D2.1 Development of an Annotation Scheme for Social Media Rumours

Arkaitz Zubiaga (University of Warwick)
Peter Tolmie (University of Warwick)
Maria Liakata (University of Warwick)
Rob Procter (University of Warwick)

Abstract.
FP7-ICT Collaborative Project ICT-2013-611233 PHEME
Deliverable D2.1 (WP2)

This document outlines a preliminary definition of an annotation scheme for rumours spread through social media, as well as the code frames that will be used to mark up the corpora collected for PHEME. It has been developed through an iterative process of revisions, and it is intended to encompass the different kinds of rumours that can be spread and discussed in the context of different events and situations. It especially considers the way conversations flow in social media, and has been developed in an interdisciplinary style by building on work in sociolinguistics, including the related approaches of Conversation Analysis [51] and Ethnomethodology [16]. The resulting annotation scheme will be used for the annotation of a larger corpora of social media rumours through a crowdsourcing platform. Annotation guidelines will be defined in following work and tested with small social media corpora before running the large annotation task.
Keyword list: rumors, veracity, annotation scheme, social media

Project	PHEME No. 611233
Delivery Date	July 15, 2014
Contractual Date	June 30, 2014
Nature	Report
Reviewed By	Petya Osenova
Web links	-
Dissemination	PU

PHEME **Consortium**

This document is part of the PHEME research project (No. 611233), partially funded by the FP7-ICT Programme.

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

MODUL University Vienna GMBH
Am Kahlenberg 1
1190 Wien
Austria
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

ATOS Spain SA
Calle de Albarracin 25
28037 Madrid
Spain
Contact person: Tomás Pariente Lobo
E-mail: tomas.parientelobo@atos.net

iHub Ltd.
NGONG, Road Bishop Magua Building
4th floor
00200 Nairobi
Kenya
Contact person: Rob Baker
E-mail: robbaker@ushahidi.com

The University of Warwick
Kirby Corner Road
University House
CV4 8UW Coventry
United Kingdom
Contact person: Rob Procter
E-mail: Rob.Procter@warwick.ac.uk

Universitaet des Saarlandes
Campus
D-66041 Saarbrücken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Georgi Georgiev
E-mail: georgiev@ontotext.com

King's College London
Strand
WC2R 2LS London
United Kingdom
Contact person: Robert Stewart
E-mail: robert.stewart@kcl.ac.uk

SwissInfo.ch
Giacomettistrasse 3
3000 Bern
Switzerland
Contact person: Peter Schibli
E-mail: Peter.Schibli@swissinfo.ch

History of Changes

Version	Date	Author	Changes
1.0	20.06.2014	Arkaitz Zubiaga	first version with initial annotation scheme
1.1	02.07.2014	Peter Tolmie	extended version with literature from CA and EM
1.2	04.07.2014	Arkaitz Zubiaga	Validation tests and revised annotation scheme
1.3	06.07.2014	Rob Procter	Minor edits to text mostly to fix grammatical issues and improve clarity
1.4	07.07.2014	Arkaitz Zubiaga	Annotation tests added, elaborated scheme revision, and 1st version of executive summary
1.5	08.07.2014	Rob Procter	Further minor edits and additions
1.6	13.07.2014	Rob Procter	Chnages following internal review

Executive Summary

This deliverable D2.1 describes the work performed at the University of Warwick within the Work Package 2 (WP2) as a member of the PHEME project. Pursuing the goal of studying the spread of rumours in social media, as well as the way users discuss them online, this document outlines the preceding efforts towards defining an annotation scheme. An annotation scheme will enable to label and categorise manually messages that are part of rumours. These manual annotations will then enable to perform computational analyses to gain insight so as to how rumours are spread in social media.

The annotation scheme described in this deliverable has been developed through an iterative process, with two rounds of validation tests and subsequent revisions. The validation tests have been carefully performed by looking at real data including rumourous conversations collected from the microblogging service Twitter. The annotation scheme is intended to encompass the wide variety of types of rumours that can be spread and discussed in the context of different events and situations.

The development of the annotation scheme is informed by findings from the related disciplines of Conversation Analysis and Ethnomethodology, but is also aware of the specific characteristics that social media possesses, which often differs from face-to-face communications.

Having tested the annotation scheme on a small sample of rumours retrieved through social media, the ongoing work at WP2 is dealing with the retrieval of additional rumourous data collections from social media. Further validation tests with these additional data collections will enable further refining of the annotation scheme to create the final revision. The final annotation scheme will then be used to crowdsource the annotation of a larger corpus of social media rumours.

The corpus that will be ultimately obtained through this process will then be used in multiple Work Packages within the PHEME project, including Work Package 3 to develop open source methods to track the flow of rumours, as well as Work Packages 7 and 8 to build rumour corpora of interest to the healthcare and journalism domains.

Contents

1 Introduction **3**
- 1.1 Relevance to PHEME . . . 4
  - 1.1.1 Relevance to project objectives . . . 4
  - 1.1.2 Relation to other work packages . . . 5
- 1.2 Outline of the Deliverable . . . 5

2 Defining Rumours **6**

3 Conversation Analysis and Ethnomethodology **8**
- 3.1 The Origins of Ethnomethodology & Conversation Analysis . . . 8
- 3.2 Respecification . . . 9
- 3.3 Respecifying Rumour . . . 11

4 Related Annotation Schemes **12**
- 4.1 Rumour Types . . . 12
- 4.2 Factuality and Sources . . . 13
- 4.3 Author Types . . . 14
- 4.4 Other Annotation Schemes for Conversations . . . 15

5 Developing an Annotation Scheme for Rumours **17**
- 5.1 Initial Annotation Scheme . . . 17
  - 5.1.1 Message Crafting . . . 19
  - 5.1.2 Spread and Reactions . . . 23
- 5.2 Validation of the Annotation Scheme . . . 25
  - 5.2.1 Data Collection from Twitter . . . 25
  - 5.2.2 Annotation Test . . . 26
- 5.3 Revised Annotation Scheme . . . 26
  - 5.3.1 Source Tweet . . . 27
  - 5.3.2 Response Tweets . . . 30
- 5.4 Validation of the Revised Annotation Scheme . . . 30
- 5.5 Work in Progress . . . 31
  - 5.5.1 Putting the Annotation Scheme into Practice . . . 32

6 Extending the Annotation Scheme **34**

6.1 Moving towards annotation grounded in microblog analysis 34
6.1.1 The turn-taking mechanism 35
6.1.2 Topic 35
6.1.3 The organization of conversation as applied to tweets and the organization of tweets when seen as conversations 36
6.1.4 The intersubjective constitution of tweeting as a phenomenon . . 42
6.1.5 Following and followers 43
6.1.6 Tweeting as a mode of communication 43
6.1.7 Looking at microblogging as its own job of work with its own grammars of action 44
6.1.8 Referential practices 46
6.1.9 The asynchronous character of microblog exchange 46
6.1.10 The organisation of rumour as a feature of microblog exchange . 47
6.1.11 Associated literatures 47
6.2 A staged and iterative process 48

7 Discussion **50**

A Acronym Definitions **57**

B Sample Rumourous Conversations **58**

C First Round of Annotations for Validating the Scheme **61**
C.1 ATM Hoax 61
C.2 Superbowl and Prostitution 65

D Second Round of Annotations for Validating the Scheme **69**
D.1 ATM Hoax 69
D.2 Superbowl and Prostitution 73

Chapter 1

Introduction

While inaccurate and questionable information has always been a reality, the emergence of the Internet and social media has increased this concern due to the ease with which such information can be spread to large communities of users [27]. This kind of information often starts as a rumour being posted by an individual on social media such as Twitter[1], Facebook[2], or Instagram[3], and subsequently being passed on through their social networks and reaching a larger audience. The spread of rumours may have undesirable consequences as they can convey wrong information to people. Not only does this affect ordinary individuals who might pass on information without verifying it, but also professional practitioners such as journalists who may pick up a story from social media and inadvertently disseminate inaccurate or false information via news media. Given that the spread of inaccurate information can have dangerous consequences for society, the analysis of rumours becomes crucial to prevent the diffusion of inaccurate information and to identify information that is well backed up and verified.

The study of the spread of rumours in social media is attracting increasing interest within the scientific community [15, 21]. However, these studies have generally focused on the virality and social network analysis of rumours and have not looked in more detail at the nature of rumours, how they are linguistically crafted, and how they are subsequently supported and/or denied by others in social media. We intend to fill this gap by first introducing an annotation scheme, a framework for systematic annotation of different aspects reflecting the content of rumours. Annotations resulting from this scheme will assist to perform content-based studies on conversations around rumours, and to develop a system that automatically processes rumorous texts in social media, as well as conversational aspects such as reactions around them.

One of the proposed ways of handling the development of an annotation scheme for Twitter feeds that moves beyond the work undertaken by Procter et al. [37] on the Lon-

[1]Twitter - http://twitter.com/

[2]Facebook - http://www.facebook.com/

[3]Instagram - http://instagram.com/

don riots, is to exploit existing work in the area of conversation analysis as a means of providing richer annotations of topics as they unfold. Here our primary concern will be to map out what a grounding of annotations in the microblogging domain might look like. In particular, we will argue that, whilst conversation analytic approaches will serve well as a source of inspiration, it is ultimately going to be necessary to respecify the interest somewhat as "microblog analysis" in order to steer around the potential dangers of missing the lived character of how people reason about tweeting as an activity in its own right.

In this document, we present a review of previous research developing annotation schemes that are relevant to PHEME, study their applicability to our context, and introduce a preliminary version of our annotation scheme which combines aspects of these schemes, creating a new framework aiming at capturing multiple aspects of rumours. This annotation scheme will be iteratively developed in the following months, leading to the production of a final version that will enable the creation of the annotated corpora that will be used in PHEME.

1.1 Relevance to PHEME

This section describes the relevance of this deliverable to the PHEME project's objectives, and how it relates to the other work packages in the project.

1.1.1 Relevance to project objectives

This document outlines preliminary efforts towards undertaking the objectives defined in the description of Work Package 2 (WP2). This work package aims to perform qualitative analyses of rumours spread through social media, considering their diffusion across different media as well as languages. These analyses will be conducted in an interdisciplinary setting, and from different perspectives, including a qualitative social science analysis, and the study and development of a methodology and tools for the linguistic analysis of rumours using natural language processing. In order to carry out this research, here we define the initial steps towards characterising social media rumours, performing an interdisciplinary analysis drawing, in particular, on Conversation Analysis and Ethnomethodology, and developing an annotation scheme. This annotation scheme will provide knowledge for human annotators to enrich the social media rumours that will be collected and put together in a corpus for the purposes of PHEME.

The development of the annotation scheme described in this deliverable is key for the subsequent creation of corpora of social media rumours, which will include annotations provided by human coders. The annotated corpora obtained through this process will then be used to conduct research on social media rumours, as defined in the WP2 of the PHEME's Description of Work. The annotation scheme is also designed with the goal to facilitate subsequent computational analysis of rumours using machine learning.

1.1.2 Relation to other work packages

The work presented here, and the annotated corpora that will be developed afterwards by making use of the annotation scheme, will be used in various work packages. In Work Package 2, it will support the ontology modelling Task 2.2. Work Package 3 will deal with the development of open source methods to track the flow of rumours, where the corpora will be used for development, parameter tuning and initial evaluation. In Work Package 4, it will support LOD-based reasoning about rumours. Work Packages 7 and 8 will also deal with the annotation of corpora in Tasks 7.2 and 8.2. These tasks will annotate rumours of interest to the healthcare and journalism use cases, respectively, making use of the annotation scheme we have defined.

1.2 Outline of the Deliverable

This deliverable is organised in the following chapters. Next, in Chapter 2 we provide a formal definition of rumours, which combines previous definitions from both scientific literature and dictionaries. The following two sections provide some background relevant to our work. Chapter 3 outlines ideas from the fields of Conversation Analysis and Ethnomethodology and their relevance to the investigation of rumour. Chapter 4 discusses existing annotation schemes that are relevant to the purposes of PHEME. Our proposed annotation scheme is then introduced in Chapter 5, beginning with a description of our initial attempt at creating the scheme. We then present our validation methodology and the revised scheme that has resulted from it. We discuss further the extension and iterative refinement of this annotation scheme in Chapter 6. Finally, we summarise the conclusions drawn from this work, and discuss its limitations and our future work programme in Chapter 7.

Chapter 2

Defining Rumours

While there is a substantial amount of research around rumours in a variety of fields ranging from psychological studies [46] to computational analyses [39], defining and differentiating them from similar phenomena remains an active topic of discussion within the scientific community. Some researchers have attempted to provide a solid definition and characterisation of rumours so as to address the lack of common understanding around the specific categorisation of what is or is not a rumour. DiFonzo and Bordia [12] emphasise the need to differentiate rumours from other similar phenomena such as gossip and urban legends. They define rumours as *"unverified and instrumentally relevant information statements in circulation that arise in contexts of ambiguity, danger or potential threat and that function to help people make sense and manage risk"*. This definition also ties in well with that given by the Oxford English Dictionary (OED): *"A currently circulating story or report of uncertain or doubtful truth"*[1]. Further, Guerin and Miyazaki [20] provide a detailed characterisation of rumours, emphasising what differentiates them from urban legends and gossip (see Figure 2.1 for the characterisation of rumors, gossip, and urban legends). From the differences posited by these authors, we highlight the following:

- It is of general interest to most listeners.
- It is of personal consequence and interest to listeners.
- The truth behind it is difficult to verify.
- It must be credible despite ambiguities.
- It can be ambiguous.
- It tends to be a short story as compared to e.g., urban legends.
- It gains attention with horror or scandal.

[1] http://www.oxforddictionaries.com/definition/english/rumour

Categorization of Rumors, Urban Legends, and Gossip into Twelve Conversational Properties That Gain Attention or Promote Relationships

	Rumors	Urban legends	Gossip	"Serious" knowledge
Is of general interest to most listeners	√	√		√
Of personal consequence and interest to listeners	√		√	
Deals with persons known to speaker or listener			√	
Truth difficult to verify	√	√	√	√
Must be credible despite ambiguities	√		√	√
Can be ambiguous	√	√		√
Short or long?	Short	Long	Short	Short
Uses a story plot		√		
Attention gained with horror or scandal	√	√	√	
New or novel	√	√	√	√
Can be humorous		√	√	
Unusual or unexpected		√	√	√

Figure 2.1: Characterisation of rumours, gossip and urban legends.

- It has to be new or novel.

In contrast, urban legends are stories that are usually not credible or of personal consequence to the listeners, but tend to be more engaging and attention grabbing. Urban legends also tend to be longer stories. The main characteristic that differentiates gossip from rumours is that the former deal with persons known to the speaker or the listener. Both urban legends and gossip can be humorous, but that is not a feature that commonly characterises rumours.

Despite attempts to categorise them as different phenomena, Guerin and Miyazaki [20] posit that all three – rumours, gossip and urban legends – *are merely ways of keeping a listener's attention, and are not independently definable in themselves except for their particular conglomerate of conversational properties*.

Summing up, here we expand on the OED's definition with additional descriptions from rumour-related research, which is richer and we argue more appropriate for our purposes within PHEME. We formally define a rumour as a **circulating story of questionable veracity, which is apparently credible but hard to verify, and produces sufficient skepticism and/or anxiety so as to motivate finding out the actual truth**.

Chapter 3

Introducing the Conversation Analytic and Ethnomethodological Approaches

The preceding section of this deliverable is concerned with arriving at a workable definition of 'rumour' for the purposes of informing the collection of tweets that are immediately identifiable as rumours for the purposes of annotation. However, a longer-term strategy that we shall also be adopting in Work Package 2 is the use of Conversation Analytic and Ethnomethodological approaches to understand how tweets and whole bodies of related tweets are organised as social accomplishments, and how rumours are therefore constituted within this as social accomplishments in some way.

3.1 The Origins of Ethnomethodology & Conversation Analysis

Ethnomethodology first arose as an approach for analyzing social phenomena in sociology in the 1950s. Its principal articulation as a programme of research can be found in the works of Harold Garfinkel, most notably his *Studies in Ethnomethodology* (Garfinkel [16]). What might be termed a radical empirical sociology, it is heavily influenced by the phenomenological writings of Edmund Husserl and the later philosophical writings of Ludwig Wittgenstein. Its fundamental concern is with how orderly social phenomena are organised by people themselves through concerted local action to produce them as being recognizably the social phenomena they are taken to be. It focuses on what people do, and how people do it as a matter of method, such that everyone else can see that that is indeed what they are doing.

Conversation Analysis was first developed through the work of Harvey Sacks (see the *Lectures on Conversation* (Sacks [51])). Sacks was one of Garfinkel's students and his studies of conversation can be seen as a practical working through of Ethnomethodology's programme by taking a readily available phenomenon within society -– namely 'talk' -–

and seeing just what its organised properties might be as features of the social order.

In a seminal paper dating from 1963, Sacks made some important observations regarding the ways in which sociological description tended to be pursued at that time (and, for the larger part, since as well). To illuminate his concerns Sacks conceived of a machine at an industrial exhibition consisting of two parts where one part is designed to undertake some particular job whilst the other part systematically and contiguously provides a narration of what the first part is doing. He suggested that a lay understanding of this machine would be something along the lines of a 'commentator machine' and that any attempt to make sense of the machine would involve being able to reconcile the relationship between the parts doing the job itself and the parts doing the narration.

Sacks's idea here is to use the machine to represent the social world where you have a whole bunch of stuff going on that together constitutes the 'doing' part of society but, at the same time, you also have a bunch of talk going on whereby people systematically narrate their lives, the ways in which their lives are organised, and through which many of the 'doing' parts get implicated or even done. For Sacks the point is that, to understand the social world you cannot split those two bits apart and make use of the narration part without first of all looking at the narration part to see just how that works as well. What he is alluding to here is the tendency within social science to make use of language imported from the commonsense, everyday world without first of all opening up to inspection the work that language does in the world. Social scientists make use of language unreflectively as a resource for doing the job of description of the social world without taking that use of language to be itself a topic for investigation. Thus social scientists are, in reality, just engaging in the same work as the other narration components within the machine.

Sacks's work, and by necessity the rest of Conversation Analysis, is therefore heavily invested in the business of taking the social production of language-based phenomena as a serious topic for investigation in its own right. As tweets are also language-based phenomena with their own organizational properties, we are therefore similarly seeking to understand tweets in this kind of way.

3.2 Respecification

Alongside of this interest of Sacks in the problematic character of sociological description, Garfinkel had already been developing what he called a foundational 'respecification' of the problem of sociology. It was founded upon a re-working of Durkheim's famous aphorism that "The objective reality of social facts is Sociology's fundamental principle" (Durkheim [13]). Grounded in his own reading of the works of Edmund Husserl, Garfinkel sought to reframe this as a matter of it not being the case that there were just social facts out there to be picked up and inspected, so to speak, but rather that the sense of something counting as a social fact was something that was accomplished itself by the ordinary members of society. The problem was therefore opening up for inspection what

this accomplishment might consist in. People take for granted that there is order in the world and they expose in everything they do just what kinds of orderly arrangements they are presuming will hold. The job of the sociologist is to uncover and bring into view these assumptions about 'the way the world works' and to explicate the ways in which they provide methodologically for the production of orderly phenomena.

The notion of respecification was absolutely central to the work of Garfinkel. It can be seen to resonate through much of his writing but he gives explicit voice to the idea in several places. In an edited transcript of a conversation between Garfinkel and Benetta Jules-Rosette recorded in the summer of 1985 he presents respecification in the following way:

> *"Our studies developed a radical, alternate technology of social analysis. Some of its policies are well known ... These and others were developed in the attempt to avoid the intractable absurdities that everywhere accompany classic methods of analytic social studies of practical action. With our alternate methods we have specified several identifying issues of the problem of social order as discoverable phenomena in and as immortal ordinary society ... These identifying issues are only discoverable. They cannot be imagined and they cannot be obtained by operating on representations of social order. Their import is that they respecify the ordinary society and do so in inspectable, detailed ties between practical action and the phenomena of order/production."*, Garfinkel and Jules-Rosette, 1986, unpublished transcript

Using the placeholder 'order*' for all possible topics of interest 'in-and-as-of-the-workings-of-ordinary-society' Garfinkel offers a further articulation of the idea in another later volume of collected works (Garfinkel [17]):

> *"Not only the topic of detail, but every topic order* is to be discovered and is discoverable, and is to be respecified and is respecifiable, as only locally and reflexively produced, naturally accountable phenomena of order*. These phenomena of order* are immortal, ordinary society's commonplace, vulgar, familiar, unavoidable, irremediable and uninteresting 'work of the streets'."*, (Garfinkel [17]: 17)

This notion of respecifying what it is we might be talking about and not taking for granted articulations of the social world as features of the social world but rather inspecting how people themselves make them a feature in some way, is central to our own longer-term strategy within PHEME.

3.3 Respecifying Rumour

So, whilst there is a pragmatic necessity involved in pinning down what phenomena count as 'rumour' for the realization of a workable annotation scheme, this should be set against a longer-term concern with not just taking these assignations for granted but rather treating them as ongoingly revisable according to what our investigations into rumour production in Twitter reveal about how people themselves reason about rumour in various ways. Thus we consider ourselves to also be involved in the job of taking what has been construed as a topic in other fields, namely 'rumour', and considering what it could amount to as a topic of investigation 'in-and-as-of-the-workings-of-ordinary-society'.

A feature of our work over time will therefore be a respecification of 'rumour' as a topic of interest for sociological investigation by focusing upon what the 'local production', 'natural accountability', and 'coherence' of phenomena conventionally glossed as [rumours] look like in praxis. To do this involves moving away from taken-for-granted assumptions about what 'rumours' might be, and towards what it might take to be able to call something a 'rumour' -– reasonably or otherwise — in lived social action. It also involves exploring what work members of society are engaged in when they articulate the proposition that something might be a rumour. What kinds of things does ascribing something the status of a 'rumour' accomplish in the world? As a programme of work this will involve setting aside taken-for-granted notions of what rumour might amount to and instead looking at: a) what kinds of features of interaction are taken by members themselves to be recognizable as 'rumour' in some way; and b) what kinds of work in interaction ascriptions of rumour to phenomena might be seen to do. In particular this will be directed towards an examination of how rumour-relevant phenomena are organised features of microblogging practices, and specifically the use of Twitter, in their own right.

This exercise will build upon an existing corpus of work in the conversation analytic and ethnomethodological literatures that already touches upon rumour-related matters in various ways. Relevant texts here include: Meehan [32], Mellinger [33], Rapley [40], Smith [58] & Wooffitt [68] with regard to the accomplishment of 'facticity'; Coulter [7], Harper [22], Jalbert [25], Sacks [51] and Sidnell [57] regarding 'belief'; Antaki [2], Bergmann [5], Goodwin [19], Parker & O'Reilly [36] and Sacks [51] with regard to 'gossip'; and Clifton [6], Heritage et al. [23] and Sacks [51] regarding 'subversion'.

Chapter 4

Related Annotation Schemes

Here we discuss the most relevant annotation schemes and corpora that are closely related to the purposes of our work on the development of an annotation scheme for rumours. We have organised the annotation schemes into the following subsections: (i) rumour types, (ii) factuality and sources, and (iii) author types.

4.1 Rumour Types

Procter et al. [37] conducted a study of tweets sent during the 2011 England riots. They grouped tweets into "information flows", which is defined as a thread of tweets that retweet and make comments on a common source tweet. They looked at popular (i.e. large) information flows, and categorised them into an introduced typology of messages – media reports, pictures, rumours and reactions – as well as of author types. The paper provides detailed lists for both types of messages and authors. In the specific cases of rumours, they include the following subtypes: (i) claim without evidence, (ii) claim with evidence, (iii) counterclaim without evidence, (iv) counterclaim with evidence, (v) appeal for more information, and (vi) comment. They identified and characterised how rumours begin with someone tweeting an alleged incident, and quickly pick up popularity as others retweet and spread them. The veracity of a rumour is eventually questioned as Twitter users subject it to various "facticity tests" (e.g. questioning evidence, applying "common sense reasoning") and over time a consensus is usually reached. However, the authors posit that even previously refuted rumours can re-surface and continue to be spread.

Qazvinian et al. [39] studied the automatic detection of rumours from tweets. They dealt both with retrieval of rumour-related tweets, as well as with identification of whether the tweet author endorsed the rumour. In the first step, they categorised a tweet as a rumour or non-rumour, whereas in the second step they categorised those deemed rumours as the author of the tweet confirming it, or denying/doubting/questioning the veracity of the rumour. They used some manually defined queries to retrieve tweets that potentially

concerned rumours (e.g., "Obama & (muslim—islam)" for the rumour on whether Barack Obama is muslim). They developed a classifier using three different types of features: content-based, network-based, and Twitter-specific features. They found that content-based features led to the best classification performance both for the rumour vs non-rumour and for the rumour support vs denial/questioning classification.

Soni et al. [59] investigated how linguistic resources and extra-linguistic factors affect perceptions of the certainty of quoted information on Twitter. They collected tweets posted by 103 American journalists and bloggers, which were identified from lists of journalists on Muckrack.com[1] and selected quoted content from those journalists by filtering tweets with source-introducing predicates (e.g., claim, say, insist) listed by Saurí and Pustejovsky [54]. Then they used Amazon Mechanical Turk to annotate a subset of 1,265 tweets with quoted content from the journalists. The turkers rated the tweets in a 5-point likert scale from "Certainly False" to "Certainly True". Using regression techniques, they studied the correlation of features with a claim being true or false. The features they analysed include: (i) cue words, (ii) cue word groups, (iii) source quoting the content, (iv) journalists as the authors of the tweet, and (v) claims as the bag-of-words of the text in the tweet. They found that cue words used to introduce the claim did correlate with the factuality perceptions, but other extra-linguistic factors such as the source and the author were not relevant.

Zubiaga and Ji [72] relied on four aspects that determine how people perceive the veracity of a piece of information: (i) authority, (ii) plausibility and support, (iii) corroboration, and (iv) presentation. They conducted a study where users rated each of these four features for tweets and found that users mostly rely on author details to determine the veracity of a tweet, even though some author details such as location and description are not readily available on Twitter and third party clients' feeds. Additionally, they found that corroboration often misleads viewers into falling for a hoax, misunderstanding that the existence of many supporting claims does not necessarily mean a rumour is true, which matches up with previous findings in Psychology research for offline information verification.

These existing annotation schemes for rumours have their merits, but are not detailed enough for our purposes. We will consider them in the development of our scheme, incorporating new factors in order to drill down further into the nature and salient features of rumours.

4.2 Factuality and Sources

Saurí and Pustejovsky [53] described the annotation scheme as well as the process they followed to annotate the existing TimeBank corpus [38] with event factuality details. While TimeBank includes temporal and event information, FactBank adds a new layer

[1] http://muckrack.com/

providing information about the factuality of those events. Event factuality considers two dimensions, polarity – positive (+), negative (-), or underspecified (u) – and modality – certain (CT), probable (PR), possible (PS), and underspecified (U). The annotation was performed by two students, who were instructed to ignore any kind of real world knowledge and annotate the content of the sentences. This annotation led to an inter-annotator agreement (computed on 40% of the corpus) of $\kappa_{COHEN} = 0.81$. They found the annotation to be skewed towards cases that were certainly positive (CT+) and underspecified (Uu), which was not surprising as the corpus was made of news articles and these types of statements would be expected to predominate. In addition, the annotation scheme also includes the *events*, which are part of the original TimeBank corpus, the *sources* mentioned in the statements and *other sources that are relevant* to the statement, such as the text author.

De Marneffe et al. [10] collected annotations through Mechanical Turk for the FactBank corpus, which in this case referred to the veridicality of the sentences, defined as the perceived likelihood of a piece of information being true, informed by context and real world knowledge. The turkers achieved a lower inter-rater agreement ($\kappa = 0.53$) than Saurí and Pustejovsky [53] did with two annotators. They then built a maximum entropy classifier to automatically determine the veridicality of the sentences.

Vlachos and Riedel [64] described the creation of a corpus of fact checked statements. Using statements PolitiFacts' Truth-O-Meter[2] and the fact checking blog of Channel 4[3] as sources, they curated a set of statements annotated as True, MostlyTrue, HalfTrue, MostlyFalse, and False (the two sources employ different categorisations of truth, which were manually combined). They removed all statements that could not be corroborated with online sources. The corpus includes 106 statements at present, which will be made available online[4].

While the above annotations have been collected for news and political statements, which we could expect to be grammatically richer and more precise in terms of the factuality expressed, the annotation scheme could also be readily applicable to social media posts like tweets. It is likely that social media posts being grammatically less comprehensive would lead to more "underspecified" statements, which we will study in detail during the annotation process. Similarly, we expect that the source of a rumour in a tweet might not be as clear as in other texts such as news.

4.3 Author Types

As in other forms of communication, the identity of the person posting ("authoring') content on social media may have a bearing on how recipients assess its likely credibility.

[2]http://www.politifact.com/truth-o-meter/statements/

[3]http://blogs.channel4.com/factcheck/

[4]https://sites.google.com/site/andreasvlachos/resources

For example, where there is knowledge of the poster's previous trustworthiness, this will influence how new postings are assessed. Similarly, where the poster is understood to be acting in a professional capacity (e.g., as a journalist), then this (and the organisation they represent) may also influence how postings are assessed.

De Choudhury et al. [9] researched the development of an automatic classification system that identifies types of users on Twitter, which can be useful to differentiate them in the context of events. They introduced a categorisation of three types of users, which included organisations, journalists/media bloggers and ordinary individuals. They used vectors represented by the following features for the classification: number of followers and followees, number of tweets posted, the fraction of tweets that are replies, the presence/absence of named entities and the topical association of the user's history from a list of 18 topics. The named entities and topics were derived using OpenCalais[5]. They use a kNN classifier, which empirically performed better than 9 other classifiers that they tried. Experimenting with tweets associated with 8 different events, their classifier performed most accurately when categorising ordinary individuals, with slightly lower performance values for journalists and organisations.

In their study on the spread of rumours in the context of the 2011 England riots, Procter et al. [37] also introduced a typology of types of authors that posted the tweets. This typology included up to 20 types of authors, which defined a fine-grained categorisation, differentiating, for instance, ordinary individuals from rioters or from researchers. While this represents an exhaustive categorisation of users, it appears to be specifically crafted for riots and it might need to be revised to generalise it to other types of events.

Both of these annotation schemes for author types are of interest for our purposes when annotating authors in rumours. However, while the first might not be specific enough to consider all the author types that we might need to differentiate in the context of rumours, the second might need to group some of the types into higher level types to make it generalisable to a wider variety of event types.

4.4 Other Annotation Schemes for Conversations

There are other annotation schemes that also analyse conversational aspects of textual communication, but significantly differ from the purposes of PHEME of annotating rumours. For instance, some have made attempts to categorise types of dialogue that occur during argumentation. One such example is the categorisation made by Walton [65], which includes seven types of dialogues that were observed in cases of argumentation: (i) persuasion, (ii) inquiry, (iii) discovery, (iv) negotiation, (v) information-seeking, (vi) deliberation and (vii) eristic. While this categorisation also deals with conversational practices, it clearly differs from rumours. Even though some types such as information-seeking can also apply to rumours (here we define it as "*appeal for more information*"

[5] opencalais.com

to code the way a statement is presented), other types like negotiation are not straightforwardly applicable to rumours. Related to this, in our own annotation scheme, described below, we initially included a feature called *presentation*, which was intended to code for the type of dialogue.

Rittel et al. [45] looked at the use of topic modelling approaches for categorisation of tweets within conversations. They identify conversations from Twitter as sets of tweets responding to each other. They list 8 types of conversational messages for Twitter: status, question to followers, reference broadcast, question, reaction, comment, answer, and response. While this is an interesting typology of conversational messages observed in Twitter dialogues, it is rather generic and does not specifically tie in within the context of social media rumours. For our annotation scheme, we define a similar typology for the specific case of rumours discussed in social media.

Chapter 5

Developing an Annotation Scheme for Rumours

Having studied existing annotation schemes and their suitability for our purposes, we set out to develop a new annotation scheme adapted to the context of conversational threads around rumours in social media. This annotation scheme needs to be as generalisable as possible to different kinds of rumours that are discussed and disputed in social media, providing annotations that will enable the study of both linguistic aspects of the conversations, as well as sociological aspects that can be observed in the behaviour of participants.

To define this annotation scheme, we have followed an iterative process where it has been progressively tested and refined. First, we defined an initial annotation scheme that was based on the aforementioned schemes, which was then tested by assessing rumourous conversations extracted from Twitter. This testing brought to light a set of strengths and weaknesses in this initial scheme, which was then refined in a new version. This new scheme was then tested again to validate the changes. This chapter describes each of the steps of this process, showing the resulting annotation scheme, as well as our plans to put it into practice with the creation of the annotated corpora that will be used in the PHEME project for the study of social media rumours.

5.1 Initial Annotation Scheme

One of the key things to take into account when developing an annotation scheme in the context of Twitter is the unit to be annotated. While users on Twitter post short messages or so-called tweets, there are a number of ways to engage in conversations on this social media service. Hence, tweets can be seen as independent messages in some cases, but are themselves part of a conversation in other cases. Conversations on Twitter are observed in the form of replies from one user to another, retweets as the practice of resharing someone else's tweet, or a modified tweet as when a comment is appended to someone else's tweet.

Grosser degrees of topical relation may also sometimes be encapsulated within the use of hashtags, which is a keyword led by a hash sign (#) and is often meant as a categorisation of the tweet into the topic expressed by the keyword in question. When it comes to the analysis of rumours as they are spread and discussed in social media, we are interested in gathering higher level conversations that include tweets from multiple users interacting with one another and contributing to one another's postings, which can include a combination of the aforementioned ways of interacting. The organisation of conversation around topics, topical coherence and shifts of topic is another central focus of the conversation analytic literature, which we will study with the development of this annotation scheme. Consequently, the annotation scheme that we introduce below considers a set of factors to be annotated, sometimes the unit of annotation being single tweets that are part of a conversation and sometimes being the whole conversation.

This initial annotation scheme builds on the aforementioned annotation schemes when they are suitable for our purposes and also includes new features that we incorporate. Existing annotation schemes provide a set of features that are straightforwardly applicable in our setting, but there is a need to consider new features as required, given that no annotation scheme has been defined before for the specific case of social media rumours. We believe that all three types of annotation schemes described in the previous section, namely rumour type, factuality and author, play an important role in social media rumours and need to be considered in some way. We begin the presentation of our initial annotation scheme for social media rumours by identifying the three main parts that affect the life cycle of a rumour:

- **Message crafting:** the first step in the life cycle of a rumour is defined by how it is written and posted by the author.
- **Spread:** once the message is written and posted on social media by the author, others can interact with the message by resharing it to their network with actions like retweets.
- **Reactions:** the message posted by the author can also spark a set of replies that contribute to the original message with statements or comments that occasionally give support or debunk it.

Considering these three parts in the life cycle of a social media rumour, we define the features that characterise a statement and the subsequent conversation around it. For the purposes of the annotation scheme, we combine *spread* and *reactions* into a single group, namely *spread and reactions*, since these often appear together. Thus, the annotation scheme below is divided into two parts (i.e., message crafting, and spread and reactions) and includes a set of features to be annotated in each of these two parts. For the first step regarding *message crafting*, we define the following features to be annotated: (i) factuality, (ii) presentation, (iii) author, (iv) plausibility, and (v) evidentiality. Whereas for the *spread and reactions* that come afterwards, we rely on the following features: (i)

acceptability, and (ii) veracity. Next, we further describe and contextualise each of these dimensions, explain how they would be annotated for tweets and provide a list of values that each of the features can take. As a visual summary, Figure 5.1 depicts the structure of the resulting annotation scheme, showing the two parts of the life cycle, the features within each of them, as well as the possible values they can take.

5.1.1 Message Crafting

The first part of the annotation scheme analyses the message itself as it originated from the source, from how the author wrote and conveyed the information as it was posted on social media. We include five major factors in this first part: (i) factuality, (ii) presentation, (iii) author, (iv) plausibility, and (v) evidentiality. Next, we describe each of these and elaborate on how they can be annotated.

Factuality

As defined by Saurí and Pustejovsky [53], factuality defines whether a statement refers to an actual situation in the world. The authors define factuality as "the level of information expressing the commitment of relevant sources towards the factual nature of events mentioned in discourse". Factuality, hence, considers the linguistic structure of the message, coding its polarity and modality, and does not take into account any real world knowledge that can affect the perception of the recipient. Similarly, the factuality of a statement is a factor also involved in rumours and thus can be similarly coded for the original message of a rumour. Here, we include a new dimension to be annotated, besides polarity and modality, which is presentation. We believe that in the specific case of rumours, the way a statement is presented – and therefore how the information is being conveyed – plays an important role also in determining the factuality of a rumourous statement. The factuality is hence annotated by the following three dimensions:

- **Polarity:** The polarity defines if the message is conveyed as a positive or negative statement. It is different from the actual veracity of the statement, and the fact of the author supporting or denying a rumour. Instead, the polarity only defines if the sentence is syntactically positive or negative, or in the absence of utterances that state its polarity, underspecified.

  1. positive.
  2. negative.
  3. underspecified.

- **Modality:** The modality measures the degree of certainty expressed by the author when posting the rumour. The author can express different degrees of certainty

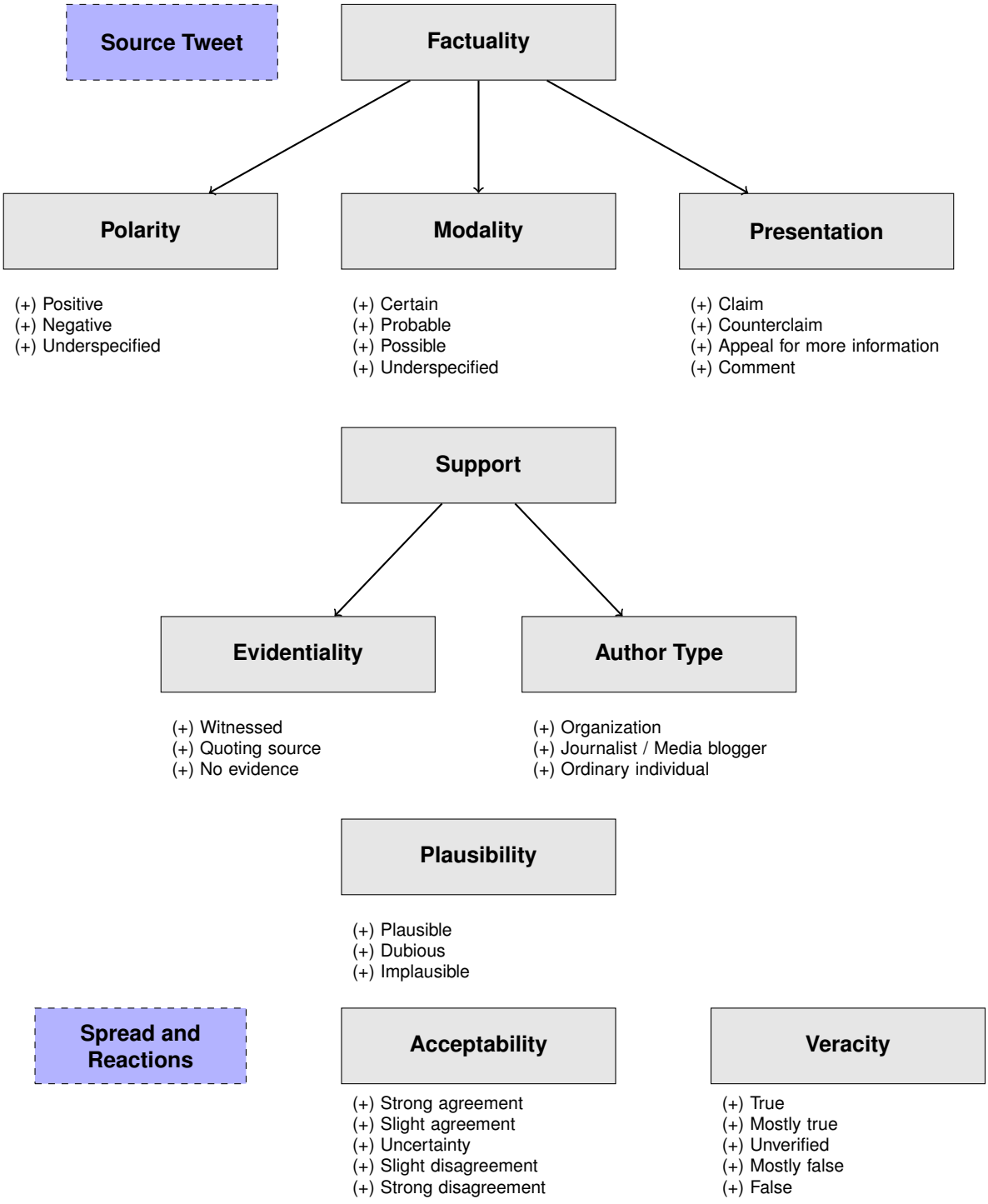

Figure 5.1: Annotation scheme for rumours

when posting a statement, from being 100% certain, to considering it as a possible occurrence. Note that the polarity has no effect in the annotation of modality, and thus it is coded regardless of the sentence being positively or negatively written. Likewise, the certainty expressed by the written language is coded here, regardless of the statement being plausible or not. Possible values that the modality can take include:

1. certain.
2. probable.
3. possible.
4. underspecified.

- **Presentation:** The presentation refers to they way that the statement is being conveyed, where the author expresses their position with respect to the rumour. This feature has been slightly adapted from Procter et al. [37], who called it "*rumour type*" and also included the fact of whether evidence was provided or not. Here we differentiate it from another feature "*evidentiality*" defined below, which is itself another type of annotation for our purposes as different types of evidence can be provided. We believe that "*presentation*" is tightly related to the "*modality*" introduced above and so they could probably be merged. Possible annotations for the presentation include:

  1. Claim.
  2. Counterclaim.
  3. Appeal for more information.
  4. Comment.

The factuality with which the author posts a message plays an important role in determining the confidence and conviction of the statement. This is also of utmost importance for ethnomethodological studies, where the certainty expressed by the author is crucial in the subsequent development of the conversation [7]. For instance, someone making a comment like 'I think', 'I believe', 'I understand', or 'I thought' might make a difference in how the others interpret it and trust or question the statement. Note that the factuality of a statement does not necessarily mean that it is more likely or true, but rather the degree to which the author is convinced of its veracity. How factuality relates to the veracity of statements is an interesting issue for study and could help understand the diffusion of rumours in social media, as well as be of help to improve a rumour classification system.

Support

The support defines how the statement is being backed up so as to assist the recipients in trusting the content and having access to the original source of the information when

applicable. Here we split the support given in a statement into two separate dimensions, one concerning the evidentiality provided to back up the statement itself and the other concerning the author posting the information.

- **Evidentiality:** The evidentiality determines whether and what kinds of evidence the author posting the statement has had access to that supports it. This is a new feature that we have not seen in previous research, but we consider key for our purposes. It can be annotated with the following values:

  1. witnessed.
  2. quoting source.
  3. no evidence.

- **Author Type:** The author that posts (and therefore backs up) the statement can be categorised into various types. Two works introduced typologies of authors that can also be relevant for our purposes. Procter et al. [37] listed 20 types of authors defined specifically for the 2011 England Riots, while De Choudhury et al. [9] provided a generalisable typology of three types of authors. Having the latter as a starting point (below), we will further expand it to come up with an equally generalisable but more detailed typology:

  1. Organisations.
  2. Journalists/media bloggers.
  3. Ordinary individuals.

We consider author and evidence to be crucial for determining how the author may have had access to the statement being claimed [14]. This will also allow us to study the importance of these factors when determining the veracity of a statement, as well as the reactions of social media users and whether they trust supporting statements or not.

Plausibility

The plausibility determines the extent to which a statement is seemingly valid, likely or acceptable considering common sense and knowledge about the real world. To define this feature, we rely on previous works from the psychological perspective [14, 72] who posit that plausibility is a strong factor that determines how a piece of information is perceived by its recipients, and subsequently supported and passed on. Possible annotation values for plausibility include:

1. plausible.
2. dubious.

3. not plausible.

An important aspect to consider when coding for the plausibility of a statement is the fact that knowledge about the world is not the same for everyone and hence different cultural backgrounds can have a significant impact on the outcome of this annotation [51]. We intend to collect the ratings from several annotators here in order to make assumptions that are as generalisable as possible and as unbiased as possible in terms of cultural backgrounds.

5.1.2 Spread and Reactions

The second part of the annotation scheme analyses the behaviour of the recipients of the message, including how they react to it as well as how they contribute to its diffusion. In contrast to the features introduced above regarding message crafting, the annotation of spread and reactions requires looking at the whole thread for each rumour, rather than doing the annotation at the tweet level. Here we consider two major factors that have an effect in the spread and reactions of a message: (i) acceptability and (ii) veracity. We describe and provide the annotation guidelines for these two factors next.

Acceptability

The recipients of a rumourous statement can reply to the author, occasionally providing additional evidence that supports or denies it, or leaning either in favour or against the statement, depending on its perceived truthfulness. To measure this, we define the acceptability as the extent to which Twitter users responding to the original author agree with the statement posted. Note that this is different from the plausibility of the message. While the plausibility aims to measure the extent to which the annotators believe that the story of the rumourous message is likely to have happened, acceptability will ask the annotators to rate the degree of agreement or disagreement of the actual Twitter users who responded.

This is a new feature introduced in this work and has not been used in previous work to the best of our knowledge. It would be annotated at the conversation level, rating the extent to which the responding users show a consensus towards agreement or disagreement, or that there are instead both agreeing and disagreeing responses. Annotators will be able to pick one of the following values to determine the degree of agreement of respondents:

1. strong agreement.
2. slight agreement.
3. uncertainty.
4. slight disagreement.

5. strong disagreement.

We can expect that plausibility will correlate with the subsequent acceptability of social media users, i.e., how plausible information will be deemed valid. However, it is worth studying if, as posited previously for other communication forms, the plausibility of a piece of information does not necessarily make it more likely to be passed on to friends, but rather that people tend to spread information they find implausible but funny or that evokes other kinds of emotions for the recipient, irrespective of the information's truth value [30]. Similarly, there is also evidence that suggests that social media users tend to share posts that involve socially deviant events, given that people tend to try to protect themselves from threatening events [56, 11]. The fact that some rumours involve threat to the reader in question might also affect the likelihood of certain rumours being passed on. The study of this form of human behavior as observed in social media in the context of rumours, makes plausibility a key factor to be considered and analysed in more detail.

Veracity

The veracity expresses whether the statement in question has been verified as true or false, or whether it still remains unverified. This is a feature that would be annotated for the whole thread as an aggregation of tweets, and sometimes needs time to analyse how it evolves and to see whether the veracity can be determined for a given rumour. In some cases, especially when the statement is very hard to verify or there is little evidence that the users involved in the discussion may have access to, the rumour can be annotated as "unverified". Instead, if the users seem to have found the actual truth of the story, it can be coded either true or false, when its veracity is clear, or mostly true or mostly false when not the whole story is true or false. To define the possible values that the veracity may take, we rely on the categories defined by Vlachos and Riedel [64] and Soni et al. [59], i.e., a 5-point likert scale ranging from true to false. Hence, possible annotation values for veracity include:

1. true.
2. mostly true.
3. unverified.
4. mostly false.
5. false.

5.2 Validation of the Annotation Scheme

Next, we define the process we followed to test the initial annotation scheme proposal and to identify potential weaknesses to be revised in a new version of the scheme. We first define the data we collected from Twitter and then explain the tests we conducted on that data.

5.2.1 Data Collection from Twitter

Identifying rumourous tweets to be collected from Twitter is a challenging task, as it is hard to characterise the content of a rumour in a such a way as to be able to retrieve it straightforwardly through an input query. An alternative way of collecting rumourous conversations on Twitter is to look at reactions to rumourous tweets. Those reactions might have certain characteristics that make it easier to identify. The original tweet can use any terms to post a rumourous tweet, without even being aware that it is an unverified rumor. However, a reaction in the form of a reply is more likely to have certain characteristics. Here we have relied on the methodology followed by Hannak et al. [21], who made use of the rumour database from Snopes.com[1]. This website puts together a set of stories that have gone viral online and discusses the veracity of these stories. The researchers then collected tweets pointing to a link for any of the rumours on this website. They focused on the tweets that were replying to others and corroborating, questioning or denying the original tweet's statement. This included about 1,300 tweets replying to other tweets.

We have therefore followed the data collection method based on snopes.com to retrieve sample rumourous conversations from Twitter. To do so, we performed the following steps:

1. We crawled the snopes.com site to list the links and underlying content for all rumours.
2. We used Topsy[2] to collect the tweet IDs for all the tweets pointing to one of the links listed in the step above. We call these "snoping tweets".
3. Having the tweet IDs of snoping tweets, we collected the content of the tweets through Twitter's API.
4. For each of the tweets collected in the previous step, we collected the related conversation. This was done in two steps:
   (a) Having the snoping tweet's content, we looked at the "in_reply_to_status_id_str" field in the JSON string of the tweet. When

[1] http://www.snopes.com/
[2] http://www.topsy.com/

the value of this field was not null (i.e., it was replying to someone else), we collected the content of the tweet being replied to by using the ID given in this field. Note that this tweet being replied to can also be a reply to another tweet, hence this step was performed iteratively until the field had a null value for the tweet collected.

(b) The first step collected tweets being replied to, or the "parent tweets". Unfortunately, Twitter neither provides a similar field in the JSON string for replying nor "child tweets" and there is no API endpoint to get those. Hence, in order to collect the child tweets we scraped each tweet's HTML web page, which gives access to a paged list of replying tweets for each tweet.

From all the conversations collected following this process, we sampled two cases that sparked a large number of replies, which we used as the examples to test, validate and refine the annotation scheme. These two conversations, as well as the underlying tweets, are in the Appendix B.

5.2.2 Annotation Test

The initial annotation scheme was then tested by two people, both of which have experience with Twitter. The annotation test was limited to the first 10 tweets in each of the two threads sampled in the data collection phase. Each of the annotators attempted to assign a value to each of the features in the annotation scheme for each of the 10 tweets. The meaning of each of the features included in the annotation scheme was clear to the annotators, as they had attended all the meetings, and had participated in the discussion and definition of the scheme.

The annotation test performed by the two annotators led to the results shown in Appendix C. There were slight annotator differences for some of the tweets, but the agreement was rather high in general. A meeting following the annotation test was held to discuss the experience of annotating the tweets with the defined scheme and to share the main issues that each of the annotators found. The two main points discussed at the meeting were the unit of annotation (i.e., do we need to annotate the whole conversation as a whole, or each responding tweet separately?) and the list of features to be annotated. We will go through the improvements proposed in this discussion in the next section where we also define the resulting annotation scheme.

5.3 Revised Annotation Scheme

The annotation test described above helped us identify both the strengths and the weaknesses of the preliminary annotation scheme. The test has helped us find out that some of the features are suitable as they are, others need to be combined as they were adding

redundant information and others need to be slightly redefined. Here we go through the changes that we have identified and suggest a revised annotation scheme that takes into account what we learned from the tests.

From the discussions after the first annotation tests, there was a strong agreement that the differentiation of two parts in a rumour – i.e., message crafting, and spread and reactions – makes sense. This is true especially when it comes to the source tweet, which is the one that introduces the rumour, and needs to be differentiated from the rest. However, there was a suggestion to redefine how the *spread and reactions* part was being treated. Instead of annotating the conversation as a whole, which was rather complicated due to the need of aggregating all the responses and providing a single annotation, the new proposal was to annotate each response tweet separately. Borrowing from conversational analysis and the concept of *adjacency pairs* and *turn taking*, we argue that each response tweet should be understood as a posting that is paired with some previous tweet – i.e., the former is either a retweet, a reply to or otherwise mentions the author of the latter. From this, we conclude that each tweet subsequent to the source tweet should be annotated in the context of the tweet to which it is paired – that is it is annotated for how it can be seen to stand in relation to that particular preceding tweet. Therefore, we relabel the two parts of the annotation scheme from *message crafting* and *spread and reactions* to the new labels defined as *source tweet* and *response tweets*.

For these two newly defined parts of the annotation scheme, we then defined the features that were deemed relevant and discarded the rest. We summarise next the features we ended up considering suitable for each of the two parts.

The resulting annotation scheme is shown in Figures 5.2 (for the annotation of the source tweet) and 5.3 (for the annotation of responses).

5.3.1 Source Tweet

The features previously defined for the source tweet were mostly deemed suitable during the annotation process by the two annotators. The main caveat is that they found in annotating the source tweet that the *presentation* did not add anything new with respect to what was already annotated with the polarity and modality features. This led to the conclusion that the *presentation* should therefore no longer be considered as a feature for annotating the source tweet. The features that were ultimately considered valid to annotate the source tweet were the following:

- **Polarity:** positive, negative, underspecified.
- **Modality:** certain, probable, possible, underspecified.
- **Evidentiality:** witnessed, quoting source, none.
- **Author Type:** organisation, journalist / media blogger, ordinary individual.

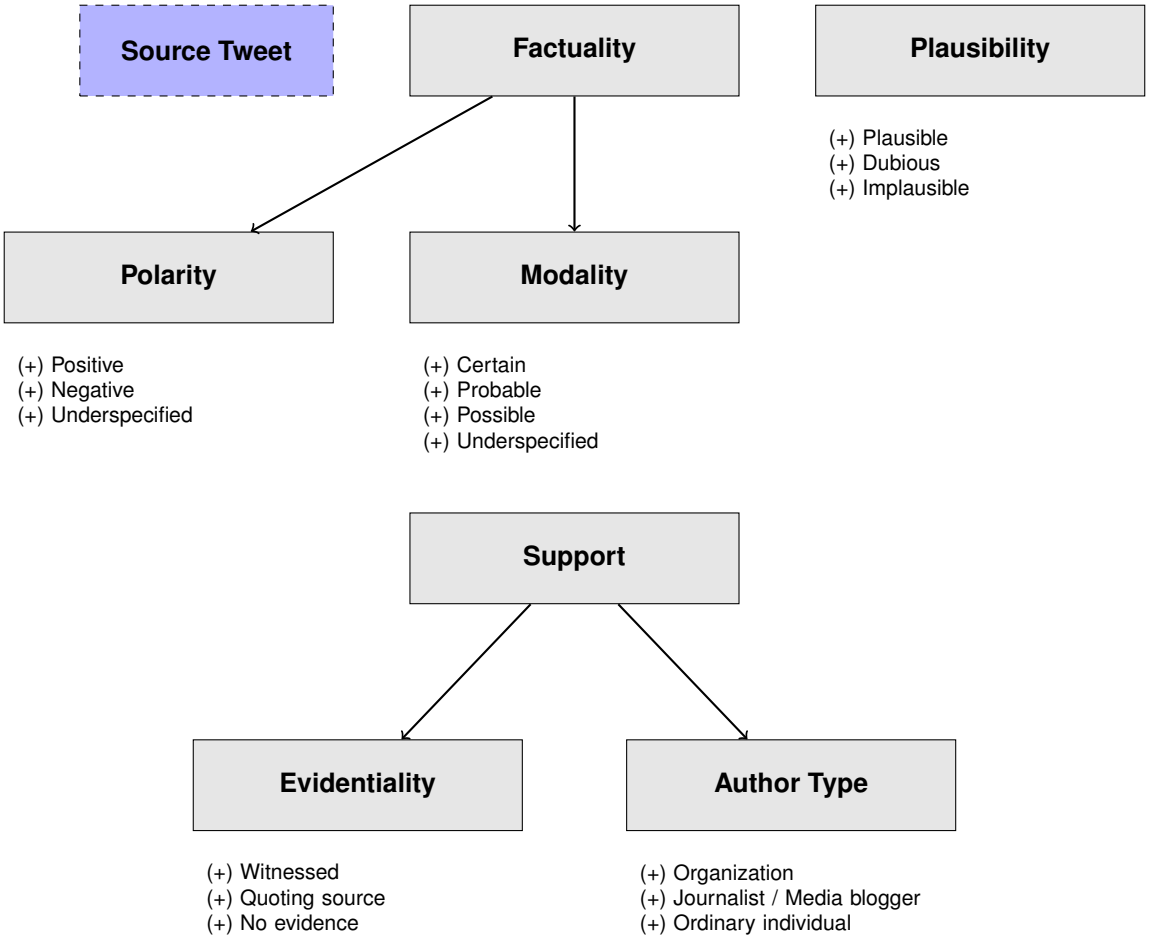

Figure 5.2: Annotation scheme for source tweets that initiate rumors

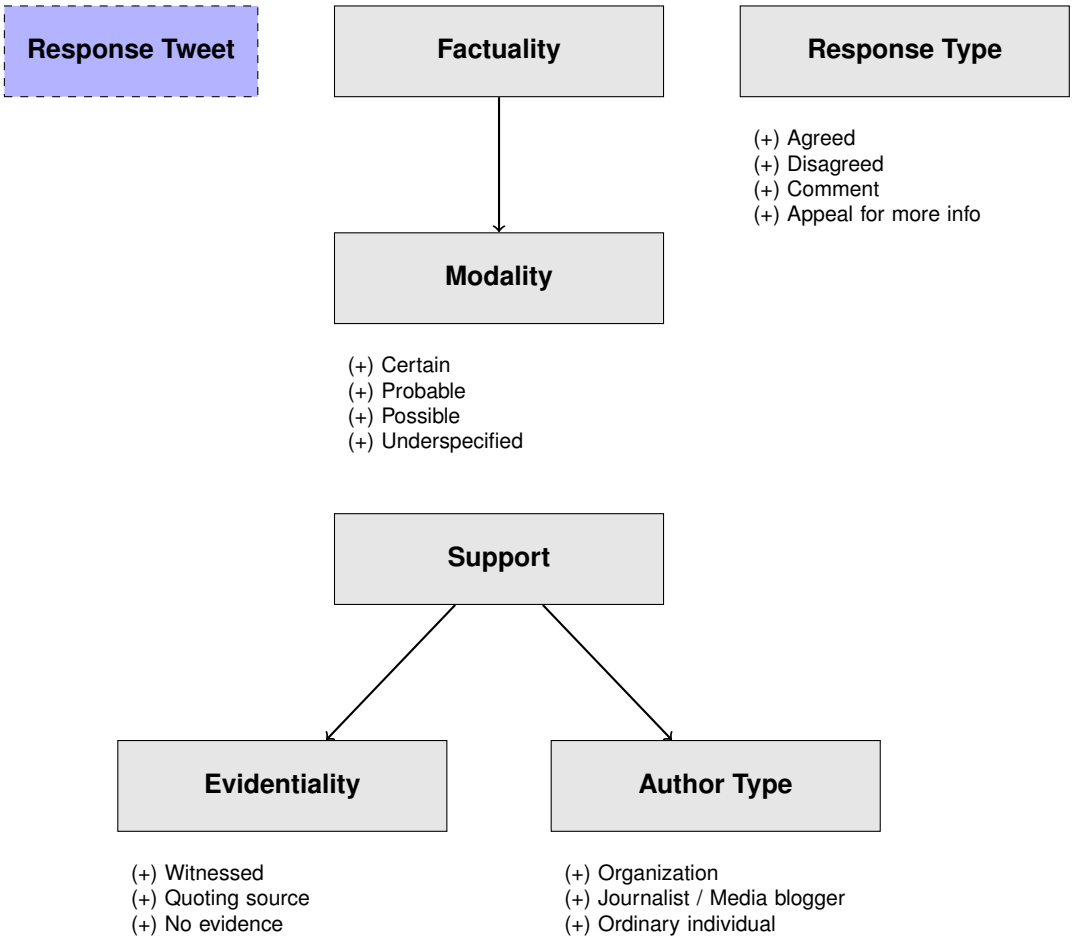

Figure 5.3: Annotation scheme for tweets responding to the initial rumourous tweets, as well as subsequent responses

- **Plausibility:** plausible, dubious, implausible, unclear.

While most of the features above remain the same as defined previously in Section 5.1, there was discussion regarding polarity. The polarity originally coded for the linguistic composition of a sentence, i.e., whether a not, none, or similar is changing the positiveness of the sentence. However, this is not very useful when the goal is to identify the veracity of a rumour. Instead, we code for the polarity as the position of the author with respect to the rumourous statement. The polarity will be positive when they are backing it up, while it will be negative when they are debunking it.

5.3.2 Response Tweets

A bigger number of changes was identified as necessary for the response tweets, especially as the unit of annotation now changes to each single tweet rather than the whole conversation in an aggregated way. One of the features that was clearly identified as one that needs to be taken out is the *plausibility*. The annotators agreed that the *plausibility* is an intrinsic characteristic of the rumour originating source tweet, as it describes the overall credibility of the rumour itself. Presentation and acceptability were found to be tightly related and mostly overlapping, and therefore they were combined into a new single feature, namely the *response type*. The *response type* defines how a response is addressing the statement in the posting with which it is paired, either agreeing or disagreeing with it, or making an alternative type of comment. The polarity was also removed to avoid confusion between the polarity of a rumour originating tweet and the polarity of responses to the rumour, which are better reflected by the response type.

- **Modality:** certain, probable, possible, underspecified.
- **Evidentiality:** witnessed, quoting source, none.
- **Author Type:** e.g., organisation, journalist / media blogger, ordinary individual.
- **Response Type:** agreed, disagreed, comment, appeal for more information.

5.4 Validation of the Revised Annotation Scheme

Annotation using the revised scheme now involves annotation of the source tweet as a single unit and thereafter annotation of pairs consisting of either adjacent response tweets (in the context of the source tweet) or response tweet and source tweet.

A similar process of validation was conducted again with the newly revised annotation scheme. This time two annotators relied on the new annotation scheme to annotate the

same rumours. The same process was followed to do the annotations, using the revised scheme. The resulting annotations are shown in Appendix D.

The annotation test on the revised annotation scheme led to a higher degree of agreement between the annotators. The annotators also confirmed that the revised annotation scheme fitted better with the requirements, and they felt more comfortable coding according to the new scheme. While this annotation scheme has fitted well with the characteristics of the sample rumours under study, our ongoing work now is dealing with the analysis of additional types of rumours to validate the applicability of the annotation scheme to other situations such as rumours that emerge with breaking news. In the next section we will outline the main issues that we are dealing with on ongoing work.

5.5 Work in Progress

While the development of the annotation scheme through the iterative process described above has enabled us to start stabilising its structure, we are now looking at some of the remaining details to tweak the scheme so as to resolve a few minor issues.

The main aspect to look at is how to reliably determine Twitter threads and how to present them to annotators, so they can follow the conversation while still keeping the the annotation relatively simple. It is crucial that each annotator annotates a complete thread, so we make sure they follow the entire conversation. However, showing the whole conversation at a time would complicate the task. An alternative that we are considering is to show the current tweet as well as the previous one being responded to, keeping always the original source tweet visible. This would be showing 3 tweets at a time to the annotators – i.e., current tweet, parent tweet responded to, source tweet – and moving the current tweet to deeper levels of the conversations as the annotations are done. One of the key tasks to do next will be to run some tests assessing this way of annotating tweets.

Another aspect to look at is how we deal with topic coherence and topic changes. Conversations can become quite long on Twitter, involving topic changes to side topics or even unrelated topics. We have seen examples where the conversation flows initially focused discussing about the rumour, but then switches to a more general topic. Similarly, a conversation could switch to a completely different topic. We are currently studying these cases and putting together alternative solutions to deal with its annotation. The final annotations scheme will include clear guidelines with regard to this aspect. A related issue is the reliable identification of tweet adjacency pairs. If a new tweet is a retweet or a reply to some previous tweet, then the new tweet's metadata will include the ID of the latter tweet. We will use this to identify candidate adjacency pairs in tweet threads. If this field is null, then the inclusion of a mention of another user will provide us with a default adjacency pair identification mechanism.

One of the caveats identified by the annotators during the second annotation test was the need for additional possible values when coding for evidentiality. While this feature

currently allows for annotation of three possible values – i.e., witnessed, quoting source, none –, other values might be necessary to further nail down the evidentiality for a tweet. Examples of additional values identified in this annotation include:

- **Reasoning:** there is no source quoted in the tweet, but the author provides their own evidence to support the statement, e.g., *I think it's true because....*
- **Quoting unverifiable source:** the author may sometimes refer to external information sources, although without specifying the source (e.g., *I read somewhere before that...*), or providing a source that is hard to have access to (e.g., a friend of mine said that...).
- **Quoting source (URL):** the author is not only quoting the source of the information, but is also giving access to it through a URL.

Similarly, the source being quoted in a tweet needs expansion to distinguish the reliability of the evidence. This could be provided by coding for the reliability of the source of *quoting source* that is selected for *evidentiality*.

This second annotation test with the revised scheme has therefore proved the structure and features suitable. The following iterations with the annotation scheme will deal with the revision of lists of values that each of the features can take, especially by looking at rumours of different types which are also spread and discussed on social media. In this continuing work, we will explore the extent to which we might apply conversational analysis and ethnomethodology as outlined in Chapter 6 below to elaborate on ways of annotating for e.g., adjacency pairs amd topic coherence.

5.5.1 Putting the Annotation Scheme into Practice

Once we come up with a tested and refined annotation scheme that enables the annotation of rumourous conversations publicly discussed in social media, we will pursue the collection of a corpus that fits with the needs of the numerous work packages of the PHEME project that will make use of it. This corpus will include Twitter conversations that begin within the context of a rumour, and will contain not only the tweets themselves, but also other Twitter metadata useful for the task such as author details, as well as the content from external links (e.g., web pages or pictures posted in tweets).

This corpus will then be annotated through a crowdsourcing platform such as Amazon Mechanical Turk[3]. To do this, first we will carefully define the guidelines that will be given to the participants of the crowdsourcing platform, also known as workers. These guidelines will need to be easy to understand by anyone, and so we will have to make sure first that they do not lead to confusion. In a preliminary step, we will run small

[3]http://www.mturk.com/

experiments with the initial guidelines, from which we will assess the interpretation and performance of workers. The guidelines will be refined as needed in an iterative way, and we will then ask the workers to annotate larger amounts of tweets once the guidelines are well defined and clear. From these larger annotation sets, where each tweet will be annotated by several workers, we will aggregate the annotations to put together the final corpus.

Chapter 6

Extending the Annotation Scheme

As already indicated in the earlier sections of this document, our goal is to build upon the annotation scheme outlined in Chapter 5 by exploiting existing work in the areas of conversation analysis and ethnomethodology. This approach offers us with the possibility of: a) uncovering richer ways to annotate how Twitter feeds and rumours around particular topics unfold across extended threads and sequences of action; and b) grounding annotations in a way that will enhance their capacity to capture features of people's own situated reasoning. In this section we look a little more closely at what pursuing this approach to grounding annotations might look like. Central to the proposition is the notion that streams of Twitter feeds around the same topic can be conceptualized in some way as conversations. However, as one begins to explore the ways in which such a suggestion might be justified, one begins to also realise that there are ways in which tweeting stands as independent phenomenon that needs to be understood on its own terms. Thus we shall be arguing that, whilst conversation analytic approaches will serve well as a point of departure, differences between spoken conversation and the organisational character of tweet-based interaction will ultimately make it necessary to respecify the approach as 'microblog analysis' in order to steer around the potential dangers of missing the lived character of how people reason about tweeting as an activity in its own right.

6.1 Moving towards annotation grounded in microblog analysis

Over the course of this section we will be looking at the conceptualization of tweets as conversations a little more closely and exploring the ways in which similarities do exist and the ways in which tweeting may be seen to present discrete phenomena that cannot easily be subsumed within conventional approaches to conversation analysis.

6.1.1 The turn-taking mechanism

At the very heart of conversation analysis, as laid out by Sacks et al. [52], is the observation that talk is organised such that only one speaker speaks at once. This is seen as a fundamental premise of social order because any other system would frequently render talk completely ineffectual. On the basis of this, and probing just how it could be that this is systematically provided for in interaction, Sacks et al. elaborated what they called the 'turn-taking mechanism'. It contains some primary features that together serve to underpin most other kinds of conversational phenomena. So there are: speakers (recognizable individuals who produce utterances); speakers who talk first, and other speakers who may also talk as a conversation unfolds; mechanisms whereby a current speaker may select who talks next; and mechanisms whereby speakers may select themselves to be the next person to produce an utterance. Despite its asynchronous character and the potential interleaving of a number of distinct sequences of tweets on Twitter there are ways in which this kind of mechanism can be seen to hold. Tweets are composed and arrive as distinct units within global Twitter feeds. With regard to any one particular topic there is a 'first speaker' in terms of there being an originator, there are subsequent parties who may be implicated as respondents within the original tweet, and there are parties who select themselves as respondents to a tweet in some way. Differences here particularly relate to other matters such as: 'co-placement', where responses to a specific tweet may not be sequentially directly adjacent to that tweet within a feed (because, in principal, all comers may respond to all tweets, so next up in a feed may be an entirely unrelated response to a different topic); and 'rights of response' in that any recipient of a tweet may respond to it or retweet it, whilst this is clearly not the case in face-to-face conversation, where just who gets to speak is a very tightly managed affair. Having said all this, however, there are ways in which some tweets clearly implicate other tweets, so extended kinds of sequential relationships should be open to being tracked.

6.1.2 Topic

A further temporal consequence of how Twitter is organised is that the time spans over which respondents may address themselves to a topic without loss of coherence are much greater in the case of Twitter than they are in face-to-face conversation. In ordinary conversation, as most speakers will readily recognise, failure to address oneself to a topic quickly enough means that another topic will be floored and addressing oneself to the original topic becomes much more difficult and accountable. Conversation analysis has looked closely at how 'change of topic markers' are handled in conversation. Part of this also relates to 'return to topic markers' such as 'but as I was saying...', 'but going back to what you were saying earlier about...', and so on. Thus, there are ways of managing topic preservation over more extended periods in spoken conversation. The temporal organisation of Twitter makes it likely that it will have certain distinct but equally systematic ways of marking out topic relationships (re-tweeting being one obvious one) that people

will use in various sophisticated ways in order to manage coherence across more extended conversational threads.

The organisation of conversation around topics, topical coherence, and shifts of topic is a central focus of the conversation analytic literature and understanding topic-based relationships may prove to be an important part of tracking the flow of rumour-type phenomena across large bodies of tweets. Clearly responding to other people's tweets, commenting upon embedded tweets being retweeted, and simple retweets all exhibit certain features of topical coherence, and Twitter itself also reflects this understanding in its grouping together of connected tweets in this way as 'conversations'. Grosser degrees of topical relation may also sometimes be encapsulated within the use of hashtags. There may be more subtle indicators of 'on topic' / 'off topic' that can be uncovered through a systematic examination of how tweeting is organised as an interactional phenomenon.

6.1.3 The organization of conversation as applied to tweets and the organization of tweets when seen as conversations

In order to make full use of the extant conversation analytic literature one of the longer-term activities we aim to undertake is to work through the principal organisational devices in play in conversation that conversation analysis has identified over the years, to explore how these devices might or might not be present in tweet-based phenomena in various ways, and to examine the extent to which they are organised in a similar fashion or otherwise. Such devices can be seen to include: adjacency pairs; change-of-state marking; correction-invitation devices; formulation; membership categorization devices; prefacing; premising; pre-sequences; receipt tokens; recipient design; repair procedures; sequential objects; speaker selection techniques; topic marking; and so on. Each of these areas of interest has a large body of literature already devoted to it. Some of the areas more evidently related to the concerns of the PHEME project have already been discussed above because of their foundational character (i.e. speaker selection and topic management). A number of others have potential relevance for the current annotation scheme and may therefore reward further investigation. Where relevant to the existing annotation scheme these are grouped under related headings, otherwise they can be seen to constitute ways in which the annotation scheme may subsequently be extended.

Factuality (Presentation/Claim)

Ambiguity: Whilst it is possible for a range of utterances to be taken as ambiguous with regard to their meaning, some research in conversation analysis also points to ways in which utterances can be ambiguous by design. This may have relevance with regard to how the factuality of claims is first presented in tweets with it being deliberately the case that people might take what is being claimed in several different ways. As an example, Sacks (1992) comments on the use of the word 'you' where it could equally mean 'one' or

it could mean 'they, e.g.': "If you're hotrodding you're bound to get caught". In this case the 'you' could be directed at the specific recipient, it could be directed at another body of people, or it could be a more general observation upon the outcomes of hotrodding. Even in the context of its production it was not clear and the recipient was obliged to settle on one understanding. With regard to rumours it should be noted that this kind of method stands as a) a resource for 'getting hold of the wrong end of the stick', but also b) a reasonable account for claiming 'they got hold of the wrong end of the stick' so that, if things turn out badly, it can be claimed that your utterance was taken the wrong way.

Framing, prefacing, premising markers: Another part of the conversation analytic literature relevant to the status of claims as presented refers to the ways in which different utterances may get framed or prefaced in order to inform recipients how to understand the speaker's orientation to what they are about to say. Many of these relate to matters of certainty, for instance 'I believe that . . . ' (see Coulter [7]), 'I think. . . ', 'I thought. . . ', 'I understand that. . . ', 'it would seem that. . . ', and so on. Utterances that may shape up to be rumours can include these kinds of prefacing words or remarks, e.g. '*Apparently the rioters are moving towards Birmingham Children's hospital*'. It is important to note that these are not just about the certainty or otherwise of a speaker producing them, but also about providing for how the speaker might be called to account for what they say.

Evidence & Inference: So one thing that is pointed to is that there is a range of methods whereby the grounds of claims are made visible, where inference is supported or resisted according to need, and where the very need for evidence is set aside. Benson & Hughes [4], for instance, explore how the work of variable analysis trades upon a range of ordinary competences and commonsense assumptions and how the recognisable adequacy of statistics as evidence trades upon these things. This can be seen to extend to ordinary everyday interactions where to produce a statistic is commonsensically seen to be providing a certain kind of claim regarding the credibility of what is said.

Evidentiality

In section 5 one of the features of the annotation scheme is evidentiality. This refers to the degree to which evidence is provided within a tweet to support the claims or propositions being made. The conversation analytic literature has also addressed itself to this kind of phenomenon and how speakers might go about producing utterances in such a way as to not be called to account for them being dubious in some sense. It explores the matter in a variety of ways:

Sacks [51], in a discussion regarding the distinction between claiming and demonstrating in conversation, looks in particular at the work that can be done by second stories. As we discuss again below with regard to motive power, a commonplace phenomenon is that when one person tells a story another person will follow it up with a similar story of some kind. If a first story is simply followed by 'I know just what you mean' or 'I agree' and nothing more this amounts to only being a claim that you are aligned with the speaker in

some sense. Telling a second story that exhibits the same point from your own experience serves to actually demonstrate your concurrence. So there are methods for making clear that you are doing more than just claiming alignment that are oriented to as acceptable ways of doing that. In rumour production, then, second story production is one way in which speakers may demonstrate whether they attach credibility to the rumour in some sense.

In another discussion about the character of story production Sacks (op cit) makes an observation that is in some ways the counterpart of the observations above regarding ambiguity, which amounts to saying that 'brevity invites inference'. Sacks' point was that where speakers are concerned that a story may lead to the wrong kinds of inferences being made they will often elaborate the story quite significantly to ward off potentially awkward judgments (the specific case discussed related to potential assumptions regarding a man's sexual preferences). This is relevant to rumour production in a variety of ways. For instance non-specificity can deliberately encourage speculation. And one way of trying to encourage acceptance can be through the production of detail.

Another relevant discussion in conversation analysis relates to how people use certain kinds of stock phrases such as "*everyone does that don't they*" as a means of setting aside all further need for account. Proverbs can also be seen to be used in the same kind of way e.g. "better the devil you know". Once again, as a response to rumours, such phrases can be seen to be doing quite definite kinds of alignment work. In particular, once produced within a stream, they make it such that to contest now is not just contestation of the rumour but also contestation of a stock body of knowledge that has been applied, something that is much harder to do. Indeed, conversation analysts have catalogued quite a range of instances where something is made to be self-evident by its association with a particular thing that just anybody knows. In other words a routine way of promoting acceptance is to work something up as being just another case of what everybody knows.

Warrants: Related but in some ways distinct to the preceding discussion but nonetheless still bound up with matters of evidentiality is the matter of warrants or rights to be able to claim certain things in certain kinds of ways such that what is said is taken for granted to be true. Discussions here lead to something we shall be discussing in greater detail below which is that just how people are categorised in talk already sets up a bunch of assumptions regarding what might be reasonably claimed about their actions (and thus never called to account in any way). Sacks [51], for instance, discusses a report of an incident where part of the report is that the sister calls the police. He points out that within the report and the response to the report the nature of the sister (is she elderly? is she prone to hysterics?) is never put into question. Part of the nature of categorisation of people is that it provides for warrantable action, e.g. as we shall be seeing below, Hell's Angels rape young girls and Hotrodders like to drive fast cars. Sacks makes the strong claim here that "*a task of socialization is to produce somebody who so behaves that those categories are enough to know something about him*". However, he also points out that these kinds of assumptions are overturn-able as assumptions by other rights of precedence, e.g. witness status or local knowledge. With regard to rumour a case in point

here is the following extract from the London riots tweets and the rumour that rioters were attacking a children's hospital and that police were massing to protect it: *May I remind clueless/hysterical birminghamriots commentators that Children's Hospital sits face-face with city's central police station.*

With regard to all of the matters we have discussed above an important element to hang on to here is that people methodically build into their utterances from the word go ways in which they might or might not be held accountable for the production of those utterances. So at least one part of the work of unpicking matters of evidentiality and plausibility and acceptability and veracity in sequences of tweets is to look at how people are systematically managing their accountability in the production of those tweets.

Plausibility

When it comes to matters of both plausibility and acceptability there are once again a variety of conversation analytic treatments worthy of further inspection for how they may assist in identifying aspects meriting annotation.

Lying: There are a number of discussions in the literature regarding lying. The central outcome of analysis here is that there are routine grounds upon which the prospective character of something as a 'lie' may be established. Sacks [51], when discussing the production of competence in the telling of a story, looks at a report of a car wreck to observe how the tellers make it evident that they have the competence to be reliable witnesses of car-wrecks such as 'we were stopped there for 25 minutes' and 'the car was smashed into such a small space'. Sacks' point is that people have a sense of what's usual for the report in play. Thus, stepping outside of that can prompt the questioning of its truthfulness e.g. 'we were stopped there for just a second'.

Subjectivity & Objectivity: Another related matter here is how people work with, on the one hand, what just anyone knows of the world, and on the other with what only certain people in certain positions might know of the world. Much of the conversation analytic literature points out that lots of tellings trade upon what just anybody knows of the world such that the claims made might, as a routine supposition, be seen to have an objective character until such a time as it might be there are grounds for thinking otherwise. Tightly bound up with this is Sacks' discussion of 'Doing Being Ordinary' [50]. His observation is that, for any activity, there is a presumed ordinariness about what is going on. People make commonsense assumptions about what the ordinary business of any state of affairs might be and only pause to remark upon things that fall outside of that. The implication of this is that there are ordinary ways of having riots, the same as anything else. Ordinary expectations about riots would include things like places being set on fire, guns being shot, policemen beating people, such that images of such things would not necessarily invite inspection. Thus the scope for spread of a rumour and the chances of it being called out trades upon there being background expectations in play such that the things being proposed fall within the scope of being the ordinary business of stuff like that. And it is

exactly when, for instance, an image falls outside of such background expectations that it is subject to remark, open to inspection and potentially rendered in need of an account.

Acceptability

Trust: A further elaboration regarding the points we have made so far is that the accountability of people also typically comes with notions of trust and rights and responsibilities built in. Of particular moment here is the matter of reporting to known others versus overhearing and getting stories from unknown others. So, for instance, saying to someone 'Well Sammy told me such and such', where the other party also knows who Sammy is, provides systematically for: the accountability of the speaker to Sammy and to the person they're talking to; for the accountability of Sammy for just what he said to these parties; and for the receiving party as well as to just what they might then be moved to report. However, overhearing falls outside of these routine arrangements of accountability and trust. So, an overhearing party can just report the such-and-such that was overheard without the need to make those accountabilities visible. They are only required to provide provenance if explicitly called to account. With regard to rumours on Twitter note that, for many Twitter retweets, people are passing on tales from unknown others so they already stand outside the routine arrangements of trust and accountability.

Membership Categorisation Devices (MCDs)

This refers to a strong orientation people display towards hearing certain things that might be heard as going together as indeed going together. The phenomenon was first described by Sacks [47] as a feature of the analysis of stories told by children. He pointed to the strong tendency of native English speakers to hear the utterance "The baby cried, the mommy picked it up" in such a way as to understand that it is the mother of the baby who picks it up, even though this is not actually specified. He elaborated upon a range of membership categorisation devices together with a set of tying rules (not actually 'rules' in fact but rather maxims) that provide for how people hear things as going together. In another discussion of MCDs Sacks (1979) discussed how different categorisations of exactly the same people might be used to do moral work. Thus teenagers might refer to one another as 'Hotrodders' (with certain 'cool' connotations), whilst adults might refer to them as 'kids in cars'. This then provides for taking quite different positions regarding the matter of driving fast. Slightly later work on MCDs has often focused upon examples closer to Sacks' Hotrodders where there is a deliberate use of 'morally contrastive categories'. Lee [29], for instance, in a paper entitled Innocent Victims and Evil-Doers, discussed in detail the newspaper headline "*Girl Guide Aged 14 Raped at Hell's Angels Convention*". Here the categorisations deliberately provide for seeing the parties involved in highly distinct ways. In the context of rumours it is likely that the latter kind of MCDS, drawing upon morally contrastive categories, are more likely to prove fruitful for inspecting how both the crafting and spread takes place. In particular the interest may be in how these provide

for naturally presumptive work such that certain kinds of tweets may go unchallenged, e,g, for the London Riots data '*Rioters set Miss Selfridges on Fire*' is altogether less remarkable and open to inspection than something like '*Grandmother sets Miss Selfridges on Fire*' would be; and amongst certain communities '*Police beat a 16-year-old girl*' is potentially more credible than something like '*Councillors beat a 16-year-old girl*' might be.

Sequential Ordering

Another aspect of conversation analysis looks specifically at how the positioning of utterances in relation to one another can serve to inform specifically the ways in which they are taken to be meaningful. Sacks [48] engages in an analysis of the telling of a dirty joke in order to illustrate this. He works through how the assembly of a set of potentially un-related utterances into a specific sequence can invite a certain reading where to get the joke is to see that reading and find it funny. This may be relevant for work on rumours in terms of how the crafting of specific messages may be taken to be implicative and also in terms of how to assess different kinds of response, e.g. (from the London Riots data): '*Apparently McDonalds stormed in tottenham. Rioters proceeded to take over and cook some burger 'n fries. Ya can tell it's school holidays*'.

Reportability & Motive Power

One potentially important aspect of conversation analytic investigations regarding how people manage topics in conversations is the matter of how topics can get presented in the first place and, in particular, the notion of 'first topic status' [51] and how certain topics may count as 'news'. Conversation analysis points to how certain topics that are somehow remarkable or worthy of note provides people with the special licence for comment and retelling without the topic having been already implied by something else. This raises the question as to what counts as mundane or remarkable in what kinds of situations with regard to different kinds of social media – especially where there is a clear licence to report the otherwise mundane in certain ways.

What may be of especial concern here is what Sacks and certain other conversation analysts term 'motive power'. This rides on the observation that for most kinds of topic-raising some kind of account is routinely required. The account may often be self-evident because of other surrounding events or preceding utterances. However, some kinds of accounts are generative in their own right. Motive power refers to the extent to which stories and accounts are open to transmission to other people. One of the matters that impacts upon motive power is what Sacks [48] terms 'investment'. Investment refers to the degree to which relationships with people carry with them certain rights and obligations. So complete strangers show very little investment in one another, work colleagues may exhibit an interest in your health or where you are going for your next holiday but are unlikely to ask detailed questions about your love life, whilst daughters of a certain

age are expected to report most things to their mothers but not necessarily vice versa, and husbands and wives are expected to tell one another pretty well everything. The upshot of this is that the number of people to whom you can report having met someone you haven't seen for a while on the way to the shops is very limited, whilst having seen a building on fire is much more widely reportable, and there are certain people who must be told certain things or trouble will surely follow, e.g., telling your mother you're getting married.

Another feature of motive power is what Sacks called 'entitlement to experience'. His observation here was that stories and jokes etc have high motive power according to the extent to which they convey experience. This is especially about the conveyance of experience that is out of the way and not otherwise available to you because you can figure the sheer remarkability of it is a thing that will make it self-evidently appropriate to report it. You are entitled to share it and other people are entitled to hear about it, which is not, of course, the case with just any experience you may wish to relate. A secondary phenomenon that relates to this that is also of interest is the commonplace expectation that a telling of a story will prompt the telling of a second story in return by the recipients. This second story is routinely understood to need to be a telling of something similar that either happened to you or that you once heard tell of. It is also a primary way in which conversationalists demonstrate alignment with one another in their views upon different topics.

Matters of reportability, motive power, re-tellings and alignment are all of significance for the spread of rumours. Some things that might be rumoured are clearly unlikely to carry relevance for anyone outside of highly constrained cohort of people (thus the potential distinction offered above regarding 'gossip'). However, other rumours convey matters that are tellable to a much broader set of people. What analysis here may provide is a sense of what stories 'have legs' so to speak, and just what kinds of features within stories different kinds of tellability may turn upon.

6.1.4 The intersubjective constitution of tweeting as a phenomenon

One of the fundamental insights coming out of both conversation analytic and ethnomethodological approaches is the way in which any body of social accomplishments is an intersubjectively constituted set of accomplishments. These are reflexively organised around the specific understandings of the parties to those accomplishments of just what it is they are in the business of accomplishing. Furthermore, any specific feature is indexical of those mutual understandings in play. This may seem rather densely expressed but what falls out of it is that, to understand what is being done with any one particular utterance (or other kind of action) by one party, you need only look to the immediately subsequent utterance (or action) by the next interactant to see what kind of an action the preceding utterance has been understood to be. And, where misunderstandings occur (which, of course, they do) one need only look on to the utterance after that to see the original party engaging in some kind of repair. Thus interactants involved in a course of action

routinely make available to others, who have the competence to see it, just exactly what is going on. Both CA and ethnomethodology trade in bringing this local reasoning into view. Thus they are occasionally called 'postanalytic' enterprises because they primarily work to make more explicit analysis that has already taken place on the part of those who originally produced the phenomena they are examining. The implication of this (and significant challenge) is that annotation schemes truly aligned with conversation analytic and ethnomethodological approaches would seek not to tag text with externally derived analytic categories but would rather seek to identify the ways in which any specific tweet (or comparable phenomenon) has been analysed by members themselves in directly subsequent tweets in order to tag it appropriately. In particular, a focus upon clusters of 2 or 3 inter-related tweets is likely to be fruitful: initial tweet, responding tweet, subsequent tweet by originator (if there is one). This is a feature already being exploited by the annotation scheme we have devised, but will require further work to unravel the different kinds of actions related tweets may be seen to be.

6.1.5 Following and followers

Something that falls out of the preceding observations is that it is going to be important to understand properly the subtle mechanics of following/follower relations on Twitter so that just how their respective activities are aligned with one another and implicative for one another can be properly explicated. In particular, drawing upon observations first made in section 6.1.1, we need to note and be able to properly handle the fact that there are, variously: i) equal part conversations between parties who are mutually following one another, but also ii) audiences of interchange, who follow but are not followed, but who can nonetheless both comment upon witnessed exchanges and re-circulate those exchanges amongst their own community of followers, and, additionally, iii) subtle understandings in play of just who is following you, who your actions might be visible to, and how you might or might not be accountable to those parties in various ways.

6.1.6 Tweeting as a mode of communication

As one works through the preceding body of materials something that becomes important to recognise is that tweeting is its own form of communication. It is not really conversation as in the sense of the classic forms of dyadic conversation that are the primary focus of conversation analysis. Nor is it good policy to simply assume that tweeting is just a specialised variant of traditional conversation in some way. Rather tweeting (or microblogging to use a slightly more formal term) should, in the first instance, be examined as a phenomenon in its own right with its own orderly characteristics that may or may not prove to be tightly aligned with other kinds of communicative practices. Thus the safest approach is to take the corpus of findings coming out of CA as a starting point for reflection because conversation is a relatively well-described phenomenon and tweeting

is not, rather than simply assuming that tweeting will operate in much the same way.

In this regard, there is a need to examine how identified 'conversational' phenomena within tweeting practices work as locally accountable features of a moral order. That is, as with any body of practice there are right and wrong ways of going about doing things and not just anything goes. Thus what happens within tweets may sometimes get explicitly called to account by other tweeters. Tweeters may themselves offer up 'accounts' for why they are proceeding in a certain fashion. Furthermore, one may test the orderly constitution of tweeting practice by deliberately exploring how it might be otherwise and what the consequences of doing things differently would be. All of these would serve to expose the socially mandated character of tweeting as a body of practice and how tweeters themselves manage it as an orderly set of affairs.

6.1.7 Looking at microblogging as its own job of work with its own grammars of action

In one of his most formative and programmatic papers called 'Notes on Methodology', Sacks makes the following methodological observations about how he first came to be working with talk and conversation:

> *"When I started to do research in sociology I figured that sociology could not be an actual science unless it was able to handle the details of actual events, handle them formally, and in the first instance be informative about them in the direct ways in which primitive sciences tend to be informative - that is, that anyone else can go and see whether what was said is so. And that is a tremendous control on seeing whether one is learning anything.*
>
> *"So the question was, could there be some way that sociology could hope to deal with the details of actual events, formally and informatively? One might figure that it had already been shown that it was perfectly possible given the vast literature, or alternatively that it was obviously impossible given the literature. For a variety of reasons I figured that it had not been shown either way, and I wanted to locate some set of materials that would permit a test; materials that would have the virtue of permitting us to see whether it was possible, and if so, whether it was interesting. The results might be positive or negative.*
>
> *"I started to work with tape-recorded conversations. Such materials had a single virtue, that I could replay them. I could transcribe them somewhat and study them extendedly - however long it might take. The tape-recorded materials constituted a "good enough" record of what happened. Other things, to be sure, happened, but at least what was on the tape had happened. It was not from any large interest in language or from some theoretical formulation of what should be studied that started with tape-recorded conversations, but*

> *simply because I could get my hands on it and I could study it again and again, and also, consequentially, because others could look at what I had studied and make of it what they could, if, for example, they wanted to be able to disagree with me...*
>
> *"Thus it is not any particular conversation, as an object, that we are primarily interested in. Our aim is to get into a position to transform, in an almost literal, physical sense, our view of "what happened," from a matter of particular interaction done by particular people, to a matter of interactions as products of a machinery. We are trying to find the machinery. In order to do so we have to get access to its products. At this point, it is conversation that provides us such access... "*, Sacks [49]: 26-7

So something else to take note of here is that, just as Sacks was able to explore the production of certain facets of social interaction in a replicable and inspectable way by using tape-recorded conversations, so we have available to us within PHEME an equally replicable and inspectable body of recordings in the shape of a stream of tweets coming out of Twitter. There are some limitations here in that we do not have available to us the specific individual situation in which people composed and received those tweets. However, Sacks' original tape recordings were similarly constrained in that a good deal of 'what was going on' was absent from the recordings as far as the specific individuals being recorded were concerned. So what we do have in the corpus of tweets is a body of live-when-recorded socially produced phenomena that are open to being examined for how they work as – just as Sacks put it – 'products of a machinery'. It is also worth noting here that, just as Sacks was concentrating on the orderly products of verbal interaction (often, it turns out, through the mediating technology of the telephone), so it is important that we focus upon the organisation of tweets in a Twitter stream as orderly products of an online interaction and focus on their own organisational properties as 'just that kind of thing', together with how those properties are made manifest and accountable within the way they are produced.

Elaborating a little on the preceding points, in that case, when you tweet you don't typically say you're just going to chat with someone, talk with someone, speak to someone, etc. How people articulate having conversations with one another and how they articulate tweeting, or even just looking on Twitter, are quite different. Throughout his work Coulter [8] makes much use of the notion of what he terms 'sequential grammars of action'. This idea in outline actually originated with Wittgenstein. Wittgenstein [67] emphasises in his Philosophical Investigations that a grammar is in no way an explanation of action. It sets aside questions regarding why people do what they do. Instead it allows for us to see what resources they have available to them in particular situations and how they use them: "Grammar does not tell us how language may be constructed in order to fulfil its purpose, in order to have such-and-such an effect on human beings. It only describes and in no way explains the use of signs." (Wittgenstein [67], PI: 496, p 138e).

In that they address questions of 'what' and 'how', the grammars of action in play when people are using Twitter are important. People 'tweet', they 'retweet', they 'look at Twitter', they 'catch up on Twitter', and so on. A job that therefore needs to be done is to pull out of both Twitter and other sources of reference these different grammatical articulations of what people understand themselves to be doing when they are using Twitter and to lay these out as the structure of a body of practice. This should then provide for specifying how the different aspects of that body of practice that are articulated through these grammars are actually accomplished, what their orderly features of production look like, how those are in turn implicative for further bodies of practice and action and what those in turn look like. In other words it provides us with an understanding of the sequential organisation of people's practices that is grounded quite specifically in the use of Twitter, rather than taking those sequences to be a species of conversation that is primarily intelligible through reference to situated talk. There is a sense in which this work is preliminary to other work towards the development of a framework. However, pragmatically it makes sense to begin with some of the insights from conversation analysis because these are already to hand, working under the proviso that there will be a process of refinement and revision over time as our understanding of Twitter use in its own right develops.

6.1.8 Referential practices

As an important example of the kinds of Twitter specific phenomena we will need to be able to handle, an immediately evident distinction between conversational activities and the ways in which people use Twitter is the kind of work people may be engaged in when they are doing things such as retweeting, providing hashtags, incorporating @someone in their tweets, tweeting images and including images in their tweets, providing links to other sources, and so on. In other words one can find a series of referential practices in the use of Twitter that are not produced in such explicit ways in spoken conversation. This immediately marks out one whole area of Twitter use that needs examining for the kinds of systematic practices and reasoning it may reveal. Amongst other things, we indicated above the commonplace character of second stories as alignment mechanisms in everyday conversation. It is important to understand the extent to which some of the referential practices one encounters within Twitter may be doing a similar kind of work.

6.1.9 The asynchronous character of microblog exchange

Another distinctive feature of Twitter and microblog exchange in general that is pointed to by a number of parties is its asynchronous character. This is obviously another important difference between Twitter and face-to-face conversation and some of the consequences of this have already been indicated, for instance, the absence of necessarily adjacent relations between related actions and the interleaving of different topics. As this constitutes such a significant difference it indicates a need to also examine closely how Twitter users

systematically provide for its coherence across asynchronous interaction within the production of their own actions. Indeed, we have seen through our testing of the initial annotation scheme how this coherence is made observable and hence recoverable through the use of Twitter messaging conventions, in particular, *retweet*, *reply* and *mention*.

6.1.10 The organisation of rumour as a feature of microblog exchange

Obviously, above and beyond all of the preceding considerations we have outlined, a central concern within PHEME will be to identify the ways in which rumour works within the context of Twitter-based interactions. In many ways this will build upon the other considerations already described but with the specific aim of uncovering the systematic features of how rumour and associated ascriptive devices are managed as a part of people's use of Twitter. How are specific Twitter feeds recognizable in specific situations as rumours? When feeds are described as rumours what kind of work is that doing quite specifically with regard to the operation of Twitter as an interactional device?

6.1.11 Associated literatures

It should be noted that the above move towards grounding the work on annotation within PHEME an understanding of Twitter use as microblog exchange does have some precedence in the conversation analytic literature. In particular there is a small but discrete body of work that explores the characteristics of text-based exchange in the context of practices such as SMS use and interactions in chatrooms. Relevant materials to be drawn upon here will include: Antaki [2]; Antaki et al. [3]; Abdallah [1]; Golato & Taleghani-Nikazm [18]; Have [61]; Hutchby & Tanna [24]; Jenks [26]; Laursen [28]; Markman [31]; Nilsen & Makitalo [34]; Ong [35]; Reed & Ashmore [41]; Rintel et al. [43, 44]; Rintel & Pittam [42]; Schonfeldt & Golato [55]; Stommel & van de Houwen [60]; Vallis [62, 63]; and Zemel et al. [69].

In addition, certain literatures in the more technical and NLP-related canon have taken an interest in the kinds of matters outlined above and an important job of work will be the articulation of findings coming out of this research in ways that can be connected to these kinds of interests, not to mention an inspection of the extent to which these literatures may also serve to inform the kinds of analysis being undertaken and annotation strategies being adopted. For instance, Ritter et al. [45] start out from a similar position to our own regarding tackling Twitter dialogue as a phenomenon in its own right. They then use an unsupervised method for modelling dialogue acts in order to try and arrive at something that can capture the sequential character of Twitter conversations. They discuss three kinds of initiating acts – 'Status', 'Reference Broadcast' and 'Question to Followers' – which they see as generating different kinds of responses. They then model chains of responses which they categorise as 'reactions', 'comments', 'questions' and 'answers'.

Zhang et al. [71] similarly look at ways of recognising speech acts in Twitter, though in their case they focus upon the problem of training data and present a semi-supervised approach. They categorise relevant speech acts as being 'statements', 'questions', 'suggestions', 'comments', and 'miscellaneous' to capture everything else. Also relevant here are: Mann & Thompson [66] who use Rhetorical Structure Theory to provide a means of describing the organisational relationship between different parts of naturally occurring text; and Zhai & Williams [70] who explore the latent structures in certain kinds of dialogue.

6.2 A staged and iterative process

The preceding material in this section of the deliverable has laid out a challenging and ambitious programme of work to be undertaken. It aims to build upon the annotation scheme set out in section 5 in a fashion that has not previously been explored beyond the associated literatures mentioned above. It will therefore involve some significant ground-up analysis and careful testing of its outputs all along the way. For this reason it will be necessary to adopt a phased approach that allows for step-by-step elaboration of the annotation scheme and associated evaluation of its outputs. Work will begin on this process over the coming months, beginning with application of the scheme already set out in section 5, and will initially make use of two existing corpora: the collection of tweets established for previous work on the 2011 London riots; and a collection of Twitter threads arising from the snopes.com website which is dedicated to providing a reference point for 'urban legends, folklore, myths, rumors, and misinformation'.

It is anticipated that the programme of work outlined in section 6.1 will progress through the following phases:

1. Making the specific social order pertaining to Twitter exchanges visible in its own particular way.
2. Pulling out the organised properties of rumour-related activities within the body of Twitter exchange practice as something both recognizable and ascribable and realizable as an orientation in systematic ways (which will include how this works as an accountable set of practices with a describable moral order in play). Some potential additional issues will need to be tackled here:
   (a) Rumour-related activities will necessarily be a constituent part of a broader set of Twitter exchange practices so the relationships between these practices will need to be identified together with how the spread of rumours is organised within and by those practices in various ways.
   (b) It will be necessary to articulate just what resources are required within the context of Twitter exchanges for rumour-related phenomena to take place.

(c) There may be a need to make clear important distinctions between different kinds of rumour-related phenomena because they will, as a result, be accountable and implicative in different ways, e.g., the difference between 'accidental' rumour instigation where others take remarks in unintended ways (characterised as 'misinformation' above) and 'deliberate' rumour instigation (characterised as 'disinformation' above) where there is a direct interest in propagating speculation etc, in some way. Findings from conversation analysis and the principles of intersubjectivity we outlined earleir would suggest that, where the former is the case, it will be accompanied by visible repair.

(d) It will be important to bear in mind that there are framing concerns within PHEME that will need to be specifically attended to when producing outputs for annotation that may not always necessarily be foregrounded by a focus upon rumour alone. This may include features such as the exact way in which tweets are recipient designed, the nature of the sources that rumours originate with, and other matters such as veracity and trust. Some of these features are already encompassed within the annotation scheme but will need to be revisited to understand how they relate to the conversation analytic approach as it unfolds.

3. Evolving the preceding materials so that they are not just handling specific instances but also expressing a machinery whereby rumours arise and get promulgated in the context of Twitter feeds. This will provide the grounds upon which larger scale annotation strategies might be formulated.

Throughout the progress of each of these phases there will be a number of concerns that will need to be iteratively tackled and tested for their adequacy:

- The development of annotation strategies that can capture the organisational properties being described.
- The refinement of the existing annotation scheme as outlined in section 5 so that it incorporates these developments.
- Testing the revised system of annotation against the constraints and requirements of the applications that will need to make use of it.
- Keeping a weather-eye to the fact that dependency on human annotation will not, in the longer term, meet the requirements of PHEME. The ultimate ambition to be looked to will be having the processing involved – from first identification of a rumour to the testing of its potential impact and its veracity – happening without need for human intervention such that rumours can be flagged and handled as they arise.

Chapter 7

Discussion

This document describes our preliminary efforts towards developing an annotation scheme for rumours spread on social media. This annotation scheme has been developed in the context of the PHEME project's Work Package 2, and will serve as guidelines for the annotation work to be undertaken in Work Packages 7 and 8. We have outlined relevant annotation schemes defined in the literature, which we take into account as starting points to integrate and adapt for our purposes. We have adapted some factors from existing annotation schemes and introduced new ones that are specifically suitable for rumours. The annotation scheme has been defined through an iterative process, which allowed for continuous revisions to come up with a suitable scheme. While this scheme provides a suitable structure of features to be considered in the annotation, we are currently looking at other types of rumours spread in social media. This will enable us to broaden the scope of rumours tested, and to extend the list of values that each of the features can take.

We have also discussed the suitability of these factors to the context of social media rumours from an interdisciplinary perspective, considering background from both social media research, as well as sociolinguistic literature including Conversational Analysis and Ethnomethodology. We have, additionally, discussed ways in which Conversation Analysis and Ethnomethodology will be used to inform further refinements and elaborations of the annotation scheme. The annotation scheme as described at present will serve as an initial base that will be assessed in subsequent steps of the PHEME project, through analysis of its suitability to different events and rumours collected from social media, making use of publicly available APIs. This analysis will enable us to find the strengths and weaknesses of the annotation scheme, and will allow us to pursue the development of the final annotation scheme, as well as the annotation itself of corpora that will be used to research social media rumours.

Bibliography

[1] Sebastian Abdallah. Online chatting in beirut: sites of occasioned identity-construction. *Ethnographic Studies*, 10:3–22, 2008.

[2] Charles Antaki. Two rhetorical uses of the description 'chat'. *M/C: a Journal of Media and Culture*, 3(4), 2000.

[3] Charles Antaki, Elisenda Ardévol, Francesc Núñez, and Agnès Vayreda. "for she who knows who she is:" managing accountability in online forum messages. *Journal of Computer-Mediated Communication*, 11(1):114–132, 2005.

[4] Douglas Benson and John Hughes. Method: evidence and inference-evidence and inference for ethnomethodology. *Ethnomethodology and the human sciences*, pages 109–136, 1991.

[5] Jörg R Bergmann. *Discreet indiscretions: The social organization of gossip*. Transaction Publishers, 1993.

[6] Jonathan Clifton. A membership categorization analysis of the waco siege: Perpetrator-victim identity as a moral discrepancy device for'doing'subversion. *Sociological Research Online*, 14(5):8, 2009.

[7] Jeff Coulter. Beliefs and practical understanding. *Everyday Language. Studies in Ethnomethodology. New York: Irvington Publishers*, 1979.

[8] Jeff Coulter. *Mind in action.* Humanities Press International, 1989.

[9] Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 241–244. ACM, 2012.

[10] M de Marneffe, Christopher D Manning, and Christopher Potts. Veridicality and utterance understanding. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 430–437. IEEE, 2011.

[11] Nicholas Diakopoulos and Arkaitz Zubiaga. Newsworthiness and network gatekeeping on twitter: The role of social deviance. In *Proc. International Conference on Weblogs and Social Media (ICWSM)*, 2014.

[12] Nicholas DiFonzo and Prashant Bordia. Rumor, gossip and urban legends. *Diogenes*, 54(1):19–35, 2007.

[13] Emile Durkheim, Sir George Edward Gordon CATLIN, and Sarah A SOLOVAY. *The Rules of Sociological Method... Translated by Sarah A. Solovay and John H. Mueller, and Edited by George EG Catlin*. 1938.

[14] Don Fallis. On verifying the accuracy of information: philosophical perspectives. 2004.

[15] Adrien Friggeri, Lada A. Adamic, Dean Eckles, and Justin Cheng. Rumor cascades. In *ICWSM*, 2014.

[16] Harold Garfinkel. Studies in ethnomethodology. 1967.

[17] Harold Garfinkel. Respecification: Evidence for locally produced, naturally accountable phenomena of order, logic, reason, meaning, method, etc. in and as of the essential haecceity of immortal ordinary society (i)—an announcement of studies. *Ethnomethodology and the human sciences*, pages 10–19, 1991.

[18] Andrea GOLATO and Carmen TALEGHANI-NIKAZM. Negotiation of face in web chats. *Multilingua*, 25(3):293–321, 2006.

[19] Marjorie Harness Goodwin. he-said-she-said: formal cultural procedures for the construction of a gossip dispute activity. *American Ethnologist*, 7(4):674–695, 1980.

[20] Bernard Guerin and Yoshihiko Miyazaki. Analyzing rumors, gossip, and urban legends through their conversational properties. *Psychological Record*, 56(1), 2006.

[21] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *ICWSM*, 2014.

[22] RHR Harper. Radicalism, beliefs and hidden agendas. *Computer Supported Cooperative Work (CSCW)*, 3(1):43–46, 1994.

[23] John Heritage, Elizabeth Boyd, and Lawrence Kleinman. Subverting criteria: the role of precedent in decisions to finance surgery. *Sociology of Health & Illness*, 23(5):701–728, 2001.

[24] Ian Hutchby and Vanita Tanna. Aspects of sequential organization in text message exchange. *Discourse and Communication*, 2:143–164, 2008.

[25] Paul L Jalbert. Categorization and beliefs: News accounts of haitian and cuban refugees. *The interactional order: New directions in the study of social order*, pages 231–248, 1989.

[26] Christopher Joseph Jenks. Getting acquainted in skypecasts: Aspects of social organization in online chat rooms. *International Journal of Applied Linguistics*, 19(1):26–46, 2009.

[27] Alex Koohang and Edward Weiss. Misinformation: toward creating a prevention framework. *Information Science*, 2003.

[28] Ditte Laursen. Sequential organization of text messages and mobile phone calls in interconnected communication sequences. *Discourse & Communication*, 6(1):83–99, 2012.

[29] John Lee. Innocent victims and evil-doers. In *Women's Studies International Forum*, volume 7, pages 69–73. Elsevier, 1984.

[30] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012.

[31] Kris M Markman. "so what shall we talk about" openings and closings in chat-based virtual meetings. *Journal of Business Communication*, 46(1):150–170, 2009.

[32] Albert J Meehan. Assessing the "police-worthiness" of citizen's complaints to the police: accountability and the negotiation of "facts"'. *The interactional order: New directions in the study of social order*, pages 116–140, 1989.

[33] Wayne Martin Mellinger. " accomplishing fact in police" dispatch packages": An analysis of the situated construction of an organizational record. In *Perspectives on social problems*, volume 4, pages 47–72. 1992.

[34] Mona Nilsen and Åsa Mäkitalo. Towards a conversational culture? how participants establish strategies for co-ordinating chat postings in the context of in-service training. *Discourse Studies*, 12(1):90–105, 2010.

[35] Kenneth Keng Wee Ong. Disagreement, confusion, disapproval, turn elicitation and floor holding: Actions as accomplished by ellipsis marks-only turns and blank turns in quasisynchronous chats. *Discourse Studies*, 13(2):211–234, 2011.

[36] Nicola Parker and Michelle O'Reilly. 'gossiping'as a social action in family therapy: The pseudo-absence and pseudo-presence of children. *Discourse Studies*, 14(4):457–475, 2012.

[37] Rob Procter, Farida Vis, and Alex 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.

[38] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40, 2003.

[39] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu 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.

[40] Mark Rapley. 'just an ordinary australian': Self-categorization and the discursive construction of facticity in 'new racist'political rhetoric. *British Journal of Social Psychology*, 37(3):325–344, 1998.

[41] Darren Reed and Malcolm Ashmore. The naturally-occurring chat machine. *M/C: A Journal of Media and Culture*, 2000.

[42] E Rintel and Jeffery Pittam. Strangers in a strange land interaction management on internet relay chat. *Human Communication Research*, 23(4):507–534, 1997.

[43] E Sean Rintel, Joan Mulholland, and Jeffery Pittam. First things first: Internet relay chat openings. *Journal of Computer-Mediated Communication*, 6(3):0–0, 2001.

[44] E Sean Rintel, Jeffery Pittam, and Joan Mulholland. Time will tell: Ambiguous non-responses on internet relay chat. *Electronic Journal of Communication*, 13(1), 2003.

[45] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of twitter conversations. In *Proc of NAACL*, 2010.

[46] Ralph L Rosnow and Eric K Foster. Rumor and gossip research. *Psychological Science Agenda*, 19(4), 2005.

[47] Harvey Sacks. On the analyzability of stories by children. *Directions in sociolinguistics: The ethnography of communication*, pages 325–345, 1972.

[48] Harvey Sacks. Some technical considerations of a dirty joke. *Studies in the organization of conversational interaction*, pages 249–270, 1978.

[49] Harvey Sacks. Notes on methodology. *Structures of social action: Studies in conversation analysis*, pages 21–27, 1984.

[50] Harvey Sacks. On doing 'being ordinary'. *Structures of social action: Studies in conversation analysis*, pages 413–429, 1984.

[51] Harvey Sacks. *Lectures on conversation*, volume 1. Blackwell Publishing, 1995.

[52] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735, 1974.

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

[54] Roser Saurí and James Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.

[55] Juliane Schönfeldt and Andrea Golato. Repair in chats: A conversation analytic approach. *Research on language and social interaction*, 36(3):241–284, 2003.

[56] Pamela J Shoemaker. News and newsworthiness: A commentary. *Communications*, 31(1):105–111, 2006.

[57] Jack Sidnell. 6 the epistemics of make-believe. *The morality of knowledge in conversation*, 29:131, 2011.

[58] Dorothy E Smith. K is mentally ill'the anatomy of a factual account. *Sociology*, 12(1):23–53, 1978.

[59] Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. Modeling factuality judgments in social media text. In *ACL*, 2014.

[60] Wyke Stommel and Fleur van der Houwen. Formulations in "trouble" chat sessions. *Language@ Internet*, 10, 2013.

[61] Paul Ten Have. Computer-mediated chat: Ways of finding chat partners. *M/C: a Journal of Media and Culture*, 3(4), 2000.

[62] Rhyll Vallis. Members' methods for entering and leaving# ircbar: A conversation analytic study of internet relay chat. 1999.

[63] Rhyll Vallis, A Mchoul, and M Rapley. Applying membership categorization analysis to chat-room talk. *How to analyse talk in institutional settings: A casebook of methods*, pages 86–99, 2002.

[64] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. *ACL 2014*, page 18, 2014.

[65] Douglas Walton. Types of dialogue and burdens of proof. In *COMMA*, pages 13–24, 2010.

[66] Mann William and Sandra Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[67] Ludwig Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 2010.

[68] Robin Wooffitt. *Telling tales of the unexpected: The organization of factual discourse*. Rowman & Littlefield, 1992.

[69] Alan Zemel, Fatos Xhafa, and Murat Cakir. What's in the mix? combining coding and conversation analysis to investigate chat-based problem solving. *Learning and Instruction*, 17(4):405–415, 2007.

[70] Ke Zhai and Jason Williams. Discovering latent structure in task-oriented dialogues. In *Proceedings of ACL 2014*. Association for Computational Linguistics, June 2014.

[71] Renxian Zhang, Dehong Gao, and Wenjie Li. Towards scalable speech act recognition in twitter: tackling insufficient training data. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 18–27. Association for Computational Linguistics, 2012.

[72] Arkaitz Zubiaga and Heng Ji. Tweet, but verify: Epistemic study of information verification on twitter. *Social Network Analysis and Mining*, 2014.

Appendix A

Acronym Definitions

Acronym	Meaning
API	Application Programming Interface
CA	Conversation Analysis
HTML	Hypertext Markup Language
ID	Tweet identifier
JSON	JavaScript Object Notation
MCD	Membership Categorisation Device
OED	Oxford English Dictionary
URL	Universal Resource Locator

Appendix B

Sample Rumourous Conversations

The following are two conversational threads collected from Twitter following the methodology described in Section 5.2.1. The first rumour (see Figure B.1) refers to a statement that says that inputting the PIN number backwards in an ATM automatically calls the police, which sparks discussion and disagreeing responses. The second rumour (see Figure B.2) suggests that there is a connection between the SuperBowl sporting event and prostitution, which also sparks many responses. The examples below include the first 10 tweets in each conversation, which were manually annotated for the validation of the annotation scheme.

Figure B.1: Rumorous conversation responding to an ATM hoax.

Figure B.2: Rumourous conversation responding to a statement that relates SuperBowl and prostitution.

Appendix C

First Round of Annotations for Validating the Scheme

The following are the annotations provided by two assessors in the first round of test annotations performed during the definition of the annotation scheme. These annotations were coded for the tweets shown in Appendix B. This annotation test was performed to validate the revised annotation scheme described in Section 5.1.

Note that the assessors provided slightly different annotations in this case. While the assessor #1 coded for acceptability and veracity at the thread level, the assessor #2 did it at the tweet level for all responding tweets. The assessor #1 also tried to annotate each single feature at the tweet level, mainly for validation purposes.

C.1 ATM Hoax

- **Tweet 1 (source):**
  - **Annotation 1:**
    - Polarity: positive.
    - Modality: certain.
    - Presentation: appeal for more info.
    - Evidentiality: none.
    - Author Type: ordinary individual.
    - Plausibility: plausible.
  - **Annotation 2:**
    - Polarity: underspecified.
    - Modality: possible.
    - Presentation: appeal for more info / comment.

        * Evidentiality: quoting source.
        * Author Type: ordinary individual.
        * Plausibility: dubious.

- **Tweet 2:**
    - **Annotation 1:**
        * Polarity: positive.
        * Modality: certain.
        * Presentation: counterclaim.
        * Evidentiality: none.
        * Author Type: ordinary individual.
    - **Annotation 2:**
        * Acceptability: Strong disagreement but not with tweeter - with original information
        * Veracity: 0

- **Tweet 3:**
    - **Annotation 1:**
        * Polarity: underspecified.
        * Modality: underspecified.
        * Presentation: comment.
        * Evidentiality: none.
        * Author Type: ordinary individual.
    - **Annotation 2:**
        * Acceptability: More a comment than a position.
        * Veracity: Again, it's a comment, not a position.

- **Tweet 4:**
    - **Annotation 1:**
        * Polarity: underspecified.
        * Modality: underspecified.
        * Presentation: comment.
        * Evidentiality: none.
        * Author Type: ordinary individual.
    - **Annotation 2:**
        * Acceptability: More a comment than a position.

        * Veracity: Again, it's a comment, not a position.

- **Tweet 5:**
  - **Annotation 1:**
    * Polarity: underspecified.
    * Modality: underspecified.
    * Presentation: appeal for more info.
    * Evidentiality: none.
    * Author Type: ordinary individual.
  - **Annotation 2:**
    * Acceptability: Uncertainty.
    * Veracity: Unverified.
- **Tweet 6:**
  - **Annotation 1:**
    * Polarity: negative.
    * Modality: certain.
    * Presentation: counterclaim.
    * Evidentiality: none.
    * Author Type: ordinary individual.
  - **Annotation 2:**
    * Acceptability: Strong disagreement.
    * Veracity: 0
- **Tweet 7:**
  - **Annotation 1:**
    * Polarity: positive.
    * Modality: certain.
    * Presentation: counterclaim.
    * Evidentiality: quoting source.
    * Author Type: ordinary individual.
  - **Annotation 2:**
    * Acceptability: Strong disagreement.
    * Veracity: 0
- **Tweet 8:**
  - **Annotation 1:**

    - Polarity: underspecified.
    - Modality: underspecified.
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: More a comment than a position.
    - Veracity: Again, it's a comment, not a position.

- **Tweet 9:**
  - **Annotation 1:**
    - Polarity: underspecified.
    - Modality: underspecified.
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: Alignment with query.
    - Veracity: Adopting same questioning stance as originator.

- **Tweet 10:**
  - **Annotation 1:**
    - Polarity: negative.
    - Modality: certain.
    - Presentation: counterclaim.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: Strong disagreement.
    - Veracity: 0

- **Thread:**
  - **Annotation 1:**
    - Acceptability: strong disagreement.
    - Veracity: false.

C.2 Superbowl and Prostitution

- **Tweet 1 (source):**
  - **Annotation 1:**
    - Polarity: positive.
    - Modality: probable.
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
    - Plausibility: plausible.
  - **Annotation 2:**
    - Polarity: negative.
    - Modality: probable.
    - Presentation: claim / comment.
    - Evidentiality: none.
    - Author Type: ordinary individual (but seems on the border of being a media blogger).
    - Plausibility: plausible.
- **Tweet 2:**
  - **Annotation 1:**
    - Polarity: -
    - Modality: -
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: Strong agreement.
    - Veracity: 1
- **Tweet 3:**
  - **Annotation 1:**
    - Polarity: -
    - Modality: -
    - Presentation: comment.
    - Evidentiality: none.

    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: -
    - Veracity: -

- **Tweet 4:**
  - **Annotation 1:**
    - Polarity: -
    - Modality: -
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: Not entirely sure whether this is a response to the original tweet or to some other obscure aspect of the last tweet from the originator.
    - Veracity: -

- **Tweet 5:**
  - **Annotation 1:**
    - Polarity: -
    - Modality: -
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: comment.
    - Veracity: not applicable.

- **Tweet 6:**
  - **Annotation 1:**
    - Polarity: -
    - Modality: -
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**

    - Acceptability: -
    - Veracity: -

- **Tweet 7:**
  - **Annotation 1:**
    - Polarity: -
    - Modality: -
    - Presentation: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: comment on elaboration.
    - Veracity: not applicable.
- **Tweet 8:**
  - **Annotation 1:**
    - Polarity: positive + negative.
    - Modality: certain.
    - Presentation: counterclaim.
    - Evidentiality: quoting source.
    - Author Type: ordinary.
  - **Annotation 2:**
    - Acceptability: strong disagreement.
    - Veracity: 0
- **Tweet 9:**
  - **Annotation 1:**
    - Polarity: negative + positive.
    - Modality: possible.
    - Presentation: comment.
    - Evidentiality: quoting source.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: -
    - Veracity: -
- **Tweet 10:**

  - **Annotation 1:**
    - Polarity: positive.
    - Modality: certain.
    - Presentation: counterclaim.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Acceptability: justification of disagreement.
    - Veracity: -

- **Thread:**
  - **Annotation 1:**
    - Acceptability: slight disagreement.
    - Veracity: false.

Appendix D

Second Round of Annotations for Validating the Scheme

The following are the annotations provided by two assessors in the second round of test annotations performed during the definition of the annotation scheme. These annotations were coded for the tweets shown in Appendix B. This annotation test was performed to validate the revised annotation scheme described in Section 5.3.

D.1 ATM Hoax

- **Tweet 1 (source):**
  - **Annotation 1:**
    - Polarity: positive.
    - Modality: possible.
    - Plausibility: plausible.
    - Evidentiality: quoting source.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Polarity: underspecified.
    - Modality: possible.
    - Plausibility: dubious.
    - Evidentiality: quoting source (URL).
    - Author Type: ordinary individual.
- **Tweet 2:**
  - **Annotation 1:**

        * Modality: certain.
        * Response Type: disagreed.
        * Evidentiality: none.
        * Author Type: ordinary individual.
    - **Annotation 2:**
        * Modality: certain.
        * Response Type: disagreed.
        * Evidentiality: none.
        * Author Type: ordinary individual.

- **Tweet 3:**
    - **Annotation 1:**
        * Modality: certain.
        * Response Type: comment.
        * Evidentiality: none.
        * Author Type: ordinary individual.
    - **Annotation 2:**
        * Modality: underspecified.
        * Response Type: comment.
        * Evidentiality: none.
        * Author Type: ordinary individual.

- **Tweet 4:**
    - **Annotation 1:**
        * Modality: certain.
        * Response Type: comment.
        * Evidentiality: none.
        * Author Type: ordinary individual.
    - **Annotation 2:**
        * Modality: certain.
        * Response Type: comment.
        * Evidentiality: none.
        * Author Type: ordinary individual.

- **Tweet 5:**
    - **Annotation 1:**

    - Modality: possible.
    - Response Type: appeal for more info.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: possible.
    - Response Type: appeal for more info.
    - Evidentiality: none.
    - Author Type: ordinary individual.

- **Tweet 6:**
  - **Annotation 1:**
    - Modality: certain.
    - Response Type: disagreed.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: certain.
    - Response Type: disagreed.
    - Evidentiality: none.
    - Author Type: ordinary individual.

- **Tweet 7:**
  - **Annotation 1:**
    - Modality: certain.
    - Response Type: disagreed.
    - Evidentiality: quoting source.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: certain.
    - Response Type: disagreed.
    - Evidentiality: quoting source (URL).
    - Author Type: ordinary individual.

- **Tweet 8:**
  - **Annotation 1:**

    - Modality: certain.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: underspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.

- **Tweet 9:**
  - **Annotation 1:**
    - Modality: certain.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: underspecified.
    - Response Type: appeal for more info.
    - Evidentiality: none.
    - Author Type: ordinary individual.

- **Tweet 10:**
  - **Annotation 1:**
    - Modality: certain.
    - Response Type: disagreed.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: certain.
    - Response Type: disagreed.
    - Evidentiality: none.
    - Author Type: ordinary individual.

D.2 Superbowl and Prostitution

- **Tweet 1 (source):**
  - **Annotation 1:**
    - Polarity: positive.
    - Modality: probable.
    - Plausibility: plausible.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Polarity: positive.
    - Modality: certain.
    - Plausibility: plausible.
    - Evidentiality: none.
    - Author Type: ordinary individual.
- **Tweet 2:**
  - **Annotation 1:**
    - Modality: certain.
    - Response Type: agreed.
    - Evidentiality: quoting source.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: certain.
    - Response Type: agreed.
    - Evidentiality: quoting unspecified source.
    - Author Type: ordinary individual.
- **Tweet 3:**
  - **Annotation 1:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: probable.

    - Response Type: agreed.
    - Evidentiality: reasoning.
    - Author Type: ordinary individual.

- **Tweet 4:**
  - **Annotation 1:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
- **Tweet 5:**
  - **Annotation 1:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
- **Tweet 6:**
  - **Annotation 1:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: certain.

    - Response Type: agreed.
    - Evidentiality: witnessed.
    - Author Type: ordinary individual.

- **Tweet 7:**
  - **Annotation 1:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: unspecified.
    - Response Type: comment.
    - Evidentiality: none.
    - Author Type: ordinary individual.
- **Tweet 8:**
  - **Annotation 1:**
    - Modality: certain.
    - Response Type: disagreed.
    - Evidentiality: quoting source.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: certain
    - Response Type: disagreed.
    - Evidentiality: quoting source.
    - Author Type: ordinary individual.
- **Tweet 9:**
  - **Annotation 1:**
    - Modality: probable.
    - Response Type: agreed.
    - Evidentiality: quoting source.
    - Author Type: ordinary individual.
  - **Annotation 2:**
    - Modality: possible.

        * Response Type: appeal for more info.
        * Evidentiality: quoting source.
        * Author Type: ordinary individual.

- **Tweet 10:**
    - **Annotation 1:**
        * Modality: certain.
        * Response Type: disagreed.
        * Evidentiality: none.
        * Author Type: ordinary individual.
    - **Annotation 2:**
        * Modality: certain.
        * Response Type: disagreed.
        * Evidentiality: none.
        * Author Type: ordinary individual.