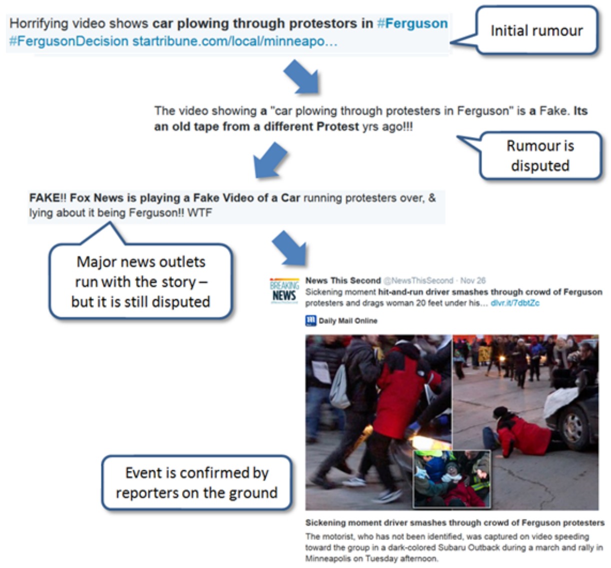
Crowdsourced Rumour Identification During Emergencies

Richard McCreadie, Craig Macdonald and Iadh Ounis
{firstname.lastname}@glasgow.ac.uk
University of Glasgow, UK

ABSTRACT

When a significant event occurs, many social media users leverage platforms such as Twitter to track that event. Moreover, emergency response agencies are increasingly looking to social media as a source of real-time information about such events. However, false information and rumours are often spread during such events, which can influence public opinion and limit the usefulness of social media for emergency management. In this paper, we present an initial study into rumour identification during emergencies using crowdsourcing. In particular, through an analysis of three tweet datasets relating to emergency events from 2014, we propose a taxonomy of tweets relating to rumours. We then perform a crowdsourced labeling experiment to determine whether crowd assessors can identify rumour-related tweets and where such labeling can fail. Our results show that overall, agreement over the tweet labels produced were high (0.7634 Fleiss κ), indicating that crowd-based rumour labeling is possible. However, not all tweets are of equal difficulty to assess. Indeed, we show that tweets containing disputed/controversial information tend to be some of the most difficult to identify.



Horrifying video shows **car plowing through protestors in** #Ferguson #FergusonDecision startribune.com/local/minneapo…

Initial rumour

The video showing **a** "car plowing through protesters in Ferguson" is **a** Fake. **Its an old tape from a different Protest** yrs ago!!!

Rumour is disputed

FAKE!! Fox News is playing a Fake Video of a Car running protesters over, & lying about it being Ferguson!! WTF

Major news outlets run with the story – but it is still disputed

News This Second @NewsThisSecond · Nov 26
Sickening moment **hit-and-run driver smashes through crowd of Ferguson** protesters and drags woman 20 feet under his… dlvr.it/7dbtZc

Daily Mail Online

Sickening moment driver smashes through crowd of Ferguson protesters
The motorist, who has not been identified, was captured on video speeding toward the group in a dark-colored Subaru Outback during a march and rally in Minneapolis on Tuesday afternoon.

Event is confirmed by reporters on the ground

Figure 1: Example rumour discussed from Twitter.

1. INTRODUCTION

Social networks (such as Twitter) are ideal for the detection and monitoring of important events. For instance, when an earthquake occurs, reports about it on social media can be observed very quickly [16]. As a result, social networks are increasingly being used during emergencies and disasters by first responders and civil protection agencies as an information source to help them better respond to those types of event [11].

However, not all posted tweets are factual, neutral or reasoned – there is often a bias, or an element of rumour in some tweets [4, 6, 13]. Meanwhile, such rumours can be quickly spread across social media. For example, during the London Riots in August 2011, Twitter users spread unsubstantiated rumours about rioters breaking into a children's hospital [12]. During emergencies, it is critical that such rumours are quickly identified and investigated since, if true, law enforcement officers need to be dispatched, while, if false, authorities need to quickly correct such misinformation to keep the public well informed. Hence, the development and evaluation of automatic or semi-automatic systems for the identification of rumours within social media is becoming increasingly important.

However, while there has been significant prior works examining rumours within social media for areas such as meme tracking [14], detecting urban legends [4] and identifying disputed and/or credible information [8, 13], there as been little investigation into rumour identification during emergencies. In this paper, we present our initial investigation into automatic identification of rumour-related posts within the Twitter social network. In particular, the main contributions of this paper are as follows. First, through a manual analysis of posts for three recent emergency events, we propose an initial taxonomy for rumour posts on Twitter. Second, we present and analyze a crowdsourced labeling experiment over Twitter data, with the aim of determining whether crowdsourcing is a viable method to identify rumours, and also to investigate the types of rumour-related posts that are difficult to identify.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742573.

The remainder of this paper is structured as follows. Section 2 discusses prior works examining rumours in social media. In Section 3, we propose a taxonomy for classifying rumour posts on Twitter. Section 4 describes the design of our crowdsourced rumour labeling experiment. In Section 5 we detail our tweet dataset and our crowdsourcing configuration, while in Section 6 we report the agreement between our crowdsourced assessors and discuss the types of tweets that are difficult to label. We summarize our conclusions in Section 7.

2. RELATED WORK

A variety of works have analyzed ‘rumours’ in different contexts previously. Within the psychology literature, understanding rumours have been extensively examined within offline environments [1, 2]. On the other hand, within the social media literature, rumours and their spread is a new topic. Early works looked at how to detect contradictions in text. For example, de Marneffe et al. [5] proposed an approach that used co-reference techniques to detect contradictions. Meanwhile, Ritter et al. [15] improved contradiction detection by using meronyms and synonyms as additional evidence. Meme tracking has also been investigated as a means to follow rumours. Leskovic et al. [9] investigated memes within social media, showing how quotations change over time and can be used to track rumours. Ratkiewicz et al. [14] developed a system named ‘Truthy’ that aimed to identify misleading political memes on Twitter. More recent literature has focused on the automatic identification of rumours. Ennals et al. [8] examined a phrase dictionary-based approach to find content that confirm, refute, question or discuss rumours. On the other hand, Qazvinian et al. [13] and Castello et al. [4] adopted machine learned approaches to the automatic identification of rumours, which use features extracted from social media posts/authors of those posts to determine whether they discuss a rumour. In contrast to these prior works, in this paper, we focus on the semi-automatic labeling of tweets using the medium of crowdsourcing. In this way, crowdsourced labels can be used either as a system in and of itself to identify rumours during an emergency, or as a method to generate training data for automatic approaches such as those described in [4] or [13].

3. DEFINING A RUMOUR TAXONOMY

Importantly, what is considered a ‘rumour’ can be quite broad. For instance, from an early study of World War 2 rumours, Knapp defined a rumour as “a proposition for belief of topical reference disseminated without official verification” [2], meanwhile in a more recent study, DiFonzo and Bordia [7] defined a rumour as an “unverified and instrumentally relevant statement of information spread among people”. Prior approaches and systems that aim to identify rumours as they appear in social media have focused broad classes of rumours, such as urban legends or celebrity news [4, 13, 14]. In these cases, posts such as:

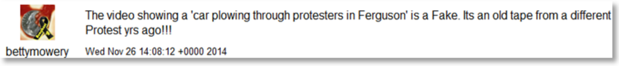
> ”RT @jonnyA99 Ann Coulter Tells Larry King Why People Think Obama is a Muslim http://bit.ly/9rs6pa #Hussein”

might be considered to be a rumour [4]. However, as we are interested in examining rumours specifically during emergencies, this would not be a rumour of interest to response services, civil protection agencies or the general public. Hence, to tackle rumours effectively, we need to define what a rumour means within the context of a crisis/emergency event.

To facilitate this, we initially analysed a sample of posts collected from the Twitter microblogging platform for three recent emergency-related events, namely: the Hong Kong Protests (October 2014); the Extreme Snowfall in Buffalo, USA (November 2014); and the Ferguson Riots (November 2014). The aim of this analysis was two-fold. First, we wanted to determine what form of relationship exists between social media posts and rumours for emergency/security events. Second, we aimed to generate a taxonomy of social media posts, which might aid when identifying rumour-related posts within social networks.

From the perspective of the relationship between social media posts and rumours, we observed that a rumour is rarely captured by a single social media post. Instead, rumours are characterized by a potentially large number of distinct tweets, all discussing a single topic. Figure 1 illustrates an example rumour as discussed within the Twitter social network for the Ferguson Riots event.

As we can see from Figure 1, a post initiates the rumour by linking to a video of a car driving through a group of protestors. This rumour is then disputed by other social media users, even when major news outlets run with the story. Finally, the rumour is confirmed to be true by a reporter on the ground. From this, we can draw two important conclusions. First, to understand and evaluate a rumour, it is often necessary to see multiple posts about that same rumour. Second, it still may be possible however to identify the existence of a rumour from a single post. For instance, given the following tweet alone:



bettymowery

The video showing a 'car plowing through protesters in Ferguson' is a Fake. Its an old tape from a different Protest yrs ago!!!

Wed Nov 26 14:08:12 +0000 2014

It is reasonable to expect that a human could identify that this relates to a rumour about a car running over a group of protesters, although much of the important information about the rumour is missing, such as when it happened or who was injured. Hence, there is an important distinction between identifying a rumour via a post discussing that rumour, and characterising the rumour and its content as a whole.

Next, given that we are interested in identifying individual posts relating to rumours, it would be valuable to create a taxonomy for social media posts to aid us during this task. Hence, from our initial analysis, we identified six distinct classes of social media post that were relevant to one or more rumours over the three events we analysed.[1] We list these classes in Table 1 below, with an example of each.

Importantly, in this paper, we aim to identify rumours over time. Social media posts that belong to the Unsubstantiated Information/Speculation class are likely to relate to a rumour, as most posts that initiate a new rumour or spread an existing rumour will belong to this category. Meanwhile, posts belonging to the Disputed/Controversial Information and Linked Dispute classes are also likely to be useful, as social media users often dispute rumours. Indeed, it is worth

[1] Of course, this taxonomy is not exhaustive, but rather an initial classification based on the types of tweets observed. We envisage that future research will identify additional classes, that encapsulate the different ways that social media users share/interact with rumours.

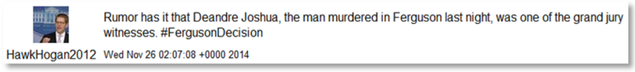
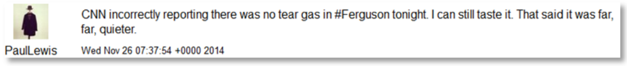
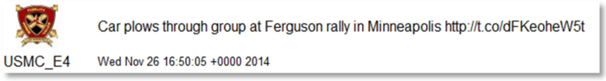
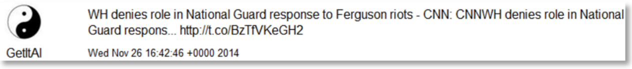
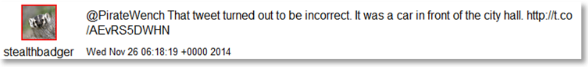
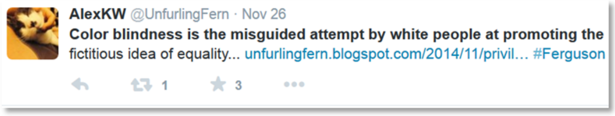
Class	Description	Example
Unsubstantiated Information/Speculation	A social media post that discusses information that is uncertain or is unsubstantiated.	 HawkHogan2012 Rumor has it that Deandre Joshua, the man murdered in Ferguson last night, was one of the grand jury witnesses. #FergusonDecision Wed Nov 26 02:07:08 +0000 2014
Disputed/Controversial Information	A post that disputes information provided in another post, article, image or video.	 PaulLewis CNN incorrectly reporting there was no tear gas in #Ferguson tonight. I can still taste it. That said it was far, far, quieter. Wed Nov 26 07:37:54 +0000 2014
Misinformation/Disinformation	A post that contains false information, misrepresents information or quotes out of context.	 USMC_E4 Car plows through group at Ferguson rally in Minneapolis http://t.co/dFKeoheW5t Wed Nov 26 16:50:05 +0000 2014
Reporting	A post that reports the occurrence of an event and supplies a secondary source, e.g. a hyperlinked news article.	 GetItAl WH denies role in National Guard response to Ferguson riots - CNN: CNNWH denies role in National Guard respons... http://t.co/BzTfVKeGH2 Wed Nov 26 16:42:46 +0000 2014
Linked Dispute	A post that attempts to deny a rumour, possibly in the form or a direct reply to a user. Like reports, corrections often supply a secondary supporting source of evidence.	 stealthbadger @PirateWench That tweet turned out to be incorrect. It was a car in front of the city hall. http://t.co/AEvRS5DWHN Wed Nov 26 06:18:19 +0000 2014
Opinionated	A post that expresses the author's opinion	 AlexKW @UnfurlingFern · Nov 26 **Color blindness is the misguided attempt by white people at promoting the** fictitious idea of equality... unfurlingfern.blogspot.com/2014/11/privil... #Ferguson 1 3

Table 1: Social media post classes identified.

noting that it is often possible to find the initial post that started a rumour given a post disputing that rumour and vice versa. Finally, note that we include the class of posts including opinions here. The idea for including opinionated posts as a possible rumour-related class is that some of these posts may discuss or be the source of rumours. Indeed, from our analysis, when relevant to a rumour, this class of tweet is similar to but distinct from the Unsubstantiated Information/Speculation class. In the next section, we describe the design of our crowdsourced labeling experiment, where we have crowd assessors categorize emergency-related tweets into these 6 categories (among others).

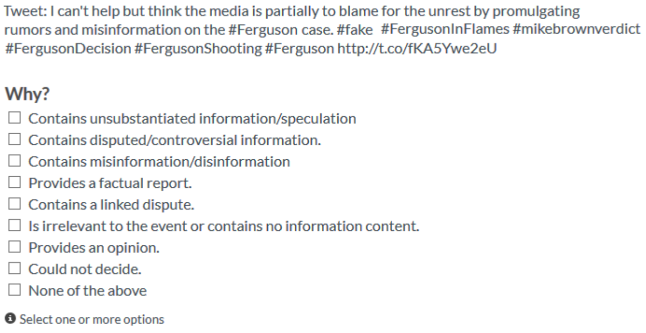
4. CROWDSOURCING RUMOUR LABELS

Having defined a taxonomy of rumour posts, we next need to determine to what extent human assessors are able to identify such posts from a generic set of posts about an event. Indeed, if this task is very difficult for humans to achieve, then we can expect similarly poor performance from automatic approaches. To evaluate this, we perform a crowdsourced labeling study. In particular, for a sample of 1000 tweets collected about the Ferguson Riots event, we have those tweets labeled by multiple crowdsourced workers. We then compare the accuracy of the crowdsourced labels produced in terms of cross-worker agreement, with the aim of identifying common sources of labeling error. Below we describe the design of our crowdsourcing experiment.

4.1 Crowdsourced Labeling Design

There are a variety of design decisions that need to be made when producing a crowdsourced experiment. First, in terms of the labeling task that we are investigating here, we need to decide on how much information we provide to the workers. Recall from Section 3 that during our analysis we noted that to fully understand a rumour, an assessor would need to see multiple posts about that rumour, but that for at least a subset of the rumour posts, it was possible to identify that the post discussed a rumour from its text in isolation. Hence, we need to decide whether to display only a single post, or provide additional supporting evidence. Importantly, recalling our motivation for examining rumour identification (see Section 1), timeliness in rumour identification is critical, since accurately identifying a rumour may lead to the deployment of resources (people, aid, etc.) during an emergency. From a experimental design perspective, this requirement naturally precludes the inclusion of additional posts about a rumour, because we want to identify a rumour starting from the very first post about it (when no other information is available). Hence, we chose to show the assessors only single tweets in isolation.

Second, we need to determine the potential labels that the assessors may select between, in addition to those defined by our proposed taxonomy (see Section 3). In particular, there are a variety of post types that are not rumour related. For instance, the tweet might be off-topic, or the assessor might simply not be able to accurately determine what class(es) it belongs to. Figure 2 shows an example our assessment interface, including the set of 9 labels/classes that the assessor can choose. For each tweet to be labeled, the assessor marks that tweet as belonging to one or more of the 9 classes. Multiple assessors label each tweet, enabling us to identify the most likely classes and also identify tweets and classes that are more difficult to label.



Tweet: I can't help but think the media is partially to blame for the unrest by promulgating rumors and misinformation on the #Ferguson case. #fake #FergusonInFlames #mikebrownverdict #FergusonDecision #FergusonShooting #Ferguson http://t.co/fKA5Ywe2eU

Why?

☐ Contains unsubstantiated information/speculation
☐ Contains disputed/controversial information.
☐ Contains misinformation/disinformation
☐ Provides a factual report.
☐ Contains a linked dispute.
☐ Is irrelevant to the event or contains no information content.
☐ Provides an opinion.
☐ Could not decide.
☐ None of the above

Select one or more options

Figure 2: Example tweet rumour label interface.

5. EXPERIMENTAL SETUP

5.1 Tweet Dataset

Dataset Crawling: We initially collected a set of tweets for the Ferguson Riots event[2] using the public Twitter API between the 26th and the 27th of November 2014. In particular, we selected the Ferguson Riots, as it saw significant discussions in social media and was subject to dispute by different groups of the populous. In particular, we identified a series of search terms that covered the main named entities relating to that event, e.g. Ferguson and #MikeBrown. For each of these search terms, we used the Twitter Search API to crawl all tweets containing that term for the stated period. This resulted in a set of 2,078,119 tweets about the Ferguson Riots event.

Post Sampling: However, it is not practical to use crowdsourcing to label millions of tweets, hence, we instead reduce the number of tweets that need to be assessed via a sampling strategy [17, 18]. Indeed, the vast majority of these tweets will not be rumour-related, but rather people generally commenting on or discussing the event. However, using a classical random sampling approach [17] would contain a very low proportion of rumorous tweets. Instead, we apply a more targeted sampling approach based on a rumour term dictionary. In particular, we manually generated a small dictionary of 24 terms that are likely indicative of the presence of a rumour, e.g. 'misleading' or 'hoax'. We then used this dictionary to filter the posts within this dataset, keeping only those tweets containing one or more of these terms. This type of sampling is similar to the sampling performed for a Cranfield-style information retrieval evaluation, where only a subset of the documents are selected for labeling based on the pooling of results from different systems [18] (in this case the words in our rumour dictionary act as the systems for the purposes of pooling). This sampling reduced the number of tweets to be labeled to 1795. For cost reasons, we assess a random 1000 tweets from this sample.

5.2 Crowdsourcing Configuration

For our crowdsourced user study, we use CrowdFlower, which is an on-demand labour website providing job creation, monitoring and analytical services on top of different crowdsourcing marketplaces. The unit of assessment is a single page, which contains 10 tweet assessments. Notably, the quality of crowdsourced work can be poor if quality-assurance (QA) techniques are not employed [10]. To mitigate this, we employ the following QA techniques. First, following best practices [3], we have 5 individual crowd workers (assessors) label each tweet. The final assessment produced is the majority vote across the 5 assessors. Second, work submitted was subject to a speed trap of 30 seconds per page, the aim being to detect automatic bots and/or users that are simply randomly selecting labels. Users that submitted a page in under 30 seconds were flagged and removed from the evaluation. Third, to avoid over-reliance on individual assessors, the maximum number of assessments that any one user can contribute was set to 200 (approximately 4% of the entire assessment workload). Additionally, since our events are largely US-centric, we restricted the geographical regions that could participate in the labeling task to only those from the United States, Canada and the United Kingdom. We paid US $0.07 for each set of 10 tweets assessed. The total number of tweet assessments is 5000, i.e. 1000 tweets * 5 unique assessors.

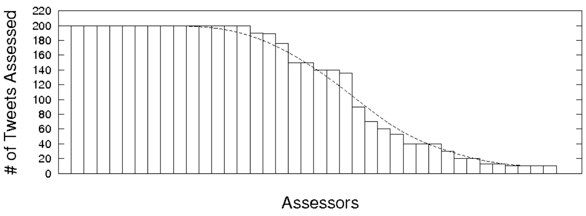


Figure 3: Crowd assessor work distribution.

5.3 Crowdsourcing Statistics

40 unique workers participated in the crowdsourced labeling task. Figure 3 shows the distribution of tweet assessments completed by each of the 40 workers. As we can see from Figure 3, about 40% of the crowd workers completed the maximum number of judgments (200 tweets/20 pages of work), followed by a 'tail' of workers who completed fewer assessments, which is typical behaviour in crowdsourced tasks [3]. The average time it took the workers to assess a tweet was 9 seconds, or 1 minute 32 seconds per page. 2.4% of assessments were dropped due to users failing the speed trap. When considering the labeling to be a set of 9 binary classification tasks, average Fleiss κ agreement across the 5 workers that assessed each tweet was 0.7634, which indicates that the labels were of high quality. We discuss inter-worker labeling agreement in more detail in the next section.

6. LABELING RESULTS

In this section, we aim to answer two main research questions, each of which we discuss in a separate section:

- Are crowd workers able to identify rumour-related tweets from their text with high agreement? (Section 6.1)
- What types of tweet are the most difficult to assess? (Section 6.2)

6.1 Rumour Labeling Agreement

We begin by examining how agreement differs between the different rumour classes within our proposed taxonomy, to determine how effectively crowd workers are able to identify rumour-related posts. A low agreement (<0.3) would

[2] see http://en.wikipedia.org/wiki/2014_Ferguson_unrest

Class	Indicative of a Rumour?	# Classified By Majority	Fleiss κ	Cohen κ
Contains unsubstantiated information/speculation.	✔	48 (5.5%)	0.6607	0.7307
Contains disputed/controversial information.	✔	66 (7.5%)	0.5626	0.6582
Contains misinformation/disinformation.	✔	0 (0%)	N/A	N/A
Provides a factual report.	✗	79 (9%)	0.8295	0.8568
Contains a linked dispute.	✔	2 (0.2%)	0.8714	0.8908
Is irrelevant to the event or contains no information content.	✗	5 (0.5%)	0.9316	0.9422
Provides an opinion.	✗	315 (35.8%)	0.6082	0.7018
Could not decide.	✗	9 (1%)	0.8794	0.8977

Table 2: Agreement statistics for classes in our taxonomy

indicate that a rumour-class is difficult to distinguish or can be mistaken easily for a different class. Meanwhile a high agreement (>0.65) would indicate that crowd workers are able to identify that type of rumour post.

Table 2 reports the number of tweets labeled as belonging to each class based on the majority vote across the 5 assessors, in addition to the agreement of those assessors in terms of both Fleiss and Cohen κ. From Table 2, we observe the following. First, looking at number of tweets that were classified as belonging to each class by the majority of workers, we see that the most common class of tweets are those expressing opinions (35.8% of the tweets). If we compare this to the other rumour-classes from our rumour post taxonomy, we see that there are many fewer rumour-related tweets. In particular, the most common type of rumour-related tweet were those containing disputed/controversial information (7.5%), followed by tweets containing unsubstantiated information/speculation (5.5%). This indicates that rumour-related posts (at least for this dataset) are quite rare. Also of note is that no tweets were labeled by the majority of assessors as belonging to the misinformation/disinformation class. This indicates that it is very difficult to identify misinformation from the text of a single tweet. Indeed, more advanced analytics such as network analysis may be needed to identify this type of information.

For the purposes of our subsequent analysis, since the prior positive (the tweet belongs to a class) and negative (the tweet does not belong to a class) class probabilities are skewed towards the negative class, we focus on Fleiss κ rather than Cohen κ, since Fleiss κ accounts for this. Furthermore, we ignore the 'could not decide' and 'linked dispute' classes, since we have too few positive classification instances for these classes to draw meaningful conclusions.

Next, examining the agreement between our assessors in Table 2, we see that in general, agreement is high, ranging between 0.9316 and 0.5626 Fleiss κ. This indicates that our crowd workers are able to accurately label tweets with respect to our rumour post taxonomy. This is an important result, since it indicates that during emergencies, crowdsourcing could be used to quickly identify rumours for crisis response agencies. Moreover, it raises the possibility that (at least some) crisis-related rumours could be identified automatically from tweet texts alone.

Examining the agreement between assessors on a per-class basis from Table 2, we see that agreement can vary by a large margin across the classes. Factual information appears to be relatively easier to identify, with an inter-worker agreement of 0.8295 Fleiss κ. In contrast, unsubstantiated information/speculation is more difficult to label (0.6607 Fleiss κ), while disputed/controversial information is even more difficult (0.5626 Fleiss κ).

To answer our first research question, crowdsourced assessors are able to identify rumour-related tweets from their text with high agreement, although the actual level of agreement can vary markedly between classes. In the next section, we will examine in more detail where disagreement between classes occurs.

6.2 Labeling Failure Analysis

We next examine in more detail where assessment errors are more common, with the aim of identifying why these errors occur and hence, what might be done to mitigate them. In particular, we begin by analyzing the most common cases where the crowd workers disagree. Table 3 reports the top 10 most common miss-classifications by the crowd workers. Here, a miss-classification is defined as a case where the majority of workers selected one class (column 1 in Table 3), but one or more workers selected another class[3] (column 2 in Table 3). The third column provides a count of the number of tweets, for which one or more workers selected the non-majority class.

From Table 3, we see that the most common errors relate to the classification of tweets as opinionated and/or containing disputed/controversial information. This is intuitive, since a tweet expressing an opinion may also be controversial. For example, consider the tweet below:

> "So the Liberal media released the Ferguson cops home address and now there are rumors that they know who the Grand Jury is- Liberals ROCK."

The majority label for this tweet was 'Provides an opinion', but two of the five assessors also labeled it as 'Contains disputed/controversial information', which is intuitive, since the assertion that personal information about an individual involved in an ongoing court case is controversial, and in this case, a rumour. Indeed, from the perspective of rumour identification, these types of disagreement are an area where further investigation is needed. The next two most common class disagreements relate to the two ambiguous tweet classes, i.e. the 'None of the above' and 'Could not decide' classes. However, also of note is that the fourth most common error was between the 'Provides an opinion' and 'Contains a linked dispute' classes. This is an interesting observation, since only 2 tweets (see Table 2) were labeled by the majority as containing a linked dispute - but a further

[3] Note that the second class needs to itself not be a majority class for the tweet - a tweet may be labeled as belonging to two or more classes, although this is rare.

Majority Class	Most Common Other Class	# Tweets
Provides an opinion.	Contains disputed/controversial information.	163
Provides an opinion.	None of the above.	161
Provides an opinion.	Could not decide.	95
Provides an opinion.	Contains a linked dispute.	63
Contains disputed/controversial information.	None of the above.	54
Contains unsubstantiated information/speculation.	Provides a factual report.	50
Contains unsubstantiated information/speculation.	Contains disputed/controversial information.	50
Provides an opinion.	Provides a factual report.	43
None of the above.	Contains disputed/controversial information.	41
Contains disputed/controversial information.	Provides an opinion.	40

Table 3: The most common miss-classifications by the crowd workers

63 tweets were labeled by one or more assessors as containing a linked dispute, such as the tweet shown below:

> "Gotta be fake RT @sawngbyrd28: YO!! RT @TheAnonMessage BREAKING: Two worlds collide; #ISIS Sends Message To #Ferguson http://t.co/wI1jINbfey"

The above tweet is an example of a linked dispute tweet, in this case calling into question a previous tweet and associated news article. Since the number of tweets for which assessors disagreed about the linked dispute class greatly exceeds the number of tweets labeled by the majority as belonging to that class, indicates that assessors were conservative when assigning tweets the linked dispute label.

To answer our second research question, the most common types of tweet that users disagree about are those that contain and opinion but also might also contain disputed/controversial information or a linked dispute.

7. CONCLUSIONS

In this paper, we presented a study into rumour identification for emergency events using the medium of crowdsourcing. In particular, we proposed a taxonomy of tweets relating to rumours based upon an analysis of three tweet datasets relating to emergency events from 2014. Then, through a crowdsourced labeling experiment, we examined whether crowd assessors could identify rumour-related tweets based upon our taxonomy, showing that overall agreement on the tweet labels produced were high. Hence, we conclude that crowd-based rumour labeling has potential as a method to automatically identify rumours in real-time from social media during an emergency. On the other hand, our failure analysis of the tweets labeled indicated not all classes of rumour-related tweet are easy to identify. For instance, tweets containing controversial information were subject to higher levels of disagreement by our assessors. Meanwhile, no tweets containing misinformation were identified.

Acknowledgments

This work has been carried out in the scope of the EC co-funded SUPER (FP7-606853) project.

8. REFERENCES

[1] F. H. Allport and M. Lepkin. Wartime rumors of waste and special privilege: why some people believe them. *Abnormal and Social Psychology*, 40(1):3, 1945.

[2] G. W. Allport and L. Postman. The psychology of rumor. 1947.

[3] O. Alonso and R. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in information retrieval*, pages 153–164. Springer, 2011.

[4] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on Twitter. In *Proc. of WWW*, 2011.

[5] M.-C. De Marneffe, A. N. Rafferty, and C. D. Manning. Finding contradictions in text. In *Proc of ACL:HLT*, 2008.

[6] L. Derczynski and K. Bontcheva. Pheme: Veracity in digital social networks. In *Proc of ISA*, 2014.

[7] N. DiFonzo and P. Bordia. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35, 2007.

[8] R. Ennals, D. Byler, J. M. Agosta, and B. Rosario. What is disputed on the web? In *Proc. of IC*, 2010.

[9] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. of SIGKDD*, 2009.

[10] R. McCreadie, C. Macdonald, and I. Ounis. Identifying top news using crowdsourcing. *Information Retrieval*, 16(2):179–209, 2013.

[11] S. Muralidharan, L. Rasmussen, D. Patterson, and J.-H. Shin. Hope for Haiti: An analysis of Facebook and Twitter usage during the earthquake relief efforts. *Public Relations Review*, 37(2):175–177, 2011.

[12] T. Newburn. Reading the riots. *British Society of Criminology Newsletter*, page 12, 2011.

[13] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proc. of EMNLP*, 2011.

[14] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. *arXiv*, 1011.3768, 2010.

[15] A. Ritter, D. Downey, S. Soderland, and O. Etzioni. It's a contradiction—no, it's not: a case study using functional relations. In *Proc. of EMNLP*, 2008.

[16] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proc. of WWW*, 2010.

[17] J. S. Vitter. Random sampling with a reservoir. *ACM Trans. on Mathematical Software*, 11(1):37–57, 1985.

[18] E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proc. of SIGIR*, 2000.