On the Wisdom of Experts vs. Crowds: Discovering Trustworthy Topical News in Microblogs

Muhammad Bilal Zafar
MPI-SWS, Germany

Parantapa Bhattacharya
IIT Kharagpur, India
MPI-SWS, Germany

Niloy Ganguly
IIT Kharagpur, India

Saptarshi Ghosh
MPI-SWS, Germany
IIEST Shibpur, India

Krishna P. Gummadi
MPI-SWS, Germany

ABSTRACT

Extracting news on specific topics from the Twitter microblogging site poses formidable challenges, which include handling millions of tweets posted daily, judging topicality and importance of tweets, and ensuring trustworthiness of results in the face of spam. To date, all scalable approaches have relied on *crowd wisdom*, i.e., keyword-matching on the global tweet stream to gather relevant tweets, and crowd-endorsements to judge the importance of tweets. We propose a fundamentally different methodology – for a given topic, we identify *trustworthy experts* on the topic, and extract news-stories that are most popular among the experts. Comparing the crowd-based and expert-based methodologies, we demonstrate that the news-stories obtained by our methodology (i) have *higher relevance* for a wide variety of topics, (ii) achieve *very high coverage* of important news-stories posted globally in Twitter, and (iii) are *far more trustworthy*. Using our methodology, we implemented and publicly deployed a topical news system for Twitter, which can extract news-stories on thousands of topics.

Author Keywords

Topical news; Topical experts; Experts vs. Crowds; Twitter Lists

ACM Classification Keywords

H.1.2 User / Machine Systems: Human information processing; H.3.5 On-line Information Services: Web-based services

INTRODUCTION

Microblogging sites like Twitter have emerged as popular platforms for exchanging real-time information on the Web. Estimates suggest that Twitter presently has more than 500 million users who post 500 million tweets daily on average [4]. These tweets contain a wide variety of information, ranging from conversational tweets to interesting content on specific topics. The users posting these tweets range from news-media sites to domain experts on various topics to spammers and malicious bots. As a result, the quality of information posted on Twitter is highly variable and *finding relevant and trustworthy news-stories on specific topics* from the massive data is a challenging research problem.

Specifically, discovering news-stories relevant to a topic[1] (such as 'politics', 'music', 'neurology' or 'physics') on Twitter poses three key challenges: (i) determining the *topical relevance* of individual tweets (which is especially difficult since tweets are limited to 140 characters) [23, 28], (ii) judging the *relative importance* of the relevant tweets, i.e., ranking the tweets [1], and (iii) ensuring *trustworthiness* of tweets in the face of the growing number and sophistication of spam attacks in Twitter [16, 33]. While several prior research studies attempted to tackle each of the above challenges in isolation, efficiently addressing all three challenges in *real-time* and at the *scale* of hundreds of millions of tweets posted daily presents a formidable barrier in building a practical topical news service. To the best of our knowledge, there are no topical news systems publicly deployed on the Twitter platform today, beyond keyword-matching services.

In this paper, we explore a fundamentally different approach towards building a trustworthy topical news system. In a nutshell, our proposal is to rely on the *wisdom of the experts rather than the crowds*. Specifically, for discovering news-stories related to a topic, we propose to analyze *only* the tweets posted by a small number of experts on the topic [13, 26] (typically, no more than a few hundred to few thousand users) rather than the global Twitter population (numbering over 500 million users [4]). Further, we propose to judge the relevance and importance of an extracted news-story to the specified topic by simply counting the number of experts (on that topic) who posted on the story. The more the

[1] Our interpretation of a 'topic' is detailed in the next section. Also, the term 'news-story' – which refers to a set of related hashtags, and the tweets which contain these hashtags – will be described in the fourth section.

CSCW '16, February 27-March 02, 2016, San Francisco, CA, USA.

ACM 978-1-4503-3592-8/16/02 $15.00
DOI: http://dx.doi.org/10.1145/2818048.2819968

experts who tweet a news-story, the more relevant and important the story is to the topic.

However, our simplistic design choices raise a number of intriguing questions:

(1) *For a given topic, can a small number (few hundreds to few thousands) of topical experts capture the most important topical news-stories that are circulating amongst the global Twitter population (over 500 million users)*? If experts do post about most of the important news-stories, then relying only on the small number of experts' tweets would eliminate the large scalability barrier for building a topical news service.

(2) *Is counting the number of topical experts that posted on a certain news-story sufficient to judge the relevance and importance of that news-story to the topic?* Does this simple technique perform as well as standard techniques based on content analysis (e.g., keyword-matching or query expansion [43]) and crowd-endorsements for ranking?

(3) *Are tweets posted by experts trustworthy by default?* After all, the experts have an incentive to maintain their reputation on Twitter and in a sense, they could be viewed as *whitelisted* users in the system. Could this obviate the need for separate and elaborate techniques to fight spammers in the crowds when discovering topical news?

To answer these questions, we implemented two methods for retrieving topical news-stories – one based on experts' tweets, and the other based on tweets posted by the global Twitter crowds – and compared their relative performance.

To extract topical news from experts, the key is to find a set of authoritative and trustworthy experts on the specified topic. For this, we rely on the methodology developed in our prior works [10,13], which leveraged the Twitter Lists feature [3] to identify topical experts in Twitter on a wide variety of topics. However, these prior works do *not* guarantee that the identified experts are trustworthy, and are vulnerable to attackers creating fake Lists [13]. So, in this work, we propose a methodology to curate (i.e., whitelist) a set of *trusted* experts, and then use the trusted set of experts to retrieve news-stories. Henceforth, we refer to this methodology as *What Is Happening* (`WIH` in short), after the name of the topical news service that we developed using this methodology (see below).

For extracting topical news from the crowds, we used the official Twitter search (**`http://search.twitter.com`**) that offers a *keyword matching* service over the global tweet stream, to gather all tweets which contain the specified topic-word (e.g., for the specified topic 'politics', we gather all tweets containing the term 'politics').[2] Since keyword-match based methods would miss tweets that are relevant to the given topic, but do not contain the specific topic-word, we also employed *query expansion techniques* [43] to improve the coverage of extracted tweets. We then judged the relative importance of the news-stories extracted from the collected tweets based on the number of users in the crowd who endorse (i.e., tweet or retweet) the news-stories. Note that, since in these crowd-based approaches, all tweets are obtained via the official Twitter search, their trustworthiness is reflective of the state-of-the-art spam defenses (as deployed by Twitter) [33]. We refer to the two crowd-based methodologies as `crowd-topic` (which uses tweets containing the topic-word) and `crowd-expanded` (which uses tweets collected through an expanded query).

Our comparative analysis of expert-based and crowd-based topical news services investigates the questions stated above, and highlights the benefits of relying on the superior wisdom of experts. We summarize our key results below:

1. Coverage of important topical news-stories: Though the expert-based methodology (`WIH`) is limited to several times fewer tweets than crowd-based methodologies (`crowd-topic` and `crowd-expanded`), the tweets posted by experts contain more than 90% of all important topical hashtags (which correspond to the important news-stories) extracted by the crowd-based methodologies. On the other hand, though the crowd-based approaches use keyword matching over the complete tweet stream, they fail to extract some important news-stories discovered by the expert-based methodology. Further, while query expansion helps to improve the coverage of news-stories drawn from crowds, it significantly degrades the relevance of the top news-stories (details in the section on Evaluation).

2. Relevance of top news-stories: More than 82% of the top 10 news-stories extracted by `WIH` are topically relevant (as judged through human feedback), compared to 73% of the top 10 news-stories returned by `crowd-topic` (and far less for `crowd-expanded`). This validates our intuition that content posted by *multiple* experts on a topic is very likely to be relevant to the topic. In fact, the crowd-based approaches (relying on keyword-match) are found to suffer from issues such as word sense disambiguation [31] and query-drift in the process of query expansion [22] (details in the Evaluation section).

3. Trustworthiness of extracted news-stories: Unlike the expert-based `WIH`, we find that the crowd-based `crowd-topic` and `crowd-expanded` occasionally extract news-stories that are promotional and spam campaigns. Despite the state-of-the-art spam defenses deployed by Twitter, we find that the tweets collected by our crowd-based methods (using official Twitter search) still contain tweets posted by thousands of spam-accounts (that are later suspended by Twitter). Our results highlight the fundamental difficulties with eliminating spam from crowd posts in real-time [16, 33]. In contrast, `WIH`, relying on a whitelisted set of trustworthy experts, does not need to employ any additional mechanisms to combat spam.

In summary, this paper makes three important contributions: (1) We adopt a fundamentally different approach of relying

[2]We considered applying advanced techniques to judge topicality of individual tweets, such as topic models [28] and mapping tweets to semantic concepts such as Wikipedia pages [23]. However, no prior work has implemented these computationally intensive techniques in real-time on the hundreds of millions of tweets posted daily.

on the wisdom of few topical experts rather than the crowds, for extracting topical content from Twitter. We show that our approach (despite its simplicity) offers as good, if not better, performance than using more advanced techniques over significantly larger data (crowd wisdom). (2) Our work contributes to the long-standing debate about the wisdom of crowds versus experts [12, 14, 27] in the context of social media. The competitive performance of WIH indicates that expert users act as curators of important information on their domains of expertise, and suggests leveraging experts' wisdom while designing future information retrieval systems in social media. (3) Finally, we implemented and publicly deployed the first (to our knowledge) system for extracting topical news from Twitter, that can return interesting news-stories on tens of thousands of topics, ranging from generic topics like 'music' and 'politics' to niche topics like 'neurology' or 'astronomy'. Interested readers are invited to test the system, named What Is Happening, at **http://twitter-app.mpi-sws.org/what-is-happening/**.

To illustrate the quality of news-stories retrieved by our system, we compared the top results of our system with the top tweets returned by the official Twitter search service. The news-stories of our system were judged by human evaluators to be at least as good as (and frequently better than) the Twitter top results for a large majority of the topics. Our deployed system also includes a *Twitter-based daily newspaper*, which retrieves news-stories on the topics covered in a traditional news media site (e.g., NYTimes).

GOAL: A TOPICAL NEWS SYSTEM

Our goal is to design a practical methodology / system for retrieving topical news-stories from Twitter. This section defines our goal in detail and contrasts our proposed news system with existing news systems, in order to bring out the challenges in designing such a system.

Comparison with existing news systems: All popular news media organizations (e.g., NYTimes, BBC, CNN) today maintain their websites as well as Twitter accounts. However, there are two important differences between these media systems and the topical news system we aim to develop. First, the news-stories reported by the websites / accounts of media organizations are *manually* curated by experienced journalists, whereas we propose to curate topical news stories *automatically*. Second, most news-stories reported by the traditional media sites are related to a few broad topics of popular interest (e.g., politics, sports, economy, science). On the contrary, our goal is to develop a news-system that can retrieve news-stories on a wider variety of topics, including topics of niche interest. For instance, we wish to retrieve news-stories not only on 'science' but also on 'neurology' and 'astrophysics' – which are rarely covered by traditional media organizations.

Note that there have been efforts towards automatic curation of 'breaking news' (i.e., news that is most popular / important at a certain point of time) in Twitter. Sankaranarayanan *et al.* [29] developed a system to capture tweets that correspond to breaking news. Also, Twitter itself periodically declares a set of 'trending topics' which are the keywords that are being most discussed in Twitter at that point of time [21]. However, these approaches do *not* attempt to provide news on any specific topic, which the present study aims to do. It can be argued that a simple way to get topical news is to infer topics for the breaking news / trending topics identified by the above approaches. However, such globally popular news-stories are almost always related to only few topics such as music, entertainment and sports (as observed in [21]). Hence, such an approach would again cover very few topics, which the media organizations already cover.

Topics that are covered by the news system: Since we propose to develop a news system over Twitter, we plan to cover any topic that is of interest to Web users. One of the most comprehensive set of such topics is available in the Open Directory Project (**http://www.dmoz.org/**), which is a topical directory of web-pages containing thousands of diverse topics arranged in a hierarchical tree structure. For instance, some of the top-level topics in this tree are 'arts', 'health', 'science', 'sports', and some of the more specialized topics in lower levels of the tree are 'dance', 'autism', 'neurology', 'softball', etc. Our goal is to develop a news-system which can return news-stories for topics such as those in ODP. Note that ODP directory has thousands of topics and is compiled by manual categorization of web-pages by thousands of human volunteers. In contrast, we show later that our proposed methodology is able to automatically retrieve news-stories for a very large fraction of the ODP topics, which is far larger than the set of topics covered by traditional media organizations.

Topics that are *not* covered by the news system: As specified above, our goal is to retrieve news-stories on relatively broad topics such as those in the ODP directory. We do *not* attempt to answer queries on specific events such as "Thorpe return in 2012 Olympics" or "phone hacking British politicians" (which are examples of queries in the TREC microblogging track [36]). In fact, our goal is to help users know about the recent and important news-stories on topics of their interest, e.g., a person interested in the topic 'politics' or 'technology' can use our methodology to know that there is a recent news-story about hacking phones of British politicians.

A prior study [32] identified that Twitter is mostly used for three types of information needs – (i) information on specific topics, ranging from popular topics such as 'technology' and 'science' to specialized topics such as 'digital forensics' and 'astronomy' [32], (ii) timely or recent news on specific topics and events, and (iii) social information, such as public opinion about a celebrity. Our proposed methodology caters to the first and partly the second types of information needs.

BACKGROUND & RELATED WORK

As Twitter is increasingly used to obtain real-time news on various topics [32], finding important and trustworthy topical news has become an important research problem. The challenges are to identify information relevant to a specific topic, to judge the importance of information, and to ensure the trustworthiness of the results in face of abuse and spam. A

number of prior studies have investigated the above aspects, some of which we discuss in this section.

Extracting topical information from Twitter: There have been prior research attempts to identify topical experts [13, 26], and infer the topicality of individual tweets [20, 23, 28, 45]. In this paper, we attempt to extract topical news-stories from the tweets posted by the topical experts, which is a non-trivial task even after identifying topical experts, because the experts often post conversational tweets that are not related to their expertise [30, 41].

Specifically, inferring the topicality of tweets is especially difficult because of the very small size of tweets. Methodologies such as topic models [28] or mapping tweets to semantic concepts [20,23] have been found to give accurate results, but it is unclear whether these computationally intensive methodologies can be scaled to characterize hundreds of millions of tweets in real-time. This is possibly why all commercially deployed information retrieval systems over Twitter [34, 37, 38] as well as most research studies [11] use simple *keyword-matching over the Twitter stream* to identify tweets that are potentially relevant to a given topic.

Inferring the importance of tweets: Another challenge is to estimate the importance of individual tweets. While the general (topic-agnostic) importance of a tweet can be judged by metrics such as number of retweets, or the popularity of the user who posted the tweet [1], it is unclear how to judge the importance of a tweet *for a particular topic* (which is crucial for a topical news system). Again, it has been observed that the most crowd-endorsed tweets may not be the most informative or important tweets on the topic [2]. This calls for novel techniques to estimate the importance of information in Twitter relative to a specific topic.

Trustworthiness of information: There exist a large number of spammers in Twitter [16, 33] who adopt intelligent techniques to pollute the results given by search systems, such as posting tweets containing trending keywords and (unrelated) spam URLs [8]. Though the spam problem in Twitter is well-known, none of the prior works on extracting topical content from Twitter [11, 25, 29] specifically attempted to ensure the trustworthiness of the results. In effect, these studies rely on the Twitter stream to be free from spam URLs; however, other studies [14] have reported large amounts of spam in the tweet streams provided by Twitter, and the limited utility of standard URL blacklists in removing such spam [16].

In this work, we ensure the trustworthiness of results by relying only on tweets posted by a relatively small set of topical experts, whom we carefully curate to ensure that no spammers are included in this set. For this curation, we start with the set of 'verified users' in Twitter, and use the TrustRank algorithm [17] (that was designed to identify trustworthy webpages) to transfer trust to other users, and thus identify other trustworthy experts. Our methodology is similar in spirit to that of the prior work [18] which identified trustworthy users based on whether trusted users communicated with them; however, we identify trusted users based on whether they have been *listed* by known trusted users.

Note that another advantage of relying on tweets posted by authoritative topical experts is that the news-stories / tweets returned by our methodology would score highly with respect to some of the factors which affect the credibility of tweets, such as the reputation of the user who posted the tweet [24].

Solving all challenges together: As stated above, each of the challenges of judging topical relevance, importance, and trustworthiness of tweets, has been addressed individually by various studies. However, no prior study has attempted to solve all the challenges *simultaneously* and *at scale* for retrieving topical news-stories from Twitter. In this work, we attempt to address all these challenges by relying on relatively few topical experts, instead of the global Twitter population. Hence, this study offers an interesting take on the long standing expert-vs-crowd debate.

Sampling experts versus crowds: Debate on the expert-versus-crowd question has been continuing for several decades, but has not been fully settled yet. Wisdom of the crowds has been found to be superior to expert wisdom in applications such as understanding trends in financial stocks [12], or taking decisions in the medical domain [27]. On the other hand, expert opinion has been found to be extremely valuable in *web-based applications*, such as designing recommendation systems for movies [5]. A recent study on social investment platforms [42] observed that the content contributed by experts helps to better predict the stock market performance than the content produced by the crowd. Another recent study [14] by us also showed that tweets posted by topical experts in Twitter contain more topical information than similar numbers of tweets posted by crowds. The present work is, in fact, motivated by [14], and the observations in this work re-affirm those of [14] on the utility of the content posted by experts. Note, however, that the prior study [14] compared tweets posted by experts with the 1% Twitter random sample, and that the comparison was *not* made in the context of any concrete application. The present work actually compares the utility of content posted by all experts and by the masses in the context of a specific application namely, the topical news system. We not only show that the content posted by the experts is better, but we also develop and publicly deploy a topical news systems based on the expert wisdom.

TOPICAL NEWS SYSTEM DESIGNS

This section presents the methodologies for utilizing the wisdom of topical experts and the crowd, to discover news-stories relevant to a specified topic.

Utilizing experts' wisdom

At a high level, our methodology (which we refer to as `WIH`) relies on two key intuitions. First, experts on a topic are more likely to post important information related to the topic, as compared to ordinary individuals (crowd). However, along with topical information, individual experts also frequently post day-to-day conversation [30, 41] (as we quantitatively

Topic	Some experts identified by List-based method
Music	Lady Gaga (48.5M), coldplay (15.1M), Dallas Martin [SVP of A&R Atlantic Records] (26.5K), TenorRyan [Opera singing road warrior] (1.8K)
Politics	Barack Obama (61.7M), Al Gore (2.8M), NPR Politics (1.9M), BristolRed [Bristol Labour Party] (3.2K), Scott Fluhr [Harrison County GOP Chairman] (625)
Environment	TreeHugger.com (323K), GreenPeace USA (136K), Dennis Dimick [environment editor @natgeomag] (6.6K)
Physics	CERN (1.2M), Institute of Physics (94K), astroparticle (24.5K), Fermilab Today (763)
Neurology	Oliver Sacks (101K), Neurology Today (19.8K), AAN Public (6.1K), MNT Neurology News (5.3K)
Geology	geosociety (20.7K), Kim Hannula [Structural geology professor] (1.6K), Dave Mayer [Planetary GIS/Data Specialist] (1.4K)

Table 1: Examples of topical experts on Twitter, identified by a List-based methodology developed in our prior works [10,13]. The experts are indicated by the real names or the screen-names as given in the profile of their Twitter accounts. The methodology identifies experts having a wide variety of popularity. For some of the less popular experts, extracts from their Twitter account bio are given within square braces. The numbers in parentheses give the approximate number of followers (K: thousand, M:million).

show later in the paper). Hence, we apply our second intuition – information that is posted by *multiple experts* on a given topic are much more likely to be relevant and important to the topic – to extract topical news-stories. Further, we observe that different experts post different tweets even when discussing the same news-story. So, to identify news-stories discussed by multiple experts, we *cluster similar tweets* into news-stories, and then rank news-stories according to their importance.[3] These steps are detailed below.

Identifying trustworthy topical experts

Methodology to identify topical experts: Given a topic, we first need to identify a set of experts on that topic. In various social media, there are different mechanisms which can be exploited to identify topical experts, such as 'Lists' in Twitter [10, 13, 41], 'Skills' in LinkedIn [7], and so on. In this study, we focus on Twitter. Prior works [13, 41] have shown that the topical expertise of individual Twitter users can be accurately inferred using the *Twitter Lists feature*. Lists are an organizational feature, by which users can group experts on topics that interest them [3]. To create a List, a user specifies a name and an optional description, and then adds other users as members of the List; for instance, a user can create a List named "Music and musicians", and add accounts such as Lady Gaga, Katy Perry, Yahoo Music to the List.

We leverage the List-based methodology developed in our prior works [10, 13] to identify experts on a given topic. Briefly, we considered a Twitter user to be an expert on a topic (e.g., politics) if and only if the user has been listed at least 10 times on that topic, i.e., if the topic-word ('politics') appears at least 10 times in the names and descriptions of the Lists containing this user as a member. We considered as topics, only unigrams (e.g., 'politics', 'music') and bigrams (e.g., 'social media', 'video game') which were identified as nouns and adjectives by a standard English parts-of-speech tagger (as done in [10, 13]).

A key advantage of this List-based methodology is that, since it relies on crowd-sourced social annotations, it helps to identify experts on a large and diverse set of topics that are of interest to users, including popular topics (such as 'politics' and 'music') as well as more specialized topics (such as 'neurology' and 'physics') [10]. Table 1 shows some example topics and some of the Twitter users identified as experts on the topic. The table also gives the number of followers of the experts in Twitter (as of July 2015). Note that the identified experts include not only globally popular users having millions of followers (e.g., Barack Obama, Lady Gaga), but also less popular ones having a few hundred to a few thousand followers. For the less known experts, we also give extracts from their Twitter account bio in Table 1. Hence, our expert-based methodology will be able to leverage content posted by authoritative experts on a given topic, independent of their popularity.

Ideally, we would like to identify all topical experts in Twitter. However, due to Twitter API restrictions, we could only gather data of the Lists created by or containing the first 50 million users to have joined Twitter. We then utilized the List-based methodology to identify the topical experts from this set of users – this resulted in 1,276,201 topical experts (who are listed 10 or more times on some topic).

Note that our prior studies [10,13] considered *any* user who is listed 10 times or more on a topic to be an expert on the topic. However, this methodology does *not* guarantee the trustworthiness of the experts – malicious users can easily create 10 or more fake Lists (List-spam) containing a certain user, and this user would be inferred as an expert. Hence, to ensure trustworthiness of the news-stories, it is essential to curate a trustworthy set of experts.

Curating a trustworthy set of experts: We now describe how we ensure the trustworthiness of this large set of experts. To establish the trustworthiness of the above set of experts, we use the TrustRank algorithm [17] that was designed to identify trustworthy web-pages and combat Web spam. TrustRank's basic assumption is that good web-pages mostly link to other good web-pages. Hence, TrustRank starts by assigning high trust scores to a seed set of known good pages, and then propagates trust to pages which are linked from the good pages, in a way similar to Pagerank.

For applying TrustRank, we constructed a *List network* among all the Twitter users whose data we could gather (as stated above). The List network is a directed graph where users are nodes, and the link $u \rightarrow v$ is present if user v is included in a List created by user u. Note that our List network satisfies the TrustRank assumption; good users mostly List other good users as experts on topics of interest to them.

[3]Even among general Twitter users, different users post different tweets on the same news-story. Hence we use the same approach while utilizing crowd wisdom (as described later in this section).

Then, as the seed set of trusted nodes, we considered *verified accounts* in Twitter [40]. Twitter verification is an exclusive badge that establishes the authenticity of highly popular Twitter accounts. We found 83,852 verified users in our List network, and we considered these as the initial set of trusted nodes. We then applied TrustRank (with decay factor 0.85) on this List network to rank all nodes according to their trustworthiness.

We observed that nearly 94% of our identified experts fall within the top 20 percentile most trusted nodes, while 99% lie within the top 38% node ranks. This suggests that there is little List spam in the Twitter network today. However, this might change in the future, when List-based methods are used to retrieve information. Since almost all of our current 1.27 million experts rank very highly in TrustRank, we decided to *whitelist* all of them for our news system.

Note that existing retrieval systems that rely on the content posted by the global Twitter population, have to continuously guard against spam. This is especially difficult because blacklisting services (e.g., URL blacklists and Twitter suspension process) are too slow and never exhaustive [16,33], and spammers also develop new techniques to evade such defenses. On the other hand, our approach is based on a small set of *a priori whitelisted* expert user-accounts, and this set could be updated periodically (say once every week or month).

Finally, the trustworthiness of a news system also depends on whether the news-stories contain *rumors or misinformation*. Relying on trustworthy experts could provide a natural defense against misinformation. First, being popular topical authorities, experts are inherently less susceptible to rumors / misinformation on their topics of expertise. Second, as we explain below, the top WIH news-stories are those that are discussed by *multiple* experts on the specified topic and they contain the opinions of the different experts; so unless multiple experts promote a false rumor on the topic (a very unlikely event), the rumor will *not* come up in the top news-stories for the topic and even if it does, it would contain the divergent opinions on the trustworthiness of the rumor.

Clustering tweets into news-stories

We gather the tweets posted by the topical experts, from which we now need to extract important news-stories for the given topic. We refer to the set of tweets posted by the experts on a given topic as the *expert digest* on that topic.

As stated earlier, our key idea is that content posted by *multiple* experts on a given topic, is more likely to be relevant and important to that topic. However, since different experts post different tweets even when discussing the same news-story, we cluster similar tweets into news-stories.[4] A number of advanced text-clustering algorithms have been proposed [9]; however, given the enormous number of tweets that a news-system on Twitter needs to handle in real-time, we opted for a simple clustering methodology. Since conversations in Twitter on particular news-stories are known to be focused around specific hashtags [44], we cluster tweets based on common hashtags contained in them.

For a given topic, we extract the hashtags from the expert digest for the topic and then cluster similar hashtags to form *hashtag-clusters*. We use a simple bottom-up approach, which starts with clusters containing one hashtag, and two clusters are merged if the Jaccard similarity of the tweet-sets containing the hashtags in the two clusters is higher than 0.5. We then group tweets on the basis of the hashtag-clusters, i.e., we group tweets containing similar hashtags, and refer to them as *tweet-clusters*. Thus a tweet-cluster, which we refer to as a 'news-story', is a 2-tuple (H, T) where H is a set of related (frequently co-occurring) hashtags, and T is the set of tweets which contain at least one hashtag from H.

Ranking news-stories by topical importance

The next step is to rank the tweet-clusters (news-stories) based on their importance relative to the specified topic. We follow the intuition that *if more experts are tweeting about the same news-story, then that news-story is more relevant and important to the community interested in the said topic*. So, for each tweet-cluster $(H_i, T_i), i = 1, 2, \ldots$, we obtain the set of distinct experts E_i who have posted the tweets in T_i, and rank the tweet-clusters based on $|E_i|$; in case of a tie, we rank them again based on the total number of tweets in that cluster, i.e., on $|T_i|$.

Note that the above intuition fails in a particular case. When a *globally important event* occurs, such as a natural calamity or a global sports tournament such as the Soccer World Cup, then many experts discuss news-stories about that event, irrespective of their individual topics of expertise. To guard against this, we decided to ignore news-stories which appear among the top results for many topics (analogous to stopword removal in information retrieval). Concretely, we ignored news-stories which appear among the top 25 results for more than 10 topics among the 25 diverse topics stated in Table 3 (which are later used for evaluation of the methodologies).

Topics covered by WIH

As stated earlier, our goal is to retrieve news-stories for topics such as those in the Open Directory Project (ODP), which is expected to include most of the topics that users of online systems are interested in. We now see whether the expert-based methodology (WIH) can indeed return news-stories for a large fraction of the topics in the ODP. The set of topics for which we could identify at least one expert – i.e., the set of topics for which the expert-based methodology can potentially return news-stories – is referred to as the *experts-topic* set.

Both the ODP and expert-topics contain a large number of non-English topics (e.g., words in other languages, names of celebrities). We decided to focus on English topics only[5], and found that out of 23,157 English topics in the ODP, as

[4]Note that clustering similar tweets into news-stories has additional advantages like reducing redundant information, and increasing the diversity in the top results.

[5]We consider only those topics which occur in the `wamerican-english-small` dictionary, or are combinations of words which appear in this dictionary (e.g., 'video game').

Topic	Additional terms in expanded query
baseball	game, #baseball, team, players, fan
golf	#golf, #golfjp, course, club, day
law	order, enforcement, #law, firm, school
politics	obama, #politics, border, money, news

Table 2: Examples of expanded queries, obtained through pseudo-relevance feedback from the original query (which is the topic-word itself, shown in the first column).

# experts	Example Topics
$>$ 20K	music, tech, politics, food, health
10K–20K	fashion, wine
5K–10K	books, government, beer, law
500–5K	environment, baseball, golf, hollywood, history, iphone, religion, psychology, astronomy
$<$ 500	astrology, theology, geography, neurology, malaria

Table 3: 25 example topics used for evaluation, along with the number of experts in each topic. The topics include both popular and niche ones.

many as 18,941 (82%) topics are included in the set of expert-topics. In particular, out of the 2,070 (relatively broad) topics in the top three levels of the ODP, 1,948 (94%) are covered by the expert-topics. The remaining 6% topics (not covered by expert-topics) are very niche topics such as 'fine-chemical intermediates' and 'pet food preparation'.

Thus, the expert-based methodology can potentially retrieve news-stories for a very large majority of the topics that are likely to be of interest to Web users.

Utilizing crowd wisdom

We wanted to compare the expert-based methodology with methodologies based on crowd wisdom, i.e., tweets posted by the global Twitter population. However, we could not find any existing methodology / system that extracts topical news-stories from Twitter, with which we could directly compare our results. So, we ourselves implemented two crowd-based strategies for retrieving topical news-stories.

Gathering the crowd wisdom for a given topic: Since it is unclear whether computationally intensive methods such as topic models [23, 28] can be used to judge topicality of hundreds of millions of tweets in real-time, all deployed mechanisms for topical retrieval from Twitter (e.g., the official Twitter search `http://search.twitter.com`) rely on keyword-matching to gather tweets relevant to a topic. Hence, we decided to use two approaches based on keyword-matching to collect the wisdom of the crowds for a given topic.

(i) `crowd-topic` – *Twitter search with topic-word:* For a given topic, we use the topic-word itself (e.g., the word 'politics' or 'neurology') as a query to the Twitter search service, and collect all tweets returned by the service.

(ii) `crowd-expanded` – *Twitter search with expanded query:* One of the primary challenges in collecting tweets relevant to a given topic is that many relevant tweets do not contain the topic-word (primarily because of the extremely small size of tweets). To counter this problem, we use *query expansion through relevance feedback* which has been used for Web mining [43] as well as in Twitter (e.g., see the TREC microblogging track [36]). Since it is infeasible to get explicit relevance feedback (via human assessors) for millions of tweets, we used pseudo (blind) relevance feedback – we first used Twitter search with the given topic-word, and selected the 5 most common terms in the resulting set of tweets. Then, we added these terms to the topic-word to get an expanded query, and used Twitter search to gather all tweets containing any of the terms in the expanded query. Table 2 shows examples of expanded queries for some of the selected topics.

Clustering tweets into news-stories and ranking news-stories: Even in the case of the crowds, different users post different tweets on the same news-story (as we had observed for experts). Hence, we apply the same clustering and ranking methodologies on the crowds' tweets as we did for the experts' tweets, to retrieve and rank news-stories. Thus, we utilize the wisdom of the crowds (general Twitter population) in both gathering the set of tweets as well as to rank the news-stories.

Note that the three methodologies – expert-based `WIH`, `crowd-topic`, and `crowd-expanded` – differ among themselves *only* in the methodology of collecting the tweets. The techniques for clustering, ranking and removing globally popular news-stories are exactly the same for all the methodologies.

EVALUATION

We now evaluate the performance of the expert-based methodology and the crowd-based methodologies on the following key aspects: (1) **Compactness:** How compact is the expert digest for a topic, as compared with the set of tweets gathered by the crowd-based methodologies? (2) **Relevance:** For a given topic, which methodology succeeds better in retrieving relevant news-stories? (3) **Coverage:** Does the expert-based methodology discover the important news-stories related to a given topic? How does its coverage compare with the coverage of tweets retrieved through keyword matching over the global tweet stream? (4) **Trustworthiness:** Do the set of tweets used by the different methodologies contain spam? (5) Additionally, we directly compare the top results returned by the various methodologies. All evaluations are done over news-stories identified by the different methodologies on a single day (July 8, 2014).

Selecting topics for the evaluation: We chose a set of 25 topics for the evaluation, which are shown in Table 3, along with the number of experts identified for each topic. Also, Figure 1 shows the distribution of the number of experts for all the topics on which we could identify experts, where the selected 25 topics are marked. It is evident that we chose the 25 topics such that they span the entire range of popularity – i.e., the selected topics include very popular topics (e.g., 'music' and 'politics') having tens of thousands of experts, as well as niche topics having few hundreds or even few tens of experts (e.g., 'malaria', 'neurology', 'geography').

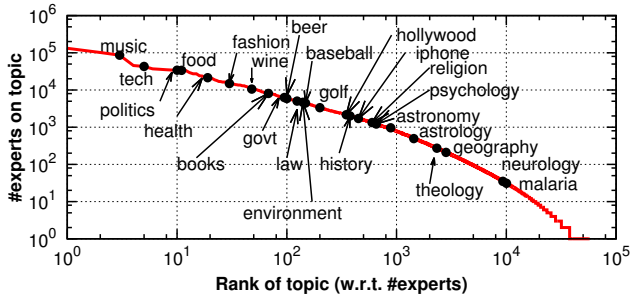

Figure 1: Distribution of the number of experts for various topics. Topics in Table 3 are shown using points.

Topic	WIH	crowd-topic	crowd-expanded
music	393K (39K)	768K (357K)	1466K (642K)
iphone	8K (0.71K)	907K (544K)	1105K (556K)
geography	0.7K (0.11K)	3K (3K)	1014K (668K)
malaria	0.03K (0.01K)	2K (1K)	1244K (609K)
neurology	0.1K (0.02K)	0.46K (0.32K)	542K (158K)

Table 4: Number of tweets and distinct users (in parentheses) who posted these tweets, in the three digests (K: unit of thousands).

Compactness

We first compare the size of the expert digest for a topic (i.e., the set of tweets posted by the experts on that topic) to that of the crowd digests (obtained via keyword matching over all tweets) for the same topic. Table 4 shows the number of tweets and the number of distinct users who posted these tweets (in thousands), in the digests for the three methodologies for some of the selected topics. It is evident that the expert digests contain tweets posted by only few hundreds to few thousands of experts, whereas the crowd-based methodologies have to often process hundreds of thousands of tweets posted by thousands of users.[6] From a practical system design perspective, the compactness of the expert digest makes it significantly easier to gather and analyze the data in real-time.

Relevance

We next analyze the relevance of the top news-stories returned by the different methodologies, where we consider the 10 top-ranked news-stories as the top results returned by each methodology. Specifically, we are interested in answering the question – *can simply ranking news-stories by the number of topical experts discussing the story (without any filtering for topical keywords) perform as well as crowd digests collected through keyword-matching and query expansion techniques?*

Methodology for evaluating relevance

Since judging the topical relevance of a tweet / news-story is inherently subjective, we judged relevance through human feedback. Similar to prior attempts to evaluate the performance of Information Retrieval systems (e.g., [15] or studies in the TREC Crowdsourcing Track [35]), we used the Amazon Mechanical Turk (AMT) service[7] to recruit human volunteers to assess the topical relevance of tweets / news-stories. To obtain more accurate judgements, we used 'AMT master workers' who are known to perform such tasks more accurately [6]. During the survey, an AMT evaluator was shown a topic and a news-story (i.e., a set of related hashtags, and an illustrative tweet containing some of these hashtags, as shown by the WIH system), and was asked to judge whether the news-story is relevant to the given topic. The evaluator gives her feedback by choosing one from the three options – Relevant / Not Relevant / Cannot say. Each tweet / news-story was evaluated by five different master workers, and we determined the relevance of a tweet / news-story based on the majority opinion amongst the five evaluators.[8]

Relevance results

Table 5 shows the relevance judgments for the top 10 news-stories returned by the three methodologies – WIH (3^{rd} column), crowd-topic (4^{th} column) and crowd-expanded(5^{th} column) – for the 25 selected topics. Also shown are relevance judgments for *50 randomly selected tweets* from the expert digest of each topic, i.e., randomly selected tweets posted by experts on a given topic (2^{nd} column). The table also gives the mean percentage of tweets judged relevant, and the nDCG (normalized discounted cumulative gain) scores for the top 10 results across all 25 topics.

Relevance of random tweets from topical experts vs. top WIH news-stories: On average, only about 53% of the tweets in the expert digest for a topic are relevant to the topic (Table 5, 2^{nd} column). This reinforces the fact that experts frequently post tweets that are not related to their topics of expertise (as also observed in [30,41]). Hence, even after identifying experts, extracting topically relevant content from their tweets remains a non-trivial challenge.

Table 5 shows that, for almost all topics (except for a few niche topics having very few experts, which are discussed later in this section), the relevance scores for top WIH news-stories are much higher than those for tweets randomly selected from the expert digest. In fact, the mean relevance score (considering all 25 topics) rises from 52.8% (random experts' tweets) to 82.6% (top WIH news-stories) showing the utility of the proposed ranking and clustering methodology.

Relevance of top WIH news-stories vs. top news-stories of crowd-based methodologies: Table 5 shows that, for 18 out of the 25 selected topics, the top WIH news-stories are at least as relevant, if not more, than the top news-stories obtained by the crowd-based methodologies. The mean relevance of the top 10 news-stories, averaged across all the 25 topics, is significantly higher for WIH (82.6%) than for the crowd-based methodologies (73.6%).

Note that, since the news-stories in crowd-topic include tweets that contain the topic-word itself, such news-stories

[6]Overall, the 1.27 million topical experts we identified post 5.1 million tweets per day on average, which amounts to only 1% of the 500 million tweets posted per day in Twitter [4].

[7]**https://www.mturk.com/mturk/welcome**

[8]These specifications remain the same for the AMT surveys described later in the paper.

Topic	Percentage judged relevant			
	Random experts' tweets	WIH top 10	crowd-topic top 10	crowd-expanded top 10
music	48.0 %	**100.0%**	**100.0%**	50.0%
tech	53.1 %	**80.0%**	70.0%	50.0%
politics	56.3 %	**90.0%**	**90.0%**	60.0%
food	46.8 %	70.0%	60.0%	**80.0%**
health	42.9 %	80.0%	**100.0%**	70.0%
fashion	45.8 %	**100.0%**	60.0%	30.0%
wine	18.4 %	**70.0%**	40.0%	40.0%
books	43.8 %	**100.0%**	**100.0%**	**100.0%**
government	51.1 %	**100.0%**	**100.0%**	20.0%
law	51.0 %	**100.0%**	80.0%	20.0%
environment	55.3 %	**100.0%**	50.0%	30.0%
beer	47.9 %	**80.0%**	70.0%	0.0%
baseball	65.3 %	**100.0%**	80.0%	20.0%
golf	50.0 %	**90.0%**	60.0%	10.0%
hollywood	46.8 %	**100.0%**	50.0%	50.0%
iphone	35.4 %	50.0%	**60.0%**	**60.0%**
history	39.6 %	**60.0%**	0.0%	20.0%
religion	**61.2** %	60.0%	60.0%	20.0%
psychology	55.6 %	**80.0%**	**80.0%**	10.0%
astronomy	55.1 %	**100.0%**	90.0%	30.0%
astrology	72.3 %	**100.0%**	80.0%	0.0%
theology	56.3 %	60.0%	**90.0%**	60.0%
geography	40.8 %	60.0%	**100.0%**	0.0%
neurology	82.0 %	**90.0%**	**90.0%**	0.0%
malaria	75.0 %	44.4%	**80.0%**	0.0%
Mean Rel	52.8 %	**82.6%**	73.6%	33.2%
nDCG	NA	**0.94**	0.92	0.73

Table 5: AMT relevance judgments for (i) 50 randomly selected tweets posted by experts on a topic, and top 10 news-stories by (ii) WIH, (iii) crowd-topic, and (iv) crowd-expanded. Also shown are mean relevance and nDCG scores for the top 10 news-stories across all topics.

can be always expected to be relevant to the topic. However, on average, 26% of the top crowd-topic results were judged to be *non-relevant* to the topic (Table 5, 4^{th} column). This surprising result will be discussed below.

We also notice that the relevance of the top news-stories is reduced drastically when the query is expanded via pseudo-relevance feedback (see Table 5, 5^{th} column). This is because many of the terms that appear frequently in the results of the original query (the topic-word), but are *not* directly related to the topic, get included in the expanded query – this problem is known as *query drift* in the IR literature [22]. These results show the limited utility of traditional IR techniques (such as keyword-matching and query expansion via pseudo-relevance feedback) on the Twitter platform.

Analyzing non-relevant results

We study the top WIH and crowd-topic results which were *judged non-relevant by the majority* of AMT evaluators. We do not consider the crowd-expanded results in this analysis because of the very low fraction of relevant news-stories.

Non-relevant results in crowd-topic: The methodology of selecting tweets from the global tweet stream (crowd) through keyword-match and then clustering / ranking them based on crowd-endorsements, implies that any group of tweets containing the specific topic-word that are posted or retweeted by many users, is likely to be included in the top news-stories.

Topic	Illustrative tweet (extract)
Keyword used in a different sense / context	
religion	A Walk track from Bad Religion album by Bad Religion #nowplaying
food	I have completed quest Food for thought in Cat Story #AndroidGames
environment	#Job Are you looking for a caring and friendly work environment
Keyword used in correct sense, but tweet unrelated	
history	Greatest team performance in football history? #Germany
wine	Now that #TheBachelorette is over, its time to drink enough wine ...
religion	Football is religion in Brazil! ... #WorldCup2014

Table 6: Examples of top crowd-topic news-stories, which were judged non-relevant by AMT evaluators.

Topic	Illustrative tweet (extract) and hashtags
psychology	Anti-HIV drug Efavirenz, marketed as Sustiva, doubles risk for #suicide
beer	9 Wine Myths That Need to Die [URL] #wine
malaria	Scabies, that affects 130M worldwide, added to @WHO list of #NTDs
iphone	Download new iOS 8 Wallpapers [URL] #iOS8 #Wallpapers [URL]

Table 7: Examples of top WIH news-stories, which were judged non-relevant by AMT evaluators. These are actually related to a broader topic than the one specified.

This often leads to non-relevant results if the particular term was used in a different sense or context (examples shown in Table 6). This is the classic word sense disambiguation problem in information retrieval [31]. Another more subtle reason for non-relevant crowd-topic results is that in some news-stories, the topic-word is used in the same sense as the specified topic, but the tweets do not contain any topical information (see Table 6 for examples).

Non-relevant results in WIH: One reason for non-relevant news-stories in top WIH results is that many of the experts we identified for a particular topic are, in fact, experts on a *broader topic*. For instance, many of the experts on topics like 'psychology' or 'malaria' are experts on the broader topic 'health', while many of the experts on the topic 'iphone' are experts on the broader topic 'technology'. These experts discuss news-stories related to their broader areas of expertise, and such stories would also be included in the top results for the more *specific* sub-topic. For instance, we observed that most of the results judged non-relevant for the topic 'iphone' were in fact related to other Apple devices and technological news in general. Table 7 shows some examples of such news-stories.

Another reason for non-relevant WIH results is that, for some niche topics (e.g., 'malaria', 'theology' etc.) for which we could collect only few tens of experts, news-stories discussed by even two or three experts get included among the top news-stories. This, at times, leads to conversational news-stories being extracted, which explains the low relevance of WIH results for topics like 'history', 'geography', 'theology', 'malaria' in Table 5. Crawling more experts for these topics is likely to solve this issue.

Coverage

We now conduct a comparative analysis of the coverage of the set of tweets gathered from topical experts and the crowd (through keyword matching). Specifically we investigate *whether the expert digest for a topic (that is several times smaller than crowd digests) can cover most of the important topical news-stories posted by the global Twitter crowd.*

Methodology to evaluate coverage

For each of the 25 topics, we collected the expert digest, the `crowd-topic` digest and the `crowd-expanded` digest on a particular day. Ideally, for a given topic, one would like to evaluate a digest by checking what fraction of all the important news-stories on that topic (on the given day) is covered by the digest. Since there is no easy way to identify *all* important news-stories on a given topic, we adopt the following comparative approach – we identify the most important / popular news-stories contained in one of the digests, and then check what fraction of these news-stories is contained in the other digests.

However, it is difficult to quantitatively compare the coverage of news-stories (a set of related hashtags and the tweets containing these hashtags), since a news-story in one digest can be only partially covered in another digest, e.g., if the other digest contains only a fraction of the hashtags or tweets in this news-story. Hence, for simplicity, we decided to consider individual hashtags are representatives of news-stories in this coverage analysis. Since conversations in Twitter on particular news-stories tend to be focused around hashtags [44], we believe considering hashtags as representatives of news-stories is a reasonable approximation.

Deciding to focus on the *important* topical news-stories, we identified the top 25 hashtags in each digest (based on the number of distinct users posting a hashtag in that digest), and then checked what fraction of the top 25 hashtags in one digest are included in the other digests for the same topic.

Comparative analysis of coverage

For 75% of the selected topics, more than 60% of the top 25 hashtags are common among all three digests. Thus the most important news-stories are likely to be covered in the digests obtained from both the experts and the crowd.

We now analyze the popular hashtags in one digest that are *not* found in the other digests (for the same topic). We also check the relevance of such hashtags to the specified topic (as judged through human volunteers).

News-stories missing in expert digests: Out of the 625 top hashtags in the `crowd-topic` digests for all 25 topics taken together, there were 192 hashtags that were not contained in corresponding expert digests. Table 8 shows some of these hashtags, along with an example tweet containing the hashtag. Manual human analysis of these hashtags and respective tweets revealed two major categories based on their relevance to the specified topic.

First, 62 out of the 192 (i.e., 32%) hashtags were actually relevant to the topic (but were not included in the corresponding expert digests). Importantly, 32 of these 62 hashtags missing in the expert digests come from only two specific topics – 'malaria' and 'neurology' – for which we have very few experts (31 and 35 respectively). Hence, gathering a larger set of topical experts would significantly improve the coverage of the expert digests for these topics.

Topic (#tag)	Illustrative tweet (extracts)
Hashtags relevant to the topic	
music (#flipagram)	flipagram with @flipagram Music: Ginuwine-Differences #flipagram I love you baby (:
hollywood (#gossip)	#Hollywood Nicole Kidman has Amnesia in Before I To Go Sleep - Filmonic #Gossip
neurology (#love)	Dr. Jordan Talks Merging His #Love of Sports and #Medicine to Practice Neurology
Hashtags non-relevant to the topic	
baseball (#thebachelorette)	Josh only talks about baseball #TheBachelorette
hollywood (#raw)	The MIzs entrance reminds me of Hollywood Rock #WWE #RAW
golf (#Volkswagen)	#Volkswagen #GolfR tested. Read: [URL]
Hashtags about promotions / games / spam	
iphone (#gameinsight)	Ive collected 16,459 gold coins! #iphonegames, #gameinsight
music (#porn)	#porn, #video, #hd, #adult Sound [URL]
wine (#KateSpade)	Win #WINE fridge #KateSpade glasses, rack and more #foodie

Table 8: Examples of top (most popular) hashtags in the `crowd-topic` or `crowd-expanded` digests, which are not included in the expert digest for the same topic.

Second, the rest 130 out of the 192 hashtags (i.e., 68%) are actually *not* related to the said topic, but were included in the `crowd-topic` digest because many tweets containing the hashtag also contained the topic-word (see Table 8). In fact, a significant fraction of these hashtags were related to spam / promotional campaigns, which shows the vulnerability of keyword-match based approaches to such campaigns.

Thus, out of the 625 top hashtags across the 25 topics, only 62 topically relevant hashtags were not contained in the expert digests. Hence, *the expert digests (despite their compactness) contain more than 90% of the important topical hashtags that are posted in the global Twitter stream*.

We also checked the top 25 hashtags in the `crowd-expanded` digests for the selected topics, that do not appear in the corresponding expert digests. We found 229 such hashtags across all topics, out of which only 9 were relevant to the topic. This is expected, given the low relevance of the top `crowd-expanded` news-stories (as reported earlier).

News-stories missing in crowd digests: Out of the 625 top-25 hashtags in the expert digests for all 25 topics taken together, (i) 88 hashtags were *not* contained in the corresponding `crowd-topic` digests, out of which, 58 (i.e., 66%) were relevant to the topic, and (ii) 37 hashtags were *not* contained in the `crowd-expanded` digests, out of which 26 (i.e., 70%) were relevant to the topic. Table 9 shows examples of topical hashtags from the expert digests, which were missing in the `crowd-topic` as well as `crowd-expanded` digests – these are

Topic (#tag)	Illustrative tweet (extracts)
hollywood (#Emmys)	#Emmys noms revealed! Chat: [URL]
psychology (#MinorityMentalHealth)	African American college students least likely to seek help for #mentalillness [URL] #MinorityMentalHealth
geography (#UrbanObservatory)	Transform Big Data into Big Understanding: [*url*] #UrbanObservatory #EsriUC
golf (#AiringOctober)	I am co-hosting the next season of @BigBreak Myrtle Beach #AiringOctober

Table 9: Examples of topically relevant popular hashtags in the expert digest, which are *not* included in `crowd-topic` or `crowd-expanded` digests for the same topic.

possibly specialized news-stories that are not discussed by the general Twitter population.

Overall, even though the expert digests contain several times fewer tweets than the corresponding crowd digests, the expert digests cover more than 90% of the important topical hashtags, which tend to correspond to the important topical news-stories. The much larger crowd digests also miss some topical hashtags, even after applying query expansion.

Trustworthiness

We now ascertain to what extent the expert-based and crowd-based methodologies are vulnerable to spam attacks in Twitter [16, 33]. For this, we look for presence of spam in the tweet digests gathered by the three methodologies.

Promotional / spam campaigns: As stated in the coverage analysis, the `crowd-topic` and `crowd-expanded` digests for many topics contain *thousands* of tweets related to promotional / spam campaigns, e.g., online games, adult content, and so on. For instance, we found 996 tweets in the `crowd-topic` digest for 'wine' containing the hashtag '#KateSpade' for a promotional campaign. Table 8 gives more examples of such news-stories. On the other hand, we could not find any such campaign in the expert digests.

Number of suspended accounts: Another measure of the (un)trustworthiness of a tweet digest is the number of spammers / malicious users whose tweets are included in the digest. Given that Twitter regularly suspends accounts involved in malicious activities [33], we attempted to see whether the three digests contain tweets from users whose accounts have later been suspended. As stated earlier, all our evaluation is based on the results obtained on July 08, 2014. We identified all users whose tweets were included in the three digests on this day, and attempted to re-crawl the profile of each user after 18 days.

For the 25 topics taken together, as many as 4,342 users whose tweets were included in the `crowd-topic` digests were later suspended. Similarly, 9,631 users from the `crowd-expanded` digests were suspended. Note that though Twitter could detect and suspend these users later, their tweets got included in Twitter search results on the given day.

These statistics show that even the crowd-based systems which apply state-of-the-art spam defenses (e.g., Twitter search) cannot guarantee the trustworthiness of the results [33]. This vulnerability is already being exploited by spammers, such as by using trending (popular) topics to promote spam [8]. On the contrary, expert digests are free from such spam / suspended users, and the issue of trustworthiness is solved at the very source of the information.

Experts vs. crowds: head-to-head comparison

Till now, we compared the expert and crowd-based methodologies in terms of relevance, coverage and trustworthiness of results. However, the overall quality of a news system depends on other factors as well, such as whether the top news-stories contain *important / interesting information* on the topic of interest. To judge the overall quality of the results, we now perform a direct head-to-head comparison between the top news-stories returned by different designs.

Methodology: Since judging the overall quality of topical news-stories is subjective, we conducted AMT surveys where human evaluators judged the quality of the news-stories. We conducted two different head-to-head comparisons – (i) `WIH` vs. `crowd-topic`, and (ii) `WIH` vs. `crowd-expanded`. In each evaluation, the evaluator was shown a topic, and the top 10 news-stories returned by two different methodologies for this topic. The results were *anonymized*, i.e., the evaluator was not told which result-set is from which methodology, in order to prevent bias in judgment. Then the evaluator indicates which set of news-stories is better for the given topic, or whether both sets are equally good or equally bad. For a particular topic, we consider the verdict – which set of news-stories is better, or a tie – based on majority agreement amongst five distinct AMT evaluators.

Results: Table 10 shows the results of the evaluation for the selected topics; only those topics are shown for which there was a majority agreement. The topics for which there was a unanimous agreement (i.e., all evaluators agreed that one set of news-stories was better) are italicized. The top `WIH` news-stories were judged better than top `crowd-topic` news-stories for 8 topics, there was a tie for 12 topics (both judged equally good), while `crowd-topic` was judged better for 2 topics (no majority decision for the other topics). The comparison between `WIH` and `crowd-expanded` was even more skewed in favor of `WIH` – `WIH` news-stories were judged better for as many as 21 topics, and there was a tie for 2 topics.

The fact that `WIH` was judged to be equally good or better than `crowd-topic` and `crowd-expanded` for a large majority of the topics, clearly brings out the superior quality of the top `WIH` news-stories. These results provide a strong validation for the proposed strategies of (i) relying on topical experts, and (ii) ranking the results based on their popularity among the expert-community, in order to extract topically relevant and important news-stories.

`WIH`: A PUBLICLY DEPLOYED TOPICAL NEWS SYSTEM

We have implemented a fully functional topical news system on Twitter using the expert-based methodology proposed in this work. The system, named 'What is Happening', is deployed at **`http://twitter-app.mpi-sws.org/what-is-`**

Comparison	WIH better	Crowd better	Tie (Both good)
WIH vs crowd-topic	(8 topics) *fashion*, baseball, hollywood, religion, environment, beer, law, history	(2 topics) *theology*, food	(12 topics) politics, *government*, iphone, tech, malaria, health, music, books, neurology, astrology, golf, astronomy
WIH vs crowd-expanded	(21 topics) *music*, politics, *psychology*, *astrology*, *baseball*, *malaria*, health, *environment*, wine, geography, *history*, *law*, *fashion*, *neurology*, hollywood, *beer*, theology, religion, *golf*, *astronomy*, *government*	(0 topics)	(2 topics) *books*, iphone

Table 10: Head-to-head comparison of top 10 news-stories returned by the different methodologies. Only those topics are shown for which there was a majority agreement among evaluators. Topics in italics are the ones for which there was a unanimous agreement.

Comparison	WIH better	twitter-top better	Tie (Both good)
WIH (tweets) vs twitter-top (tweets)	(14) *beer*, *food*, *neurology*, golf, history, environment, psychology, health, hollywood, music, books, astronomy, tech, politics	(4) government, theology, fashion, malaria	(2) astrology, religion
WIH (news-stories) vs twitter-top (tweets)	(18) *beer*, *neurology*, golf, history, environment, psychology, iphone, health, hollywood, music, food, books, astronomy, law, religion, tech, wine, politics	(0 topics)	(1) astrology

Table 11: Head-to-head comparison of top 10 tweets / news-stories returned by WIH (proposed expert-based methodology) and twitter-top (top tweets returned by Twitter search). Only those topics are shown for which there was a majority agreement among evaluators. Topics in italics are the ones for which there was a unanimous agreement.

happening/. The system crawls the recent tweets posted by the 1.27 million experts (identified as described earlier) once every 15 minutes, and uses the tweets to retrieve topical news-stories. Anyone can use the system to enter a topic, and view news-stories relevant to that topic, which were posted within the last day (relative to the time instant when the query was asked). We encourage the readers to try out the service.

In the service, the main result page for a given topic shows the 25 top-ranked news-stories (tweet-clusters); for each news-story, the front-page shows the hashtags and one illustrative tweet in the cluster. This tweet is selected such that it is posted by that expert who is listed the most number of times on the given topic (i.e., for whom the given topic is a primary topic of expertise). Additionally, each news-story also has a "Similar tweets" URL, by accessing which one can see all the tweets in this tweet-cluster.

We keep our database of experts and their topics of expertise up-to-date by periodically re-crawling the Lists created by the Twitter users. This helps us to discover new experts as well as figure out if a known expert develops some new topic of expertise. For instance, in addition to the 1.27 million experts that we identified from among the first 50 million users to have joined Twitter (as stated earlier), we have recently discovered more than half a million *new* topical experts. The topics of expertise of these new experts include several topics which have recently become popular, such as '3d printing', 'html 5', and 'crowdfunding'.

Apart from news-stories on various topics, users are also known to search Twitter for temporally relevant information and information related to specific persons [32]. For this, the queries issued are mostly keywords that are popular on a given day, or names of celebrities. For such queries, our service retrieves relevant tweets via keyword-matching over the tweets posted by experts (during the last day), and then clusters the matched tweets using the methodology described earlier.

WIH vs. official Twitter search

To illustrate the quality of news-stories retrieved by our system, we conducted a head-to-head comparison between the top WIH results and the top results returned by the official Twitter search system. Since the algorithm used by Twitter search is not publicly known, we treat the Twitter search system as a black box, and consider only the top results. We collected the top 10 Twitter search results for the 25 selected topics, using the Twitter Search API [39]; we refer to these results as twitter-top.[9] We then used the same methodology for the comparing the top 10 twitter-top results with the top 10 WIH results (generated on the same day) as described earlier in the head-to-head comparison between WIH and the crowd-based methodologies.

Comparing WIH results and twitter-top results is not straightforward, since Twitter search does not cluster the returned tweets into news-stories. Hence, we performed two types of comparisons – (i) we compared the top 10 tweets shown on the WIH main results page (one from each of the top 10 news-stories) with the top 10 tweets returned by Twitter search, and (ii) we compared the top 10 news-stories (i.e., tweet and hashtag clusters) returned by WIH with the top 10 tweets returned by Twitter search.

Table 11 shows the results of the comparison, where the verdict is decided by majority agreement among five distinct AMT workers. Only those topics are shown in Table 11 for which there was a majority agreement. In the comparison between the top 10 *tweets* from WIH and twitter-top, WIH results were judged to be better for as many as 14 topics, whereas twitter-top results were judged better for only 4 topics (tie for 2 topics). In the comparison between WIH news-stories and twitter-top tweets, the results were more in favor of WIH which was judged better for as many as 18

[9]Note that some of the Twitter top results include photos along with the tweet text. Since WIH processes only the text of tweets, we consider only the text of the tweets returned by Twitter search for a fair comparison.

topics. The fact that the WIH system, developed by a small group of researchers, performs at least as well as the official Twitter search system for many topics, indicates the power of the proposed expert-based methodology.

A Twitter-based newspaper

To illustrate a potential application of our system, we present news-stories curated by our system for all topics covered by NYTimes (**http://www.nytimes.com/**) at **http://twitter-app.mpi-sws.org/what-is-happening/**. Compared to traditional news media sites, where news-stories are manually curated by a few experienced journalists, our system retrieves news-stories crowd-sourced from 1.27 million topical experts on Twitter. Consequently, our topical news system can automatically retrieve recent news-stories on thousands of topics, which is far greater than the few broad topics covered by any traditional news media site (which was one of our primary goals). Note that some of the topics covered by traditional news media sites are geographical / regional in nature, e.g., 'New York', 'Chicago', 'Africa', in case of NYTimes. To return news-stories on such topics (such as political news from a specific region), we can limit the search to the tweets posted by the experts from the desired region. The geographical locations of experts can be inferred either from their profile [19], or from the Lists whose names / descriptions often include names of cities and countries. We leave the development of location-specific news systems as future work.

CONCLUSION

This paper explores a fundamentally novel and simpler design for a topical news system on the Twitter microblogging site. While all prior approaches for topical retrieval in Twitter rely on the wisdom of the crowd (i.e., the tweets posted by the entire Twitter population), we rely only on tweets posted by a relatively small set of topical experts. We show that the proposed methodology can retrieve most of the important news-stories relevant to a wide variety of topics; additionally, the results are far more trustworthy. We also implement and publicly deploy a fully functional topical news system over Twitter (the first of its kind, to our knowledge). The top news-stories returned by this system are at least as good as (and frequently better than) the top results returned by similar crowd-based designs, as well as the top tweets of the official Twitter search system.

Our work is an exploration of the potential of the wisdom of experts for retrieving topical news in the blogosphere. While we focused on contrasting the wisdom of experts against the wisdom of crowds, there is nothing to prevent one from exploiting the wisdom of *both experts and crowds*. In fact, we envisage that future topical news systems would be hybrid systems that would leverage tweets posted by both the general crowds and the topical experts.

Finally, though our present study focuses on Twitter, the research question we address – comparing the utility of expert wisdom and crowd wisdom for applications such as topical search / topical news – can be investigated on any other social media which provides some mechanism to identify experts on a variety of topics, such as LinkedIn Skills [7] or Facebook group memberships, or up-votes and sub-reddit participations in Reddit, and so on.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers whose suggestions greatly helped to improve the paper. This research was supported in part by a grant from the Indo-German Max Planck Centre for Computer Science (IMPECS). Additionally, P. Bhattacharya was supported by a fellowship grant from Tata Consultancy Services, and S. Ghosh was supported by a Post-doctoral fellowship from the Alexander von Humboldt Foundation.

REFERENCES

1. 2010. How Google Ranks Tweets. (2010). **http://www.technologyreview.in/web/24353/**.
2. 2014. How Justin Bieber's Troubles Exposed Twitter's Achilles' Heel. **http://bit.ly/twitter-achilles-heel**. (2014).
3. 2015. Twitter Help Centre — Using Twitter lists. **http://bit.ly/how-to-use-twitter-lists**. (2015).
4. 500M-tweets-daily 2013. Twitter in numbers. **http://bit.ly/twitter-in-numbers**. (2013).
5. Xavier Amatriain, Neal Lathia, Josep M. Pujol, Haewoon Kwak, and Nuria Oliver. 2009. The Wisdom of the Few: A Collaborative Filtering Approach Based on Expert Opinions from the Web. In *ACM SIGIR*.
6. amt-masters 2011. Get better results with less effort with Mechanical Turk Masters – The Mechanical Turk blog. **http://bit.ly/amt-masters**. (2011).
7. Mathieu Bastian, Matthew Hayes, William Vaughan, Sam Shah, Peter Skomoroch, Hyungjin Kim, Sal Uryasev, and Christopher Lloyd. 2014. LinkedIn Skills: Large-scale Topic Extraction and Inference. In *ACM RecSys*.
8. Fabrício Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgílio Almeida. 2010. Detecting spammers on Twitter. In *CEAS*.
9. Michael W. Berry. 2004. *Survey of text Mining - clustering, classification and retrieval*. Springer.
10. Parantapa Bhattacharya, Saptarshi Ghosh, Juhi Kulshrestha, Mainack Mondal, Muhammad Bilal Zafar, Niloy Ganguly, and Krishna P. Gummadi. 2014. Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale. In *ACM CSCW*.
11. Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011. Find Me the Right Content! Diversity-Based Sampling of Social Media Spaces for Topic-Centric Search. In *AAAI ICWSM*.
12. Eugene F. Fama. 1970. Efficient capital markets: a review of theory and empirical work. *The Journal of Finance* (1970).
13. Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012. Cognos: Crowdsourcing Search for Topic Experts in Microblogs. In *ACM SIGIR*.
14. Saptarshi Ghosh, Muhammad Bilal Zafar, Parantapa Bhattacharya, Naveen Sharma, Niloy Ganguly, and Krishna Gummadi. 2013. On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream. In *ACM CIKM*.
15. Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with Mechanical Turk. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.
16. Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @Spam: the underground on 140 characters or less. In *CCS*.
17. Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In *VLDB*.
18. Martin Hentschel, Omar Alonso, Scott Counts, and Vasilis Kandylas. 2014. Finding Users we Trust: Scaling up Verified Twitter Users Using their Communication Patterns. In *Proceedings of AAAI International Conference on Weblogs and Social Media*.

19. Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. 2012. Geographic Dissection of the Twitter Network. In *Proc. AAAI International Conference on Weblogs and Social Media (ICWSM)*. Dublin, Ireland.
20. Juhi Kulshrestha, Muhammad Bilal Zafar, Lisette Espin Noboa, Krishna P. Gummadi, and Saptarshi Ghosh. 2015. Characterizing Information Diets of Social Media Users. In *AAAI ICWSM*.
21. Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. 2011. Twitter Trending Topic Classification. In *IEEE ICDM Workshops*.
22. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
23. Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding semantics to microblog posts. In *ACM WSDM*.
24. Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is Believing?: Understanding Microblog Credibility Perceptions. In *ACM CSCW*.
25. Brendan O'Connor and others. 2010. TweetMotif: exploratory search and topic summarization for Twitter. In *AAAI ICWSM*.
26. Aditya Pal and Scott Counts. 2011. Identifying topical authorities in microblogs. In *ACM WSDM*.
27. R. M. Poses, C. Bekes, R. L. Winkler, W. E. Scott, and F. J. Copare. 1990. Are two (inexperienced) heads better than one (experienced) head? Averaging house officers' prognostic judgments for critically ill patients. *Archives of Internal Medicine* (1990).
28. Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *AAAI ICWSM*.
29. Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: news in tweets. In *ACM SIGSPATIAL GIS*.
30. Naveen Sharma and others. 2012. Inferring Who-is-Who in the Twitter Social Network. In *Workshop on Online Social Networks*.
31. Mark Stevenson and Yorick Wilks. 2003. Word-Sense Disambiguation. In *The Oxford Handbook of Computational Linguistics*.
32. Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. 2011. #TwitterSearch: A comparison of microblog search and web search. In *ACM WSDM*.
33. Kurt Thomas and others. 2013. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *USENIX Security Symposium*.
34. topsy 2015. Twitter Search, Monitoring, Analysis — Topsy.com. (2015). **topsy.com**.
35. trec-crowdsourcing-track 2013. TREC Crowdsourcing Track. (2013). **https://sites.google.com/site/treccrowd/**.
36. trec-microblog 2013. TREC 2013 Proceedings. (2013). **1.usa.gov/1pYamLX**.
37. twinitor 2015. Twinitor: Twitter search and monitoring. (2015). **www.twinitor.com**.
38. twitter-search 2015. Twitter Search. (2015). **https://twitter.com/search-home**.
39. twitter-search-api 2015. The Search API — Twitter Developers. (2015). **https://dev.twitter.com/rest/public/search**.
40. twitter-verification 2015. Twitter verified accounts. (2015). **bit.ly/1bJ8HoL**.
41. Claudia Wagner, Vera Liao, Peter Pirolli, Les Nelson, and Markus Strohmaier. 2012. It's not in their tweets: modeling topical expertise of Twitter users. In *IEEE SocialCom*.
42. Gang Wang, Tianyi Wang, Bolun Wang, Divya Sambasivan, Zengbin Zhang, Haitao Zheng, and Ben Y. Zhao. 2015. Crowds on Wall Street: Extracting Value from Collaborative Investing Platforms. In *ACM CSCW*.
43. Jinxi Xu and W. Bruce Croft. 1996. Query Expansion Using Local and Global Document Analysis. In *ACM SIGIR*.
44. Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. We know what @you #tag: does the dual role affect hashtag adoption?. In *WWW*.
45. Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. 2014. Large-scale High-precision Topic Modeling on Twitter. In *ACM SIGKDD*.