pompeu Fabra
310 PUBLICATIONS **20,629** CITATIONS
SEE PROFILE

Fernando Diaz
Microsoft
54 PUBLICATIONS **3,850** CITATIONS
SEE PROFILE

Extracting Information Nuggets from Disaster-Related Messages in Social Media

Muhammad Imran[1] University of Trento Imran@disi.unitn.it
Shady Elbassuoni American Univ. of Beirut se58@aub.lb.edu
Carlos Castillo QCRI chato@acm.org
Fernando Diaz Microsoft Research fdiaz@microsoft.com
Patrick Meier QCRI pmeier@qf.org.qa

ABSTRACT

Microblogging sites such as Twitter can play a vital role in spreading information during "natural" or man-made disasters. But the volume and velocity of tweets posted during crises today tend to be extremely high, making it hard for disaster-affected communities and professional emergency responders to process the information in a timely manner. Furthermore, posts tend to vary highly in terms of their subjects and usefulness; from messages that are entirely off-topic or personal in nature, to messages containing critical information that augments situational awareness. Finding actionable information can accelerate disaster response and alleviate both property and human losses. In this paper, we describe automatic methods for extracting information from microblog posts. Specifically, we focus on extracting valuable "information nuggets", brief, self-contained information items relevant to disaster response. Our methods leverage machine learning methods for classifying posts and information extraction. Our results, validated over one large disaster-related dataset, reveal that a careful design can yield an effective system, paving the way for more sophisticated data analysis and visualization systems.

Keywords

Supervised classification, Information Extraction, Social Media, Twitter

INTRODUCTION

Microblogging platforms have become an important way to share information on the Web, especially during time-critical events such as "natural" and man-made disasters. In recent years, Twitter[2] has been used to spread news about casualties and damages, donation efforts and alerts, including multimedia information such as videos and photos (Balana, 2012; Pew 2012; Blanchard, Carvin, Whittaker, Fitzgerald, Herman and Humphrey, 2010). Given the importance of on-topic tweets for time-critical situational awareness, disaster-affected communities and professional responders may benefit from using an automatic system to extract relevant information from the Twitter Firehose.[3] An automatic system for disaster-related information extraction requires two components: Classification of tweets and Extraction from tweets. First, because the messages generated during a disaster vary greatly in value, an automatic system needs to filter out messages that do not contribute to situational awareness. These include those that are of personal nature and those not relevant to the disaster. As a result, we design a system for detecting *informative* messages. Once a system has detected tweets likely to contain relevant information, it must analyze candidate tweets to decide the *type of information* to extract (e.g. donation offers, casualty reports). The final system output consists of *information nuggets*, brief, self-contained pieces of information most likely to augment situational awareness[4].

This paper is organized as follows. First, a short overview of the dataset is provided. Next, the ontology and process for generating training data for the automatic classifiers and extractors is described. The latter are then evaluated on a real-world dataset. The paper concludes by comparing the findings with that previous research.

THE JOPLIN DATASET

The dataset consists of tweets posted during the Joplin 2011 tornado that struck Joplin, Missouri in the late

[1] Work done while the author was at QCRI.
[2] An online microblogging service that enables millions of users to share text-based short messages.
[3] http://iRevolution.net/2012/12/17/debating-tweets-disaster
[4] While we describe our system for the case of tweets, it can be applied to any sort of social media without any fundamental changes to the system components.

afternoon of Sunday, May 22, 2011. The dataset was originally constructed by researchers at the University of Colorado at Boulder[5]. The 206,764 unique tweets were selected by monitoring the Twitter Streaming API using the hashtag *#joplin* a few hours after the tornado hit. This monitoring process continued until the number of tweets about the tornado became particularly sparse [6].

DISASTER-RELATED MESSAGE ONTOLOGY

The system needs to detect messages that may add situational awareness information—that is, tweets that provide "tactical, actionable information that can aid people in making decisions, advise others on how to obtain specific information from various sources, or offer immediate post-impact help to those affected by the mass emergency" (Vieweg et al., 2012). To this end, the categories of messages we considered were,

- *Personal Only*: if a message is only of interest to its author and her immediate circle of family/friends and does not convey any useful information to other people who do not know the author.
- *Informative (Direct)*: if the message is of interest to other people beyond the author's immediate circle, and seems to be written by a person who is a *direct eyewitness* of what is taking place.
- *Informative (Indirect)*: if the message is of interest to other people beyond the author's immediate circle, and seems to be seen/heard by the person on the radio, TV, newspaper, or other source. The message must specify the source.
- *Informative (Direct or Indirect)*: if the message is of interest to other people beyond the author's immediate circle, but there is not enough information to tell if it is a direct report or a repetition of something from another source.
- *Other*: if the message is not in English, or if it cannot be classified.

Only informative messages were selected for subsequent study since this class is most likely to contain information valuable for disaster recovery and rescue. The ontology developed in the thesis by (Vieweg et al., 2012) forms the basis for the categories below. Vieweg et al., created a comprehensive list of several dozen "Information Types" using inductive, data-driven analysis of twitter communications, which she combined with findings from the disaster literature and official government procedures for disaster response. In total, Vieweg et al., identified 32 specific types of information that contribute to situational awareness. These included Caution, Advice, Fatality, Injury, Offers of Help, Missing and General Population Information. The ontology used for this study was,

- *Caution and advice*: if a message conveys/reports information about some warning or a piece of advice about a possible hazard of an incident.
- *Casualties and damage*: if a message reports the information about casualties or damage done by an incident.
- *Donations of money, goods or services*: if a message speaks about money raised, donation offers, goods/services offered or asked by the victims of an incident.
- *People missing, found, or seen*: if a message reports about the missing or found person effected by an incident or seen a celebrity visit on ground zero.
- *Information source*: if a message conveys/contains some information sources like photo, footage, video, or mentions other sources like TV, radio related to an incident.

These classes, taken together, provide a principled method for accumulating—either manually or automatically—tweets likely to contain relevant information nuggets.

[5] Sincere thanks to Kate Starbird and Project EPIC at University of Boulder, Colorado, for sharing Tweet IDs.
[6] Due to Twitter's data-sharing policy, the entire Joplin dataset could not be directly shared with us. Instead, we were provided by the tweet ids which were used to retrieve the tweets from the Twitter API. We could not construct the entire Joplin dataset because some user accounts were either deleted or inaccessible after the construction of the original Joplin dataset.

MANUAL CLASSIFICATION AND EXTRACTION WITH CROWDSOURCING

This section describes the manual classification process carried out by crowdsourcing workers. The labeling process was performed using CrowdFlower[7], a crowdsourcing platform that works across multiple crowdsourcing services (including Amazon's Mechanical Turk). Each worker was presented with a set of instructions and a few items to label. A small number of the items to label belong to a gold standard dataset. The gold standard dataset is a small set of items selected by the authors of this paper, whose labels are uncontroversial; for each item in the gold standard, a message is included which is shown to workers who did not agree with the target label. Workers that do not agree substantially with the golden standard are removed from the pool, and their labels are ignored. The design of the labeling task was driven by an effort to keep each task as simple as possible. In our opinion asking simple and straightforward questions increases the overall quality of a task, which is considered to be a major concern in crowdsourcing. In total, we asked crowdsourcing workers to annotate 4,406 tweets sampled uniformly at random form the Joplin dataset.

Task 1: Filtering Informative Messages

The first task corresponds to an annotation of tweets according to whether they are entirely of personal nature, informative ("direct", "indirect", or "direct or indirect"), or other, according to the ontology described above. We created 190 gold standard messages.

Task 2: Classifying Messages By Type

The second task, which comprised 1,233 messages, asked workers to examine an individual tweet carefully and assign an appropriate label from the given five categories. We created 62 gold standard messages for this task, equally distributed among the categories. Only the messages labeled as informative (1,233 in total) are considered in the rest of the tasks.

Task set 3: Classifying Messages By Sub-Type and Extracting Information Nuggets

For each of the four types, a separate crowdsourcing task was created.[8] As in previous tasks, workers were asked to classify a message by sub-type. Additionally, they were asked to extract a sub-string from the message containing a given piece of factual information.

MANUAL CLASSIFICATION RESULTS

The bulk of the classification was done during October 2012. Each task typically involved from 50 to 100 different contributors. Contributions from 3 trusted (agreeing sufficiently with the gold standard) workers per item was required. The total cost of the labeling effort did not exceed 200 US dollars.

Quantitative Assessment

In the first task, which is used to filter only informative messages, 32% of the messages were labeled as "Personal", 19% were labeled as "Informative (direct)", 24% labeled as "informative (indirect)", 17% labeled as "Informative (direct-indirect)" and the remaining 8% were labeled as "Other". In our case the most important categories were "Informative (direct), Informative (indirect) & Information (direct-indirect)" and as an overall 60% of the messages were labeled to these categories. For the first task the inter-annotator agreement was 55.94%. The inter annotator agreement value shows the level of agreement among workers on an assessable unit (i.e., in our case a tweet). High agreement indicates that different workers frequently gave the same response for the same tweet message.

In Table 1, the first two columns show the categories and the corresponding distribution of the messages classified for that category. Of the 1,233 representative informative messages half (i.e., 50%) of them were labeled as "Caution & advice". As an overall 94% of all the messages were assigned a category and only 6% were labeled as unknown. The inter-annotator agreement for this task was 74.16%.

In Table 2, the last two columns show the sub-types categories and their corresponding distribution of messages.

[7] http://www.crowdflower.com

[8] Each category in the second task became a new task except the "People missing, found, or seen" category, as no messages were labeled for this category.

In figure 4, we show inter-annotator agreement for all the questions asked in all the tasks. High inter-annotator agreement values validate high quality of the results.

Types	Percentage	Sub-types	Percentage
Caution and advice	50%	A siren heard	12%
		A tornado/thunderstorm warning issued/lifted	42%
		A tornado sighting/touchdown	30%
		Other	16%
Information source	18%	Webpage information source	44%
		Photo information source	16%
		Video information source	20%
		Other	18%
Donation	16%	Money	38%
		Equipment	2%
		Shelter	2%
		Volunteers	2%
		Blood	2%
		Other	54%
Causalities & damage	10%	People dead	44%
		Infrastructure damage	10%
		People injured	2%
		Both people and infrastructure damage	2%
		Not specified but damage	34%
		Unknown	8%
Unknown	6%		

Table 1. Crowdsourcing results of types and sub-types.

Qualitative Assessment

Personal/Informative

Several notable patterns emerge from the first round of manual coding for types of disaster related tweets. Tweets conveying "Personal Messages" are frequently miscoded as "Informative (Direct)," while Tweets discussing official tornado watches also tend to get coded as "Informative (Direct)." A number of tweets also covey both personal sentiment as well as personal observation, which may be one source of confusion among coders. One example of this is the following tweet: "Informative (Direct) = Thankful we have thunder & lightning w/o tornados today. Our thoughts & prayers are with the folks in #Joplin today." Finally, some of the multimedia content—particularly pictures—shared about the Joblin tornado are actually of hurricanes from previous years. This was also an issue during October 2012 when Hurricane Sandy struck the Northeastern US. Overall, the quality of the coding was relatively high.

Informative Direct/Indirect

Tweets conveying "Personal Messages" were frequently miscoded as "Informative (Direct)," and vice versa. Take this tweet, for example: "Personal only: Carrie & Em are ok -- house wasn't touched but lots of other damage around #tornado." While the first half of the message is indeed personal, the latter is Informative (Direct). The introduction of the new category "Informative (Direct or Indirect)" also resulted in some patterns that suggested confusion among coders. In particular, "Informative (Direct)" was often coded as "Informative (Direct or Indirect)." In addition, tweets that reported headline news (or were in the form and style of headline news) were often coded as "Informative (Direct or Indirect)". A different pattern that emerged had to do with the "Other" category, which often tended to include information important to disaster response but not otherwise captured by another category.

Caution/Advice/Damage/Donations

The manual codification of tweets in this dataset focused on a different set of categories. Again, some patterns tend to emerge. For example, messages that are personal in nature appear to throw many coders off.

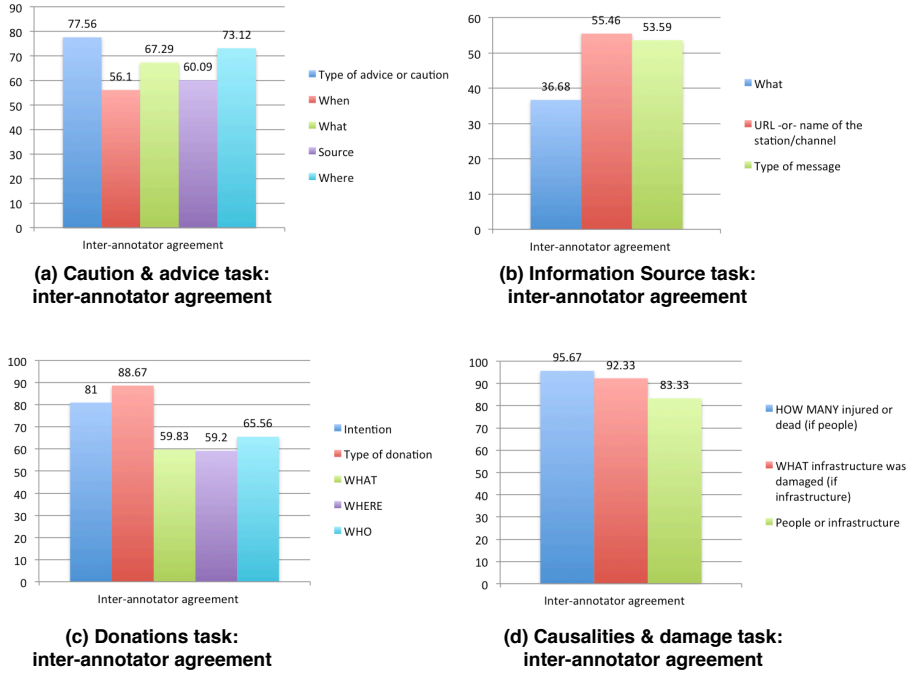

Figure 4. Charts showing inter-annotator agreement for all the sub-types.

In addition, there is frequent miscoding and confusion vis-a-vis the following two categories: "Casualties/Damage" and "Information Source". This confusion is mostly the result of our codification rules, however, since the two categories just listed are not mutually exclusive while our rules require coders to choose between one or the other. For example, "@andersoncooper: I took this earlier. Cross still stands above destroyed church #Joplin http://yfrog.com/h2d7rgrj" is coded as "Casualties/Damage" but the tweet also conveys/contains an information source—i.e., a picture. Still, the most robust coding for this dataset appear to be for the Casualties/Damage category and the "Donation of money, goods or services" category. Overall, the intercoder reliability for this dataset and category set is relatively more robust than that of the Informative Direct vs Indirect categories.

Sub-Cluster Extractors

The coding of "What," "Where," and "Who" segments appear to be relatively robust, including intercoder reliability. An interesting pattern worth noting, however, is the coding of the "What" segment which at times results in the codifying of references to people while at other times is used to capture the disaster itself. In terms of categorization, the option "Other" is at times miscoded despite seemingly obvious references to various warnings. There is miscoding of the "Warning" category. For example, this tweet: "Emergency crews flying down highway toward Springfield at an alarming rate. Wish I stayed for the second #tornado" was wrongly coded as a "Warning."

AUTOMATIC CLASSIFICATION AND EXTRACTION WITH MACHINE LEARNING

Classification

A set of multi-label classifiers were trained to automatically classify a tweet into one or more of the classes identified in the previous section. Naïve Bayesian classifiers were used as implemented in Weka (Hall, Frank, Holmes, Pfahringer, Reutemann and Witten, 2009). To this end, a number of binary, scalar, and text features were employed. The *binary features* consisted of: (1) whether a tweet contains the @ symbol or not; (2) whether

a tweet contains a URL or not; (3) whether a tweet contains a hashtag or not; (4) whether a tweet contains an emoticon or not; and (5) whether a tweet contains a number or not. The *scalar features* only consisted of a numeric feature indicating the tweet length. The *text features* consisted of sparse linguistic features. Specifically, all non-words were removed (i.e., Twitter handlers, URLs, hashtags, emoticons and numbers) along with Twitter specific stopwords. The remaining text was stemmed after which the following textual features were extracted: (1) unigrams; (2) bigrams; (3) Part of Speech (POS) tags for which the Stanford POS tagger (Toutanova, Klein, Manning and Singer, 2003) was used; (4) Part of Speech tag-bigrams; and (5) Verbnet classes to extract the set of Verbnet classes for each verb that appears in the tweet. Verbnet (Kipper, Dang, Palmer, 2000) is an ontology for verbs that organizes a set of similar verbs into classes and consists of a hierarchy of verb classes that include hypernyms, synonyms, etc.

Filtering Informative Tweets

As noted above, the first classification problem consists of classifying the tweet into one of three classes: Informative, Personal or Other. A classifier was trained to do so using the features set articulated above. The training data was obtained using crowdsourcing as explained in the previous section. Since the purpose of this study is to identify informative tweets, the AUC of the Informative Class was used as an indicator of the performance of a classifier. The first classifier was able to identify Informative tweets with an AUC of 0.79 and a precision of 0.79 at recall 0.77 after a 10-fold cross validation.

unigrams	Bigrams	pos	pos bigrams	Tweet Length	hashtag	url	emoticon	@	number	verbnet	AUC
Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	0.828
Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	No	0.827
Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	No	No	0.827
Yes	Yes	Yes	No	Yes	No	No	Yes	No	Yes	No	0.826
Yes	Yes	Yes	No	Yes	No	No	No	No	Yes	No	0.826

Table 3. Top-5 best performing feature combinations & their performance in terms of AUC of the Informative Class.

A feature selection experiment was carried out to test the effect of various features on the performance of the classifier. Table 3 shows the performance of top-5 feature combinations where "Yes" indicates the feature was included and "No" indicates it was not used by the classifier. The last column indicates the AUC of the Informative Class. Unigrams, bigrams, POS tags and Tweet Length excluding POS bigrams and the Verbnet classes are the common features in all top-5 best performing feature combinations. The rest do not appear to have a clear effect on the performance of the classifier.

Identifying Eye-Witness Tweets

The second classification problem is a binary classification problem concerned with classifying informative tweets into either Eye Witness or not (i.e., classifying tweets into Informative Direct or Indirect as described in the previous section). This information is crucial in differentiating between tweets that are aimed to share information reported in news outlets and those that are first-hand experience which are only communicated through word of mouth or through microblogs. Again, a classifier using the same set of features described in the beginning of this section was developed and the training data obtained by using crowdsourcing as described in earlier. The performance of the classifier after a 10-fold cross validation resulted in the following: identification of eyewitness tweets with precision 0.57 at recall 0.57.

Classifying Informative Tweets into Caution, Donation, Casualty or Information Source

Class	AUC	Precision	Recall	F-Measure
Caution	0.911	0.859	0.765	0.809
Donation	0.890	0.726	0.716	0.721
Casualty	0.871	0.526	0.652	0.583
Information Source	0.763	0.545	0.581	0.562

Table 4. 10-fold cross validation evaluation of the classifier classifying an Informative tweet into Caution/Advice, Donation, Casualty/Damage or Information Source.

The third and final classification problem was to classify informative tweets into one of the following four categories: Caution/Advice; Donation; Casualty/Damage; and Information Source. Again, training data was generated using crowdsourcing as per the two previous classifiers. The results are listed in Table 4.

Information Extraction

Various types of information were extracted from informative tweets depending on their class, which we refer to as information nuggets. For each class, we describe the information extracted and how we extract them next.

Caution and Advice Nuggets

For Caution tweets, the following information was extracted: 1) Location references; 2) Time references; 3) Caution/Advice; 4) Source; and 5) Type of Caution.

For location and time references, the Stanford Named Entity Recognizer (Finkel, Grenager and Manning, 2005) was used. For the Caution/Advice part, the Stanford Part of Speech Tagger (Toutanova and Manning, 2000) was used to tag the tweet. The tagged tweets in the training data were then reviewed to identify the set of consecutive tags that used to tag any manually extracted Caution/Advice part. For example, assuming the manually extracted caution/advice part of a given tweet was "Tornado Warning" and that after applying POS tagging, the part "Tornado Warning" was tagged with the following tag sequence: "NN NN". Then, "NN NN" would be extracted as a possible tag sequence. The same procedure was used for every manually extracted Caution / Advice part to get a set of tag sequences. Those tag sequences were then mined to retrieve the most common ones among them that appear in most of the tweets in the training data. Finally, given the tag sequences extracted, the Caution/Advice extractor works by applying part of POS tagging to a given tweet and extracting the part of the tweet tagged by an tag sequence in the set.

For sources, all Twitter Handlers were extracted (i.e., words starting with the @ symbol and URLs from the tweet). These typically contain the source from which the information conveyed in the tweets was retrieved. Finally, for the type of caution, a classifier was trained to automatically classify an "Advice and Caution" tweet into the following classes: 1) Warning issued or lifted; 2) Siren heard; 3) Shelter open or available; and 4) Disaster sighting or touchdown.

Casualty and Damage Nuggets

For Casualty tweets, the following was extracted: 1) Location references; 2) Time references; 3) Number of Casualties; 4) Damaged Object; 5) Source; and 6) Type of Casualty/Damage.

Like Caution tweets, the Stanford Named Entity Recognizer was used to extract locations, times and number of casualties. For the damaged object, the Wordnet (Miller, 1995) classes (Synsets) were identified that cover all the damaged objects that manually extracted. That is, in most of the casualty and damage tweets, these were either: 1) a thing constructed; a complex entity constructed of many parts (Sysnet ID 4290455) such a house or a church; or 2) a large indefinite location on the surface of the Earth (Synset ID 8150527) such as a neighborhood; or 3) a living organism lacking the power of locomotion (Synset ID 16858) such as a tree. Given a damage tweet, all the nouns in the tweet was extracted (using Stanford POS tagger) along with any nouns whose hypernym is one of the above identified Wordnet classes.

The sources were extracted in the exact same way as in the case of Advice and Caution tweets. For the type of casualty/damage, a classifier was trained to automatically classify a casualty as: 1) Infrastructure; 2) Death; 3) Injury; 4) Unspecified; 5) No Damage; and 6) Both Infrastructure and People.

Donation and Offer Nuggets

For donation tweets, the following information was extracted: 1) Location references; 2) Time references; 3) Intention of Tweet; 4) Source; and 5) Type of Donation. The location and time references and the sources was extracted in the exact same way as in the previous two cases. For the Intention of the tweet, a classifier was trained to classify a donation tweet to whether it is reporting a donation effort or requesting one. The training data was obtained by manual labeling of *204* tweets on CrowdFlower. For the type of donation, another classifier was trained to automatically classify a donation tweet into the following classes: 1) Money; 2) Blood; 3) Voluntary Work; 4) Food; 5) Equipment; 6) Shelter; 7) Discounts; and 8) Other. The training data was obtained by manually labeling the Donation tweets on CrowdFlower.

Information Source Nuggets

For information source tweets, the following was extracted: 1) Location references; 2) Time references; 3) Source; and 4) Type of Information. Location and time references and sources were extracted in the exact same way as in the previous two cases. For the type of information, a classifier was developed to classify an information source tweet into the following classes: 1) Photo; 2) Video; 3) Website; 4) TV Channel; 5) Radio Station; and 6) Unspecified. The training data was obtained by manually labeling *279* tweets on CrowdFlower.

Evaluation

Classification Task	Weighted AUC	Weighted Precision	Weighted Recall	Weighted F-Measure
Predicting Caution	0.771	0.618	0.598	0.605
Predicting Casualty	0.770	0.578	0.645	0.610
Predicting Donation	0.668	0.546	0.632	0.585
Predicting Information Source	0.656	0.476	0.434	0.435

Table 5. Weighted AUC, precision, recall and F-measure for the various classification tasks.

The performances of the various classifiers described in the previous subsections are shown in Table 5. To measure the quality of our information extractors, a set of training tweets was used to ask CrowdFlower workers to manually extract the desired information. We then tried to automatically extract the same information from the same tweets using our extractors. Finally, the hit ratio was computed, which is defined as follows:

$$hit - ratio = \frac{\sum_{i=1}^{|tweets|} hit_i}{|tweets|}$$

where |tweets| is the total number of tweets for which we have manually extracted items and

$$hit_i = \begin{cases} 1 & \text{if } editdistance(manual_i, auto_i) \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

$manual_i$ is the manually extracted information from tweet t_i and $auto_i$ is the automatically extracted information from tweet t_i. Note that here we assume that from each tweet, only one possible piece of information (of certain) type is extracted from the tweet. The above definition can be easily generalized to the case where there are multiple pieces of information extracted per tweet but we avoid this here for simplicity and since that case occurred very rarely in our experiment.

Extractor	# Tweets	# Tweets with Extracted Items	Hit Ratio	Expected Hit Ratio	Precision
Damaged Object	18	24	0.500	0.861	0.47
# Casualties	60	61	0.933	0.983	0.79
Caution/Advice	283	143	0.265	0.709	0.71
Location	395	157	0.192	0.827	0.93
Time	176	73	0.216	0.781	0.85
Source	514	512	0.819	0.870	0.83

Table 6. The performance of the various extractors.

Table 6 reports the performance of the different extractors. The first column contains the number of tweets for which we had manually extracted information and the second column contains the number of tweets for which we managed to automatically extract corresponding information. As can be seen from Table 6, our Source and Number of Casualties extractors have a hit ratio (column 3) close to 1 indicating that they manage to extract most of the human extracted information. This is also evident in the fact that the Source extractor managed to extract information for 512 tweets whereas the human annotators managed to extract information for 514 tweets.

Similarly, for the Number of Casualties extractor, it extracted information for 61 tweets and there were only 60 tweets for which manually extracted information exist. Note that any tweets for which no manually extracted information existed were excluded from our hit ratio computation.

To better interpret the hit ratios reported in Table 6, the expected hit ratio of the human annotators (column 4) was also computed. The expected hit ratio of a human annotator is the average hit ratio as defined above when taking one annotation as a test value and the rest as training ones and then averaging over the number of annotations per tweet. As can be seen from columns 3 and 4, the extractors for Sources and Number of Casualties have a hit-ratio very close to the expected one. However, for the rest of the tasks our extractors seem to perform poorly in comparison to the human annotators. This indicates the difficulty of such tasks and the need for more sophisticated extraction algorithms for these difficult tasks.

Finally, the quality of extractors was also assessed by asking CrowdFlower workers to rate the quality of the extracted segments. Each worker was presented a tweet, an extracted segment, and a question that was a function of the type and sub-type of the tweet and the class of segment extracted. For instance, if type was "Donation", sub-type was "Money", and class "Location", we asked "In <tweet>, is <Location> the name of a place where money is donated, received, or needed?" The last column in Table 6 shows the precision of the various extractors (i.e., the percentage of extracted items that were perceived as useful by the human judges). We labeled 1,488 tweets providing 46 items in the gold standard. As can be seen from this column, the precision of all the extractors are above 0.7 except for the Damaged Object Extractor. This indicates that while some of classifiers do have a low ratio, this is due to a recall issue—meaning that the extractors are over conservative, extracting only information that is very likely to be correct. In the future, we plan to develop more sophisticated extractors that use complex Natural Language Processing (NLP) techniques in order to improve the recall of the extractors without unduly sacrificing their high precision.

RELATED WORK

Twitter has been used extensively during emergency situations in recent years. In November 2012, Twitter revealed that over 2 million tweets had been posted during Hurricane Sandy. In 2011, over 5,500 tweets were posted every *second* following the tsunami and earthquake in Japan (Anderson, 2012). Users posted over 2 million tweets about Haiti following the earthquake in January 2010 (Sakaki, Okazaki and Matsuo, 2010). In Terpstra et al. (2012), authors show a real-time analysis of Twitter data. They use predefined geographical displays, twitter message content etc. to analyze and filter crisis-related information.

Most event-detection methods, however, are based on keyword (Mathioudakis and Koudas, 2010). Sakaki et al. (2010) analyze keywords used in tweets during disaster events in order to classify messages as real-time event reports, using a support vector machine (SVM). As we show on this paper, our methods compares favorably against a supervised SVM. Other related research includes Vieweg et al., (2012) methodology and disaster ontology to identify tweets that provide situational awareness. Kongthon et al., (2012) develop classifiers to analyze tweets from the Thailand floods in 2011. (Starbird, Palen, Hughes and Vieweg, 2010) and (Vieweg, Hughes, Starbird and Palen, 2010) analyzed microblog usage and information lifecycles during crisis situations.

CONCLUSION

In this paper, we presented a system to automatically extract information nuggets from microblogging messages during disaster times. Our system utilizes state-of-the-art machine learning techniques to classify messages into set of fine-grained classes and to extract short self-contained structured information that can be leveraged for complex data analysis and integration beyond plain text. The system was tested on a real-world disaster-related dataset consisting of hundreds of thousands of microblogging messages. The training data for our machine learning techniques was generated using crowdsourcing and our techniques were then evaluated on the same dataset. The results of our experiments show that indeed machine learning can be utilized to extract structured information nuggets from unstructured text-based microblogging messages with good precision and recall. Next, we plan to test our techniques on different disaster-related datasets to gain insight on how dataset tailored our techniques should be.

The results of this applied and ongoing research will continue to inform the development of the Qatar Computing Research Institute's (QCRI) Twitter Dashboard for Disaster Response.9 The purpose of this experimental dashboard is to provide humanitarian organizations like the United Nations Office for the

[9] http://iRevolution.net/2013/02/11/update-twitter-dashboard

Coordination of Humanitarian Affairs (UN OCHA) with a platform that will enable them to create their own automatic classifiers on the fly—real-time supervised learning combined with machine learning.

REFERENCES

1. Anderson, C. (2012) Japan Earthquake Social Media Coverage: Disaster By The Numbers. *Mashable*.
2. Balana, C.D. (2012) Social media: major tool in disaster response. *Inquirer Technology*, 5 pp.
3. Blanchard, H., A. Carvin, M.E. Whittaker, M. Fitzgerald, W. Harman and B. Humphrey. (2012) The case for integrating crisis response with social media. *White Paper, American Red Cross*, Washington, DC, 32 pp.
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B. and P, Reutemann, and Ian H. Witten. (2009) The Weka Data Mining Software: An Update. *SIGKDD Explor. Newsl*. 11, 1, 10-18.
5. Hughesm A.L. and L. Palen. (2009) Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6, 3, 248-260.
6. Kipper, K., Dang, H. and M. Palmer. (2000) Class-Based Construction of a Verb Lexicon. *In Proceedings of the Seventeenth National Conference on Artificial Intelligence,* Austin, TX, 691-696.
7. Kongthon, A., Haruechaiyasak, C., Pailai, J. and S. Kongyoung. (2012) The Role of Twitter During a Natural Disaster: Case Study of 2011 Thai Floods. *2012 PICMET*, Technology Management for Emerging Technologies.
8. Kristina, T. and C.D. Manning. (2000) Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora,* 63-70.
9. Mathioudakis, M. and N, Koudas. (2010) Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 1155-1158, New York, USA.
10. Miller G. (1995) Wordnet: a lexical database for English. *Communication of the ACM*, 38, 11, 39-41.
11. Pew Research. (2012) Hurricane Sandy and Twitter. *Pew Research Center for Excellence in Journalism*, New York.
12. Rose, J. and T.G. Finkel, and C. Manning. (2005) Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*. 363-370.
13. Starbird, K., L. Palen, A.L. Hughes, and S. Vieweg. (2010) Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 241-250.
14. Toutanova, K., Klein, D., Manning, D. and Y. Singer. (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology,* 1, 03, 173-180, Stroudsburg, PA, USA.
15. Vieweg, S., A.L. Hughes, K. Starbird, and L. Palen. (2010) Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, 1079-1088.
16. Terpstra, Teun, R. Stronkman, A. de Vries, and G. L. Paradies. (2012) Towards a realtime Twitter analysis during crises for operational crisis management. In *Proceedings of the 9th International ISCRAM Conference*. Vancouver, Canada.