Curating and Contextualizing Twitter Stories to Assist with Social Newsgathering

Arkaitz Zubiaga
Computer Science Department
Queens College and Graduate Center
City University of New York
New York, NY, USA
arkaitz@zubiaga.org

Heng Ji
Computer Science Department
Queens College and Graduate Center
City University of New York
New York, NY, USA
hengjicuny@gmail.com

Kevin Knight
Computer Science Department
Information Sciences Institute
Univ. of Southern California
Marina del Rey, CA, USA
knight@isi.edu

ABSTRACT

While journalism is evolving toward a rather open-minded participatory paradigm, social media presents overwhelming streams of data that make it difficult to identify the information of a journalist's interest. Given the increasing interest of journalists in broadening and democratizing news by incorporating social media sources, we have developed TweetGathering, a prototype tool that provides curated and contextualized access to news stories on Twitter. This tool was built with the aim of assisting journalists both with gathering and with researching news stories as users comment on them. Five journalism professionals who tested the tool found helpful characteristics that could assist them with gathering additional facts on breaking news, as well as facilitating discovery of potential information sources such as witnesses in the geographical locations of news.

Author Keywords

newsgathering, user-interface, social media, trends, scoop

ACM Classification Keywords

H.5.m Information Systems: Information Interfaces and Presentation - Miscellaneous

INTRODUCTION

Social media is steadily growing as it gains importance as an information source to learn about current affairs. More specifically, the microblogging service Twitter, where users can post short messages known as tweets, is becoming the *par excellence* social media to catch up on recent news and events. The brevity of tweets and the increasing usage of mobile devices are facilitating to quickly share information from anywhere at anytime, and thus make Twitter the first venue where news break in many cases, anticipating news media [19]. This phenomenon is giving rise to an ever stronger citizen journalism, where millions of users share about, discuss, comment on and illustrate recent happenings on a daily basis. The participation of a large community of users enables access to an unprecedented volume of real-time information, constituting a gold mine for journalists to research the news and search for sources for ideal news reporting.

Journalists constantly track numerous sources with the aim of discovering and researching breaking news stories as soon as they happen. This process is known as *newsgathering*, by which they try to collect as much information and sources as they can for a subsequent well-researched news reporting step. While journalists have traditionally relied on authoritative information sources for newsgathering, the growth of social media is picking up steam as an additional participatory information source fed by citizens. In addition, collecting information by citizens in social media allows to approach to the readers by reporting news in a broader, more democratic, and culturally more relevant way [29]. Furthermore, taking advantage of information and media shared by citizens enables extensive coverage of news as never seen before due to limited reach of journalists themselves, both geographically and culturally. However, this open participation often produces an overwhelming amount of informal and noisy information, which makes it difficult to identify and grasp the main contents. Despite the potential existence of valuable and sometimes unique information in that mess, some journalists are still skeptical about using social media as a source, given the endeavor needed to curate its contents.

With the aim of making newsgathering from social media an easier experience, especially for journalists, we have developed TweetGathering, a tool for curated and contextualized access to Twitter stories early on as news break. Distilling Twitter, an open platform where users can share and comment on all kinds of matters, often riddled with pointless chatter, we provide enhanced access to trending stories that are likely to be newsworthy. The enhancements include curation of contents for faster access to salient information, addition of context for easier understanding of events that are not familiar to the reader, and improved access to contributors for the discovery of potential sources such as witnesses.

Five journalism professionals were invited to test the tool. By interviewing them and evaluating their user experience, we have found numerous features and visualizations that could be helpful to journalists in the social newsgathering process. TweetGathering is expected to assist collecting additional facts on breaking news, as well as facilitating discovery of potential information sources such as witnesses in the geographical locations of news.

IUI'13, March 19–22, 2013, Santa Monica, California, USA.

TWITTER, NEWS, AND JOURNALISM

Twitter is a hodgepodge that, among many other kinds of contents, includes a bonanza of news [14]. It can be considered as an ambient media system where users get information from both established media and each other, but where not everything is an act of journalism [11]. It has shown the ability to quickly spread the stories of community's interest, which the site reflects as *trending topics*. These trending topics are often produced by breaking news [38], whereupon users can contribute, comment and spread. This often produces a great deal of citizen journalism contents that build on and enrich news. Popular examples of citizen journalism on Twitter include the Hudson River plane crash photo in 2009, or the witness photo in the 2012 Empire State shooting, among many others.

The emergence and success of the participation of users in news sharing through Twitter has strengthened citizen journalism. Journalism is evolving toward a participatory paradigm [35] as the participation of readers in the distribution of news and supply of information is becoming fundamental [27, 36]. News is becoming a participatory activity, as people contribute their own stories and experiences and post their reactions to events [31]. This quickly attracted journalists and news organizations to be part of the social media service, both for distributing breaking news quickly and concisely, and for soliciting story ideas, sources and facts to a community of users [12]. News organizations have stepped into social media to tell about breaking news and recent events, conceiving it as a new medium akin to traditional sources [29]. In fact, the recent trend of news organizations is to be anywhere, anytime, on any platform [30]. It enables to be closer to the community by engaging new readers, and to be on top of news and events as they occur.

Besides a powerful platform for the dissemination of breaking news, social media presents an unprecedented opportunity to catch the scoop from citizen's updates, to quickly learn more about recent news, and to find sources for further news research. In this work, we present our efforts toward developing a friendly interface for easier newsgathering from the social media jumble.

RELATED WORK

We discuss relevant research on the identification of news stories from social media, and visualizations that assist journalists with the optimal utilization of social media.

Sankaranarayanan et al. [32] presented one of the first approaches to identify whether or not a tweet is related to a news. They use a set of 2,000 handpicked Twitter users as a seed, and assume that they are always sharing news to identify tweets by other users. Unfortunately, their approach does not identify news stories initialized by unknown users, and do not consider that those users might be sharing tweets that are not related to news. Phelan et al. [28] focused on recommending news to Twitter users by assuming the news identification utilizing external sources such as RSS feeds from news media. Zubiaga et al. [38] presented a real-time approach to categorize emerging trending topics in a topology of four types of trend triggers: news, current events, memes, and commemoratives. They present a set of language-independent features that accurately identify the type of trigger each trending topic belongs to. In this work, we use the social features suggested by them as one of the inputs to our news ranking system. Similarly, Naaman et al. [25] introduced a taxonomy of trending topics, but they did not deal with its classification. Other studies have focused on analyzing the credibility of tweets [5], studying the evolution of trending topics [1, 6], performing behavioral studies of tweets [19], and studying the spread of news in social media [22, 26].

There have been some studies on the development of user interfaces to assist journalists with the use of social media. Marcus et al. [24] presented TwitInfo, a tool that identifies peaks in the tweeting activity as sub-events that occur in events, with the aim of assisting with the summarization of events. Diakopoulos and Shamma [9] developed a tool that visualizes temporal dynamics of sentiments in tweets during a political debate. More recently, Diakopoulos et al. [8] presented a tool to help journalists identify eyewitnesses in the context of an event. We believe that the present is the first research work on a user interface to assist journalists in curating and contextualizing social media stories for newsgathering.

There are online services that leave the description of a trending topic in hands of users. WhatTheTrend[1] stores Twitter's trending topics, and collects users' definitions. Being maintained by users, definitions often delay, and it does not provide an alternative to automatic identification of news. Alternatively, [34, 33] automate the summarization of topics in a single definition by looking at common phrases found in tweets, while [37] look at spikes in the tweeting activity to automatically create real-time summaries of events.

Despite the interest in mining conversations and trends from Twitter, and the increasing interest of journalists and news organizations in the power of social media as a news source, as far as we know no previous research dealt with the facilitation of an effective newsgathering from social media sources. This work attempts to fill this gap by studying the development of a social newsgathering tool.

SOCIAL MEDIA AS A SOURCE FOR NEWSGATHERING

On a recent study, Knight [17] analyzed the use of social media sources to research the news related to the Iran protests in 2009. He found that social media sources are sometimes hidden in news reporting, but social media, and especially Twitter, is often used and quoted. The meta-journalistic discussions about the use of Twitter imply that journalists are much more reliant on social media as a source than is apparent from an analysis of the articles. Social media is changing the way news are gathered and researched. Traditionally, the newsgathering has been known as a *gatekeeping* process [3], i.e., news organizations receive news stories while they choose which of them should be ultimately published. First with collaborative news websites like Digg, and now with other social media services like Twitter, it is becoming a *gatewatching* process [4], where news organizations have the chance to observe what the community is considering as newsworthy, and what they are contributing to the news. Twitter being a

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

rich source of breaking news, the need for *gatewatching* increases. However, social media poses new challenges, where a great many individuals post no end of updates, which aggregate to the mess. Successfully exploiting this information source provides an unprecedented way of covering news.

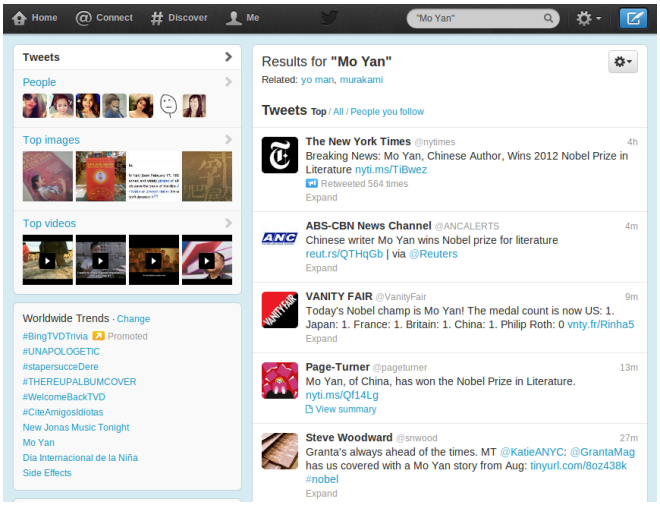

Figure 1. Interface of Twitter search.

Figure 1 shows the interface of Twitter's search. The interface is similar to a traditional web search engine, where a selected set of tweets are displayed, sorted by relevance. Although the sorting algorithm used by Twitter is not known, it appears to be limited to a combination of number of retweets and recency. Along with a set of tweets in the right column, the interface shows some media items on the left column: people, top videos, and top images. What Twitter labels as "people", for instance, is just a set of profile images of the top users for a story, that help neither find who these users are, nor what these users have posted about. To look for more details about any of these users, one has to click on them to open a new page, which leads to leaving the search page being browsed. In addition, the new page might not show the tweets in the conversation being analyzed, but the whole timeline of the user, which may contain irrelevant information.

Having analyzed Twitter's interface, and the needs to make it easier to understand trending stories, to grasp the story being discussed in a conversation, and to research it, we describe the two top shortcomings we detected: (i) the overwhelming amounts of content, and (ii) lack of context.

Shortcoming #1: Overwhelming Amounts of Content

On one hand, given the interest that a trending story arouses in the community of users, the number of tweets about the story increases, which in turn creates a big amount of tweets being shared about it. This produces overwhelming amounts of tweets about a story, often mixed with tweets that, instead of sharing or commenting on a news, are making fun of it or introducing spam. This makes it difficult to identify and understand the main information within the story, and to discover users who are involved as sources to be followed. Thus, we believe that curating these tweets would benefit the task of following the story and identifying the main contents.

Shortcoming #2: Lack of Context

On the other hand, a trending story presents several issues that are not easy to understand without further help. Breaking news being mentioned in a Twitter story may have occurred anywhere in the world, and have anyone involved in it. It is common to have news breaking in the country where it occurs, but later becoming international news as it spreads. Depending on the prior knowledge of the user analyzing the story, it becomes challenging to understand what is going on when the majority of the tweets are written in an unknown language, or the people and organizations involved are unknown. Geographical and cultural differences play an important role in understanding certain news. This suggests the necessity to translate tweets when needed, as well as to incorporate some context to help understand who or what are involved. Having tweets translated and contextualized would in turn make the user feel closer to the story and better understand it.

INTERFACE DESIGN

The user interface of TweetGathering aims to provide easy navigation through recent stories being discussed on Twitter. The tool is designed as a standalone web application that meets the standards to be rendered by all major browsers. It is also designed so that everything can be browsed within a single tab of the browser, so there is no need to open additional tabs or windows. Besides the contents collected from Twitter, and what the tool itself displays, it provides access to external URLs included in tweets. This is done by loading external URLs in a floating iframe on top of the tool, to avoid opening new windows or tabs. To facilitate access to and run-through of recent stories at all times, the prototype is composed of a navigational list of stories on the left, and a bigger panel on the right to explore the story of the user's choice.

Overall, the interface aims to be as easy to adapt as possible for a user who is familiar with Twitter. For this, tweets retain the original format, with the profile image on the left side, and the username, the text of the tweet, the timestamp, and the ability to perform a set of actions such as reply and retweet on the right. Figure 2 shows the interface of the TweetGathering prototype, with the main tab for a sample story loaded on the right panel.

The right panel includes all the information related to the selected story, with the main information on the initial tab, and having chance to access additional information and features from the other tabs. The initial tab is mainly composed of a box where tweets are loaded, and a set of features around to navigate through tweets. These features aim to help contextualize and characterize a story, which we describe with details in the next section. The tweet box is initially loaded with a set of representative tweets, with the aim of curating the contents and providing the most salient tweets. When clicking on the features surrounding the box, the tweets relevant to the selected feature will be visualized in the box.

TECHNICAL DEVELOPMENT

In this section, we describe the data gathered from Twitter to conduct the user studies, and elaborate the processing we did

Figure 2. Interface of TweetGathering. 1) List of recently trending stories, sorted by newsworthiness, 2) Title, descriptions, and start time of the story being explored, 3) Tabs for additional navigation, 4) Summary of URLs, 5) Search engine, 6) Users, hashtags, named entities and events found in the story, 7) Access to selected tweets and all tweets, 8) Tweet box.

on this data to provide the user with enhanced information.

Dataset

In order to build and test the TweetGathering prototype, we relied on Twitter's worldwide trending topics as a source of potential breaking news. Twitter shows, at all times, a list of top 10 trending terms that are being mentioned outstandingly at the very moment. We tracked these trending topics from February 1 to 28, 2012, and randomly sampled a set of five topics per hour. For each of these trending topics that we collected, we also gathered up to 1,500 tweets[2] that contained the term, that is, the most recent tweets that gave rise to the trending topic. This gathering process resulted in a final set of 2,593 trending topics, and a total of 3,629,089 tweets associated with them. As we rely on worldwide trending topics, the stories are related to issues from several countries, which produces an ideal input for the purposes of this research of helping understand stories that not always are culturally or geographically close to the user. The collected data is also diverse in terms of languages used, where English amounts to a 57.8% of the tweets, but other languages contain large amounts of tweets: Spanish (14.9%), Portuguese (7.6%), Dutch (4.2%), Indonesian (3.1%), Italian (3.1%), etc.

For evaluation purposes, each of the trending topics was manually annotated as newsworthy or not by 3 annotators. Determining whether a story is newsworthy is not an easy task, since it can depend on many factors that are ultimately an editor's decision [10]. To make it easier to decide what is newsworthy, the annotators were instructed to annotate as newsworthy those trending topics whose underlying story was later covered by major news media. We asked the annotators to provide the URL of a news story, and the domain of the URL provided by them had to be listed on The Paper Boy[3], a comprehensive directory of online worldwide news media, to be considered as valid. A trending topic would not be deemed newsworthy if the URL provided as a proof did not match the requirements. The annotation resulted in a final set of 358 newsworthy trending topics, with a 92.3% inter-annotator agreement. We split the dataset into a development set, composed of the first three weeks –i.e., February 1 to 21–, and a test set with the last week –from February 22 to 28.

Since we collected tweets for stories that recently trended on Twitter, it is not always likely that newsworthy stories will contain links pointing to news media. By labeling as news media all the links with a domain listed on The Paper Boy, and otherwise not, we analyzed the characteristics of newsworthy and unnewsworthy stories. Figure 3 shows the ratio of links pointing to news media from tweets in newsworthy and unnewsworthy stories. In both cases, the number of links pointing to news media is usually low. Although the number of links to news media is more likely to be smaller for unnewsworthy stories, there is no clear difference. Note that for half of the newsworthy stories, less than 0.05 of the links point to news media. This reveals that the early discovery of news from Twitter is not as easy as following or tracking links

[2]Due to the limitation of Twitter's search API to the last 1,500 tweets for a given query.

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

to news media, but also requires more complex searching and processing.

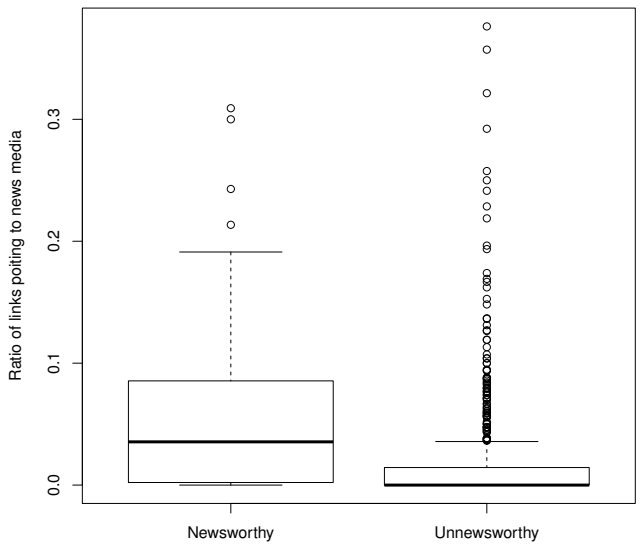

Figure 3. Ratio of links pointing to news media.

Feature Extraction

The newsgathering prototype includes a number of additional features to help with curation and contextualization of trending topics. Next, we enumerate and describe the technical details of these features:

Machine translation of tweets. We used automatic translation software to translate non-English tweets into English. After we translate these tweets, we can extract English-based features from them, and users can read them. Automatic translations were provided by SDL plc[4], whose statistical translation engines processed 98.1% of the non-English tweets in our collection. Of the top 20 non-English languages, SDL's engines covered 19 (the exception being Icelandic), while overall, tweets in 27 of the 46 represented languages were translated.

Ranking trending topics by newsworthiness. Having a large number of stories trending every day on Twitter, it is convenient to provide the chance to more easily identify top stories. With this purpose, we rank the trending stories as they come out, so that we can include a new story at its corresponding position in the ranking. We use Support Vector Machines (SVM [7]) to create the rankings, specifically *SVM-light* [16]. Although SVM is originally designed for text classification, the output provided in the form of margins that refer to confidence can be used as values to rank. We split the development set into groups of two weeks for training, and the remaining week for testing. We evaluate with three-fold cross-validation.

[4]http://www.sdl.com/

We use two different kinds of features for the ranking: (i) vectors of term frequencies (TF) from tweets translated into English, and (ii) social features, as suggested by Zubiaga et al. [38], including 15 features concerning diversity and sharing characteristics, which were demonstrated to accurately classify trending topics in a typology that includes news as one of the categories. Besides the comparison of these two kinds of features separately, we also combine them by using classifier committees [21]. Classifier committees add up the outputs of both classifiers, getting the most of them, and often improving the results of each of them.

Figure 4 shows the averaged results for top positions of ranking trending stories by newsworthiness. We use the Normalized Discounted Cumulative Gain (nDCG [13]) as a measure, which evaluates the effectiveness of a ranking considering that higher positions will have a higher influence on the measure. The results show that the TF approach on translated tweets does better than the social features for the top positions, but the social features improve in positions after 10. However, combining them solves this issue, outperforming both of the independent approaches, and high level performance is achieved, especially for the top positions. Thus, we rely on the combined approach, as inferred from the development set, when ranking new stories from the test set.

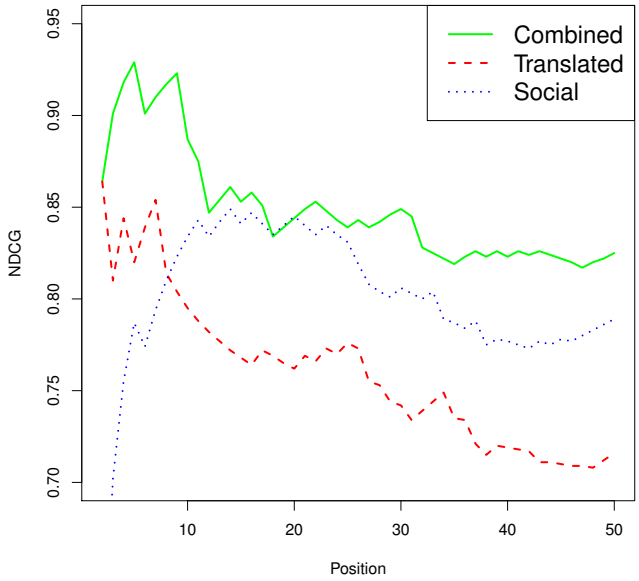

Figure 4. nDCG values for top 50 stories ranked by newsworthiness.

URL lengthening and statistics. When tweets contain links, they tend to be shortened to save characters. The short URLs hide the actual URL to the users, and so they need to click on it to check where it actually points to. Moreover, different short URLs often redirect to the same URL, so it is necessary to resolve the final URLs to get actual statistics of the use of URLs, and to find out the most popular URLs. Therefore, we resolved all the short URLs to their final URLs. With those,

we can create a ranking of the most popular URLs, the top of which is shown on the interface. On the other hand, we categorize the final URLs as news or not. A URL was considered as news when its domain belongs to news media, i.e., when its domain is listed as a news media on The Paper Boy. The tool visualizes the number of URLs contained in the story as well as the fraction that belongs to news media, and the top five URLs categorized as news are shown. These top five URLs are accompanied by the title of the page as gathered from the HTML tag `<title>`, so it is easy to know what it is about without clicking on it. The user can click on any of the URLs to preview its contents, and can also switch to the "URLs" tab for the full list of URLs in the story. The number of news URLs does not mean to refer to the newsworthiness of the story, since newsworthy stories frequently lack news URLs early on when they break. Instead, it aims to estimate the extent to which the news has already been covered by news media, or it recently broke on Twitter with no media coverage yet.

Selection of Representative Tweets. We aim to provide the user with a set of representative tweets that outline the main contents of the story being analyzed. We extract the following four types of tweets as the most representative:

- *Most retweeted:* the most retweeted tweet is straightforward to get, as the one that has been re-shared most.

- *Most frequent:* we select the most frequent tweet as the one that best matches with the vocabulary used in the whole conversation. To find this tweet, we weigh each tweet in the story by adding up the frequency of each term in the whole story. The tweet with most common terms is assigned the highest weight.

- *Trendiest:* we call the trendiest tweet that making the biggest difference from the tweets seen on previous days. The idea of selecting this tweet is to get rid of those tweets that are frequent in many trending stories on Twitter, e.g., *Hey, let's make this trend, please RT!*. To do this, we rely on a language modeling approach using the Kullback-Leibler Divergence (KLD [18]). By computing the divergence of terms in the story being analyzed to the terms seen in previous stories (i.e., those in the development set), we can extract the tweet that is more diverse from usual tweets. This tweet will be the one that maximizes the sum of its terms' diversity values.

- *Earliest:* It is not easy to determine which is the earliest tweet that starts a story. In some cases, there might be mentions before the story broke, which are not actually related to the story. After performing a set of empirical experiments on the development set, we found that finding the earliest tweet with common vocabulary compared to later tweets does well. Thus, we select the tweet containing one of the top 3 terms in the story as the earliest tweet. Besides showing the earliest tweet as one of the selected representative tweets, this also allows us to show the start time of the story on Twitter, which is shown on the header, next to the title of the story.

On the interface, we group them into two types of tweets: (i) *breaking tweets*, which include the most retweeted, the most frequent, and the trendiest, and (ii) the *earliest tweet*. We group the first three tweets since the differences are mainly technical, but they have similar meanings in practice for the user.

Twitter features. There are a number of features that are inherent to Twitter, which we preprocess to facilitate navigation through the tool. We show lists that contain the top of these features, accompanied by the number of appearances in the story, while clicking on them shows the tweets that contain the selected feature:

- *Users*. We extract the top users who tweeted or are mentioned by others in a conversation. This enables us to generate the list of top users that are relevant in the story. The user information not only includes the username, but also additional information including the description and location of the user, when available.

- *Hashtags*. Hashtags are terms preceded by #, referring to an important concept within a tweet, usually meant to categorize the tweet in a conversation. We extract the top hashtags within a story, as a summary of representative concepts or categories that users chose for the story. Hashtags are often a combination of several words into a single string. For instance, users can utilize the hashtag `#unitedstates` to refer to the *United States* without explicit tokenization. Without prior normalization of these hashtags, it might be difficult to read, and especially to process and contextualize since it is unlikely that Wikipedia will have an entry with that name when looking up its meaning. To alleviate this, we analyze all the occurrences of the hashtag in the story. When the hashtag has been capitalized in some cases, we rely on the most frequent capitalization as the tokenization of the hashtag. Thus, if the most frequent capitalization is `#UnitedStates`, we tokenize it by starting a new word on each capital letter, for easier understanding.

- *Images and videos (Media)*. On a separate tab, we show the thumbnails for top images and videos that appear in the conversation, sorted by the number of occurrences. Each media is accompanied by the tweet(s) that include it, as well as the top users associated with it. Different from Twitter's original interface, which is limited to thumbnails, the added context aims to describe the media, as well as to identify the top sources.

Named entities. Using information extraction techniques, we identify the named entities that are mentioned in the story. Types of named entities we identify include person, location, and organization names. We apply a high-performing name tagger [23] to extract these names. The tagger adopts the linear-chain CRFs [20] as the learning method, and incorporates multi-layer local features based on ngrams, case and shape, Part-of-Speech, chunking, gazetteers, Brown Cluster [2], as well as sentence-level and document-level features such as the position and the frequency of the name candidates. The interface shows a ranked list of top names mentioned in

a story, as well as their number of occurrences, while clicking on them loads associated tweets in the tweet box.

Descriptions of names. In some cases, the user of the tool might be unfamiliar with the names, locations, or organizations involved in the story, especially when the story originates from a different country, language or culture. To provide background on the names extracted in the story, we gather their descriptions from the structural knowledge base DBpedia[5]. As a description, DBpedia provides the initial sentences of the corresponding article on Wikipedia. In case this information is not enough, additional alternatives are given to preview the whole Wikipedia article, as well as to look for the term on the Bing[6] search engine.

Events. Events play an important role in the context of news. With events, we aim to provide the user with the key happenings that are mentioned in a story, e.g., death, resignation, or attack. We apply a state-of-the-art English event extraction system [15] to extract events from tweets. The system combines pattern matching with a set of Maximum Entropy classifiers incorporating diverse lexical, syntactic, semantic and ontological knowledge. It takes raw documents as input and conducts some pre-processing steps. The texts are automatically annotated with part-of-speech tags, parsing structures, entities, time expressions, and relations. The annotated documents are then sent to the following classifiers: to distinguish events from non-events; to classify events by type and subtype; to distinguish arguments from non-arguments; to classify arguments by argument role; and given a trigger, an event type, and a set of arguments, to determine whether there is a reportable event mention. Each component can produce reliable confidence values. As for named entities, the top events are shown with their number of occurrences, and clicking on each of them loads associated tweets.

Locations of users. It is often difficult to identify the location of users, since many of them do not provide any details. In this work, we rely on the location of users as specified in their settings. However, the location field in the settings is a string that can be set to anything like "somewhere in the world" or "at home" instead of a location that can be mapped. 20.2% of users in the development set have this field empty, so we cannot easily guess where they are. Also, users specify locations with different levels of granularity, e.g., a few users give the specific coordinates, while some users say they are in *New York City*, and others only say they are in the *USA*.

Search engine. The interface incorporates a query-based search engine that allows the user to look for the terms of their interest, especially to explore relevant subtopics found in tweets. The search enables manually curate the contents by filtering with the term(s) of their choice. The search engine will return the tweets matching the input query as an AND search of the input query terms, so that more specific and restricted searches can be done by using more terms. The match between query terms and tweets is not limited to full tokens, so that searching for a term also retrieves usernames and hashtags with that name.

Access to all tweets. The interface also provides access to all the tweets in the story, so the user can explore the whole conversation when needed.

Some of the features described above rely on methods that might be inaccurate in some cases, such as NLP techniques. Tweets being short and often with an incomplete grammar present an extra difficulty to get accurate results when using NLP techniques. However, relying on stories of tweets rather than single tweets considerably alleviates this issue by putting together the outputs and relying on the top items of combined results. A quantitative evaluation of NLP outputs would be needed to precisily measure the accuracy of such techniques, which in this case has been qualitatively evaluated through journalists' undertanding during user studies.

EVALUATION

In this section, we describe the user studies we conducted to assess the TweetGathering prototype, and present comments and findings collected from these studies.

User Study

The evaluation of TweetGathering was carried out with five journalism professionals who fit the characteristics of actual users of a social newsgathering tool. All of them are native English speakers, have a Twitter account and are familiar with social media, and have long been following the microblogging system to catch the scoop and research news in numerous situations. The studies were conducted in a newsroom, one by one, where they had the chance to play around with the tool while they were thinking aloud about anything they did or found. The prototype was fed by the trending topics in the test set, and each of the interviewees used a specific snapshot of the prototype in different dates and times within the week of the test set. Therefore, each of them used the tool in a real scenario with the stories that trended recently in that specific moment, so they could analyze the utility of the tool to discover news, as well as the utility of curation and contextualization provided for a newsgathering process. We interviewed them about their prior experiences before they used the tool, and also afterward to learn about their feelings while using the tool. The stories shown to them in the user study were previously known to us, but not necessarily to journalists –they did remember a few of the news shown, but they were unaware of most of the news. While journalists were thinking aloud while using the tool and commenting on their findings, we verified that their understanding of the story was the correct one.

Even though two of the interviewees admitted that they never or rarely tweet, all of them follow a number of Twitter accounts on a near-daily basis. They do believe that Twitter opens up a tremendous oportunity to access information shared by users as well as to identify potential sources and witnesses to follow. However, they find that the number of tweets sent during breaking news is overwhelming, and it is difficult to identify the useful information and relevant users. Actually, one of them pointed out that *"I know journalists*

[5]http://dbpedia.org/
[6]http://www.bing.com/

who are not willing to use Twitter, because they feel overwhelmed by all the information, and that requires a big effort to follow" (P2). At this point, they feel the need of a curation tool to make easier the task of researching breaking news.

News Discovery

Given a real scenario of recently trending stories, the participants could explore them in search of breaking news that they would be interested in learning more. They followed different strategies to navigate through the trending stories provided in the left menu. While two of them looked at the top ranked stories, and clicked on the one of their top interest, the other three read through all the list of stories before clicking on one. One of them noted that *"the ranking is very helpful, since the stories in the bottom of the list are mostly memes and pointless conversations, so I can easily get rid of them"* (P1).

After clicking on and exploring some stories, we asked each of them about the main features they were relying on to check whether a story is likely to be about a newsworthy event. They were quite unanimous on this, commenting that they were mostly relying on the selected subset of tweets, as a good summary of the story. Three of them, however, added that they rather looked at the summary of URLs first. When the URLs point to news, only the titles of the URLs are self-explanatory but, otherwise, they looked at the selected tweets. One of them added that *"the navigation through the tool is really easy, as it conveniently enables you to catch the scoop as you click on a news story and find out whether the story is worth exploring in more detail for later news reporting"* (P5).

Curation of Stories

One of the main characteristics introduced in the prototype is the curation of contents, due to the overwhelming amounts of tweets shared in a story. The participants made exhaustive use of the selected tweets that appear initially in the tweet box, which is the main feature of the tool in terms of curation. *"I usually find that Twitter conversations are full of spam and people making fun instead of talking about the actual news, and I need to skip many of the tweets since they are off-topic to me. Having a few tweets selected really helps me discover salient tweets instead of doing it manually"* (P3). Overall, they did not find a great difference among the three tweets labeled as "breaking", but they believe they are similarly helpful and complementary in some cases. However, the "earliest tweet" arouses much more interest in them, since they believe that it might be the source, the tweet that broke the news, so it can be worth following what that user says next, or even contact the user to ask for additional information or an interview.

The search engine is another feature that enables curation, although this requires prior selection of the terms. They did not use the search engine very frequently, but they did find some possible utility in it: *"I would use it to look for users that are in a specific location, for instance. This is something that I miss on Twitter when exploring a story"* (P2).

Finally, the machine translation of tweets also contributes to curation of contents. One of them did not realize that the tweets were originally written in another language for a specific story, but she noticed it having seen the location of the story; *"it is very helpful to have the tweets translated into English, so I don't have to wait to later tweets written in English. This is certainly a story that will soon become internationally popular, and it is important to quickly catch the scoop"* (P4).

Contextualization of Stories

The other main characteristic introduced in the prototype is the contextualization of the story, which we accomplish by extracting a set of features from tweets, and describing some of them. The participants did not doubt that having context besides tweets is of utmost help.

The feature they liked most for contextualization is the summary of URLs in the story. Besides helping to consider the story as newsworthy or not, they also agree that it is helpful to find context. They find it really helpful to see the final URLs instead of the shortened ones, and to have the chance to open them in the tool without switching tabs. *"It all depends on whether the story contains links to news media. When it does, I can research the news written by others, while in the event that there are no links to news media, I can explore the links to check where they are pointing to and what they are about"* (P2).

The next features they used most are the lists of users, named entities, hashtags, and events. *"Those really help place the story into context and see who is involved. It's not easy to get all that information at first glance from tweets. And they have counts associated, which somewhat reflects the importance of each item. I was looking for the highest numbers, and found interesting things that you don't realize from reading through tweets."* (P2). They also found interesting utilities of these four features:

- Users: *"I'd say that the list of users is the most interesting of these four features. We often look into the contributing users to find sources, or even witnesses that might be reporting on the news. We rely on those as information sources, and sometimes contact them to learn more"* (P5).

- Named entities: *"The list of names reveals, in some cases, where is the story possibly happening, and so makes it easier to locate it. Likewise, being able to easily identify involved people and organizations, with the additional feature of describing them, rapidly places the news into context, especially when it is culturally new to me"* (P1).

- Hashtags: *"Although hashtags don't seem to be as useful as other features, it is sometimes worth exploring the ones that are co-occurring with the story. I was really wondering why #ows appears so frequently in a story about #wellsfargo, that's something I needed to look into more detail"* (P3).

- Events: *"Events can be good in some cases, to see what people are commenting on the story. It's not only about the story itself. I found in a story about a political announcement that a popular event being mentioned in the tweets was resignation, this must be a strong community that wants this politician to step down"* (P4).

Among the stories they had to analyze, there were stories from other countries. Besides being written in foreign language, some of the people and organizations involved were unknown to them. They had to use the external sources provided to describe them. *"I really had no idea who this guy was, so I was unable to say how newsworthy this story could become. Then I saw that he is a participant from reality show. I'm no longer interested in the story, that's not of interest to me"* (P5).

None of them really needed to go to the "all tweets" section, because they could find the essential information from the rest of the features. Two of them tried it out to see how it would look like without having all the enhanced information, and one of them said: *"this is what I usually see, and it is really overwhelming. It takes much longer to grasp the story and find the people, organizations, and locations involved in the story. I'd say it's fine to have it somewhere in the tool, as an alternative if it ever is needed, but it's secondary. I would rarely use it"* (P1).

Finally, the start time of the story helped find how old a story is. *"It's hard to know when a story started, that's not available anywhere. I'd rely on that time to have an idea"* (P3).

Overall Utility and Final Remarks

Overall, the participants liked the features provided to curate and contextualize the tweets of a story. *"I find it very easy to use. Everything is well organized on a single page, and I can explore all the features from here, without leaving the tool"* (P1). They would appreciate having the chance to customize the stories listed, for instance, by including the topics of their interest, the trending topics of specific areas, or stories from the users they follow, which altogether becomes overwhelming to follow on a daily basis. *"I need something like this for my work. With a few customizations to follow the information of my interest, and having it running in real-time, I would switch to this tool"* (P4).

The participants also mentioned some features that they would like to see on the tool:

- Two of them commented that they would like to know what is the location of a story. Currently, the name of the location might appear among the named entities listed by the tool. However, the use of more sophisticated techniques could help determine the location, mapping the stories for easier browsing.
- One of them suggested that, besides having a ranking of tweets, a categorization of the stories by topic would be interesting to follow specific topics.
- Other suggestions were related to customization and having an account to log in on the tool. These suggestions include being able to ban certain topics/users as spam they no longer want to hear from, or being able to bookmark a story to check it again later.

DISCUSSION

This paper explores the identification of news from stories that appear in Twitter's lists of trending topics. Trending topics are a convenient way of identifying recently emerging conversations on Twitter's stream. However, the exploration of stories does not need to be limited to trending topics as presented by Twitter. Alternatively, the identification of emerging conversations from Twitter streams would allow to customize desired stories to be presented, which is not within the scope of this work. Additionally, the system can straightforwardly be applied to specific events, so that it makes easier to find breaking news associated with those specific situations. For instance, the presented prototype could also be suitable for emergency journalism, curating and contextualizing events such as protests, hurricanes or quakes, where the user could input a query, getting the information for the associated story. Also, we believe that the ability to have an account and to log in on the tool would benefit the user experience, having access to stories of the user's interest in a customized way.

CONCLUSION

In this paper, we have presented TweetGathering, a prototype tool that facilitates social newsgathering from Twitter for journalists. The tool aims to make it easier for journalists to find out and research breaking news, by curating and adding context to stories made up by short and often contextless tweets. We have conducted user studies with five journalism professionals to evaluate and get feedback on the tool. These studies brought to light numerous features of the tool that may assist journalists with gathering additional facts on breaking news, alleviating the difficulty of understanding stories from other countries, cultures, and languages, and facilitating discovery of potential information sources.

The feedback received from journalists has also brought to light some future challenges for further improvements on the tool. First, it would benefit from incorporating customized contents, so that it provides personalized lists of stories to each user. Second, the limited features of Twitter in terms of geographical locations of users requires the analysis of sophisticated techniques to determine the locations of both stories and users. And third, we aim to roll out an updated version of the tool to journalists for its use in newsrooms in real-time. This will enable us to quantify the gains provided by the newsgathering tool as compared to using Twitter's original interface.

Acknowledgments

Thanks to Daniel Marcu of SDL for assistance with translations, and to Nicholas Diakopoulos for discussion. This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF- 09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. NSF EAGER Award under Grant No. IIS-1144111, the U.S. DARPA Deep Exploration and Filtering of Text program and CUNY Junior Faculty Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

1. Asur, S., Huberman, B., Szabo, G., and Wang, C. Trends in social media: Persistence and decay. In *5th International AAAI Conference on Weblogs and Social Media* (2011).
2. Brown, P., Desouza, P., Mercer, R., Pietra, V., and Lai, J. Class-based n-gram models of natural language. *Computational linguistics 18*, 4 (1992), 467–479.
3. Bruns, A. Gatewatching, not gatekeeping: Collaborative online news. *Media International Australia Incorporating Culture and Policy: quarterly journal of media research and resources 107* (2003), 31–44.
4. Bruns, A. Gatekeeping, gatewatching, real-time feedback: new challenges for journalism. *Brazilian Journalism Research 7*, 2 (2011), 117–136.
5. Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, ACM (New York, NY, USA, 2011), 675–684.
6. Cheong, M., and Lee, V. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *Proceedings of the 2nd ACM workshop on Social Web Search and Mining* (2009), 1–8.
7. Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.
8. Diakopoulos, N., De Choudhury, M., and Naaman, M. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 2451–2460.
9. Diakopoulos, N., and Shamma, D. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, ACM (2010), 1195–1198.
10. Harcup, T., and O'neill, D. What is news? galtung and ruge revisited. *Journalism studies 2*, 2 (2001), 261–280.
11. Hermida, A. From tv to twitter: How ambient news became ambient journalism. *M/C Journal 13*, 2 (2010).
12. Hermida, A. Twittering the news. *Journalism Practice 4*, 3 (2010), 297–308.
13. Järvelin, K., and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS) 20*, 4 (2002), 422–446.
14. Java, A., Song, X., Finin, T., and Tseng, B. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM (2007), 56–65.
15. Ji, H., and Grishman, R. Refining event extraction through unsupervised cross-document inference. In *Proc. of ACL, the Annual Meeting of the Association of Computational Linguistics* (2012).
16. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* (1998), 137–142.
17. Knight, M. Journalism as usual: The use of social media as a newsgathering tool in the coverage of the iranian elections in 2009. *Journal of Media Practice 13*, 1 (2012), 61–74.
18. Kullback, S., and Leibler, R. On information and sufficiency. *The Annals of Mathematical Statistics 22*, 1 (1951), 79–86.
19. Kwak, H., Lee, C., Park, H., and Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, ACM (New York, NY, USA, 2010), 591–600.
20. Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML* (2001), 282–289.
21. Larkey, L., and Croft, W. Combining classifiers in text categorization. In *SIGIR*, ACM (1996), 289–297.
22. Lerman, K., and Ghosh, R. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *ICWSM* (2010).
23. Li, Q., Li, H., Ji, H., Wang, W., Zheng, J., and Huang, F. Joint bilingual name tagging for parallel corpora. In *Proc. The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)* (2012).
24. Marcus, A., Bernstein, M., Badar, O., Karger, D., Madden, S., and Miller, R. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, ACM (2011), 227–236.
25. Naaman, M., Becker, H., and Gravano, L. Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology 62*, 5 (2011), 902–918.
26. Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci '11: Proceedings of the 3rd International Conference on Web Science* (2011).
27. Paulussen, S., Heinonen, A., Domingo, D., and Quandt, T. Doing it together: Citizen participation in the professional news making process. *Observatorio (OBS*) 1*, 3 (2007).
28. Phelan, O., McCarthy, K., and Smyth, B. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*,

RecSys '09, ACM (New York, NY, USA, 2009), 385–388.

29. Phillips, A., et al. Old sources: New bottles. *New media, old news: Journalism and democracy in the digital age* (2010), 87–101.

30. Picard, R. Blogs, tweets, social media, and the news business. *Nieman Reports 63*, 3 (2009), 10.

31. Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T., and Olmstead, K. Understanding the participatory news consumer. *Pew Internet and American Life Project 1* (2010).

32. Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M., and Sperling, J. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM (2009), 42–51.

33. Sharifi, B., Hutton, M., and Kalita, J. Experiments in microblog summarization. In *Proc. of IEEE Second International Conference on Social Computing* (2010).

34. Sharifi, B., Hutton, M., and Kalita, J. Summarizing microblogs automatically. In *Proc. of NAACL* (2010).

35. Singer, J., Domingo, D., Heinonen, A., Hermida, A., Paulussen, S., Quandt, T., Reich, Z., and Vujnovic, M. *Participatory Journalism: Guarding Open Gates at Online Newspapers*. Wiley, 2011.

36. Stassen, W. Your news in 140 characters: exploring the role of social media in journalism. *Global Media Journal African Edition 4*, 1 (2011).

37. Zubiaga, A., Spina, D., Amigó, E., and Gonzalo, J. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, ACM (2012), 319–320.

38. Zubiaga, A., Spina, D., Fresno, V., and Martínez, R. Classifying trending topics: a typology of conversation triggers on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, ACM (New York, NY, USA, 2011), 2461–2464.