Proceedings of the Ninth International AAAI Conference on Web and Social Media

CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations

Tanushree Mitra and **Eric Gilbert**
School of Interactive Computing & GVU Center
Georgia Institute of Technology
{tmitra3, gilbert}@cc.gatech.edu

Abstract

Social media has quickly risen to prominence as a news source, yet lingering doubts remain about its ability to spread rumor and misinformation. Systematically studying this phenomenon, however, has been difficult due to the need to collect large-scale, unbiased data along with in-situ judgements of its accuracy. In this paper we present CREDBANK, a corpus designed to bridge this gap by systematically combining machine and human computation. Specifically, CREDBANK is a corpus of tweets, topics, events and associated human credibility judgements. It is based on the real-time tracking of more than 1 billion streaming tweets over a period of more than three months, computational summarizations of those tweets, and intelligent routings of the tweet streams to human annotators—within a few hours of those events unfolding on Twitter. In total CREDBANK comprises more than 60 million tweets grouped into 1049 real-world events, each annotated by 30 human annotators. As an example, with CREDBANK one can quickly calculate that roughly 24% of the events in the global tweet stream are not perceived as credible. We have made CREDBANK publicly available, and hope it will enable new research questions related to online information credibility in fields such as social science, data mining and health.

Introduction

When the Ebola virus arrived in the United States, a satirical website claimed that two Ebola victims in Africa had risen from the dead[1]. This led to widespread panic about a potential "Ebola zombie apocalypse", eventually flooding social media streams with inaccurate information about the disease. There is a popular narrative that social media is full of inaccurate information like this. But how much? Do these rumors have structure—temporal, linguistic or otherwise? In this paper, we introduce a new corpus called CREDBANK designed to help answer questions like these. CREDBANK systematically weaves together machine computation with the judgements of human annotators to produce credibility annotations of the global tweet stream. As an illustration of what can be done with CREDBANK, we can quickly answer the "How much?" question: a moderate threshold of inter-rater agreement (70% of human annotators agree on the credibility assessment) suggests that an alarming 23.46% of the global tweet stream is not credible.

The domain within which CREDBANK falls, *credibility assessment*, has received significant attention. Earlier work has looked at automatic detection of credibility (Castillo, Mendoza, and Poblete 2011; Qazvinian et al. 2011), the diffusion patterns of rumors (Friggeri et al. 2014), building interactive tools to allow investigation of these patterns (Resnick et al. 2014) and exploring the factuality of various claims (Soni et al. 2014). CREDBANK extends the state of the art in two primary ways. First, because of the inherent difficulty associated with collecting large-scale rumor data, previous work has had to *select on the dependent variable* (Tufekci 2014) — presuming a priori what rumors look like (i.e., constructing retrieval queries) or working from a known set of rumors or post hoc investigation of prominent events with known disputed information or credibility judgements of specific topics trending on social media. Importantly, CREDBANK overcomes this *sampling bias* by asking human raters to assess the credibility of *all* social media events that cross our data pipeline. Second, we have made CREDBANK available to the research community[2]; the hope is that it will spark new questions about rumors in social media.

In this paper, we describe the development and validation of the framework on which CREDBANK is based, as well as a brief statistical overview of the corpus. We tracked more than 1 billion streaming tweets over a period of over three months, computationally summarizing those tweets into events, and routed the events to crowd workers for credibility annotation. By guiding annotators through a framework inspired by theoretical work, we show that crowd workers can approximate the credibility judgements of University-employed reference librarians, the gold standard used in this paper. In total, CREDBANK comprises more than 60M tweets grouped into 1049 real-world events, each annotated by 30 Amazon Mechanical Turk workers for credibility (along with their rationales for choosing their annotations). The primary contribution of CREDBANK is a unique dataset compiled to link social media event streams with human credibility judgements in a systematic and comprehensive

[1] http://huzlers.com/breaking-news-ebola-victim-rises-dead-africa-fear-zombie-apocalypse/

[2] http://comp.social.gatech.edu/papers/#icwsm15.credbank.mitra

way. This is the first attempt we are aware of to do so at such a scale and in a naturalistic setting, that is in an environment which closely resembles the way a person would search Twitter for event information. CREDBANK enables a set of new research questions. For example, social scientists might explore what role does the mainstream media plays in online rumors; a data mining researcher might explore how the temporal pattens of rumors differ from highly credible information or study the interplay between highly disputed event and other less disputed ones occurring in the same time span; a health researcher could investigate how folk theories of a new disease (the emergence of Ebola is captured in CREDBANK) diffuse through a population.

Related Work

With social media's growth as a news resource (Caumont 2013) and it's purported role in spreading false rumors and misinformation, there have been several organized attempts to study social media credibility. One trend in this domain is to study specific events that were subjects of misinformation. For example, studies have tracked the spread of rumors during the 2011 Great East Japan earthquake (Liu, Burton-Jones, and Xu 2014), provided descriptive analysis of information and misinformation shared during the 2013 Boston bombings (Maddock et al. 2015) and reported a case study of rumor dynamics in a Chinese microblogging community (Liao and Shi 2013). Together, these studies suggest the importance of anxiety, personal involvement and informational ambiguity in spreading rumors. However, their findings are based on these extreme, hand-selected cases of rumor.

On the other hand, tracking less well-known disputed information along with the widespread instances of newsworthy content is challenging. It requires sifting through massive amounts of social media posts, followed by a labor intensive task of content evaluation for credibility assessment. There has been some preliminary journalistic research on identifying, tracking and logging misinformation[3]. However, most of this work seems to be an arduous effort by a small number of journalists screening most of the data, or relying on externally reported instances of misinformation. Our work addresses these challenges by systematically combining machine computation with multiple independent micro-labor annotations.

Another line of research in this domain involves performing extensive quantitative analysis on social media traces—traces corresponding to historically reported cases of rumors. For example, researchers have collected documented rumor cases from popular urban legend websites (i.e., Snopes) and have analyzed their corresponding Facebook posts to find new insights on rumor mutations (Friggeri et al. 2014). Similar techniques have been used to identify temporal and structural features of rumor in Twitter (Kwon et al. 2013).

Predictive analysis of information credibility is also a popular trend in this area, such as building classifiers to detect whether tweets are factual or not (Castillo, Mendoza, and Poblete 2011), automatically assessing the credibility level of a set of tweets (Qazvinian et al. 2011) or assigning credibility score to tweets in a user's timeline based on learning from historically prominent event tweets (Gupta et al. 2014). A common theme in these work is to treat credibility assessment as a two step process. The first step is to extract newsworthy content, while the next step is to assess the credibility of the retrieved content. Our work builds on this basic two-step approach. However, we extend it by going beyond the traditional setting of post-hoc investigation of historical events, investigating in real time every social media event.

CREDBANK's Construction

CREDBANK is built on an iterative framework of five main phases, combining machine computation (MC) and human computation (HC) in an efficient way. Figure 1 presents the construction of CREDBANK graphically, along with the descriptive statistics of the data item generated at each phase.

Phase 1: Streaming Data and Preprocessing (MC)

We used the Twitter streaming API to collect a continuous 1% sample of global tweets[4] (Priedhorsky, Culotta, and Del Valle 2014). For every group of 1 million streaming tweets, we first filter out English only tweets identified by Twitter's machine language detection algorithm (annotated in the tweet's metadata returned from the API).

Spam Removal. Next we apply standard spam-removal techniques to filter out tweets which might have escaped Twitter's spam detection algorithms. Adopting the practices of Firefox's Clean Tweets[5] add on, we filter out tweets from accounts less than a day old and eliminate tweets containing three or more hashtags. Additionally, we also remove tweets where the entire text is in upper case. Next, following (Chu et al. 2012)'s approach we check whether text patterns within a tweet correspond to spam content. This step detects tweets which look like prototypical spam, such as *buy Viagra online without a prescription* or *get car loan with bad credit*. We first compiled a list of textual spam patterns based on existing lists of spam trigger words (WordPress.org ; Rubin 2012; Shivajappa). We then mark a tweet as spam if it contains a URL and a spam phrase—a standard technique used by spammers to direct users to spam websites (Chu et al. 2012).

Tokenization. We tokenized tweets using a regex-based Twitter tokenizer (O'Connor, Krieger, and Ahn 2010). While Traditional Penn Treebank style tokenizers work well on structured texts (like news articles), they perform poorly on social media text, often breaking punctuation, emoticons and unicode characters into a single token per character.

Stop-word Removal. Next, we employ multi-stage stopword removal. The purpose of doing this is to select a vocabulary for topic modeling (discussed shortly) that eliminates overly generic terms while retaining terms which are frequent in the corpus, but infrequent among general Twitter messages (O'Connor, Krieger, and Ahn 2010). Specifically, we had the following stages of stop word removal:

[3] http://www.craigsilverman.ca/2014/09/02/researching- rumors-and-debunking-for-the-tow-center-at-columbia-university/

[4] http://github.com/reidpr/quac

[5] https://addons.mozilla.org/en-US/firefox/addon/clean-tweets/

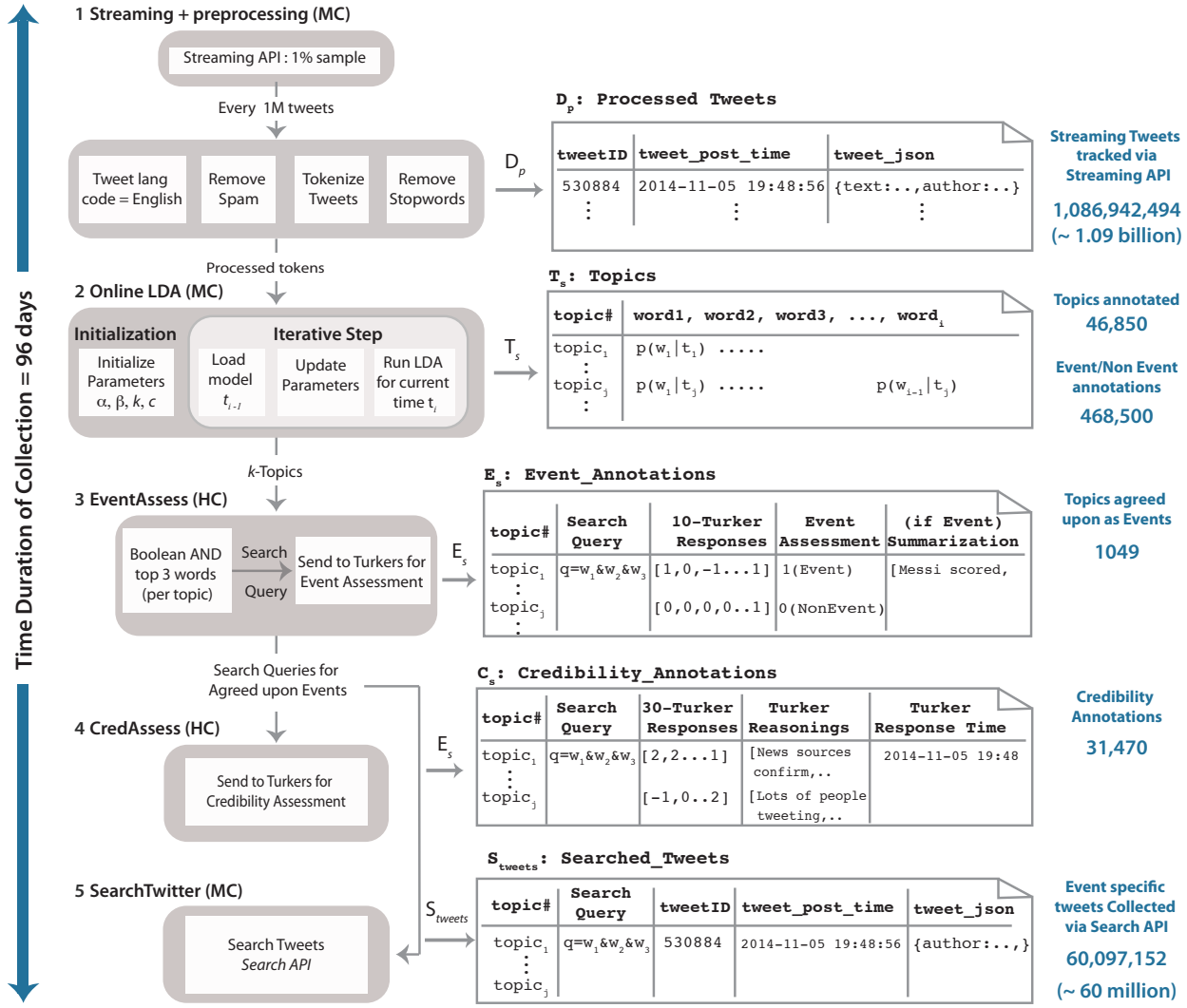

Figure 1: Map of the steps taken to create the corpus (left side). The right side shows the corpus schema resulting from each step along with the summary statistics of the data items generated at these steps.

1. First, we remove a list of function words based on the standard SMART stop list dictionary (Lewis et al. 2004).
2. Next, we filter out Twitter specific stop words, like *RT*, *follow*, *follows*, *@username*.
3. The earlier two approaches of removing stop words based on a static list is often plagued by being too generic. To overcome this limitation we adapt a method to automatically generate stop words–a method based on Zipf's law (Zipf 1949). Inspired by Zipf's observation that in a data collection the frequency of a term (TF) is inversely proportional to its rank, stop words can correspond to both top (TF-High) and low ranked words (TF-Low) (Makrehchi and Kamel 2008). Thus, in addition to the static stop-word list, we remove both: 1) the most frequent words (TF-High), corresponding to the top 50 words at every iteration, and 2) words that occur fewer than 10 times (TF-Low). Overall, this results in a dynamic vocabulary generated at each iteration that we send to the event detection phase.

Phase 2: Event Candidates using Online LDA (MC)

The basic premise of this phase is rooted in our main objective: gathering credibility assessments of real-world events as reported by social media streams. In this regard, the concept of events defined by the information retrieval community fits our study objective (Allan 2002; Becker, Naaman, and Gravano 2011). As per their definition, an event is a real-world occurrence happening at a specific time and associated with a time-ordered stream of messages. Hence, the primary goal of this phase is identifying such real-world events.

We opted for an online approach over traditional retrospective methods of event detection because of its ability to provide real-time responsiveness. While deciding on the specific technique for online event detection we considered

a range of existing online approaches, ranging from simple keyword-based methods (Culotta 2010; Weng and Lee 2011), bursty term analysis techniques (Osborne et al. 2012; Petrović, Osborne, and Lavrenko 2010) to more sophisticated topic-modeling based methods (Lau, Collier, and Baldwin 2012). The disadvantage of keyword-based approaches using a pre-defined set of keywords is that it fails to capture rapidly evolving real-world events. Bursty term analysis is based on bursts in term frequencies to overcome the limitations of pre-defined keyword-based techniques. However, it still cannot capture multiple co-occurring terms associated with an event (Zanzotto, Pennacchiotti, and Tsioutsiouliklis 2011). On the other hand, topic models can learn term co-occurrences associated with an event, making them a much better choice to capture quickly evolving real-world events. It also does not make any assumption (unlike keyword based methods) of what an event will look a priori. Thus alleviating the risks associated with selection bias (Tufekci 2014).

We ran the online LDA model (Lau, Collier, and Baldwin 2012) iteratively for every set of 1 million streaming tweets. The input to the LDA is a bag-of-words representation of individual tweets that passed the preprocessing steps. The model output is a set of latent topics represented as a set of related words which tend to co-occur in similar tweets. In other words, the topic modeling step segregates a collection of tweets into sets of coherent topics, where each topic can be interpreted by the top N terms with the highest marginal probability $\mathrm{p}(\mathrm{w_j} \mid \phi_\mathrm{k})$—the probability associated with each term $\mathrm{w_j}$ in a given topic ϕ_k.

One important step in our process is setting the model parameters. The model is initialized with Dirichlet prior values $\alpha = 0.001$ and $\beta = 0.01$, where α controls the sparsity of document-topic distribution and β determines the sparsity of topic-word distribution. A low value of α is preferred, because it produces a sparse distribution, leading to very few topic assignments per tweet. This intuitively makes sense, because it is almost impossible to mention large number of topics in a 140 character long tweet. Likewise lower values of β favor fewer words per topic. We opted for $\beta = 0.01$ to ensure that we do not end up with only single word topics. Multi-word topics provide more context for the human computation tasks employed in the later phases of our pipeline. Additionally, the LDA model also takes the number of topics K as an input parameter. We empirically evaluated the sensitivity of our model against a range of K-value settings, while keeping the other parameters constant. We tested this on a dataset of 1 million tweets. Recall, that this is the size of our streaming data at every iteration of our workflow. We converged on $\mathrm{K} = 50$. Such a high value of K also allows us to capture more granular topics rather than some high level general topics. Another input parameter to the model is the contribution factor c, which determines the contribution of the model learnt in the previous iteration (i-1^{th}). Following a related approach (Lau, Collier, and Baldwin 2012), we set $\mathrm{c} = 0.5$, so as to give equal weighting to tweets collected in successive iterations. The set of topics from each iteration form our set of candidate events.

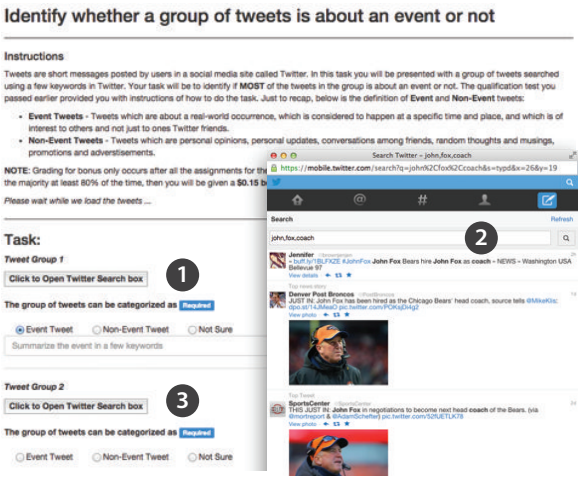

Figure 2: Turker interface for event assessment. The numbers correspond to a Turker's workflow in annotating each item. 1) Click the search box launching a new window of tweets corresponding to a search query, 2) Read tweets from the pop-up Twitter search window, 3) Select one of the annotation options. If Turker selects 'Event', then she is required to enter a short event summarization. Validation checks are put in place to ensure adherence to this workflow.

Phase 3: Filtering Event-specific Topics (HC)

One potential problem of a purely computational approach for event detection is occasional false positives in our set of candidate events. In other words, non-event activities such as conversations or opinions may be marked as events. To prevent these from being distractors in our main credibility assessment task we turn to human judgments. We recruited independent human annotators from Amazon Mechanical Turk (AMT) to decide whether a tweet is truly about an event or not. AMT is a widely used micro-labor market for crowd-sourcing annotation tasks.

Task Design (Event Annotation): Recent work on recruiting crowd workers for data annotations found that selectively screening and training workers, followed by offering financial incentives, is the best strategy for obtaining quality data annotations (Mitra, Hutto, and Gilbert 2015). Thus, we designed our crowd event annotation framework to first selectively screen and train workers via a qualification test. The test first provides rubrics to differentiate between events and non-events, followed by accompanying examples describing how to do the task. To come up with a definition of events, we traced back to research done for topic detection and tracking tasks (Allan 2002) and presented the following definition to our human judges:

> **Event Tweets:** Tweets which are about a real-world occurrence, which is considered to happen at a specific time and place, and which is of interest to others and not just to one's Twitter friends.
> **Non-Event Tweets:** Tweets which are personal opinions, personal updates, conversations among friends, random thoughts and musings, promotions and advertisements.

Overall, this served as a training phase for the workers. Next the workers were screened on the basis of their score on the qualification test. The qualification test questions were specifically designed for performing the event annotation task. The purpose was to provide task-specific orientation to the workers. Only those who scored a minimum of 75% were allowed to work on the annotation task (also called Human Intelligence Tasks, or HITs).

We designed our annotation task such that it closely mimics the way a person would search Twitter to see information related to an event. The advantage of this approach is that our data collection closely reflects everyday practices. For each of the K = 50 topics, we first take the top 3 terms and create a Twitter search query by using a boolean AND over all three terms. Our choice of top 3 terms was based on the following factors. Firstly, the top terms are the most representative terms of a topic. Secondly, usability studies show that people prefer to read short topical headlines, scanning only the first 3 and last 3 words[6] . Thirdly, the average word length of tweet is argued to be 10.7 words (Soanes 2011). Thus limiting tweets containing all top 3 topical terms provides a reasonable balance between being too generic (by including fewer terms) and too specific (by including more than 3 terms). Next, for each of these K queries corresponding to the K topics, we create an annotation item by embedding a Twitter search query box inside a HIT (see Figure 2). While annotating an item in a HIT, the worker has to first click the search box, see a set of real-time tweets and choose to label the tweet stream as representing: 1) Event Tweets, 2) Non-Event Tweets, or 3) Not Sure. If she selects the **Event** option, she is further prompted to summarize the event in a free-form text field. There were two factors which guided our design decision to elicit such subjective responses. First, subjective responses tend to improve rating quality and minimize random clicking (Kittur, Chi, and Suh 2008). Second, it allows us to filter out any ill-formed search queries—queries which are too broad to surface any specific Twitter event. The intuition is that such a query will likely return non-coherent tweets, making it difficult for humans to find a coherent theme and come up with a meaningful summary.

Determining Events: An initial pilot study confirmed this intuition. During the pilot phase, we also determined the number of Turkers needed to correctly annotate an item as an event. We manually checked the Turker annotations and found that if 6 out of 10 independent Turkers agreed that the topic is an event, then the majority label matched expert annotation. Note that our approach here is conservative in the sense that we are only considering topics to be true events if there is an absolute majority agreement among annotators. The pilot study also helped us determine the number of items to post per HIT so that a single HIT is completed within a short duration (under 2 mins), ensuring lower cognitive load per HIT. We posted 10 items per HIT and paid $0.15, adhering to minimum wage requirements. We also offered an additional $0.15 bonus to workers whose responses matched the modal response of the crowd. For each item we asked 10 Turkers for their independent ratings, and if a majority agreed that the topic is an event, then we added the topic to the queue for credibility assessment. We purposely choose a conservative inter-rater agreement (6 out of 10 agreements) to ensure maximum precision in the event detection step.

[6]https://blog.kissmetrics.com/how-to-write-headlines/

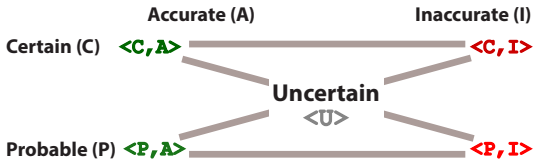

Figure 3: Credibility scale, adapted from Saurí et al. (2009)

Phase 4: Credibility Assessment of Topical Events (HC)

In this step we gather all the topics which were agreed upon as **Events** in the previous phase and recruit Turkers to rate their accuracy. We need to address two important factors while designing the credibility annotation task: 1) the scale of annotation, and 2), the correct number of workers to recruit per item so as to get reliable and quality annotations.

Determining the Credibility Scale: Credibility assessment bears close resemblance to research done on 'Event Factuality' by the linguistic community (de Marneffe, Manning, and Potts 2012; Saurí and Pustejovsky 2009), where factuality of an utterance is expressed in terms of polarity and modality markers. Building on their work, we represent event credibility as an interaction between two dimensions: *Polarity* which distinguishes among 'Accurate', 'Inaccurate', and 'Uncertain', and *Degree of certainty* which differentiates among 'Certainly','Probably' and 'Uncertain'. Figure 3 shows the interaction between the two dimensions by mapping this into a Square of Opposition (SO) diagram—a diagrammatic representation of a set of logical relationships (Parsons 1999). Thus, event credibility is represented as a tuple <*degree, polarity*>, forming a set of tuple sequences as: {<*Certainly Accurate*>, <*Probably Accurate*>, <*Uncertain*>, <*Probably Inaccurate*>, <*Certainly Inaccurate*>}. In others words, credibility assessment is based on a 5-point Likert scale ranging from '[-2] Certainly Inaccurate' to '[+2] Certainly Accurate.'

Determining Number of AMT Workers Necessary for Quality Annotations: In order for the credibility annotations to be useful and trustworthy, an important criteria is collecting high quality annotations—annotations which are at par with expert level responses. While fact-checking services have successfully recruited librarians as expert information providers (Kriplean et al. 2014), their limited time and availability makes it impossible to scale real-time expert annotation tasks. Moreover, with a small pool of in-house experts and a constant update of Twitter streams, near real-time annotation is infeasible. Crowd-sourced micro-labor markets like AMT are a promising option for addressing these challenges. But can Turkers provide credibility annotations roughly equivalent to those provided by experts? One standard way of addressing the issue of collecting quality re-

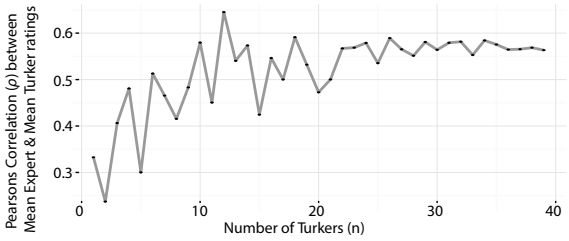

Figure 4: Correlation between expert and Turker ratings. For each number of Turkers ($n \in [1, 40]$) we compute the Pearson correlation ρ between Turker mean responses and expert mean responses. The plot shows average correlations after 10,000 resamplings.

sponses is by redundancy—taking the majority or averaging over multiple independent responses (Surowiecki 2005).

Naturally a question arises as to how many Turker responses will closely approximate an expert's judgment. To determine this number, we piloted our system over a span of 5 days until we finished collecting and annotating 50 events both by Turkers and expert annotators. We recruited reference librarians from our University library as a our expert raters. For each event, we collected 40 independent Turker ratings and 3 independent expert ratings. The web interface shown to librarians were similar to the one shown to Turkers. To estimate the effect of using 'n' Turkers, we randomly sampled 'n' ratings for each annotation item ($n \in \{1, 40\}$). We then took the mean of these n ratings and computed Pearson correlation between Turker mean responses and expert mean responses. We controlled for sampling error via bootstrapping—recalculating the correlations 10,000 times for each n and then averaging over the 10,000 re-computations (Figure 4). The correlation keeps increasing and finally levels off at 30. Hence we fixed the number of workers to 30 to obtain reliable and quality credibility annotations.

Task Design (Credibility Annotation): For designing the annotation tasks (or HITs), we follow the same principles as in the earlier phase. Each HIT had items corresponding to only those topics which were determined as events in the previous step. A Twitter search query box corresponding to the topic is embedded in the HIT (see Figure 5). A worker performing the annotation task has to first click the search box to see the real-time tweets and then choose one of the options from the 5-point Likert scale. Next, the Turker is prompted to enter a reason behind their choice. Asking workers for such free-form subjective responses, while on one hand improves annotation quality, on the other hand adds an extra dimension to our annotation framework. The implications of the collected worker reasonings are discussed in the next section.

Similar to the previous phase, we selectively screen and train workers through a task specific qualification test, requiring workers to score at least 75%. An initial pilot study helped us determine the number of items to post per HIT so as to ensure shorter time durations per HIT (under 3 mins) and lower per-HIT cognitive load. We allowed a maximum of 5 items to be posted in a single credibility assessment HIT.

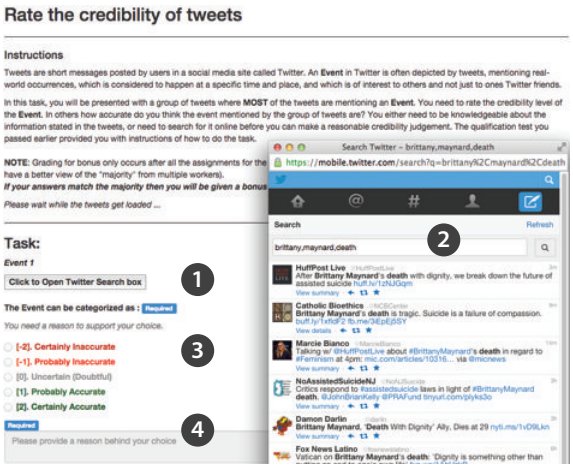

Figure 5: Turker interface for credibility assessment. The numbers correspond to a Turker's workflow. 1. Click the search box. 2. Read tweets from the pop-up Twitter search window. 3. Select one of the credibility scale options. 4. Provide a reason for the selection. Validation checks within the HIT ensure adherence to this workflow.

Phase 5: Collection of Topical Event Streams (MC)

In the final phase, we collected tweets specific to each of the topical events returned by `EventAssess` in Phase 3. Using the Twitter Search API with search queries corresponding to each of these topical events, we collect the most recent tweets, going as far back as the last 7 days—the limit imposed by the search API. Intuitively, this doesn't seem to be a severe limitation because our method tracks recent events as they appear in the stream, followed by their annotation.

CREDBANK Overview

While the purpose of CREDBANK is to enable new research questions to be asked, we next briefly overview the CREDBANK corpus, considering aspects such as agreement among raters, events that group into similar credibility distributions, as well as a brief exploration of the reasons some Turkers gave for their responses.

Agreement among Raters

A total of 1736 unique Turkers participated in our credibility annotation task. We used intraclass correlation (ICC) to quantify the extent of agreement among raters (Shrout and Fleiss 1979). ICC is argued to be a better measure compared to chance corrected measures (e.g., Cohen and Fleiss Kappa) because unlike chance-corrected measures, ICC does not rely on the notion of perfect agreement. Instead, it measures the reliability of ratings by comparing the portion of variation in the data that is due to the item being rated and the variation that is due to raters. If the rater-induced variation exceeds the item-induced variation then the raters are said to have low-inter rater reliability. Moreover, ICC is flexible enough to adapt to different rater agreement study designs. In our study design we have a random sample of 'r' raters rating each event. Hence, we use the Class 2 ICC measure (Shrout

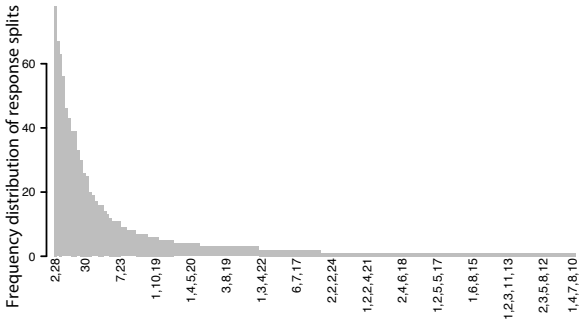

Figure 6: Frequencies of response distributions. The labels on the x-axis correspond to the splits in the annotation categories. For example, the bar at label '1,2,2,4,21' corresponds to the count of events where the annotations are split five ways over the 5-point Likert credibility scale, while the label '30' correspond to the total number of events where all 30 Turkers agreed on a credibility label.

and Fleiss 1979) and obtain a fairly high ICC coefficient (Average Random Raters ICC = 0.77, 95% C.I. = [0.77, 0.81]) indicating high reliability of the collected annotations.

Annotation Response distribution

How often did Turkers agree with one another on a credibility label? To explore this question we study the entire distribution of credibility annotations. We plot the frequency distribution of the splits in the annotation categories (see Figure 6). Hereafter we refer to these as *response splits*; the labels on the x-axis correspond to the different *response splits*. For example, the label '6,7,17' groups the items for which the ratings are split three ways on the 5-point Likert scale.

The long tail of the distribution suggests the difficulty associated in converging on a single credibility rating. In fact, there are a significant number of events where the response splits span the entire 5-point credibility scale. For further examination of these cases, we first divide annotated events into two sets—those where more than 50% of Turkers agreed on a credibility label (*Majority set*) and those where there is no clear majority (*No Majority set*). We next provide closer examination of these groups.

Majority Agreement

We explore the *Majority set* by varying the majority agreement threshold and plotting the percentage of event annotations falling within that threshold. More than 95% of events had 50% Turkers agreeing on a single credibility label, 'Certainly Accurate'. Increasing the majority threshold results in rapid drop in the agreement percentages, with 76.54% of events having 70% Turker agreement, while only 55% of events had 80% Turker agreement. All 30 Turkers agreed on only 2% of events being 'Certainly Accurate'. In other words, considering moderate threshold of inter-rater agreement (70% majority threshold), an alarming 23.46% of events were not perceived to be credible. An important implication of this finding is the presence of Twitter events where credibility assessment did not converge on 'Certainly Accurate,' hinting at the presence of non-trivial percentages of inaccurate information in Twitter. Figure 7 summarizes these results.

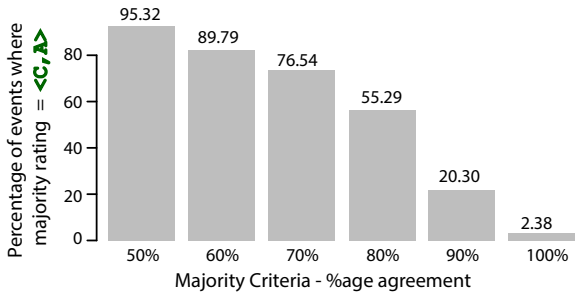

Figure 7: Percentage of events with majority Turker agreement. The majority agreement threshold is varied from 50% to 100% in steps of 10%.

No Majority Agreement

We next examine the cases where 50% of the Turkers did not converge on a single credibility score. In our dataset there are 49 such events, each with 30 annotations, resulting in a total of 1470 ratings. To compare the variations in these ratings we introduce the concept of a *credibility curve*—a histogram of rating frequencies at each credibility score. Are there different types of curves? Is it possible to group these based on their shape and magnitude? We turn to clustering techniques to look for meaningful groups based on shape and magnitude.

Credibility Clusters: In the absence of any prior hypothesis concerning the number of clusters, we apply hierarchical agglomerative clustering to group these credibility curves. The clustering process starts with each annotation item as a cluster and then merges pairs of clusters based on a similarity metric. We used the Euclidean distance similarity metric and Ward's fusion strategy for merging. Ward's method is a preferred strategy because it uses an analysis of variance approach and is very efficient (Ward Jr 1963). Figure 8 shows the complete cluster dendogram. Though hierarchical clustering does not require a predefined number of clusters, it might be useful to partition it into disjoint sets for deeper qualitative comparisons. A preferred rule for partitioning is to set number of clusters to $k \sim \sqrt{(n/2)}$, where n the number of data points (Mardia, Bookstein, and Moreton 2000). Applying this rule, results in approximately four clusters. We qualitatively compare them by plotting their corresponding *credibility curves*. The credibility curve of each cluster is a normalized plot of the rating counts at each credibility label for all events in that cluster. Figure 8 illustrates these curves and their associated event clusters. Here we focus mainly on the trends among these four groups and also highlight a few events in these clusters.

Step Curve The credibility ratings of the events in this group are spread all over the 5-point scale. The shape of the credibility curve further suggests the even split between the two categories: 'Uncertain' and 'Probably Accurate'. A closer examination of these events reveal instances

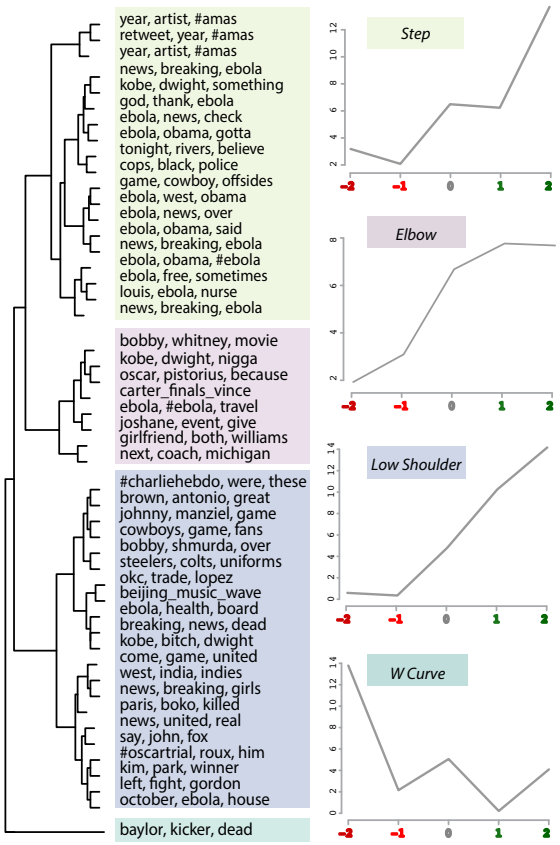

Figure 8: Cluster dendograms for a sample of events and their corresponding average credibility curves.

with high degree of speculation and uncertainty. For example, the topic *kobe, dwight, something* refer to the verbal altercation between the basketball players Kobe Bryant and Dwight Howard in an NBA game, followed by attempts to lip read the words exchanged[7]. We find that the reasons provided by the Turkers also reflect this speculation.

> "Although we see him mouthing what looks liek soft there is no audio and no confirmation from either person."*(Tuker rating: [0] Uncertain)*

More seriously, other examples of events in this group refer to the Ebola pandemic. During the time when we collected data, Twitter witnessed a massive increase in conversations related to the Ebola virus along with several rumors about how the virus spreads, as well as false reports of potential cases of the disease[8]. Our corpus annotations capture these posts along with human judgments of credibility as this event was unfolding on Twitter. Although a third of the Turkers rated the event as accurate, the bump around the 'Uncertain' and 'Probably True' categories suggests the uncertainty associated with this event. Most reasons for the choice of 'Uncertain' were circulating conflicting reports about an event.

> "Conflicting stories on Ebola and what is happening." *(Turker rating: [0] Uncertain)*

Elbow Curve Events in this group had credibility ratings spanning the entire 5-point scale. In this sense, the group has similar characteristics to the *Step Curve* group. Closer investigation reveals events marked with considerable speculations and dynamically evolving over time. For example, the event *oscar, pistorius, becaue* in this group refers to the trial of Oscar Pistorius for the murder of his girlfriend[9]—a seven-month long, emotional trial which was accompanied by speculations regarding the prosecution.

Low Shoulder Curve Events in this group had very few ratings in the 'Certainly Inaccurate' and 'Probably Inaccurate' category, while the other three categories had progressively increasing membership. Closer examination of the reasons entered by Turkers for rating as 'Certainly Inaccurate' revealed that sometimes when they were unable to find a coherent theme in the tweets they would pick this category.

> "All more or less about the same thing with credible sources, but not on any specific event. Scattered reports with different related topics." *(Turker rating: [-2] Certainly Inaccurate)*

However, there were only a few these instances. One such case in our data set relates to the topic *breaking,news,dead*. The time during which these terms surfaced, it referred to the killing of sixteen people at a concert accident in Korea[10]. While the terms are fairly general, it was encouraging to see that most of the Turkers were able to relate it to the Korean concert accident. A few of the annotations, however, reflected the confusion arising from the generality of terms. An interesting observation here are the different strategies used to assess event accuracy. Consider the event *news, breaking, girls*. It is about the kidnapping of 200 Nigerian school girls by an Islamist militant group. The event sparked a global outrage, with speculations of their release after a truce between the Nigerian government and the militant group[11]. Looking at the reasons provided by Turkers while rating this event, we see assessment strategies ranging from simply trusting the source of the information (Lewandowsky et al. 2012) to using suspicion as a cue for judgement (Kim and Levine 2011).

> "219 Kidnapped Nigerian girls to be released, BBC reports "cautiously optimistic." *(Turker rating: [1] Probably Accurate)*

W Curve This was a sparse group with only one event. The event *baylor, kicker, dead* refers to a turning point in the college football match between Baylor and Michigan State when the Baylor's kicker got hit so hard by his opponent that he appeared to lose consciousness. Subsequently Twitter

[7] http://espn.go.com/los-angeles/nba/story/_/id/11783332
[8] http://time.com/3478452/ebola-Twitter/
[9] http://www.washingtontimes.com/news/2014/oct/17/pistorius-sentencing-final-arguments-begin
[10] http://www.aljazeera.com/news/asia-pacific/2014/10/deaths-reported-s-korea-concert-accident-20141017112748873969.html
[11] http://www.bbc.com/news/world-africa-29665165

was flooded with messages expressing concerns whether the kicker had survived the hit. We find that most Turkers were able to confirm this as a false report and a few expressed doubts on its credibility. The event had the most number of ratings in the "Certainly Inaccurate" category, followed by the next highest in the "Uncertain" category.

> Baylor kicker is not dead, he just got hit very hard. *(Turker rating: [-2] Certainly Inaccurate)*
>
> It looks like "Baylor kicker dead" may have been a joke about the kicker's performance. *(Turker rating: [0] Uncertain)*

Future Research Implications

Our central goal in this paper was to systematically study social media credibility and generate a corpus with event information and their associated credibility scores. We envision that this will be a useful data resource for the community to further the study of social media credibility. Next, we sketch a handful of possible future CREDBANK research directions and the contributions they would make to existing work.

Temporal dynamics of credibility. In our dataset we find that certain events in the *No Majority* set re-occur (e.g., events related to Ebola, sporting events, etc.). Is there a difference in the temporal dynamics of these less credible events from the highly credible ones? Our corpus contains timing information at every step of its building process and allows exploration of this question. Are rumors burstier than credible information? Moreover, we can also look for correlations between the temporally ordered annotations and the state of the events around that timestamp. How do annotations vary over time for a rapidly evolving event? Is there any qualitative difference in credibility between temporally persistent and peaky events (Shamma, Kennedy, and Churchill 2011)? CREDBANK should enable investigations like these.

Social and structural dynamics of events across credibility. What role do users of a social network play in spreading information varying in credibility level? How does audience size and the level of information credibility affect information propagation? Our corpus enables delving into these questions. Investigating the following and follower graphs of sets of user posts and their corresponding credibility ratings might be a first step in this direction.

What role does the mainstream media play in online rumors? Studies have demonstrated social media's importance in news production, highlighting several instances where news surfaced in social media before mainstream media reports (Newman 2011). With this in mind, it seems very worthwhile to investigate the role played by mainstream media in the propagation of online misinformation. With the available user profile information in Twitter posts, CREDBANK allows unpacking of these questions.

Could credibility be modeled as a distribution? The long tail of our credibility rating distributions (Figure 6) suggests the nuances associated with finding a single unique credibility label for an item. Perhaps we need to rethink the widely held assumption of the existence of distinct, single-valued credibility labels. CREDBANK's large set of per-item credibility ratings allows future work on probabilistic modeling of credibility.

What are the strategies used to evaluate credibility? Individuals use a wide variety of strategies when assessing information accuracy. Research focused on mental models have found that people often look for coherence in the story, as it is difficult to interpret a single piece of information in isolation (Johnson-Laird, Gawronski, and Strack 2012). We see echoes of this strategy in the reasons provided by CREDBANK's Turkers. We think that this corpus allows a systematic study of strategies used for credibility assessment via CREDBANK's rationales.

Studies in cognitive psychology have demonstrated the tendency of individuals to estimate the likelihood of an event "by the ease with which instances or associations come to mind" (Tversky and Kahneman 1973) (also known as *availability heuristics*). This might result in judgement biases, with people attaching more value to information they can easily recall, such as information that is more emotionally resonant or more recent (Schwarz et al. 1991). Are there any repeated biases associated with human annotators' credibility assessments?

Supplementary data. A byproduct of the corpus-building process are the event annotations—groups of tweets annotated as events or non-events, along with short keyword-based summarizations upon being judged as events. We envision one use of this data may be in event-extraction and summarization systems. Imagine an automatic system which needs to reason about whether a set of posts is coherent enough to be considered as an event, or a system which generates short summaries from a set of event-based posts. CREDBANK could provide valuable ground truth.

References

Allan, J. 2002. Introduction to topic detection and tracking. In *Topic detection and tracking*. Springer. 1–16.

Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on twitter. In *Proc. ICWSM*.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proc. WWW*.

Caumont, A. 2013. 12 trends shaping digital news. *Pew Research*.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2012. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Trans. Dependable Sec. Comput.*

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*.

de Marneffe, M.-C.; Manning, C. D.; and Potts, C. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics* 38(2):301–333.

Friggeri, A.; Adamic, L.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. In *Proc. ICWSM*.

Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. 2014.

Tweetcred: Real-time credibility assessment of content on twitter. In *Social Informatics*. Springer. 228–243.

Johnson-Laird, P.; Gawronski, B.; and Strack, F. 2012. Mental models and consistency. *Cognitive consistency: A fundamental principle in social cognition* 225–244.

Kim, R. K., and Levine, T. R. 2011. The effect of suspicion on deception detection accuracy: Optimal level or opposing effects? *Communication Reports* 24(2):51–62.

Kittur, A.; Chi, E. H.; and Suh, B. 2008. Crowdsourcing user studies with mechanical turk. In *Proc. CHI*.

Kriplean, T.; Bonnar, C.; Borning, A.; Kinney, B.; and Gill, B. 2014. Integrating on-demand fact-checking with public dialogue. In *Proc. CSCW*.

Kwon, S.; Cha, M.; Jung, K.; Chen, W.; and Wang, Y. 2013. Aspects of rumor spreading on a microblog network. In *Social Informatics*. Springer. 299–308.

Lau, J. H.; Collier, N.; and Baldwin, T. 2012. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *COLING*, 1519–1534.

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

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research* 5:361–397.

Liao, Q., and Shi, L. 2013. She gets a sports car from our donation: rumor transmission in a chinese microblogging community. In *Proc. CSCW*.

Liu, F.; Burton-Jones, A.; and Xu, D. 2014. Rumors on social media in disasters: Extending transmission to retransmission. In *Proc. PACIS*.

Maddock, J.; Starbird, K.; Al-Hassani, H.; Sandoval, D. E.; Orand, M.; and Mason, R. M. 2015. Characterizing online rumoring behavior using multi-dimensional signatures. In *Proc. CSCW*.

Makrehchi, M., and Kamel, M. S. 2008. Automatic extraction of domain-specific stopwords from labeled documents. In *Advances in information retrieval*. Springer. 222–233.

Mardia, K. V.; Bookstein, F. L.; and Moreton, I. J. 2000. Statistical assessment of bilateral symmetry of shapes. *Biometrika* 285–300.

Mitra, T.; Hutto, C.; and Gilbert, E. 2015. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proc. CHI*.

Newman, N. 2011. Mainstream media and the distribution of news in the age of social discovery. *Reuters Institute for the Study of Journalism, University of Oxford*.

O'Connor, B.; Krieger, M.; and Ahn, D. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proc. ICWSM*.

Osborne, M.; Petrovic, S.; McCreadie, R.; Macdonald, C.; and Ounis, I. 2012. Bieber no more: First story detection using twitter and wikipedia. In *Proceedings of the Workshop on Time-aware Information Access. TAIA*.

Parsons, T. 1999. The traditional square of opposition.

Petrović, S.; Osborne, M.; and Lavrenko, V. 2010. Streaming first story detection with application to twitter. In *Proc. HLT-NAACL*.

Priedhorsky, R.; Culotta, A.; and Del Valle, S. Y. 2014. Inferring the origin locations of tweets with quantitative confidence. In *Proc. CSCW*.

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

Resnick, P.; Carton, S.; Park, S.; Shen, Y.; and Zeffer, N. 2014. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. In *Symposium on Computation + Journalism*.

Rubin, K. 2012. The ultimate list of email spam trigger words. http://blog.hubspot.com/blog/tabid/6307/bid/30684/The-Ultimate-List-of-Email-SPAM-Trigger-Words.aspx.

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

Schwarz, N.; Bless, H.; Strack, F.; Klumpp, G.; Rittenauer-Schatka, H.; and Simons, A. 1991. Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social psychology* 61(2):195.

Shamma, D. A.; Kennedy, L.; and Churchill, E. F. 2011. Peaks and persistence: modeling the shape of microblog conversations. In *Proc. CSCW*.

Shivajappa, A. N. Top 100 spam trigger words and phrases to avoid. http://www.leadformix.com/blog/2013/09/top-100-spam-trigger-words-and-phrases-to-avoid/.

Shrout, P. E., and Fleiss, J. L. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86(2):420.

Soanes, C. 2011. Keeping it short and tweet. http://blog.oxforddictionaries.com/2011/05/short-and-tweet/.

Soni, S.; Mitra, T.; Gilbert, E.; and Eisenstein, J. 2014. Modeling factuality judgments in social media text. In *Proc. ACL*.

Surowiecki, J. 2005. *The wisdom of crowds*. Random House LLC.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proc. ICWSM*.

Tversky, A., and Kahneman, D. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology* 5(2):207–232.

Ward Jr, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58(301):236–244.

Weng, J., and Lee, B.-S. 2011. Event detection in twitter. In *Proc. ICWSM*.

WordPress.org. Spam words by wordpress. https://web.archive.org/web/20130116191953/http://codex.wordpress.org/Spam_Words.

Zanzotto, F. M.; Pennacchiotti, M.; and Tsioutsiouliklis, K. 2011. Linguistic redundancy in twitter. In *Proc. EMNLP*.

Zipf, G. K. 1949. *Human behavior and the principle of least effort.* addison-wesley press.