Information Processing and Management 57 (2020) 102107

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Automatic identification of eyewitness messages on twitter during disasters

Kiran Zahra[a,*], Muhammad Imran[b], Frank O. Ostermann[c]

[a] University of Zurich, Switzerland
[b] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[c] University of Twente, The Netherlands

ARTICLE INFO

Keywords:
Social media
Eyewitness identification
Machine learning
Disaster response

ABSTRACT

Social media platforms such as Twitter provide convenient ways to share and consume important information during disasters and emergencies. Information from bystanders and eyewitnesses can be useful for law enforcement agencies and humanitarian organizations to get firsthand and credible information about an ongoing situation to gain situational awareness among other potential uses. However, the identification of eyewitness reports on Twitter is a challenging task. This work investigates different types of sources on tweets related to eyewitnesses and classifies them into three types (i) direct eyewitnesses, (ii) indirect eyewitnesses, and (iii) vulnerable eyewitnesses. Moreover, we investigate various characteristics associated with each kind of eyewitness type. We observe that words related to perceptual senses (feeling, seeing, hearing) tend to be present in direct eyewitness messages, whereas emotions, thoughts, and prayers are more common in indirect witnesses. We use these characteristics and labeled data to train several machine learning classifiers. Our results performed on several real-world Twitter datasets reveal that textual features (bag-of-words) when combined with domain-expert features achieve better classification performance. Our approach contributes a successful example for combining crowdsourced and machine learning analysis, and increases our understanding and capability of identifying valuable eyewitness reports during disasters.

1. Introduction

At times of disasters caused by natural and anthropogenic hazards, people use social media platforms such as Twitter and Facebook to share information (Imran, Castillo, Diaz, & Vieweg, 2015; Vieweg, Hughes, Starbird, & Palen, 2010) that can potentially be useful for disaster response. This information includes reports of injured and dead people, urgent needs of affected people, reports of missing and found people, and reports of unrest and looting, among others (Imran et al., 2015). Social media not only contains useful information, it also breaks stories and events faster than many other traditional information or news sources such as TV. For instance, the first report of the Westgate Mall attack[1] in Nairobi, Kenya in 2013 was published on Twitter, almost 33 minutes before a local TV channel reported the event. Similarly, the news about the Boston bombing incident[2] appeared on Twitter before any other

* Corresponding author.
E-mail addresses: kiran.zahra@geo.uzh.ch (K. Zahra), mimran@hbku.edu.qa (M. Imran), f.o.ostermann@utwente.nl (F.O. Ostermann).

[1] https://en.wikipedia.org/wiki/Westgate_shopping_mall_attack .
[2] https://en.wikipedia.org/wiki/Boston_Marathon_bombing .

https://doi.org/10.1016/j.ipm.2019.102107
Received 8 April 2019; Received in revised form 17 July 2019; Accepted 26 August 2019
Available online 27 September 2019

news channel reported the event. Likewise, in the case of the California earthquake[3] it was observed that the first half dozen tweets were recorded by Twitter about a minute earlier than the recorded time of the event according to the USGS. These first-hand reports come from eyewitnesses and bystanders, i.e. people who directly observe the occurrence of an event (Diakopoulos, De Choudhury, & Naaman, 2012; Zahra, Imran, & Ostermann, 2018).

At the onset of a disaster event, people share massive amounts of data such as damage reports, casualties, but much of that data has redundant information, e.g. through sharing the same news article or video. For instance, millions of messages were posted on Twitter during Hurricane Harvey in 2017, many containing similar information.[4] Nevertheless, studies have revealed that the information sources also include many local citizens, bystanders, and eyewitnesses. From the perspective of an information seeker (affected citizen or institutional response agency), information from eyewitness reports is preferred over other types of information sources (e.g. people outside the disaster area). Law enforcement agencies and first responders always look for first-hand and credible information for decision-making. Humanitarian organizations look for timely and trustworthy information that is directly observed from the disaster-hit areas to better estimate the severity and scale of damage, and the amount of aid required to help save lives and fulfill the urgent needs of affected people.

Gaining rapid access to the information shared by eyewitness reports, especially during an ongoing disaster event, is thus useful but challenging to obtain (Imran, Mitra, & Srivastava, 2016). The most straightforward approach to identify local residents in disaster-hit areas is through geotagged information, e.g. Twitter messages (called tweets) with attached coordinates from a global navigation satellite system (e.g. the Global Positioning System, or GPS). However, given that only 1–3% of tweets are geotagged, relying solely on those to identify local residents may not provide enough data required for decision-making. Moreover, not all tweets from the disaster-struck area automatically come from eyewitnesses. Many social media platforms allow the user to enter manually a home location in the user profile, but research has shown that at least in the case of Twitter it is very noisy and inaccurate, and it does not indicate location of the source at the time when a tweet is made (Lee, Ganti, Srivatsa, & Liu, 2014).

Given the above issues, it remains a challenge to process potentially millions of tweets and identify eyewitnesses reliably. In Doggett and Cantarero (2016), the authors identify a set of eyewitness and non-eyewitness linguistic features to categorize eyewitness news-worthy events on human-induced disasters such as protests, shooting, and police activities. Likewise, in Fang, Nourbakhsh, Liu, Shah, and Li (2016), the authors highlight a similar set of linguistic (e.g. personal or impersonal expressions, time awareness) and meta-features (e.g. client application) to identify witness reports on various natural and human-induced disasters. They also used the topic (e.g. accidents, crimes and disasters) of tweets as a feature to automatically classify tweets as witness reports. The work presented in Tanev, Zavarella, and Steinberger (2017) identified a set of eyewitness features from several dimensions and categorized stylistic, lexical Twitter metadata and semantic features, and Truelove, Vasardani, and Winter (2014) developed a generalized conceptual model of different types of eyewitness reports for several events such as concerts, shark sightings, cyclones, and protests. However, none of the works (i) develop their classifiers through a combination of expert-driven and data-driven feature engineering, and (ii) differentiate between particular types of eyewitnesses found during natural disasters, and (iii) operationalize different characteristics associated with those types, and finally (iv) validate the results for different disaster types using crowdsourced annotations.

This paper aims to address this research gap by designing an eyewitness reports taxonomy focusing on the needs of disaster response agencies during natural disasters. Moreover, we explore different types of features associated with tweets to train machine learning models for the automatic classification of tweets. For this purpose, we first manually learn different characteristics (mainly language based) from tweets posted by eyewitnesses. We use these characteristics as independent (domain-experts) features together with content-based features from tweets to train several machine learning models. We establish that when domain-expert features are combined with the text-based features of tweets, these models outperform those which are trained on independent features. Since creating a balanced labeled dataset from social media labeled data is a challenging task, we employee a state-of-the-art class balancing technique and demonstrate that a model trained on a balanced data can achieve even higher results. Lastly, we contribute a successful example of combined crowdsourced training of machine learning classification (Imran, Lykourentzou, Naudet, & Castillo, 2013; Ostermann, Garcia-Chapeton, Kraak, & Zurita-Milla, 2018). The contributions of this work are summarized as follows:

- Designing a taxonomy consisting of different sub-types of eyewitnesses i.e., direct eyewitness, indirect eyewitness, vulnerable direct eyewitness.
- A generalized methodology, which can be applied on textual data from other domains to extract eyewitness reports, that uses textual content-based features and domain-expert features without relying on platform-specific features such as Twitter metadata.
- Combining textual and domain-expert features to train and evaluate automatic machine learning classifiers on real-world disaster-related Twitter datasets.
- Last but not least, we offer all the labeled data obtained through crowdsourcing and manual analysis to the research community to further develop and extend this line of research. The dataset will be shared at the CrisisNLP repository: https://crisisnlp.qcri.org/.

The following section gives a brief overview of related work, before we describe in more detail the methodology and obtained results in their respective sections. Before concluding, we then discuss lessons learnt and evaluate our work.

[3] http://latimesblogs.latimes.com/technology/2008/07/twitter-earthqu.html .
[4] https://crisiscomputing.qcri.org/2017/09/27/hurricane_harvey_and_the_role_ai/ .

2. Literature review

Our study uses a Twitter dataset. Twitter is a well established source to harvest opportunistically information during crisis events. Twitter messages are called tweets, and are micro blog posts, i.e. small packets of information of originally 140 characters (now doubled to 280). By default, all posts are public and can be found by everyone using the proper search parameters. Tweets can use hashtags to facilitate this search. Additionally, users can retweet any tweet and follow each other to see posts within their network on their timeline. Therefore, sharing information has zero marginal cost for the user: Posting a tweet requires only a smartphone, a Twitter app or log-in through the web interface, and a comparatively narrow network bandwidth (more if videos are to be shared). According to Twitter usage statistics,[5] around 500 million tweets are posted per day. A challenge for utilizing this massive volume of information is the often informal, unstructured, and noisy nature of Twitter posts and communication.

Twitter also has a well-establish history in breaking news in real-time. These news are often generated/started by a person who has witnessed the event. A very recent example is the eyewitness report of an emergency landing of Delta Aircraft due to its engine failure in Alaska. This eyewitness report has enough credibility to become the news source for a traditional news agency.[6] Another classic example of an eyewitness report on Twitter is the New York airplane crash in Hudson bay in 2009. This tweet also became the headline of The Daily Telegraph.[7] In Kwak, Lee, Park, and Moon (2010), the authors argue that Twitter serves also as news source and not only as social media platform. Their research reveals that over 85% of trending topics are news headlines.

Because many Twitter users post their personal experiences during a natural disaster, people are motivated to search for breaking news and real-time content on Twitter (Teevan, Ramage, & Morris, 2011) as in case of disasters (Allen, 2014; Amaratunga, 2014; Kryvasheyeu et al., 2016; Schnebele et al., 2013). In Oh, Agrawal, and Rao (2013), the authors explore the use of Twitter during social crisis situations. Academic research into Twitter and disaster management has mainly focused on user contributed data in disaster response (Haworth & Bruce, 2015) and relief phase (Landwehr & Carley, 2014) such as the Haiti earthquake in 2010 (Meier, 2012) or during forest fires (Ostermann & Spinsanti, 2012).

In the case of emergency events, extraction of relevant and reliable information from noise and redundant information is critical. Originally, researchers and relief organizations alike attempted to crowdsource the curation of information processing during disasters. Early systems such as CrisisCamp[8] and Ushahidi[9] are well-established platforms to gather and network volunteers from all over the globe to help solving different problems using a collective wisdom. Imran et al. (2014) used Standby Task Force[10] volunteers to label if the tweet belongs to information category. The work presented in Purohit et al. (2014) used Crowdflower[11] (now figure-eight) to classify tweets in categories such as requests for help or offers to provide help, among others. They also used crowdsourcing to label the resources available during the crisis. However, their results focused on the evaluation of their machine learning model. They did not evaluate the performance of their crowdsourcing task itself. In Snow et al. (2008), the authors used Amazon Mechanical Turk[12] (AMT) to evaluate the effectiveness of tasks performed by various “expert” and “non-expert volunteers” and suggested a technique to remove bias. In Callison-Burch (2009), the authors used AMT to evaluate the translation of different texts a much faster and cheaper than the conventional ways.

To facilitate and support these tasks, humanitarian and disaster relief organizations also developed real-time tweet crawler applications such as TweetTracker (Kumar, Barbier, Abbasi, & Liu, 2011), Artificial Intelligence for Disaster Response (AIDR) (Imran, Castillo, Lucas, Meier, & Vieweg, 2014), Twitcident (Abel, Hauff, Houben, Stronkman, & Tao, 2012), ScatterBlogs for situational awareness (Thom et al., 2015), cross-language aspects on Twitter (Imran et al., 2016), or during a particular disaster such as Typhoon Haiyan in the Philippines (Takahashi, Tandoc, & Carmichael, 2015). In Zahra, Ostermann, and Purves (2017), the authors discussed a time-saving and efficient technique to filter the noise from informative tweets. However, crowdsourced social media data generated by often anonymous users suffers from an absence of quality assurance (Goodchild & Li, 2012) on the truthfulness, objectivity, and credibility of the information.

Disaster response organizations search for eyewitness reports as those are considered more credible (Truelove, Vasardani, & Winter, 2015). Researchers have studied possibilities to identify eyewitness reports out of millions of tweets for journalism (Diakopoulos et al., 2012), criminal justice, and natural disasters (Olteanu, Vieweg, & Castillo, 2015). The work in Morstatter, Lubold, Pon-Barry, Pfeffer, and Liu (2014) relates the identification of eyewitness tweets to the use of language and linguistic patterns within the region during different crisis events. They also identified a set of features to automatically classify eyewitness reports. In Kumar, Morstatter, Zafarani, and Liu (2013), the authors use location information of the users to assess local users and remote users on crisis reports.

However, both research strands have worked mostly in isolation until now, with the potential of location information not fully exploited for establishing whether a source is an eyewitness, who might also be vulnerable and at risk. This paper aims to develop a holistic categorization of characteristics of eyewitness reports for frequent types of natural disasters.

[5] http://www.internetlivestats.com/twitter-statistics/ .
[6] https://channelnewsasia.com/news/world/delta-flight-middle-of-the-ocean-seattle-beijing-emergency-land-11062706 .
[7] https://www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html .
[8] https://crisiscommons.org/crisiscamp/ .
[9] https://www.ushahidi.com/ .
[10] http://www.standbytaskforce.org/about-us/our-history/ .
[11] https://www.figure-eight.com/ .
[12] https://www.mturk.com/ .

3. Methodology & experimental framework

This work identifies eyewitnesses messages from Twitter data using the following four steps:

1. **Disaster-related data collection:** First, we collect Twitter data related to four different types of natural disasters. Specifically, several months of data about earthquakes, floods, wildfires, hurricanes are collected and used in this work. We then retrieved two data samples from our corpus. Our first sample is used for an initial manual analysis (see next step) and includes data from three natural disasters types: earthquake, hurricane, and floods. Our second sample is used for crowdsourced annotation and includes data on four natural disaster types: earthquake, hurricane, floods, and wildfire. Data collection details are described in the next section.
2. **Manual Analysis to determine eyewitnesses types and characteristics:**To determine different types of eyewitness reports, we first analyse our first data sample taken from the collected data for three disaster types. Next, tweets which are identified as posted by eyewitnesses are further analyzed to understand different linguistic characteristics using their content (i.e., message text). The following annotation guidelines were developed and followed for the manual analysis:
   - **Identify tweet source:** This task aims to determine the source of a given tweet i.e., whether it is posted by an eyewitness or not, using only the message content. For this purpose, we consider the following three categories for the analysis:
     (i) **Eyewitness:** if the message is posted by an eyewitness
     (ii) **Non-eyewitness:** if the message is posted by anyone else other than an eyewitness
     (iii) **Don't know:** if it is not possible to determine which of the above two categories
   - **Identify eyewitness type:** If the previous task identifies a tweet as posted by an eyewitness, then this task aims to further determine whether the author of the tweet is a direct eyewitness, or passes on information he/she learned from a familiar source or is otherwise familiar with. We term that second type of eyewitness an *indirect eyewitness*. We also categorize a group of eyewitness tweets as *vulnerable eyewitness* where people were anticipating a disaster and were present in the region for which disaster warnings were issued.
   - **Identify eyewitness message characteristics:** This task aims to identify various linguistic characteristics and clues from the contents of eyewitness tweets. In the remainder of the paper, we refer to these as domain-expert features.
3. **Crowdsourcing to obtain labeled data:**The types of eyewitnesses and content characteristics identified in the previous step are then used to obtain labeled data using a paid crowdsourcing platform on our second data sample which includes tweets from four natural disasters i.e. earthquake, hurricane, flood, and wildfire.
4. **Training supervised machine learning models:**We consider the task of determining whether a tweet is posted by an eyewitness or not as a classification task. We use supervised machine learning techniques to train models on the crowdsourced labeled data using the following steps:
   - **Automatic feature extraction**: we first extract textual features from the textual content of the labeled tweets. For this purpose, we use two types of textual features (i) uni-grams and (ii) bi-grams and compute their TF-IDF scores Hong, Dan, and Davison (2011).
   - **Feature selection**: Feature selection techniques help identify features which help classifiers discriminate well between different classes and also help in generalization. We use the information gain feature selection technique to choose the top performing features for each class.
   - **Supervised models learning**: Among different learning schemes, Random Forest is considered best for the classification of textual data (Xu, Guo, Ye, & Cheng, 2012). We use Random Forest to train our classification models. Specifically, for each event type (e.g., earthquake), we train four different models as follows:
     (i) **Training using textual features (baseline):** In this case, we only use Bag-of-Words (BOW) based features extracted from the content of the labeled tweets.
     (ii) **Training using domain-expert features:** We extract features from tweets content using the characteristics identified by the domain-experts and use them to train new models.
     (iii) **Training using text and domain-expert features:** In this case, both text-based and domain-experts features are combined to train models.
     (iv) **Training using text and domain-expert features with class balancing:** Models trained on imbalanced classes always suffer performance issues. To tackle this problem, we used SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), which is a well-known class balancing technique. We retrained our models on reasonably balanced classes using the combined features from text and domain-experts.

The models performance is evaluated using the cross-validation (10-fold) technique and presented using standard performance evaluation metrics such as precision, recall, and F-measure.

4. Data, manual analysis and crowdsourcing

In this section, we provide details of the data collection, manual analysis, and crowdsourced annotation.

Table 1
Frequency of eyewitness, non-eyewitness, and unknown reports.

Event type	Sampled Tweets	Eyewitness	Non eyewitness	Don't know
Floods	2000	148	113	1739
Earthquakes	2000	367	321	1312
Hurricanes	2000	296	100	1604

4.1. Data collection

We used the Twitter Streaming API to collect data from July 2016 to May 2018 using a methodology described in this paper (Zahra et al., 2017). Specifically, we used *earthquake, foreshock, aftershock, flood, inundation, extensive rain, heavy rain, hurricane, cloud-burst, forest fire*, and *wildfire* keywords to collect in total 25 million tweets related to earthquakes, hurricanes, floods, and forestfires. For the manual analysis, we used two samples from this tweet corpus. Our first sample was retrieved from 1 to 28 August 2017 from three disaster types i.e., earthquake, flood, and hurricane. We chose this time period because of the occurrence of several such disaster events during that period. Our second sample was retrieved from July 2016 to May 2018 - but excluding the first sample's time period for earthquake - for the same three disaster types, and wildfire. Each sample consisted of 2000 randomly selected tweets from each disaster type. Our first data sample was used for the manual analysis and our second data sample was used for crowdsourcing.

4.2. Manual analysis results

Following the *Identify tweet source* annotation task guidelines described in the previous section, two authors of this paper manually analysed every tweet in the sample. Table 1 shows the results of this manual analysis. The number of tweets posted by eyewitnesses is very limited. A total of 148, 367, and 296 messages were found as posted by an eyewitness for floods, earthquakes, and hurricanes respectively. For many tweets it was not possible to determine with sufficient reliability whether they were posted by eyewitnesses or not, i.e., the *don't know* cases in the last column of Table 1. This difficulty already hints at the challenge automated classification may face.

Next, by following the *Identify eyewitness type* annotation task guidelines, tweets posted by eyewitnesses are analyzed further to determine if there are different types of eyewitness reports. Table 2 shows the results of this analysis. Mainly, three types of eyewitnesses are identified namely (i) direct eyewitness, (ii) indirect eyewitness, and (iii) vulnerable direct eyewitness. We provide details for each of the type in the following subsections.

4.2.1. Direct eyewitness

A direct eyewitness report represents first-hand knowledge and experience of an event. There are different ways in which direct eyewitness reports can provide information on events. Table 3 shows some examples of direct eyewitness reports taken from the manually analyzed data from all three disaster types.

The first message on floods reports the personal experience of the author about a flood situation. The second message is even more interesting since the author not only reports the event, he/she also complains about a lack of notifications or flood warnings in his/her area. Similarly, in the third message the author reports about high flood waters and that he/ she has got stuck due to it. The fourth message is also about a personal experience of a flash flood situation. All the earthquake-related messages in Table 3 express personal experiences of the authors about some earthquake events. We observe that in most of the earthquake cases, people express or relate their messages to the sense of feeling such as "just felt" or "feeling shaking".

Regarding hurricane-related messages, the first and second example report about winds and heavy rain, which are obvious signs of a direct personal experience. Both authors experienced the situation and reported it, while the author of the third message not only reported a flood situation but also gave an indication that the situation could get worse. The last example is again a personal experience of an event where the author is also reporting a power outage.

One common observation from the analysis of direct eyewitness reports is that eyewitnesses often mention the severity of situation they are in. Moreover, people associate their messages to different senses like "seeing", "feeling", "hearing" or "smelling". For example, in case of an earthquake they relate it to the sense of "feeling" such as, *Just felt an earthquake...* Likewise, in case of a storm or floods, tweets are related to the sense of seeing or hearing such as; *I've never seen or heard such a violent thunder/hail/rain storm as the one we've just experienced.*

Table 2
Frequency of different types of eyewitness reports.

Event type	Direct eyewitness	Indirect eyewitness	Vulnerable direct eyewitness
Floods	62	2	84
Earthquakes	354	13	0
Hurricanes	95	16	185

Table 3
Direct eyewitness reports from manual analysis .

No.	Floods direct eyewitness reports
(1)	I almost died driving home from work because it started to downpour and flood on the freeway and lightning and its 99 f**king degrees out
(2)	No one even notified me that this flood in our area has reached almost 3 feet. but atleast i was able to reach home safely.
(3)	Stuck in New Brunswick. High flood waters near Rutgers. Rt 1 south #Avoid
(4)	I just experienced a flash flood. they're intense
	Earthquakes direct eyewitnesses reports
(1)	Most intense earthquake i've experienced in japan so far... that is
(2)	Big midnight earthquake and aftershocks now
(3)	Just felt the house shaking in Tokyo. Been awhile since I felt an earthquake. I hope it wasn't a bad one anywhere on the island.
	Hurricanes direct eyewitness reports
(1)	Please pray for us right now, the winds and rain is heavy and the hurricane hasn't even hit us yet. #hurricaneharvey2017
(2)	This hurricane ain't no joke, the rain and winds are heavy right now. #hurricaneharvey2017
(3)	It's starting to flood in our area (hurricane Harvey) so if I don't respond back within a 7++ days expect for the worse hope we'll be safe
(4)	first time is street is starting to flood and the power went out, hurricane harvey finally hit us

4.2.2. Indirect eyewitness

During our manual analysis, we found several tweets where the author was sharing information from direct witnesses. Most often, they were sharing valuable information received from friends, relatives, and their social circle. Although we found only a small number of tweets from this category in our dataset, they are an interesting and potentially useful category, as they also allow information to cross platforms or communication channels (e.g. someone tweets about a story heard over the phone). Table 4 shows some examples of indirect eyewitness reports taken from the manually analyzed data from all three disaster types.

There were only two messages found in the flood dataset where indirect eyewitnesses were reporting about disasters by referring to their family members, while for earthquakes, one indirect eyewitness was reporting about an ongoing earthquake with emotions of worry for his/her family who were direct witnesses of this earthquake. The second example shows a unique case where an indirect eyewitness is reporting about an earthquake he/she was experiencing live but distantly during a video call with one of their family members. The last example in this section is reporting about the safety of direct witnesses from a relative.

Finally, indirect eyewitness reports on hurricanes were very interesting. The first and second examples are about an indirect eyewitness' hometown conditions mixed with emotions of worry. The third example shows concern of the indirect eyewitnesses about their relatives' property due to the prospective hazard. In the last example, the indirect eyewitness is sharing the direct eyewitness report of his friend.

4.2.3. Vulnerable direct eyewitness

During our manual analysis of sampled tweets we noticed tweets where users were anticipating a disaster and were reporting warnings and alerts they received from local authorities on their cell phones. This type of messages was only found in the floods and hurricanes datasets, probably due to the more predictable nature of those events. These tweets constitute an interesting subgroup of direct eyewitness information, because identifying people at risk is important information for crisis managers to allocate resources effectively. Table 5 shows some examples of vulnerable direct eyewitness reports.

The first message in the floods section reports the personal experience of the author about a flood warning where he relates the

Table 4
Indirect eyewitness reports from manual analysis .

No.	Floods indirect eyewitness reports
(1)	Some days in Thailand has been insane, there has been massive flood on the road to the city (only have image on my dad's phone)
(2)	The hsm school and my uncles house are right behind eachother and they were ruined in the flash flood)):
	Earthquakes indirect eyewitnesses reports
(1)	F*cking hell... my wife and kids are in Tokyo and they're in the middle of an earthquake Jesus Murphy just how crap can one day get?
(2)	Was Facetiming my brother in Tokyo when an earthquake. It wasn't strong but took a long time. Glad that he's ok. #tokyo #earthquake
(3)	Finally able to hear from my uncle and know that he and his daughters are safe, the earthquake did not affect them to much #bless #mexico
	Hurricanes indirect eyewitness reports
(1)	Texas has me going for a spin....my hometown was evacuated for the hurricane then an earthquake in Dallas where my entire family is
(2)	My city is getting a rain storm from the hurricane and hella winds but that's nothing compared to what's going on god i'm so worried
(3)	So this hurricane is heading for my brother and sister-in-law's brand new winery. Hope it doesn't get flooded before https://t.co/VBfAchRplM
(4)	Heard from friends in Houston, Austin and San Antonio. High winds and heavy rain last night. Everyone is safe. #hurricaneharvey

Table 5
Vulnerable direct eyewitness reports from manual analysis.

No.	**Floods vulnerable direct eyewitness reports**
(1)	Flash flood warning yet it's not even raining
(2)	Why am I always napping when a flash flood warning comes on to my phone? #scared
(3)	Those flash flood alerts will kill me one day, they scare the f**k out of me
(4)	Ima throw my phone if I get another flood warning
	Hurricanes vulnerable direct eyewitness reports
(1)	Hurricane Harvey is approaching.Dun dun dun. first hurricane I will experience in Texas in my new home omg I hope my area doesn't flood
(2)	Staying home for the hurricane,hopefully it doesn't flood
(3)	I'm so scared I hope this hurricane don't flood my apartment or my car
(4)	Big hurricane is supposed to hit the area tonight and i live in one of the flood zones...

alert with the current weather situation. The second and third messages depict emotions of fear created because of a hazard warning. On the other hand, in the fourth example the eyewitness is angry because of so many flood warnings. On the same note, in the hurricane section the first, second, and third examples are showing mixed emotions of hope and fear while reporting about an approaching hurricane. In last example, the eyewitness is relating their vulnerability to the intensity of approaching hazard.

In this particular category of messages, we also noticed a mix of different types of emotions written in words (not emojis) such as hate, disgust, fear, anger, and humor.

4.2.4. Non-eyewitness reports

In our dataset, tweets which did not possess any explicit eyewitness characteristics, but possessed non eyewitness characteristics (Doggett & Cantarero, 2016). However, these tweets were nevertheless reporting about disasters and were categorized as non eyewitness reports. These reports were sharing disaster related information primarily from news media sources.

4.2.5. "Don't know" cases and noise

There were a number of tweets where disaster related keywords were used as metaphors, such as *Troll army will then flood social media with press cuttings, naughty headlines, whatsapp distortions to offset growing positive opinion*. Such messages were categorized as noise along with any messages containing disaster-related keywords in URL's instead of text body. Furthermore, there were several tweets which were possibly eyewitness reports but were too ambiguous to classify them as such. We put these tweets also in this category.

4.3. Characteristics of different types of eyewitness reports

This section describes the manual analysis task to extract common characteristics of reports posted by the three types of eyewitnesses identified in the previous tasks. We performed this step on our two data samples. However, we found the same characteristics in both datasets.

Information posting on social media platforms are usually restricted by several constraints, e.g. a length limit of originally 140 (now 280) characters on Twitter. Such constraints force social media users to apply creative ways to shorten their messages while conveying their actual intent. As a consequence, Twitter communications differ from usual daily life communications such as emails, blogs etc. We believe that the identification of characteristics associated with each type of eyewitness report will help (i) differentiate among eyewitness types and (ii) also, more importantly, to build automatic computational methods and systems to automatically identify and categorize eyewitness messages.

Table 6
Direct eyewitness characteristics .

No.	Characteristic	Examples
(1)	Reporting small details of surroundings	window shaking, water in basement
(2)	Words indicating perceptual senses	seeing, hearing, feeling
(3)	Reporting impact of disaster	raining, school canceled, flight delayed
(4)	Words indicating intensity of disaster	intense, strong, dangerous, big
(5)	First person pronouns and adjectives	I, we, me
(6)	Personalized location markers	my office, our area
(7)	Exclamation and question marks	!, ?
(8)	Expletives	wtf, omg, s**t
(9)	Mention of a routine activity	sleeping, watching a movie
(10)	Time indicating words	now, at the moment, just
(11)	Short tweet length	one or two words
(12)	Caution and advice for others	watch out, be careful
(13)	Mention of disaster locations	area and street name, directions

Table 7
Indirect eyewitness characteristics .

No.	Characteristic	Examples
(1)	Mention of locations or people the author knows	mom, dad, hometown
(2)	First person adjective	my, our
(3)	Expressing emotions	thoughts, worry, relief
(4)	Reporting safety, damage, missing	missing, safe

4.3.1. Direct eyewitness characteristics

Table 6 lists all the characteristics we have observed in the direct eyewitness messages in earthquake, hurricane, flood, and wildfire types along with examples. As social media communications are short and to the point, users usually skip writing first person pronouns and adjectives. However, if a message has first person pronouns and adjectives, we observe that it is a strong indication of a direct eyewitness report (Fang et al., 2016).

Moreover, we observed that words related to perceptual senses such as *seeing, hearing, feeling* are also strong indications that a message originates from a direct eyewitness. Likewise, words indicating the intensity of a disaster situation such as intense, heavy, strong are extensively found in eyewitness messages posted during all four types of disasters. We suggest that the presence of intensity words is also a strong signal that the message is from an eyewitness, as a person far from the disaster area cannot describe the intensity of the situation. Eyewitnesses tend to mention more about their personalized locations such as my office, our area than non-eyewitnesses. Among other characteristics that are shared across all disaster types include use of exclamation and question marks and special/swear words like “wtf”, “omg”, and “s**t”. However, on the contrary, the characteristic# 11 i.e., *short tweet length* was only found in the earthquake dataset. Many examples were found where users shared tweets consisting of only one or two words to report an earthquake such as *earthquake*!. One probable cause could be the sudden and unpredictable nature of earthquake events. Furthermore, we noticed *caution and advice* and *mention of precise disaster locations* characteristics specifically in flood, hurricane, and wildfire disasters. One possible reason for this observation can be the relatively predictable and long-term nature of these disaster types.

4.3.2. Indirect eyewitness characteristics

Table 7 shows characteristics learnt from indirect eyewitness messages for all disasters. We observed that indirect eyewitness reports either mention a person or a place the contributing user already knows. The social circle of a user tends to be credible and so the indirect eyewitness is also considered credible. If an indirect eyewitness report is about the hometown of a user, then it is assumed that they know the geography of disaster hit region very well and can provide useful information if required. It was also observed that indirect eyewitness reports were either about emotions of worry or sense of relief. Indirect eyewitness reports were also about damage, safety or missing people/property.

4.3.3. Vulnerable direct eyewitness characteristics

Table 8 shows distinct characteristics of vulnerable direct eyewitness reports. This category was only found in flood, hurricane, and wildfire datasets due to their predictable nature. Characteristics 5 to 9 in Table 6 were common in both categories. Users were mostly found reporting about hazard warnings and associating it with current weather situations. As hazard alerts were often sudden in nature, they provoked different types of emotions due to sudden disruptions in user’s routine activities.

4.4. Crowdsourcing

We performed crowdsourced labeling on our second data sample for primarily two reasons: First, to acquire more training data so we can use machine learning algorithms to automatically classify the reports. Second, to validate our own eyewitness reports taxonomy developed from the manual analysis with our first data sample.

We used the Figure-eight platform to crowdsource the categorization of tweets into the eyewitness reports types identified during the manual analysis. Figure-eight is a paid and well-known crowdsourcing platform. We shared the messages with the crowd workers and asked them to categorize them according to the developed taxonomy. The crowdsourcing platform provides various quality control measures to evaluate the process and its results. The first measure is that the contributors themselves are categorized into three levels. Level one is comprised of all qualified contributors, and it got the fastest throughput. Level two is comprised of a smaller group of more experienced contributors delivering a higher accuracy of results. Level three is the smallest group of most experienced

Table 8
Vulnerable direct eyewitness characteristics.

No.	Characteristic	Examples
(1)	Warnings and alerts about expected disasters	flash flood warnings
(2)	Associating warnings with current weather situation	flash flood alert with rain
(3)	Expressing emotions	hate, disgust, anger, scare

Table 9
Crowdsourcing results for second data sample.

Event type	Direct	Indirect	Vulnerable	Non	Don't know
Floods	320	85	222	551	822
Earthquakes	1557	43	–	200	200
Hurricanes	321	67	77	1199	336
Wildfire	122	44	23	1379	432

contributors. We selected contributors from level two because of budget constraints, as the cost increases according with the level of the contributors. Moreover, the crowdsourcing platform provides additional quality control options such as posing initial test questions to annotators before authorizing them to contribute. If the annotator does not pass a threshold percentage of the test questions, he/she may not complete the job. A limit of minimum time spent on the task (in seconds) can also be set to make it sure that the annotator spent enough time in reading and understanding the task. We set a minimum limit of 80 percent accuracy of test questions which means users with 80 percent accuracy or higher of test questions will qualify for the job. Initially we created eight test questions to initiate a quiz and later we added another 13 test questions suggested by the platform based on trusted judgments. We also set a minimum time of 50 seconds for user to spend on reading and understanding the messages before moving on to the next set of messages. We asked for three judgments for every message to be able to assess the inter-rated agreement. We repeated the same step of identifying eyewitness reports features as in manual analysis from three classes (direct eyewitness, indirect eyewitness, and vulnerable direct eyewitness) from crowdsourced labelled dataset.

Crowdsourcing results are shown in Table 9. To evaluate the accuracy of crowdsourced annotation, we conducted an audit on the results. We select 50 sample messages from each dataset and two authors of the paper annotated them again. This annotation was later compared with the crowd annotation and showed 90% accuracy for floods, 92% for earthquake, 82% for hurricane, and 94% percent for wildfire datasets.

5. Experiments and results

The highly varying volume of Twitter streams makes it almost impossible for human analysts to identify eyewitness reports during peaks, e.g. during an evolving disaster. Therefore, we propose to use a supervised machine learning algorithm to automatically identify eyewitness reports.

5.1. Feature engineering

The characteristics of eyewitness messages identified by humans (described in Section 4.3) are operationalized into features to learn automatic classifiers. We refer to these features as domain-expert features.

For characteristics 2, 4, 8, and 12 from Table 6 and characteristic 1 from Table 8, we developed lists of words with the help of dictionaries and thesauri. For example, *words indication perceptual sense* uses words related to hearing and seeing, *indicating intensity of an event* has terms such as intense, small, dangerous, and big as basis and was then expanded with synonyms, while the expletives were build on Wiktionary[13] enriched by various slangs[14], since people often use various slangs on social media. *Caution and advice* words and their synonyms were searched in the dictionary and thesaurus. For characteristic 9 in Table 6 *mention of daily routines*, a list of daily routine activities[15] was used in the present continuous tense. Other characteristics (3, 10) in Table 6 were operationalized directly from message content: *reporting impact of disaster* and *time indicating words* were generated from direct eyewitness messages.

The remaining characteristics (5, 7, and 11) in Table 6, i.e. *first person pronouns and adjectives, exclamation and question marks*, and *short message length* were straightforward to implement by adding the corresponding terms (I, me, ...), characters (exclamation and question marks), and counting the words in the message.

Personalized location markers in Table 6 was overlapping with characteristic 5 and dropped. Likewise, characteristic 2 in Table 7 is overlapping with characteristic 5 in Table 6. Similarly, characteristic 2 in Table 8 is overlapping with characteristic 3 in Table 7. Characteristic 3 and 4 in Table 7 comprised a list of words extracted from indirect eyewitness reports. Stop words[16] were excluded. For each of the operationalized characteristics, we counted the number of occurrences of matching words (only uni-grams) or, in case of *short message length*, the binary absence or presence.

The first characteristic *reporting small details of surroundings* in Table 6 and the second characteristic *associating warnings with current weather situations* in Table 8 proved too abstract to operationalize and were not implemented. Likewise, the last characteristic *mentions of disaster locations* in Table 6 and first in Table 7 were not used, because tweets are often too short and informal at times and location identification as well as pronouns from social media data is another aspect of extensive research.

In addition to the domain-expert features, the second type of features we use are BoW-based. Specifically, uni-grams and bi-grams

[13] https://en.wiktionary.org/wiki/Category:English_swear_words .
[14] https://www.speakconfidentenglish.com/english-internet-slang/ .
[15] http://www.vocabulary.cl/Lists/Daily_Routines.htm .
[16] http://www.lextek.com/manuals/onix/stopwords1.html .

Table 10
Floods results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.584	0.488	0.532	627
Non-eyewitness	0.706	0.575	0.634	551
Don't know	0.656	0.820	0.729	822
Domain-expert features				
Eyewitness	0.638	0.478	0.547	627
Non-eyewitness	0.642	0.664	0.653	551
Don't know	0.635	0.742	0.685	822
Domain-expert + text features				
Eyewitness	0.717	0.469	0.567	627
Non-eyewitness	0.748	0.653	0.698	551
Don't know	0.648	0.875	**0.745**	822
Domain-expert + text features with class balancing				
Eyewitness	0.760	0.648	**0.699**	815 (+30%)
Non-eyewitness	0.774	0.763	**0.768**	716 (+30%)
Don't know	0.688	0.798	0.739	822

features are extracted from the textual content of tweets and their TF-IDF scores are used to train machine learning models.

5.2. Classification results

Using the methodological steps described in Section 3 and the features described in the previous subsection, we train several machine learning classifiers. We used the labeled data obtained from crowdsourcing. However, the indirect and vulnerable eyewitness classes were small, i.e. having fewer labeled messages. Further, the manual classification showed that even for a human these classes can be difficult to differentiate. For these reasons, we combined all three types of eyewitness classes, i.e. direct, indirect, and vulnerable into one class namely *"Eyewitness"*. Consequently, our classification task consists of three classes: (i) *Eyewitness*, (ii) *Non-eyewitness*, and (iii) *Don't know*.

Tables 10 –13 show the results of our analysis for floods, hurricanes, earthquakes, and wildfires respectively. The last columns of the tables show class distributions. The last three rows present class distribution after applying the class balancing technique (i.e. SMOTE) where the number in parentheses indicates how many artificially labeled instances we added to that class using the SMOTE technique.

The floods results (Table 10) show slightly better performance (e.g. F-scores) when using domain expert features compared to text features (baseline). However, even better results are obtained upon combining both text and domain features. However, the minority classes *eyewitness* and *non-eyewitness* still suffer compared to the *don't know* classes which achieved an F-score of 0.745. To balance the minority classes, we add 30% more instances, which clearly seem to help achieve better results for the both classes.

Table 11
Hurricanes results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.646	0.419	0.508	465
Non-eyewitness	0.773	0.852	0.810	1199
Don't know	0.605	0.679	0.640	336
Domain-expert features				
Eyewitness	0.655	0.546	0.596	465
Non-eyewitness	0.776	0.881	0.825	1199
Don't know	0.645	0.482	0.552	336
Domain-expert + text features				
Eyewitness	0.734	0.503	0.597	465
Non-eyewitness	0.788	0.910	**0.844**	1199
Don't know	0.686	0.604	0.642	336
Domain-expert + text features with class balancing				
Eyewitness	0.816	0.796	**0.806**	930 (+100%)
Non-eyewitness	0.838	0.843	0.841	1199
Don't know	0.801	0.820	**0.810**	672 (+100%)

Table 12
Earthquakes results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.878	0.977	0.925	1600
Non-eyewitness	0.893	0.585	0.707	200
Don't know	0.629	0.280	0.388	200
Domain-expert features				
Eyewitness	0.871	0.969	0.917	1600
Non-eyewitness	0.787	0.645	0.709	200
Don't know	0.333	0.095	0.148	200
Domain-expert + text features				
Eyewitness	0.865	0.987	0.922	1600
Non-eyewitness	0.912	0.620	0.738	200
Don't know	0.641	0.125	0.209	200
Domain-expert + text features with class balancing				
Eyewitness	0.892	0.966	**0.927**	1600
Non-eyewitness	0.932	0.793	**0.857**	400 (+100%)
Don't know	0.801	0.653	**0.719**	400 (+100%)

In the case of hurricanes (Table 11), the domain features seem to give good advantage over the plain text features for the eyewitness class. However, for the other two classes the difference is not significant. Furthermore, better performance was observed upon combining both text and domain features. However, the minority classes seem to suffer again. We added 100% more labeled instances to both minority classes, i.e. eyewitness and don't know, and obtained better performance, which outperforms all three models.

The earthquakes results are shown in Table 12. Surprisingly, in this case the domain features do not seem to help much. In fact, a significant drop is observed in the *don't know* class. On combining domain and text features, the performance seems to increase a bit. These experiments were challenging as there is a big difference in the class distributions. Just to highlight the significance of balanced classes, we added 100% more labeled instances to both minority classes, i.e. eyewitness and don't know, and obtained better results.

Table 13 shows the results of wildfires. We observe a good improvement in the performance when using domain features compared to using text features. However, when domain and text features are combined, the don't know class seems to perform better than the other two classes. To tackle with the class imbalance issue, we increased the labeled instance of the eyewitness class by 100%. Even after the dataset is not balanced, but it starts showing positive indication that more labeled data can achieve better results.

Overall, we observed that in most cases domain features seem to help achieve better performance compared to text-only features. Moreover, a combination of both text and domain features seem to significantly gain the classifiers performance. Specifically, in the case of earthquakes, *tweet-length, magnitude token*, a bi-gram consisting of *earthquake and magnitude, felt* etc. were among the top features. In the case of floods and forestfires, *personal possessive, reporting impact of disasters (e.g., raining, burning), words indicating*

Table 13
Wildfires results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.649	0.265	0.376	189
Non-eyewitness	0.857	0.941	0.897	1379
Don't know	0.748	0.708	0.728	432
Domain-expert features				
Eyewitness	0.703	0.339	0.457	189
Non-eyewitness	0.863	0.943	0.901	1379
Don't know	0.737	0.688	0.711	432
Domain-expert + text features				
Eyewitness	0.794	0.265	0.397	189
Non-eyewitness	0.867	0.946	**0.905**	1379
Don't know	0.730	0.731	0.731	432
Domain-expert + text features with class balancing				
Eyewitness	0.897	0.714	**0.795**	378 (+100%)
Non-eyewitness	0.753	0.727	0.740	432
Don't know	0.876	0.935	**0.905**	1379

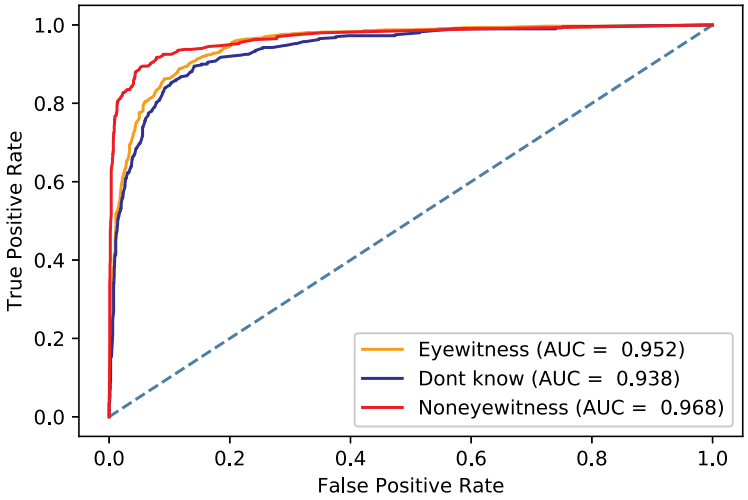

Fig. 1. Earthquake: ROC curves of all three classes of the best model.

intensity (e.g., heavy, intense), mention of locations, flash flood, etc. were among the most useful features. Furthermore, in the case of hurricanes, features including *intensity of disaster, caution and advice for others (e.g., watch out, warning, be careful), time indicating words, perceptual senses*, etc. were identified as useful ones.

Given all of our datasets are hugely imbalanced, we evaluated classifiers performance after adding more labeled data to minority classes. This approach seems to outperform all of our model training variations. To further understand the performance of our classifiers, we draw AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curves of the best models, which in most of the cases are the models that rely on domain-expert, textual features, and class balancing support. Generally, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the model. Fig. 1 shows AUC-ROC curves of the earthquake model. All classes obtained a reasonable AUC values i.e., & > 0.90&. Fig. 2 shows AUC-ROC curves of the flood model where the under performing class i.e., *eyewitness* can be easily noticed. This highlights the need of more labeled data as well as more distinguishable features among all three classes.

Fig. 3 shows the AUC-ROC curves of the forest fire model. In this particular case, the *eyewitness* class clearly outperforms and achieves an $AUC = 0.968$. The other two classes slight suffer, but still obtained a decent AUC. Fig. 4 shows the AUC-ROC curves of the hurricane model. The *eyewitness* class shows slightly lower performance compared to the other two classes. However, it achieves an $AUC = 0.938$, which is reasonable.

6. Discussion

Given the volatility of social media, our objective was to identify and engineer features for an automated classification that would be identifiable in similar social media platforms. Due to the ready availability of Twitter data, our study relies on tweets like most

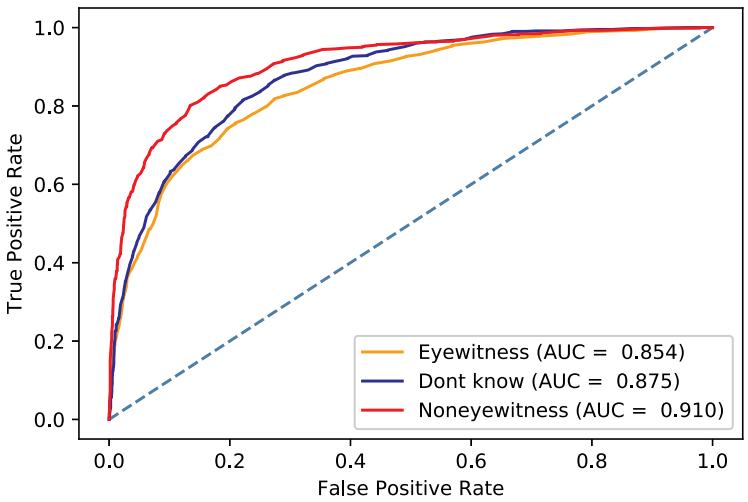

Fig. 2. Flood: ROC curves of all three classes of the best model.

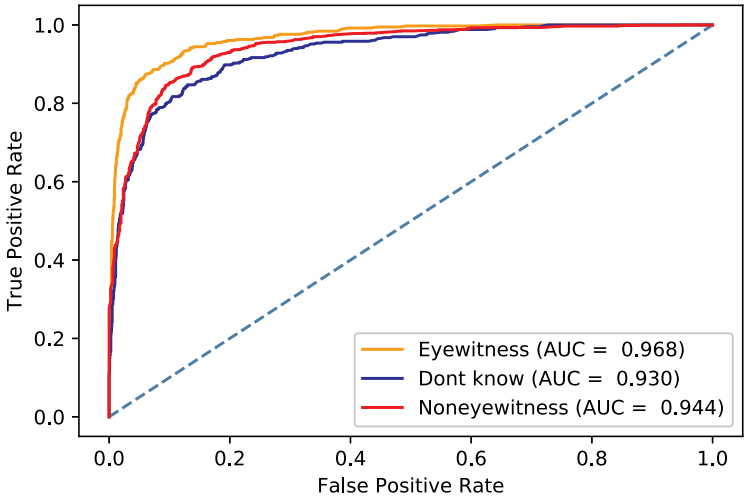

Fig. 3. Forestfire: ROC curves of all three classes of the best model.

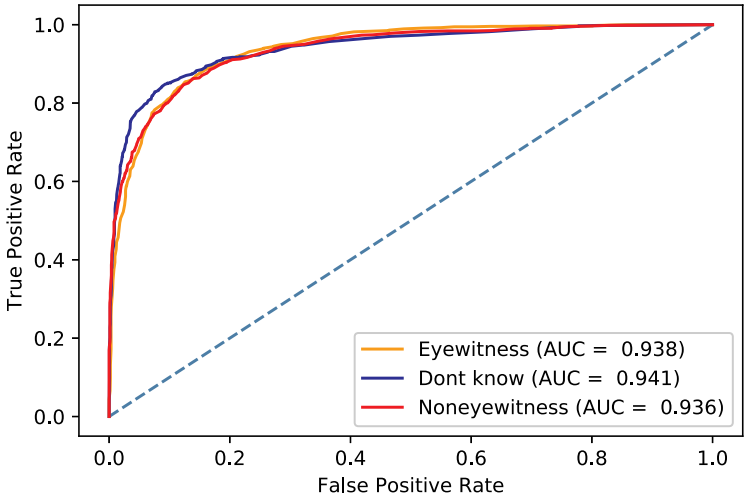

Fig. 4. Hurricane: ROC curves of all three classes of the best model.

other related studies do. However, we focused on extracting features from the text content (the messages), instead of metadata fields or the individual social network characteristics, which might differ structurally from other social media and social networks. Despite our previous observation that Twitter communication differs from other communication channels, we are therefore confident that validation studies using our approach with different social media data sources will be able to replicate our results.

Another important objective was to achieve good classification results for a variety of disaster types. We observed during manual analysis of our data for different disaster types that due to the different nature of disasters, some of the characteristics found in messages were different from each other. For example, because of unpredictable, sudden, and short nature of earthquakes, short tweets consisting of one or two words were only found in earthquake dataset. It was also observed that the frequency of words like *just, now, suddenly* was higher in the earthquake dataset as compared to other disasters which are relatively predictable and long-term in nature. Moreover, words related to *Caution and advice* were found more frequently in the flood, hurricane and wildfire dataset compared to others.

The feasibility to operationalize manual features varies, but most of them are very feasible to learn. However, some features such as *small details of vicinity* and *associating current weather conditions to the disaster* proved to be too abstract to operationalize well. Moreover, while many features can be kept or safely translated automatically after changing the language of the input data, others clearly are language dependent. For example, expletives features differ between languages, and some languages have very different grammar structures, so features depending on personal or possessive pronouns will require adjustment.

One can observe variations in classifiers performance. For instance, most earthquake-related classes scoring the highest F1-measures and floods the lowest, the performance is acceptable for all disaster types, possibly allowing our approach to be used on anthropogenic disasters as well.

A potential source for lower performance is the class imbalance that we had to deal with. We chose to use the SMOTE approach to

increase the availability of minority classes and improve balancing. Realistically, obtaining balanced labeled data from social media during an ongoing event is almost impossible. However, given our promising results obtained from nearly-balanced datasets motivates to put an effort during labeling to obtain balanced classes. During time-critical situations, one simple approach is to restrict adding new labels for a class which has majority while only allowing labels for minority classes. This is particularly suitable for human-in-the-loop systems.

On the positive side, our study provides another, reproducible example of the feasibility of crowdsourced annotation and labeling. Following best practices, we were able to increase our training dataset substantially with a modest investment of funds, while ensuring high quality of results and high inter-rater reliability. This is a further step towards a better integrated human-machine processing approach, with human validation of ”don't know” classifications a potential way to increase recall further.

Another innovative aspect of our research is that we focused on building an eyewitness reports taxonomy exclusive for disaster response agencies during natural disasters. Our taxonomy has three types of eyewitness reports: direct, indirect, and vulnerable eyewitness. Direct eyewitness reports are generated from the people who felt (hear, smell, saw) the disaster or its impacts by themselves. A direct eyewitness report restricts the geographic location of users e.g. reports originating from disaster-hit region. However, in this taxonomy we not only consider reports generated by the people who are present in disaster-hit region but also the reports generating from anywhere outside the disaster location about their family and friends who are present in disaster-hit region known as indirect eyewitness reports. Those people can be a useful resource to give more information about the whereabouts of missing people. Another interesting type of our taxonomy is vulnerable direct eyewitness. In this type we consider reports coming from the people who are anticipating a disaster and who are also present in the region for which disaster warning has been issued. The rationale of adding this type is that such reports can help disaster response agencies to launch precise rescue operations if situation gets worse in that region.

While our manual analysis identified several useful subclasses of eyewitnesses (direct, indirect, vulnerable), early experimentation with training the models showed that performance was low. The two main reasons were the semantic ambiguity of many instances and the even lower number of available instances. The semantic ambiguity made it difficult occasionally even for experienced human annotators to decide based on 140 characters of text whether the source of the message was directly observing or reporting someone else's observations, and whether danger was imminent (vulnerable). Coupled with the low number of example instances, we decided to combine all subclasses for the automated analysis.

One important limitation of our study is language. For practical reasons, we have limited our analysis to English language tweets. Depending on the language structure, some of our features, e.g. first person pronouns, might not work.

7. Conclusions

Finding firsthand and credible information during disasters and emergencies is an important task for relief organizations and law enforcement agencies. The extensive use of social media platforms during disasters provides numerous opportunities for humanitarian organizations to enhance their response. Among them, identification of bystanders and eyewitnesses can help to get important information. In this work, we presented an analysis of tweets collected related to four types of disasters to understand different types of eyewitness reports. Our manual analysis results show that we can categorize eyewitness reports into direct, indirect, and vulnerable direct eyewitnesses. Moreover, an important contribution of this work is to determine various characteristics associated with each type of eyewitness report. We observed that direct eyewitnesses use words related to perceptual senses such as seeing, hearing, feeling. Whereas, indirect eyewitness mainly express emotions such as thoughts, prayers, worry. And, the vulnerable category mostly share warnings and alerts about an expected disaster situation. We use these characteristics and manually labeled data obtained to perform extensive experimentation. Our results revealed that domain-expert features when combined with textual features outperform models which are only trained on text-based features. Moreover, we apply a class balancing technique to tackle the class-imbalance problem, which most of our datasets suffer with. The results obtained after applying class balancing reveal even better results.

Acknowledgement

This research has funded by Swiss government excellence scholarship (ESKAS), Einrichtungskredit, and Forschungskredit grant number K-75130-02 at the University of Zurich.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). *Twitcident: Fighting fire with information from social web streams. Proceedings of the 21st international conference on world wide web*. ACM305–308.

Allen, C. (2014). A resource for those preparing for and responding to natural disasters, humanitarian crises, and major healthcare emergencies. *Journal of Evidence-Based Medicine, 7*(4), 234–237.

Amaratunga, C. (2014). Building community disaster resilience through a virtual community of practice (VCOP). *International Journal of Disaster Resilience in the Built Environment, 5*(1), 66–78.

Callison-Burch, C. (2009). *Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1*. Association for Computational Linguistics286–295.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). *Finding and assessing social media information sources in the context of journalism. Proceedings of the SIGCHI*

conference on human factors in computing systems. ACM2451–2460.
Doggett, E. V., & Cantarero, A. (2016). *Identifying eyewitness news-worthy events on twitter. Conference on empirical methods in natural language processing7*.
Fang, R., Nourbakhsh, A., Liu, X., Shah, S., & Li, Q. (2016). *Witness identification in twitter. Procceedings of the fourth international workshop on natural language processing for social media, Austin, TX, USA*65–73.
Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*, 110–120.
Haworth, B., & Bruce, E. (2015). A review of volunteered geographic information for disaster management. *Geography Compass, 9*(5), 237–250.
Hong, L., Dan, O., & Davison, B. D. (2011). *Predicting popular messages in twitter. Proceedings of the 20th international conference companion on world wide web*. ACM57–58.
Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR), 47*(4), 67.
Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). *AIDR: Artificial intelligence for disaster response. Proceedings of the 23rd international conference on world wide web*. ACM159–162.
Imran, M., Lykourentzou, I., Naudet, Y., & Castillo, C. (2013). Engineering crowdsourced stream processing systems. arXiv preprint arXiv:1310.5463.
Imran, M., Mitra, P., & Srivastava, J. (2016). *Cross-language domain adaptation for classifying crisis-related short messages. Proceedings of the 13th international conference on information systems for crisis response and management (ISCRAM)*.
Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances, 2*(3), e1500779.
Kumar, S., Barbier, G., Abbasi, M. A., & Liu, H. (2011). *Tweettracker: An analysis tool for humanitarian and disaster relief. ICWSM*.
Kumar, S., Morstatter, F., Zafarani, R., & Liu, H. (2013). *Whom should I follow?: Identifying relevant users during crises. Proceedings of the 24th ACM conference on hypertext and social media*. ACM139–147.
Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is twitter, a social network or a news media? Proceedings of the 19th international conference on world wide web*. ACM591–600.
Landwehr, P. M., & Carley, K. M. (2014). *Social media in disaster relief. Data mining and knowledge discovery for big data*. Springer225–257.
Lee, K., Ganti, R. K., Srivatsa, M., & Liu, L. (2014). *When twitter meets foursquare: Tweet location prediction using foursquare. Proceedings of the 11th international conference on mobile and ubiquitous systems: Computing, networking and services*. ICST (Institute for Computer Sciences, Social-Informatics and âQ198–207.
Meier, P. (2012). Crisis mapping in action: How open source software and global volunteer networks are changing the world, one map at a time. *Journal of Map & Geography Libraries, 8*(2), 89–100.
Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J., & Liu, H. (2014). Finding eyewitness tweets during crises. arXiv preprint arXiv:1403.1773.
Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly, 37*(2).
Olteanu, A., Vieweg, S., & Castillo, C. (2015). *What to expect when the unexpected happens: Social media communications across crises. Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM994–1009.
Ostermann, F., Garcia-Chapeton, G., Kraak, M., & Zurita-Milla, R. (2018). *Towards a crowdsourced supervision of the analysis of user-generated geographic content: Engaging citizens in discovering urban places*.
Ostermann, F., & Spinsanti, L. (2012). Context analysis of volunteered geographic information from social media networks to support disaster management: A case study on forest fires. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM), 4*(4), 16–37.
Purohit, H., Castillo, C., Diaz, F., Sheth, A., & Meier, P. (2014). Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday, 19*(1).
Schnebele, E., et al. (2013). Improving remote sensing flood assessment using volunteered geographical data. *Natural Hazards and Earth System Sciences, 13*(3), 669.
Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). *Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics254–263.
Takahashi, B., Tandoc, E. C., & Carmichael, C. (2015). Communicating on twitter during a disaster: An analysis of tweets during typhoon Haiyan in the Philippines. *Computers in Human Behavior, 50*, 392–398.
Tanev, H., Zavarella, V., & Steinberger, J. (2017). *Monitoring disaster impact: Detecting micro-events and eyewitness reports in mainstream and social media*.
Teevan, J., Ramage, D., & Morris, M. R. (2011). *# twittersearch: A comparison of microblog search and web search. Proceedings of the fourth ACM international conference on web search and data mining*. ACM35–44.
Thom, D., Krüger, R., Ertl, T., Bechstedt, U., Platz, A., Zisgen, J., & Volland, B. (2015). *Can twitter really save your life? A case study of visual social media analytics for situation awareness. Visualization symposium (PACIFICVIS), 2015 IEEE pacific*. IEEE183–190.
Truelove, M., Vasardani, M., & Winter, S. (2014). *Testing a model of witness accounts in social media. Proceedings of the 8th workshop on geographic information retrieval*. ACM10.
Truelove, M., Vasardani, M., & Winter, S. (2015). Towards credibility of micro-blogs: Characterising witness accounts. *GeoJournal, 80*(3), 339–359.
Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). *Microblogging during two natural hazards events: What twitter may contribute to situational awareness. Proceedings of the SIGCHI conference on human factors in computing systems*. ACM1079–1088.
Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *JCP, 7*(12), 2913–2920.
Zahra, K., Imran, M., & Ostermann, F. O. (2018). *Understanding eyewitness reports on twitter during disasters. Proceedings of the 15th international conference on information systems for crisis response and management, Rochester, NY, USA, May 20–23*.
Zahra, K., Ostermann, F. O., & Purves, R. S. (2017). Geographic variability of twitter usage characteristics during disaster events. *Geo-Spatial Information Science, 20*(3), 231–240.