Original citation:
Zubiaga, Arkaitz, Liakata, Maria, Procter, Robert N., Bontcheva, Kalina and Tolmie, Peter (2015) Crowdsourcing the annotation of rumourous conversations in social media. In: World Wide Web Conference, Florence, Italy, 18-22 May 2015. Published in: WWW 2015 Companion.ACM. (accepted)
Permanent WRAP url:
http://wrap.warwick.ac.uk/66848

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work of researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.
Publisher statement:
'This is a preprint of an Article accepted for publication in WWW 2015 Companion, May 18–22, 2015, Florence, Italy © 2015. International World Wide Web Conference Committee'.

A note on versions:
The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: publications@warwick.ac.uk

warwick**publications**wrap

highlight your research

http://wrap.warwick.ac.uk/

Crowdsourcing the Annotation of Rumourous Conversations in Social Media

Arkaitz Zubiaga[1], Maria Liakata[1], Rob Procter[1], Kalina Bontcheva[2], Peter Tolmie[1]
[1]University of Warwick, Coventry, UK
[2]University of Sheffield, Sheffield, UK
{a.zubiaga,m.liakata,rob.procter}@warwick.ac.uk
k.bontcheva@sheffield.ac.uk,peter.tolmie@gmail.com

ABSTRACT

Social media are frequently rife with rumours, and the study of rumour conversational aspects can provide valuable knowledge about how rumours evolve over time and are discussed by others who support or deny them. In this work, we present a new annotation scheme for capturing rumour-bearing conversational threads, as well as the crowdsourcing methodology used to create high quality, human annotated datasets of rumourous conversations from social media. The rumour annotation scheme is validated through comparison between crowdsourced and reference annotations. We also found that only a third of the tweets in rumourous conversations contribute towards determining the veracity of rumours, which reinforces the need for developing methods to extract the relevant pieces of information automatically.

Categories and Subject Descriptors

I.2 [**Applications and Expert Systems**]: Natural language interfaces; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology, Natural language*

General Terms

Experimentation

1. INTRODUCTION

Social media has become a ubiquitous platform that everyday more and more people use to communicate with one another, and to stay abreast of current affairs and breaking news as they unfold [2]. Popular social media, such as Twitter or Facebook, become even more important in emergency situations, such as shootings or social upheavals, where information is often posted and shared first [7], before even news media [9]. However, the streams of posts associated with these crisis events are often riddled with rumours that are not verified and corroborated [20] and hence need to be handled carefully.

The ease with which uncorroborated information can be propagated in social media, as well as the potential resulting damage to society from the dissemination of inaccurate information [10], highlights the need to study and understand how rumours spread. However, little is known today about the way rumours propagate in social media, as well as how to mitigate their undesirable effects. Controversial information introduced by unconfirmed rumours usually sparks discussion among the user community. Users contribute their opinions and provide more evidence that either backs or denies the rumour [14]. The conversations produced in these discussions can provide valuable information to help manage rumours posted in social media [12] and assess their veracity. Previous research in this area has investigated the way rumours are spread, by looking at whether individual social media posts support or deny a rumour [15, 14, 3, 18]. However, to the best of our knowledge the present work is the first to look into the conversational aspects of rumours, as a more thorough analysis of social media posts in the context of rumours.

We describe our method to identify and collect rumourous conversations on Twitter performed in collaboration with journalists, which improves on the shortcomings of previous rumour collection methods in the literature. We introduce an annotation scheme that enables the annotation of tweets participating in rumourous conversations, making it possible to track the trajectory of these conversations, e.g., towards a consensus clarifying the veracity of the rumourous story. We assess the validity of the annotation scheme through experiments using a crowdsourcing platform, and compare it to our reference annotations. The novelty of our approach consists in providing a methodology for obtaining rumour datasets semi-automatically from social media and providing a framework for annotating and analysing rumourous conversations.

2. RELATED WORK

The emergence and spread of rumours has been studied for decades in different fields, primarily from a psychological perspective [1, 16, 5]. However, the advent of the Internet and social media offers opportunities to transform the way we communicate, giving rise to new forms of communicating rumours to a broad community of users [10]. Previous work on the study and classification of rumourous tweets has addressed the task of establishing veracity by looking only at whether each individual tweet supports or denies a rumour [15, 3, 14, 18]. While this is an important part of such analysis, it fails to capture the interaction between tweets that

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2743052.

leads to the eventual verification or discrediting of a rumour. Additionally, some of this work led to contradictory conclusions regarding the extent to which the community of social media users manage to debunk inaccurate statements and buttress truthful facts. While [3] found a similar number of tweets supporting and denying false rumours during the 2010 Chilean earthquake, [18] reported that the majority of tweets supported false rumours during the 2013 Boston Marathon bombings. Factors such as people's inherent trust in information they see repeatedly from multiple sources, irrespective of its actual validity, seems to play an important role [19]. We have collected a diverse set of rumours with more detailed annotations, to help shed light on conversational aspects of tweets. We have also defined a more detailed annotation scheme which enables the annotation of additional features playing an important role in determining the veracity of rumours and capturing the interaction between tweets around a rumourous story.

We adopt the definition of a rumour given in [20], as *"a circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety"*. We collect rumours in a process guided by journalists, and develop an annotation scheme for social media rumours, which we test and validate through crowdsourcing. While similar efforts have been made for the annotation of social media messages posted during crises [8], to the best of our knowledge this is the first work to delve into the data collection and sampling method, as well as to set forth and validate an annotation scheme for social media messages posted in the context of rumours.

3. DATA HARVESTING AND CURATION

The process we followed to collect rumourous datasets from Twitter is semi-automatic, from accessing the Twitter API through to collecting manual annotations.

3.1 Harvesting Source Tweets

We collect tweets from Twitter's streaming API. Contrary to existing approaches [15], we do not necessarily know in advance what specific rumours we are collecting data for. Instead, we track events associated with breaking news that are likely to be rife with rumours. By following this approach, we aim to emulate the scenario faced by journalists who are tracking breaking news in social media. That is, journalists follow an ongoing event and new rumours emerge constantly as the event unfolds. Hence, for the collection of events, we ask journalists for keywords and hashtags associated with relevant events and collect associated tweets as they are being posted. We applied this process to three different events. Two of them were breaking news events expected to generate rumours: (1) the Ferguson unrest in August 2014 (tracking "#ferguson"), where many citizens protested after the killing of a black adolescent by the police in the United States; (2) the Ottawa shootings in October 2014 (tracking "#ottawashooting" and "#ottawashootings", and relevant keywords like "ottawa shooting" and "parliament hill"), where a soldier was shot by a gunman in Canada. The third event is a specific rumour identified by the journalists beforehand: on 12th October, 2014, a rumour circulated that AC Milan footballer *Michael Essien had contracted Ebola* (tracking "ebola AND essien"), which was later denied.

3.2 Obtaining Conversation Threads

Once the tweets for each event were collected, the next step involved identifying rumour-bearing tweets and associated conversations (tweet threads). This includes three steps: (i) as the datasets were originally very large (i.e., 8.7 million tweets for Ferguson, and 1.1 million tweets for the Ottawa shootings), we automatically sampled the data based on certain criteria to make manual annotation possible, (ii) we collected conversations sparked by the sampled set of tweets, where conversations include all tweets which replied to the source tweets, directly or indirectly. These conversations/threads provide additional context for manual annotation. Finally in (iii) we separated rumours and non-rumours using manual annotation.

To sample the large sets of tweets collected for each of the events, we relied on the number of retweets as a human-sourced signal that a tweet has produced interest, in line with our definition of rumours. This filtered set of tweets was manually curated by journalists, distinguishing between rumours and non-rumours for the most eventful days of the 3 events under study. We refer to the latter sampled set of tweets as source tweets, for which associated conversations were also collected. Given the very different nature of the datasets, we used 100 as the retweet threshold for the Ferguson unrest and the Ottawa shootings, which were very popular, and we used 2 as the threshold for the story of Michael Essien having contracted Ebola.

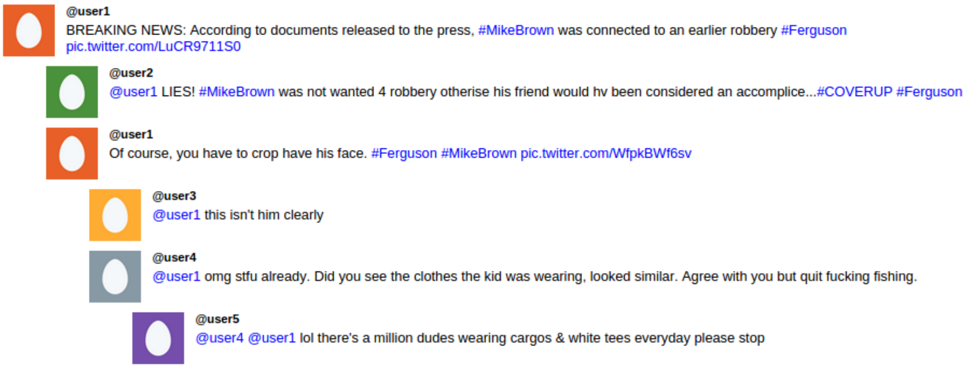
Conversations associated with each of the source tweets were obtained by retrieving all the tweets that replied to the source tweets. Due to the absence in Twitter's API endpoint of a way to retrieve such replying tweets directly, we scraped the web page of each of the source tweets to retrieve the IDs of the replying tweets, which were then collected from Twitter's API by tweet ID. This allows us to form the complete conversation sparked by a source tweet, which we have used for: (i) assisting the manual annotation of rumours vs non-rumours with additional context, and (ii) studying how conversations evolve around rumours. Figure 1 shows an example of a conversation collected from Twitter and visualised as a forum-like thread.

3.3 Manual Annotation of Source Tweets

Finally, having sampled the source tweets and collected the conversations associated with them, we used manual annotation to distinguish between rumours and non-rumours. This process was performed by a team of journalists at swissinfo.ch, following our definition of rumours. To facilitate the process, we developed an annotation tool that visualises a timeline of tweets. This tool makes it easier to visualise the threads associated with source tweets and provide source tweet annotations. Each annotation is assigned a title which allows us to group together different conversations on the same category or story. More details are available in [20]. The annotation of the 3 events mentioned above led to the identification of 291 rumourous conversations for the Ferguson unrest, 475 for the Ottawa shootings, and 18 for the Ebola rumour.

4. ANNOTATION SCHEME FOR SOCIAL MEDIA RUMOURS

In order to manually annotate tweets with respect to how they contribute to rumourous conversations, we have devel-



@user1
BREAKING NEWS: According to documents released to the press, #MikeBrown was connected to an earlier robbery #Ferguson pic.twitter.com/LuCR9711S0

@user2
@user1 LIES! #MikeBrown was not wanted 4 robbery otherise his friend would hv been considered an accomplice...#COVERUP #Ferguson

@user1
Of course, you have to crop have his face. #Ferguson #MikeBrown pic.twitter.com/WfpkBWf6sv

@user3
@user1 this isn't him clearly

@user4
@user1 omg stfu already. Did you see the clothes the kid was wearing, looked similar. Agree with you but quit fucking fishing.

@user5
@user4 @user1 lol there's a million dudes wearing cargos & white tees everyday please stop

Figure 1: Example of conversation sparked by a rumourous tweet in the context of Ferguson.

oped an annotation scheme which addresses different aspects of rumours. The datasets produced using human annotation will then be used to train machine learning classifiers to recognise patterns characteristic of rumours observed in new events and associated conversations. The annotation scheme we introduce here has been developed through an iterative process of rounds of annotation and evaluation with PhD students experienced in conversation analysis (we omit details due to lack of space).

The annotation scheme has been designed to annotate Twitter conversations both in terms of properties of each tweet and the relation to its parent tweet, that is a tweet it is linked to through a reply. The conversation is thus modelled as a tree structure. Within the conversation tree, we distinguish two types of tweets: (i) the source tweet, which is the first tweet in the tree which has triggered the entire conversation, (ii) replies, which are tweets that respond to other tweets in the conversation tree. We make a distinction between first-level replies, tweets that directly reply to the source tweet, and deep replies, which are all subsequent replies in the conversation which are not replying directly to the source tweet but to other replies.

The main goal of the annotation scheme has been to cover three aspects of the tweet conversation that are key in determining the veracity of a story: (i) whether a post supports or denies a story, as also used in previous work [14], (ii) the certainty with which the author of a post presents their view, and (iii) the evidence that is being given along with the post/statement to back up the author's view. In this work, we introduce the latter two to enable a more thorough analysis of posts in the context of rumours, beyond the fact that they support or deny a story. Also, we extend support to include responses to all tweets within the conversation, not just the story in the source tweet, as done in previous work (response type). Our annotation scheme therefore includes four different features, two of which apply to both source tweets and responses (certainty and evidentiality) and two variations of support that are dependent on the tweet type (i.e. support for the story in the source tweets, and response type for the relation between replies or replies and the source tweet). Figure 2 shows a diagram depicting the annotation scheme designed for rumourous conversations in social media. In what follows we present in detail each of the features forming the annotation schema.

Support: Support is only annotated for source tweets. It defines if the message in the source tweet is conveyed as a statement that supports or denies the rumour. It is hence different from the rumour's truth value, and intends to reflect what the tweet suggests is author's view towards the rumour's veracity. The support given by the author of a tweet can be deemed as: (1) supporting the rumour, (2) denying it, or (3) underspecified, when the author's view is unclear. This feature is related to the "Polarity" feature in the factuality scheme by Saurí et al. [17].

Response Type: Response type is used to designate support for the replying tweets. Given a source tweet that introduces a rumourous story, other users can reply to the author, leaning for instance in favour or against the statement. Some replies can be very helpful to determine the veracity of the rumour, and thus we annotate the type of reply with one of the following four values: (1) *agreed*, when the author of the reply supports the statement they are replying to, (2) *disagreeing*, when they deny it, (3) *appeal for more information*, when they ask for additional evidence to back up the original statement, or (4) *comment*, when the author of the reply makes their own comment without adding anything to the veracity of the story. Note that the response type is annotated twice for deep replies, i.e., tweets that are not directly replying to the source tweet. In these cases, the response type is annotated for a tweet determining two different aspects: (i) how the tweet is replying with respect to the rumour in the source tweet, and (ii) how the tweet is replying to the parent tweet, the one it is directly replying to. This double annotation allows us to better analyse the way conversations flow, and how opinions evolve with respect to veracity. The inclusion of this feature in the annotation scheme was inspired by Procter et al. [14], who originally introduced these four types of responses for rumours.

Certainty: Certainty measures the degree of confidence expressed by the author when posting a statement in the context of a rumour and applies to both source tweets and replies. The author can express different degrees of certainty when posting a tweet, from being 100% certain, to considering it as a dubious or unlikely occurrence. Note that the value annotated for either support or response type has no effect on the annotation of certainty, and thus it is coded regardless of the statement supporting or denying the rumour. The values for certainty include: (1) *certain*, when the author is fully confident or the author is not showing any kind of doubt, (2) *somewhat certain*, when they are not fully confident, and (3) *uncertain*, when the author is clearly unsure. This feature and the possible values were inspired

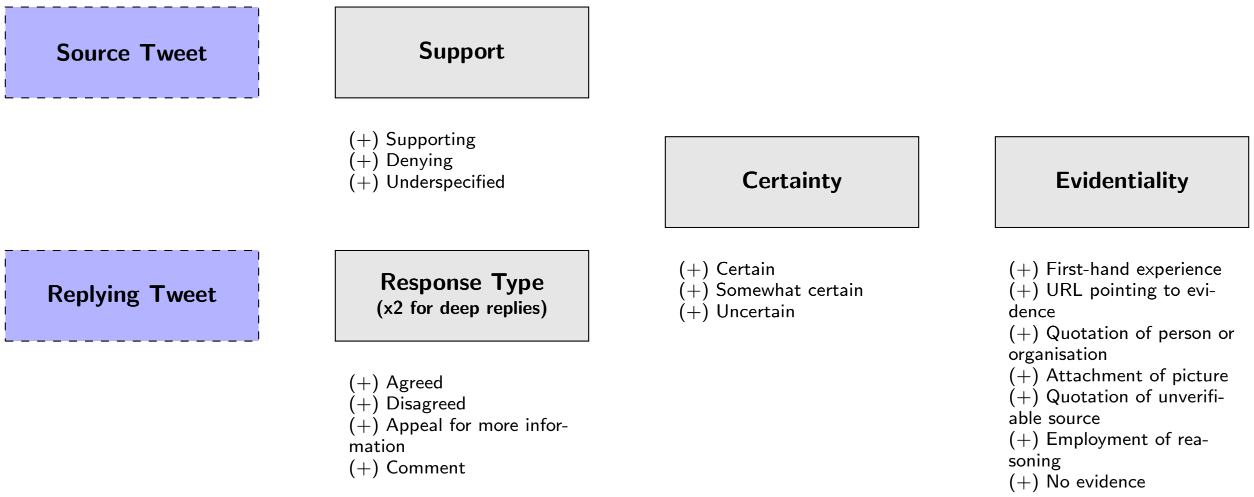


Figure 2: Annotation scheme for rumourous social media conversations.

by Saurí et al. [17], who referred to it as "modality" when annotating the factuality of news headlines.

Evidentiality: Evidentiality determines the type of evidence (if any) provided by an author and applies to both source tweets and replying tweets. It is important to note that the evidence has to be directly related to the rumour being discussed in the conversation, and any other kind of evidence that is irrelevant in that context should not be annotated here. Evidentiality can have the following values: (1) *first-hand experience*, when the author claims to have witnessed events associated with the rumour (2) *attachment of a URL* pointing to evidence, (3) *quotation* of a person or organisation, when an accessible source is being quoted as a source of evidence, (4) *attachment of a picture*, (5) quotation of an *unverifiable source*, when the source being mentioned is not accessible, such as "my friend said that...", (6) *employment of reasoning*, when the author explains the reasoning behind their view, and (7) *lack of evidence*, when none of the other types of evidence is given in the tweet. Contrary to the rest of the features, more than one value can be picked for evidentiality, except when "lack of evidence" is selected. Hence, we cater for the fact that a tweet can provide more than one type of evidence, e.g. quoting a news organisation while also attaching a picture that provides evidence.

5. CROWDSOURCING ANNOTATION OF RUMOUROUS CONVERSATIONS

In order to annotate the harvested Twitter conversations with the annotation scheme described above, we used crowdsourcing, so as to maximise speed [13]. Crowdsourcing has been used extensively for the annotation of Twitter corpora [6, 11]. We have used CrowdFlower[1] as the platform for crowdsourcing as it provides a flexible interface and has fewer restrictions than Amazon Mechanical Turk. None commercial crowdsourcing platforms were not a viable alternative at this stage. To validate and assess the viability of crowdsourcing annotations using our scheme, we sampled 8 different source tweets and their associated conversations from the 784 rumours identified for the 3 events. This includes 4 source tweets for the Ferguson unrest, and 2 source tweets each for the Ottawa shootings and the story of Essien having contracted Ebola (Table 1 shows the number of source tweets and replies included in each case).

Event	Src. tweets	1st rep.	2nd rep.
Ferguson unrest	4	63	58
Ottawa shootings	2	20	35
Ebola	2	22	10
TOTAL	8	105	103

Table 1: Tweets sampled for annotation.

In order to have a set of reference annotations to compare the crowdsourced annotations against, the whole annotation task was also performed by one of the authors of this paper, which we use as a reference annotation (REF). A second annotator, one of the journalists that contributed to the manual annotation of source tweets, annotated one third of the whole (REF2). The inter-annotator agreement between REF and REF2 was 78.57% measured as the overlap. This serves as a reference to assess the performance of the crowdsourced annotations in subsequent steps.

5.1 Disaggregating Annotation Task into Microtasks

The annotation of an entire rumourous conversation can become time-consuming and cumbersome as it involves the annotation of all four features for all tweets in a conversation. As a first step we split the conversation into triples, where each triple consists of a tweet, which replies to the source tweet either directly or indirectly, its parent tweet (the tweet it replies directly to) and the source tweet. If the tweet replies directly to the source tweet and no other previous tweet in the conversation then this is a tuple rather than a triple. Where the objective is to annotate the source tweet, this will appear on its own. Along with these tweets, we also show annotators the title assigned to the conversation during the rumour identification phase (see section 3.3), which facilitates crowdsourced annotation of conversations by keeping in focus what the rumour is about.

To facilitate the task of the annotators further [4], we nar-

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

rowed down the annotation unit to a single feature for each tweet triple, i.e., an annotator that accepts to take up a microtask would be able to focus on a single feature (e.g. Response Type) without having to switch to other features. This can significantly speed up the process of annotating the same feature across triples or even different conversation threads. An alternative way of narrowing down the annotation unit would be to ask each worker to annotate all the features for a single tweet. However, this would involve having to focus on different features, understanding the annotation guidelines for all of them at the same time, and requires more effort and concentration. Instead, our approach lets workers focus on a single feature, which makes the task guidelines easier to read and understand well. The disaggregation produced a total of 10 different microtasks that we set up in the crowdsourcing platform. These 10 microtasks include 3 tasks for source tweets (annotation of each of support, certainty, evidentiality), 3 tasks for first-level replies (annotation of response type wrt the source tweet, certainty, evidentiality), and 4 for deep replies (annotation of response type wrt the source tweet, response type wrt the previous tweet, certainty, evidentiality). Each of these represent a separate job on the crowdsourcing platform.

5.2 Crowdsourcing Tasks Parameters

Our annotation units consist of either a triple/tuple of tweets or a single source tweet, annotated for a particular feature. For each annotation unit we collected annotations from at least 5 different workers. Each CrowdFlower job consists of 10 annotation units as described above. Thus this is the minimum an annotator commits to when accepting a job. We paid $.15 for the annotation of each set of 10 units. In order to make sure that the annotators had a good command of English, we restricted participants to those from the United States and the United Kingdom.

We performed an initial test on CrowdFlower to evaluate these parameters, which allowed further optimisation for the final crowdsourcing task. All 8 threads were annotated in this initial test which lead us to optimise the following two aspects. Firstly, we identified that having always 5 annotators (as was our initial configuration) was not optimal, as often more annotators were needed to reach agreement (defined below) in difficult cases. Thus, we enabled the *variable judgments mode* which allows us to have at least 5 annotators per unit, and occasionally more, up to a maximum number of annotators until a confidence value of 0.65 is reached. In most cases it was sufficient to set the maximum number of annotators to 7, apart from evidentiality where it was set to 10. Evidentiality is more challenging as one can assign 7 different values and more than one option can be picked, thus increasing the chance for a diverse set of annotations. Secondly, we noticed that some annotators were completing the task too fast, annotating a set of 10 units in a few seconds. To avoid this, we changed the settings to force the annotators to spend at least 60 seconds annotating sets of 10 source tweets, and at least 90 seconds annotating sets of 10 units of replying tweets. With the revised settings, the cost for the annotation of all 8 threads amounted to $102.78.

5.3 Crowdsourcing Task Results

The Crowdsourcing task involved the annotation of 216 different tweets, which are part of the 8 tweet threads sampled for the three events. This amounts to the annotation of 4,974 units (tweet triple+feature combination), and was performed by 98 different contributors. The final set of annotations was obtained by combining annotations by all workers through majority voting for each annotation unit. In order to report inter-annotator agreements, we rely on the percent of overlap between annotators, as the ratio of annotations that they agreed upon. When we compare the decisions of each of the annotators against the majority vote, we observe an overall inter-annotator agreement of 60.2%. When we compare the majority vote against our reference annotations, REF, they achieved an overall agreement of 68.84%. While this agreement is somewhat lower than the 78.57% agreement between REF and REF2, it is only worse when annotating for ”certainty”, as we will show later. This also represents a significant increase from earlier crowdsourcing tests performed before revising the settings, where the annotators achieved a lower agreement rate of 62.5%. When breaking down the agreement rate for each of the features (see Table 2), we see that the agreement values range from 58.17% for certainty in reply tweets, to 100% for support in source tweets. The agreement rates are significantly higher for source tweets, given that the annotation is easier as there is only the need to look at one tweet, instead of tuples/triples. This analysis also allows us to compare the agreement by feature between the crowdsourced annotations (CS) and REF, as well as between REF and REF2. We observe that agreements are comparable in most cases, except for the agreement on certainty, which is significantly higher between REF and REF2. The latter represents the major concern here, where the crowdsourcing annotators performed worse, which we explain later.

Source tweets			
	Support	**Certainty**	**Evident.**
CS vs REF	100%	87.5%	87.5%
REF vs REF2	100%	62.5%	87.5%

Replying tweets			
	Resp. type	**Certainty**	**Evident.**
CS vs REF	70.42%	58.17%	74.52%
REF vs REF2	71.82%	87.14%	78.89%

Table 2: Inter-annotator agreement by feature.

In more detail, Table 3 shows the distribution of annotated categories, as well as the agreement rates for each feature when compared to the reference annotations, REF. Looking at the agreement rates, annotators agreed substantially with the reference annotations for source tweets (100% agreement). For replying tweets, as discussed above, the depth of the conversation and the additional context lead to lower agreement rates, especially for some of the categories. The agreement rates are above 60% for the most frequent types of values, including response types that are ”comments” (67.69%), authors that are ”certain” (60.22%), and tweets with ”no evidence” (85.37%). The agreement is lower for the other annotations, which appear less frequently. This certainly proves that the annotation of replies is harder than the annotation of source tweets, as the conversation gets deeper and occasionally deviates from the topic discussed in the source tweet. One of the cases with a low agreement rate is when the evidence provided is “reasoning”. This shows the need to emphasise even more in subsequent crowdsourcing tasks the way this type of evidence should be annotated, by remarking that the reasoning that is being given in a tweet must be related to the rumourous story and

Source tweets					
Support		**Certainty**		**Evidentiality**	
% of times	agreem.	% of times	agreem.	% of times	agreem.
supporting (100%)	100%	certain (75%)	100%	no evidence (37.5%)	100%
denying, underspecified (0%)	–	somewhat certain (12.5%)	100%	author quoted (37.5%)	100%
	–	uncertain (75%)	100%	picture attached (25%)	50%
				URL given, unverifiable source, witnessed, reasoning (0%)	– –

Replying tweets					
Response type		**Certainty**		**Evidentiality**	
% of times	agreem.	% of times	agreem.	% of times	agreem.
comment (66.56%)	67.69%	certain (54.33%)	60.22%	no evidence (79.81%)	85.37%
disagreed (15.43%)	53.70%	somewhat certain (25.96%)	40%	reasoning (9.62%)	29.17%
agreed (10.61%)	50%	uncertain (19.71%)	41.18%	author quoted (3.37%)	62.5%
appeal for more info (7.40%)	33.33%			URL given (3.37%)	50%
				picture attached (2.89%)	33.33%
				witnessed (0.48%)	0%
				unverifiable source (0.48%)	0%

Table 3: Distribution of annotations: percent of times that each category was picked, and the agreement with respect to our reference annotations (CS vs REF).

not another type of reasoning.

When we look at the distribution of values the annotators chose, we observe an imbalance in most cases. For response type, we see that as many as 66.5% of the replies are comments, which shows that only the remainder 33.5% provide any information that adds something to the veracity of the story. The evidentiality is even more skewed towards tweets that provide no evidence at all, which amount to 85.4% of the cases. Both the abundance of comments, and the dearth of evidence, emphasise the need for carefully analysing these conversations when building machine learning tools to pick out content that is useful to determine the veracity of rumourous stories. The certainty feature is slightly better distributed, but still skewed towards more than 54% cases of certain statements; this could be due to the fact that many users do not express uncertainty in short, written texts even when they are not 100% sure.

To better understand how the different features that have been annotated fit together, we investigated the combinations of values selected for the replying tweets. Interestingly, we observe that among the replying tweets annotated as comments as many as 80.3% were annotated as having no evidence, and 47.5% were annotated as being certain. Given that comments do not add anything to the veracity of the rumour, it is to be expected that there would be no evidence. We also investigated several cases to understand how certainty was being annotated for comments; we observed that different degrees of certainty were being assigned to comments where certainty can hardly be determined as it does not seem to apply, e.g., in the tweet "My heart goes out to his family". This also helped us understand the low agreement rate between CS and REF for certainty, which may drop due to the comments with an unclear value of certainty. For these two reasons, together with the fact that comments represent tweets that do not add anything to the veracity of the story, we consider revising the annotation scheme so that these two features should not be annotated for comments. This, in turn, reduces significantly the cost of running the crowdsourcing tasks, given that for as many as 66.5% replying tweets that represent comments, we would avoid the need for two annotation tasks.

6. DISCUSSION

We have described a novel method to collect and annotate rumourous conversations from Twitter, and introduced an annotation scheme specifically designed for the annotation of tweets taking part in these rumourous conversations. This scheme has been revised iteratively and used for the crowdsourced categorisation of rumour-bearing messages. As far as we know, this is the first conversation-based annotation scheme specifically designed for social media rumours and complements related annotation work in the literature. Earlier work only considered whether a message supported or denied a certain rumour. Our annotation scheme is able to annotate the certainty with which those messages are posted, the evidence that accompanies them, as well as the flow of support for a rumour within a conversation, which are all key additional aspects when considering the veracity of a story. The agreement achieved between the crowdsourced annotations and our reference annotation, which is comparable to and occasionally better than the agreement between our own reference annotations, has enabled us to validate both the crowdsourcing process and the annotation scheme. While the annotation scheme has so far been applied to a relatively small data sample, it reveals some interesting patterns, especially suggesting that to a great extent conversations around rumours in social media mostly involve comments, which do not add anything to the veracity of the story. This reinforces the motivation of our work of categorising tweets in these conversations so as to identify the tweets that do provide useful knowledge and evidence to determine the veracity of a rumourous story. The annotation tests have also helped identify suitable settings for the crowdsourcing tasks, and have ultimately revealed a form of simplifying the scheme while keeping the main, required annotations and reducing the cost of running the task.

Our next plan is to apply the annotation scheme to a larger dataset of social media rumours, collected for a broader set of events and including tweets in other languages besides English. The creation of this dataset will then enable us to perform a conversation analysis study, as well as to develop machine learning tools to deal with social media rumours.

Acknowledgments

We would like to thank the team of journalists at swissinfo.ch for the annotation work. This work has been supported by the PHEME FP7 project (grant No. 611233).

7. REFERENCES

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

[2] M. Anderson and A. Caumont. How social media is reshaping news. `http://www.pewresearch.org/fact-tank/2014/09/24/how-social-media-is-reshaping-news/`, 2014.

[3] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.

[4] J. Cheng, J. Teevan, S. T. Iqbal, and M. S. Bernstein. Break it down: A comparison of macro- and microtasks. In *Proceedings of CHI*, 2015.

[5] L. Festinger, D. Cartwright, K. Barber, J. Fleischl, J. Gottsdanker, A. Keysen, and G. Leavitt. A study of a rumor: its origin and spread. *Human Relations*, 1948.

[6] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. Association for Computational Linguistics, 2010.

[7] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. Processing social media messages in mass emergency: A survey. *arXiv preprint arXiv:1407.7071*, 2014.

[8] M. Imran, C. Castillo, J. Lucas, M. Patrick, and J. Rogstadius. Coordinating human and machine intelligence to classify microblog communications in crises. *Proc. of ISCRAM*, 2014.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[10] S. Lewandowsky, U. Ecker, C. Seifert, N. Schwarz, and J. Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012.

[11] S. A. Paul, L. Hong, and E. Chi. What is a question? crowdsourcing tweet categorization. *CHI 2011*, 2011.

[12] R. Procter, J. Crump, S. Karstedt, A. Voss, and M. Cantijoch. Reading the riots: What were the police doing on twitter. *Policing and Society*, 23(4):413–436, 2013.

[13] R. Procter, W. Housley, M. Williams, A. Edwards, P. Burnap, J. Morgan, O. Rana, E. Klein, M. Taylor, A. Voss, et al. Enabling social media research through citizen social science. In *ECSCW 2013 Adjunct Proceedings*, 2013.

[14] R. Procter, F. Vis, and A. Voss. Reading the riots on twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214, 2013.

[15] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[16] R. L. Rosnow and G. A. Fine. *Rumor and gossip: The social psychology of hearsay.* Elsevier, 1976.

[17] R. Saurí and J. Pustejovsky. Factbank: A corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227–268, 2009.

[18] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. In *Proceedings of iConference.* iSchools, 2014.

[19] A. Zubiaga and H. Ji. Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, 4(1):1–12, 2014.

[20] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Towards detecting rumours in social media. In *AAAI Workshop on AI for Cities*, 2015.