From Puzzles to Portraits: Enhancing Situation Awareness during Natural Disasters Using a Design Science Approach

Charles Z. Liu, Naga Vemprala, and Kim-Kwang Raymond Choo

Department of Information Systems and Cyber Security, University of Texas at San Antonio, 1 UTSA Circle, San Antonio, TX, USA

Email: {charles.liu; naga.vemprala; raymond.choo@utsa.edu}
From Puzzles to Portraits: Enhancing Situation Awareness during Natural Disasters Using a Design Science Approach

Charles Z. Liu, Naga Vemprala, and Kim-Kwang Raymond Choo

Department of Information Systems and Cyber Security,
University of Texas at San Antonio,
1 UTSA Circle, San Antonio, TX, USA
Email: {charles.liu; naga.vemprala; raymond.choo@utsa.edu}

Abstract

During emergency situations, a quick and concise summary from the deluge of messages improves “Situation Awareness” and enables informed decisions. However, this is challenging due to the volume, variety and veracity of information. Grounded in the Situation Awareness Theory, this study presents a streamlined protocol to process millions of social media messages and render them in an easily comprehensible format for various stakeholders to gain deeper insights. Specifically, using a design science approach, we develop a set of artifacts using incremental hierarchical clustering and enhanced text summarization algorithm to produce informative summaries under crisis situations. We implemented our protocol on 2.5 and 3 million tweets collected during the two major hurricanes in 2017 and 2018. The results show that our protocol can derive critical information not captured by Twitter’s search tools and mainstream news media and significantly improve on accuracy and efficiency when compared to other contemporary tools.

Keywords: Situation Awareness, Crisis Response and Management, Design Science, Social Media Analysis, Text Summarization, Hierarchical Clustering, Crisis Lexicon.

JEL Classification Codes: H12, H84, C80, D70.

INTRODUCTION

The series of hurricanes in 2017 and 2018 have brought unparalleled devastation to several US coastal states and the Caribbean Islands, displacing thousands of families and individuals. In Houston, Texas, for example, many neighborhoods were inundated by record-setting floodwater,
and residents were trapped on the roof desperately waiting for rescue. However, the city’s 911 system was overwhelmed by numerous emergency calls that far exceeded its handling capacity. In fact, the city had to urge citizens to call only if there was imminent danger. During the disaster, social media emerged as an effective communication tool that allowed stranded residents to seek help from friends and volunteers. Some web-savvy volunteers vigilantly picked up critical information and organized rescue groups through social media. Unlike TV and radio, which broadcast information unidirectionally, social media allows help seekers and providers to exchange the much needed situational and geographical information, and enables first responders to have real-time communication with the victims and provide timely comfort to them (Zhao and Rosson, 2009). However, the use of social media has largely been self-directed and disconnected. In fact, federal agencies had to urge people not to solely rely on social media for help as they simply did not have enough manpower to monitor and process the large strings of plea for help online. In light of the critical role social media played during and after the recent societal and natural events and vast amount of information circulated on these platforms (Tim et al., 2017), this research attempts to develop a set of data analytic tools to systematically analyze and process social media data to support decision making and facilitate effective coordination across various stakeholders.

Given the tremendous potential of social media in providing valuable information, researchers have adopted various techniques to extract social media data (Tim et al., 2017, Xu et al., 2017). These approaches include topic modelling using sentiment analysis (Deng et al., 2017, Wu and Cui, 2018, Pournarakis et al., 2017), social network analysis using semantics and emotion mining (Garcia-Crespo et al., 2010), message tracking through a viral diffusion model (Goel et al., 2015), query-based text extraction using location-based entity models (Aker and
Gaiauskas, 2015, Chakrabarti and Punera, 2011), and generating spatial temporal graphs of semantic relations between concepts (Xu et al., 2017). Each of these approaches provides a unique perspective and a set of tools to derive useful information from a given type of social media data. Nevertheless, most of these approaches rely on ad hoc modeling techniques that require strong modeling assumptions. For example, many text extraction algorithms provide better summaries when the text input follows a well-structured format. However, the format of the user-generated content on social media, especially at times of emergency, does not always follow the linguistic structure of news articles, online blogs and other traditional webpages (Wu and Cui, 2018). Moreover, many studies adopt sentiment analysis and clustering techniques to derive situational awareness, but such techniques often lack the contextual information and only offer high-level information (Chua et al., 2019). In essence, the existing methods are largely ad hoc, data structure dependent, and address individual issues in isolation, which calls for a more flexible and holistic approach that systematically covers various aspects of information processing during emerging social events, as advocated by the “design science research” (DSR) (Sonnenberg and Vom Brocke, 2011). Therefore, in this study we focus on developing a streamlined protocol that covers data collection, text pre-processing, text filtering, topics extraction, data visualization, and the ultimate task of text summarization. Following the principles advocated by the DSR literature, we also develop a set of functionally-related artifacts. These artifacts provide the basis for constructing an automated data analytic tool capable of processing large amount of unstructured social media data on a near real-time basis, and deriving useful information that can enhance the situation awareness for victims, first responders and policy makers. Specifically, the protocol and artifacts aim at addressing the following research questions:
RQ1: How to detect emerging topics as they appear and develop on social media, especially when the messages are characterized with unstructured data and a variety of data noise?

RQ2: Based on the topic/event identified, can we provide an extended summary of the evolution of the topic/event over a given period of time using analytic tools?

RQ3: Can we visualize the dynamics of a social/natural event from social media data on a real-time or near real-time basis?

As is well established in the DSR literature, an artifact can be a construct, methods (algorithms and practices) or instantiation (implemented and prototype systems) (Hevner et al., 2004). Follow this definition, we developed three new methods and a visualization tool (instantiation) as artifacts. Our first (method) artifact is an extended text filtering lexicon (keywords dictionary) specific to crisis and natural disaster situations such as hurricanes, adapted from (Olteanu et al., 2014). The second artifact is a trending topics extraction methodology, developed using the N-gram hierarchical clustering model (Aiello et al., 2013). The third (method) artifact is a text summarization technique specific to social media messages, and the fourth artifact is a visualization tool for situation awareness which effectively complements the textual outputs produced by our analytic tool. To test our protocol and the artifacts, we compile two comprehensive datasets that consists of 2.5 million and 3 million tweets during the hurricane seasons in 2017 and 2018, respectively. The results demonstrated that our analytic tools have made significant improvements over traditional methods in four key aspects: 1) Instead of adopting an existing general lexicon to filter data, we develop a customized algorithm that iteratively analyzes the segregated messages to identify the relevant keywords for our study context. This resulted in an expanded and much improved lexicon with higher filtering accuracy. 2) We incorporate various social media data characteristics as filtering criteria in our multinomial text summarization algorithm to analytically determine the relevance of the inputs, which addresses the unstructured and sparse nature of social media data and yields informative
summaries. Our approach led to significant improvement beyond the enhancements in recent studies (Ifrim et al., 2014). 3) We develop a data visualization tool to present and summarize the emerging topics from social media data on a dynamic basis, which helps to convey the results in a very intuitive manner. 4) To allow replication and establish a generalizable benchmarking method, we conduct a thorough comparison between our algorithm and two other popular relevance ranking algorithms (i.e. Cosine Similarity and Opaki BM25 (Robertson and Zaragoza, 2009)), as well as a representative set of leading news media. We show that our algorithm can significantly reduce the processing time and improve the accuracy.

The outcomes of this study have important theoretical and practical implications. Due to the improvements over existing methods, our unsupervised machine learning algorithms allow us to discover a large number of critical messages hidden in both under- and over-circulated messages and filter the noise to derive the contextual information. For time-sensitive events, our analytic tool can process thousands of unstructured messages on near real-time basis and provide users with comprehensive visualizations that present highly relevant sentences within a given time frame (e.g. 18,000 hurricane rescue-related tweets are summarized into 80 highly significant tweets and visualized in charts). The tool also allows users to adjust the input and output parameters (e.g., time interval, frequency, and number of messages monitored). These capabilities significantly enhance the situation awareness of various stakeholders, and provide strong and timely support for decision makers during times of emergency. Other possible applications of the tools developed by our study include detection of the rise of derivative disasters, scam and rumor identification, traffic monitoring, damage evaluation, and trend forecasts during and after natural disasters.

RELATED WORK
Situation Awareness

Situation Awareness (SA) seeks to answer the questions of “what is going on around you”, “what has happened”, and “what most likely will occur next” (Endsley and Garland, 2000). SA has been studied for more than three decades in various fields ranging from air traffic control, power plant operations, military operations and many others (Jones and Endsley, 1996, Yin et al., 2012, Sharma and Nazir, 2017, Wu and Cui, 2018). However, there is relatively little research in the Information Systems (IS) field to effectively apply it to extract information, especially from social media data. SA consists of three stages that starts from “the perception of the elements in the environment within a volume of time and space”, to “the comprehension of their meaning”, and to “the projection of their status in the near future” (Endsley, 1995). The perception of an event serves as a prerequisite to establish SA and is often studied by extracting topics, topics or event bursts (O'Connor et al., 2010, Aiello et al., 2013, Chakrabarti and Punera, 2011). For example, the temporal aspect of the situation has been utilized as an important dimension to study the nature of microblogging during a crisis event (Petrović et al., 2010, Wu and Cui, 2018), which is also a salient aspect throughout the other two phases of SA.

Comprehension of SA is the basis for subsequent projection and is often realized through summarization of a series of sub-events or topics (Nichols et al., 2012, Aiello et al., 2013). Projection of SA is the forecast of the possible occurrence of future events and is often achieved by presenting the events over timelines or based on geo tagging of locations to uncover any hidden trends or ongoing patterns (Wu et al., 2018). Even though some prior studies have applied the SA theory to examine emerging topics on social media (Zhou et al., 2015, Zhao and Rosson, 2009), they do not fully embrace the three-stage framework and only evaluate SA in an isolated or static way. The most SA relevant study analyzes the message using content analysis, where
messages are labeled manually and ad hoc to derive information (Vieweg et al., 2010). However, no automatic and real time solution is proposed, which reduces the usability of the findings in emergent context. Therefore, this study attempts to adopt a more holistic framework from the SA literature to more accurately capture the unique characteristics of the different stages of a major event such as natural disasters on a near real-time basis. The perception of the rise of an event is captured by topics automatically extracted through the linguistic aspects of the messages shortly after they are posted, the situation is comprehended through the extended summaries constructed using an unsupervised algorithm, and the projection of a situation is achieved via a SA dashboard with event trends and the discovery of the momentum shifts around discussions over time.

**Information Extraction and Hierarchical Clustering:**

Social media data analytics has become a popular topic for SA researchers over the last decade (Zhao and Rosson, 2009). A variety of analytic techniques have been proposed and they largely consist of three steps: message filtering, topic extraction, and text summarization. Filtering is a fundamental but critical step as irrelevant data not only prolong the processing time, but also introduce noises and yield inaccurate results. The majority of the filtering techniques utilize lexicon based filtering (Zhou et al., 2015, Olteanu et al., 2014, Deng et al., 2017). Lexicons are dictionaries that contain situation specific contextual keywords. For example, CrisisLex is a dictionary that contains 384 keywords specific to crisis situations. The use of lexicons has been proved to be more effective than a general keyword search, as lexicons can effectively reduce the number of irrelevant messages, yet still retain the context information of event. After filtering, the next step is topic extraction. The most popular extraction method is the Latent Dirichlet Allocation (LDA) probabilistic model (Blei et al., 2003) and its derivative models. For example, the Collapsed Variant Bayesian inference for LDA (Teh et al., 2007) has
been adopted along with other text classification techniques to improve performance and reliability (Aiello et al., 2013). In general, LDA and its variants yield relatively consistent results if the events focus on a small set of structured topics with predictable patterns (e.g. extracting scores, players, outcomes from sports events) (Chakrabarti and Punera, 2011, Yuan et al., 2016), but are much less robust when there are a variety of divergent sub-events that revolved around a major topic (e.g. natural disasters) (Aiello et al., 2013). For example, Aiello et al., (Aiello et al., 2013) show that the performance of LDA is comparably better only in the case of structured social media discussions. In contrast, hierarchical clustering using the N-grams (i.e. combination of multiple keywords) outperforms other five alternative algorithms in extracting trending topics from all three datasets (Aiello et al., 2013). Hence, we adopt the Hierarchical clustering with the N-grams method in this study. The last step is to summarize the messages related to the topics derived from the previous step. Relative to topic extraction, very little research has been devoted to concisely presenting a sub-event through text summarization. One of the relatively naive techniques is to present all or a number of representative social media messages (e.g. tweets) that contain all (or most of) the topic words (O'Connor et al., 2010). However, the drawback is evident as it defers the summarization task to readers, not to mention the selected tweets may not be truly representative. Chakrabarti et al. (Chakrabarti and Punera, 2011) address the issue by considering prior probability of a message to be relevant to the sub-event, and include Hidden Markov Model (HMM) as a supervised machine learning method. Their model calculates prior probability of a message appearing in a given situation and uses this probability to determine what messages to extract in the current context. The model can be trained and refined with the addition of new data over time. Lately, sequence-based rule mining is proposed to minimize the labor involved in the manual labelling of historical messages for calculating the prior
probabilities required for HMM (Bontcheva et al., 2013). These sequence rules are created using POS (Parts of Speech) and NER (Named Entity Recognition) tags. The messages containing these tags are given a higher score and carry more weight in reconstructing a sub-event. However, the rules created for a given situation are not exhaustive and need to be manually updated based on the varying characteristics of the new data. Hence, we propose to address this limitation by developing an unsupervised extraction technique for text summarization, which can adapt to a variety of social media data characteristics.

**Design Science Research (DSR):**

Prior studies for developing SA through social media data have largely focused on extracting trending topics and summarizing messages at an event level. One common issue among these studies is that there is no systematic way for other researchers to evaluate and replicate the findings. This issue calls for a more structured methodology through a set of well-established design guidelines. To this end, DSR offers a set of great guidelines to establish the scientific merit of this research. DSR mainly constitutes seven guiding principles covering design of artifacts, problem relevance, design evaluation, research contribution towards knowledge, research rigor, design science as a search process, and communication of research (Hevner et al., 2004). To improve the rigor of the design work, the DSR-based study should include a demonstration and evaluation activity while benchmarking the robustness of the techniques against the prior studies (March and Smith, 1995, Peffers et al., 2007). The artifacts should withstand the fluctuations of environment and be extendable to a wide variety of circumstances. The evaluation of the DSR project should be backed by case studies demonstrating the usefulness and scope of future development, and showcase the utility in a real-world situation. So in essence, the DSR is a thorough and carefully designed process, which addresses a business problem to its core and keeps improvising through the iterative steps to constantly adapt to the
new data and environmental changes (Peffers et al., 2007).

**RESEARCH METHODOLOGY**

The thematic objective of our study is to provide a solution that allows all stakeholders to gain first-hand information during a crisis event, develop situation awareness as the event unfolds, and offer insights for subsequent decision making. To this end, we have developed a protocol that integrates a set of artifacts to fulfill our objective. Our methodology follows the six key phases (Identify problem & motivation, define objectives of a solution, design and develop artifacts, demonstration, evaluation and communication) as advocated in (Peffers et al., 2007, Gregor and Hevner, 2013), and provides a comprehensive solution that covers data collection, preprocessing, exploratory data analysis, topic extraction, generating summaries and a visualization dashboard, and evaluation of results. The specific steps are summarized in Figure 1 below and discussed in later section.

**Figure 1. A Data Analytic Protocol for Processing Social Media Data to Detect Emerging Events**

*The box with dash border for steps 3 and 5b indicates that only one of the two steps is needed.*
1) Select the target social network platform and analyze the API to develop data collection strategies. In Twitter’s case, the strategy is to capture the tweets related to the hurricanes using a list of keywords and hashtags for both existing and on-going tweets.

2) Visualize the data over a temporal scale to detect spikes in messages and momentum shifts in different time intervals. Gain insights on the properties and patterns of the data (This task may need to be performed periodically due to the dynamic nature of social media).

3) Filter relevant tweets using a lexicon if a well-established lexicon for the event exists. If such a lexicon does not exist, then skip step 3 and implement steps 4 and 5 to derive (or update) a lexicon.

4) Cleanse and filter data (i.e. remove unwanted characters and short tweets not specific to hurricane).

5) Extract emerging topics and creation/update lexicon.
   5a) Extract the topics using Hierarchical Clustering as the event unfolds and derive event specific keywords based on the extracted topics through forward and reverse boosting (Artifact 1).
   5b) Create a new (or update an existing) lexicon (Artifact 2) with the new keywords for subsequent filtering (e.g. upcoming hurricanes or similar natural disasters).
   5c) Select the topics of interest to generate summaries in the next step.

6) Generate summaries using an Extractive Summarization process (Artifact 3).

7) Create a SA dashboard (Artifact 4) with summaries classified into three SA levels for selected topics, and visualize the ongoing trends of the selected topics over time.

Data Collection

The recent hurricanes and the enormous amount of information exchanged on social media have provided a good testbed for our protocol. Hence, we constructed two datasets by collecting tweets posted on Twitter during the hurricanes in 2017 and 2018, to demonstrate the applicability of our protocol. We utilized the Twitter REST APIs and developed a set of web scraping tools to extract tweets from Twitter. Our two datasets include around 2.5 million and 3 million tweets for Hurricanes Harvey, Irma and Maria (August 8th to September 25th, 2017), and Hurricane Florence (September 4th to 21st, 2018), respectively.

Initial Data Visualization

Next, we present the raw data in various graphs to identify any possible patterns and gain an
overall understanding of the dataset. Figures 2a and 2b visualize the most frequent terms from the first and second week of data. Then, we categorize the data based on their various characteristics (favorites, shared, URLs to news media, blogs, images, videos, etc.) and visualize the distribution of these variables on the temporal scale in Figure 3, where multiple peaks represent the popular social events widely discussed during this sample time frame. At any hourly interval, 80% of the tweets have at least one of the high frequency words, from which we can compile sets of logically connected keywords to uncover the underlying social events.

Figure 3 also shows that the retweets, favorite tags, and the presence of media links have a significant impact on the visibility of the message and identifying the event context of, relative to those that do not have these characteristics. Around 40% of the tweets have media URLs and around 20% of the tweets attracted proactive actions by other users (i.e. liked or shared more than once). These summary statistics help us understand the characteristics and patterns of the data and focus on events and topics that occur around these spikes. The significant correlations among these characteristics also provide strong support for incorporating these properties as new parameters into our revised algorithms to be discussed next.

**Figure 2:** Most frequent words by week for first two weeks

**Figure 3:** Distribution of Tweets With At Least One Favorites/Retweet/Media Link
Initial Data Filtering Using an Existing Lexicon

Due to the large amount of data in a crisis situation, data filtering primarily involves excluding tweets not related to the topic of interest (e.g. a sports news circulated during the hurricane), or tweets that are collected due to the synonyms (e.g. tweets related to people or events with the same names as Harvey, Irma and Maria). Two approaches can be used: a lexicon/keyword-based approach and classifier-based approach (Deng et al., 2017). For events that occur frequently, a lexicon that consists of a set of event-related keywords can be compiled a priori and used to filter relevant messages (Olteanu et al., 2014). For example, for crisis events, the CrisisLex lexicon was initially compiled by having respondents evaluating the relevance of the tweets and keywords. Additional keywords were added by applying the Bayesian probability rules to labeling the tweets as crisis or non-crisis related. In contrast, classifier-based approach separates the tweets into two groups using binary classification algorithm, which falls under the category of supervised learning as it requires a pre-trained dataset that contains binary outputs of ‘yes’ or ‘no’ for a prior crisis event. To avoid the manual efforts involved in a supervised approach, we adopt the CrisisLex lexicon in our filtering. Since it is created for general crisis, we also derive additional keywords to create a hurricane specific lexicon (see §3.5).
Data Cleansing and Further Filtering

The social media data often contain unstructured data such as emoji, emoticons, URLs, image short links, other web generated content, and certain unwanted characters. All these undesirable inputs should be trimmed so that individual words or tokens can be more efficiently extracted.

Extracting Topics (Artifact 1) and Updating/Creating the Lexicon (Artifact 2)

Once the data are filtered and cleansed, the next step in our protocol is to extract topics using the Hierarchical Clustering method. As mentioned earlier, conventional clustering algorithms such as LDA and K-means are not appropriate for social media data as a regular clustering algorithm based on the TF-IDF technique (term frequency - inverse document frequency) considers all repetitive messages (i.e. retweets) as a single large cluster and may ignore the emerging topics from new incoming messages (since the old, repetitive topics still carry a large weight), making it difficult to identify the development of new topics (Sharifi et al., 2010). To address this limitation of over-weighing repetitive messages, we implement two additional filtering techniques. Specifically, we use reverse boosting to reduce the weight assigned to topics already identified and repeatedly extracted in the most recent four hours and use enhanced boosting to assign more weights to newly extracted topics (e.g. last hour). This disproportional weighing scheme can be achieved through the DF-IDF (document frequency - inverse document frequency) technique (Aiello et al., 2013), and has improved the throughput of Hierarchical Clustering by 1.5 times in our analysis.

The generated topics are used to create new CrisisLex lexicon words, which will facilitate detection and summarization of future crisis events. Specifically, topics generated by the

---

1 TF-IDF is a statistic that captures how important a word is to a document in a collection or corpus. TF-IDF value increases with the number of times a word appears in a document and decreases with the number of documents that contain the word.
clustering algorithm are labeled as either crisis-related or not, and passed as input to the CrisisLex pseudo relevance ranking process. The process starts by selecting crisis-related bigrams and unigrams and then scoring the terms based on their distribution within the tweets. A sigmoid function is used to limit the threshold value between 0 and 1.0 while evaluating the mean score. The terms that score greater than the commonly recommended threshold of 0.6 are considered relevant and retained for further processing (Olteanu et al., 2014). The remaining keywords are ranked using a graphical model based on their scores. The first iteration of CrisisLex consists of a total of 380 crisis related terms. Table 1 summarizes the 36 new keywords that we derived and added to the updated lexicon.

Table 1: New keywords added to the CrisisLex Lexicon

<table>
<thead>
<tr>
<th>animals</th>
<th>displaced dogs</th>
<th>insurance claims</th>
<th>shelter list</th>
</tr>
</thead>
<tbody>
<tr>
<td>animals stranded</td>
<td>dogs</td>
<td>move prisoners</td>
<td>shelters</td>
</tr>
<tr>
<td>beach residents</td>
<td>donations scam</td>
<td>people suffering</td>
<td>storm shelter</td>
</tr>
<tr>
<td>charity fraud</td>
<td>emergency responders</td>
<td>prisoners died</td>
<td>storm warning</td>
</tr>
<tr>
<td>church</td>
<td>evacuate prisoners</td>
<td>recovery efforts</td>
<td>supplies</td>
</tr>
<tr>
<td>church hit</td>
<td>food truck</td>
<td>recovery plan</td>
<td>supply camps</td>
</tr>
<tr>
<td>contaminated</td>
<td>gas</td>
<td>road closure</td>
<td>tesla</td>
</tr>
<tr>
<td>devastation</td>
<td>gas lines</td>
<td>roads closed</td>
<td>tesla battery</td>
</tr>
<tr>
<td>displaced animals</td>
<td>gas shortage</td>
<td>sewer water</td>
<td>wind gust</td>
</tr>
</tbody>
</table>

Such an updated lexicon can always be used as the benchmark tool in further iteration of our protocol or other related studies. Steps used in updating the lexicon are outlined below, which applies to a situation where a lexicon is not available or needs update.

Step 1: Use an existing lexicon (i.e. CrisisLex) or a set of event-related keywords (if a lexicon is not available) to filter the tweets.

Step 2: Generate the trending topics using the Hierarchical Clustering method.

Step 3: Classify tweets as related and not-related based on the generated topics.

Step 4: Apply the Pseudo Relevance Feedback (PRF) process proposed in (Olteanu et al., 2014) and extract Unigrams and Bi-Grams for the new keywords.
Step 5: Filter and exclude keywords that are specific to a given event - i.e. name of hurricane (e.g. Harvey, Irma), organizations, people and location (e.g. Rockport, Houston, Florida).  

Step 6: Update the lexicon with the new keywords that are not used in step 1.

Artifact 3: SMMS – Social Media Message Summarization Based on TextRank

Decision makers require concise and accurate summaries to comprehend the situation and take proper actions. In the Information Retrieval literature, the task of generating succinct and informative summaries is called extractive summarization. Before generating the summary, the topics of interest can be selected by users (i.e. through a dashboard to be discussed in section 3.6) and a set of keywords representing each topic (derived during the topic extraction stage) will be passed to the summarization algorithm as inputs to produce the summaries. To generate summaries, the graphical network-based approach that uses word co-occurrence and counts has recently emerged as a viable tool (Sharifi et al., 2010). A graphical model consists of nodes (i.e. individual words or sentences) and edges (i.e. the relationship between nodes). Based on the word co-occurrence relationship, the closest and similar context words can be identified. Then, the sentences with most number of these similar words are extracted as summary. It is worth noting that using single words as nodes may generate a lot of noise in the summary as the context of the words is not preserved and some words may be totally irreverent. Therefore, in our study we adopt TextRank (Mihalcea and Tarau, 2004), a graph-based text summarization algorithm which uses a complete sentence, instead of individual words to examine word co-occurrence. TextRank has been shown to provide better results in extracting the most representative information among a variety of distinct sentences or messages (Petrović et al., 2010).

---

2 These names are only relevant for a given crisis. Future studies can always add any event specific names to the lexicon.
Summarizing tweets for the selected topic is carried out in two steps: a relevance/similarity score that measures how relevant each sentence is to another sentence with a similar context will be computed, and is used to rank the sentences and extract the summary (Nichols et al., 2012, Gong and Liu, 2001). Relevance of a message to another is a primary measure for the syntactic and semantic meaning of a summary, which is often evaluated through common words, certain statistical measures (e.g., pairwise similarity, longest common words, cosine similarity using TF-IDF, etc.) and semantic concepts (e.g. association link network). The relevance score helps researchers uncover the relationship between messages and derive the underlying topics. However, when it comes to twitter, relevance score calculated based on common words is not feasible as each tweet is independently written by unique user and the same meaning can be conveyed in different forms and words, and some messages may contain sarcastic comments that deviate from the textual meaning.

In addition to these limitations, previous research (Barrios et al., 2016, Fang et al., 2017) and our subsequent validation using hurricane data show that the TF-IDF-based similarity metrics may lead to five major challenges/limitations for text summarization in social media context, namely, similar context and words, contextually similar words, unrelated messages with common words, short texts, short vs long messages. To overcome these limitations and accommodate the nature of social media data, we develop a modified ranking algorithm that explicitly incorporates various social media specific characteristics (Vieweg et al., 2010) such as the use of contextual keywords extracted from trending topics, numbers of retweets and favorites (Starbird and Palen, 2010), the presence of URLs that link to news media, blogs, images, videos, etc. (Starbird and Palen, 2010), along with the cosine similarity to collectively measure the relevance of the message. The importance of adopting such a joint measure can be illustrated in the following scenarios, as presented in Table 2.
### Table 2: Sample Tweets and Key Measures of Tweet Characteristics

<table>
<thead>
<tr>
<th>Tweet Content</th>
<th>Number of similar tweets in entire sample / hourly interval</th>
<th>Total number of shared or liked</th>
<th>Presence of contextual keywords</th>
<th>Should be included in Summary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you want to help hurricane survivors, but don't know where to donate...give to @TeamRubicon!</td>
<td>87,981 / 2,416</td>
<td>0</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Here I am. Rock you like a hurricane. pic.twitter.com/gXVexhaDaj</td>
<td>101 / 0</td>
<td>1</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>George W. Bush want to denounce racism.... as if he didn't allow a city full of Black people to drown during Hurricane Katrina</td>
<td>1,504 / 0</td>
<td>813</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Hurricane Gert AKA Hurricane GRRRRRT is further confirmation that the brand is strong. pic.twitter.com/MhUHgQW4DP</td>
<td>666 / 0</td>
<td>283</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>If you’re in hurricane-affected areas &amp; need relief (or want to help), here are some resources: <a href="http://harveyrelief.handiworks.co/relief-map">http://harveyrelief.handiworks.co/relief-map</a></td>
<td>56,413 / 2,168</td>
<td>3,063</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

In the first tweet, even though there are many similar tweets and the content of the tweet seems to be closely related to hurricane (i.e. it has the contextual keywords “hurricane” and “donate”), it was never shared by other users because such a tweet does not come from an authentic source and may be deemed a phishing message. Hence these tweets should be excluded from the summary (unless the study is to analyze spams and phishing attempts). We consider the case of having many tweets with similar contextual keywords but no circulation as the “similar context and words but untrustworthy” scenario. The second, third and fourth tweets are examples that contain keywords seemingly relevant to the context but are not referring to the current hurricanes, we refer to this scenario as “unrelated messages with common words”. Even though they were retweeted multiple times in the entire sample period, there was no similar tweet (second number in the first column) during the specific hours where the hurricane occurred or the related discussions prevail. Hence, these tweets are excluded from the summary. This is not the case with the fifth tweet, which has a large number of similar tweets both within the entire
sample period (56,413) and during the peak hurricane hours (2,168). It was also retweeted 3,063 times in a short period of time and meets the presence of contextual keyword requirement. Therefore, it falls into a “contextually similar words” category and should be considered as an important and highly relevant message to be included in the summary.

To incorporate these new parameters, we modify the algorithms derived from TextRank as adopted in (Barrios et al., 2016, Fang et al., 2017). The original similarity between sentences $i$ ($s_i$) and $j$ ($s_j$) is computed based on common words $w_k$:

$$\text{Similarity}(s_i, s_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{(\log_{10}(|S_i|) + \log_{10}(|S_j|))}$$  

(1)

Some recent studies (Barrios et al., 2016) have attempted to address the unequal length of sentences using Opaki BM25 as shown in Formula (2), which is considered the state of the art similarity measure formula for structured text.

$$\text{Score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgDL}})} + z$$  

(2)

In the above formula, $Q$ is the query strings containing keywords set $\{q_1, q_2, ..., q_n\}$, and in the context of social media messages, the set of topic words. $D$ is the part of document set used for summarization, which is the set of tweets in our context. $|D|$ is the total number of words in document $D$. Parameter $z$ captures the presence or absence of training data. $k_1$ and $b$ are constants that are context-dependent and need to be optimized during application.$^3$ $\text{AvgDL}$ is the average length of the sentences in the text corpus. $\text{IDF}(q_i)$ denotes the Inverse Document Frequency weight of the query term $q_i$, and is computed using Formula (3).

---

$^3$ Following the recommended values (1.2 to 2.0 for $k_1$ and 0.75 for $b$) from [41], we found that the values of 1.3 for $k$ and 0.75 for $b$ yield the most optimal results. In Figure 7, we compared our results against summaries generated by BM25 with these values.
\[
    \text{IDF}(q_i) = \log \left( \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \right)
\]

(3)

In the above formula, \( N \) and \( n(q_i) \) respectively denote the number of tweets and documents containing \( q_i \).

From Formulae (2) and (3), it is clear that even though the Opaki BM25 ranking algorithm takes into account the unequal length of inputs, it still does not adequately address several of the limitations discussed earlier (i.e. cannot handle unrelated messages with common words, unable to rank contextually similar tweets, etc.). Therefore, after carefully evaluating the traditional similarity measure and the Opaki BM25 algorithm, we choose to modify Formula (1) to measure the pairwise similarity between every other tweets, and we name this new formula inter-sentence similarity score (ISSS) as shown below:

\[
    \text{ISSS}(S_i, S_j) = \frac{|\{w_k \mid w_k \in S_i \& w_k \in S_j\}| + w_t + \rho}{(\log_{10}(|S_i|) + \log_{10}(|S_j|) + \log_{10}(|S_t|) + \log_{10}(|\rho|))}
\]

(4)

In Formula (4), \( w_t \) comes from the set of words produced by the clustering algorithm for topic \( t \), and \( \rho \) denotes a normalized real number calculated using parameters that capture input properties (e.g. number of retweets, favorites and the presence of media).

To a large extent, the inclusion of these parameters addresses the first two limitations discussed earlier as we will not solely depend on the contextual words that result in the first two limitations. With regard to the third limitation, even though there are relatively fewer sarcastic comments in our hurricane dataset, our similarity measure still minimizes the possibility of incorporating unrelated messages by discarding messages that do not attract the attention of other users. Next, we address the fourth limitation by explicitly incorporating short texts that generate a lot of circulation (e.g. being retweeted or liked a lot). Finally, by assigning different weights to
sentences with different lengths and different tweet characteristics, we address the last limitation by giving more weights to long sentences and sentences that have tweet characteristics (e.g. URLs, retweets). Through these extensions, our approach is expected to yield better summarization results than the previous methods. In Section 4, we will present a comparative summary in which we validated our proposed method along with these popular methods using our social media dataset.

Once the sentence similarity scores are calculated, the sentences for summarization will be ranked using the graphical model in which the edges are formed by connecting every two sentences (vertices) in the dataset. The similarity between these sentences is defined as the edge score in graphical representation. To rank these sentences, we used the TextRank algorithm which is based on the popular page ranking algorithm (Page et al., 1999) developed to rank web pages. TextRank relies on both the local context of a sentence (vertex) and the similarity between the sentence (edge score). It uses the combined weights of similarities drawn from the edge scores with the current vertex as a reference. Thus, the weighted sum of each vertex provides the overall vertex score and can be calculated using Formula (5):

\[
WS(V_i) = (1 - d) + d \sum_{V_l \in In(V_i)} \frac{w_{jl}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)
\]  

(5)

In the above formula, \(w_{jl}\) is the similarity score calculated for the corresponding edge and \(d\) is the damping factor that can be set between 0 and 1 (we use the recommended value of 0.85 following (Mihalcea and Tarau, 2004)). For a given vertex \(V_i\), \(In(V_i)\) represents the set of vertices that point to it (predecessors) and \(Out(V_i)\) are the set of vertices that \(V_i\) points to (successors). Once the vertex scores have been calculated, the input texts can be ranked based on the vertex scores and summaries can be generated for the selected topics from the most relevant
sentences. Figure 4 shows a snapshot representation of how the protocol will be carried out and how the 2.5 million raw tweets shown on the left are processed to detect the social events shown in the middle column, which in turn are summarized and presented in the right column.

**Figure 4: Keys phases to summarize thousands of social media messages**

**Situation Awareness (SA) Dashboard (Artifact 4) and Visualizing Key Events**

During crisis situations, classifying topics and events based on the three levels of SA (Endsley, 1995) and visualizing the development of these events on an interactive interface help decision makers detect the rise of a critical event, monitor its ongoing development, and predict its direction and the possible occurrence of a related event. To demonstrate the utility and applicability of the developed artifact, which is one of the key aspects of DSR (Peffers et al., 2007), we develop a SA dashboard and present several sample screenshots in Figures 5a, 5b, and 5c. The dashboard is developed using Tableau, a Business Intelligence software that specializes in interactive data visualization. Using this widely used software allows us to demonstrate the
applicability of our protocol in a relatively fast and effective manner.

**Figure 5a. Situation Awareness Dashboard**

In Figure 5a, the bottom portion of the dashboard provides an initial understanding on the trending topics through a line graph, with each line representing the development of a topic in a customizable time frame, and different colors and labels differentiating the various topics derived from our analytic tools. The spikes in a line indicates the bursts of tweets. If a user wants to gain more information about a given topic, she can click on the associated trend line (e.g. the one next to the red arrow in Figure 5a) and the summary for the selected topic in a given time period (e.g. hour) will be shown on the upper left portion of the dashboard in the ascending order of time. These summary messages provide more details on the selected topic such as the message posting time, the actual message content (when hovering over a message – see the red rectangle box in Figure 5a), and the relevance score computed by our summarization algorithm. Note that by default
the dashboard displays all the summary messages from the time when the topic is detected. If a user wants to focus on examining the information related to a specific stage of SA for the chosen topic, the dropdown menu (highlighted by a red circle in Figure 5a) will classify the messages based on either the perception, comprehension, or projection stage. We use qualitative content analysis method recommended in (Vieweg et al., 2010, Verma et al., 2011) (based on a set of keywords selected for each stage of SA theory as presented in Table 3) to segregate the messages into the three stages of SA.

**Figures 5b and 5c.** First topic specific messages by hour (Left) Detailed hourly topic specific summary (Right)

<table>
<thead>
<tr>
<th>Method used for Classifying SA messages</th>
<th>Stage 1: Perceived Elements</th>
<th>Stage 2: Topics Comprehension</th>
<th>Stage 3: Projection of Future Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Based on keywords derived from content analysis done in prior studies</td>
<td>Warning, location specific, time specific, and visibility</td>
<td>Preparatory activity, Volunteer information, emergency helpline, and message sharing</td>
<td>Forecasting, Animal management, and damage/Injury reports.</td>
</tr>
<tr>
<td>Timeline based categorization of SA levels</td>
<td>Tweets during the important phases of an event</td>
<td>Visualization of the trending topics within the last 24 hours.</td>
<td>Drill-down visualization of trending topic within the last one hour.</td>
</tr>
</tbody>
</table>

Table 3. Ways to Classify Messages into the Three Stages of SA *

---

* Only a selective number of sample keywords are shown here for illustration purposes.

---

4 For example, a tweet that reads "anticipating heavy rains, traffic will be diverted to ..." will be classified as a "projection" stage message as both the words "anticipating" and "will be" indicate the expected occurrence of a future event.
In addition to using the classification tool, users can also better comprehend an event through the visualized trends of the topic-related messages in the upper right portion of the dashboard. For example, users can switch to graphical views displayed in Figures 6a through 6f which show the distribution of tweets associated with a social event/topic in an adjustable time frame. These graphs document the momentum shift in topic discussions which can be used to forecast the future evolution of a given topic/event.

**Figure 6:** Weekly tweets distribution by topic
In Figure 6a, the initial discussions revolved around the possibility of storm in Florida and Hurricane Gert, towards the middle of second week all the discussions have shifted to Hurricane Harvey and Houston being flooded, which are shown in Figure 6b, and to Rockport, Texas devastation, and Oil shortage in other figures. These graphs provide a high-level understanding on the development of topics along with their duration and intensity. If more insights are needed, user can click on the individual message to switch to a full screen details view that displays all of the most representative tweets in each hour as shown in Figure 5b. If a user wants to see all the tweets that are summarized by our text summarization algorithm during a specific hour, the user can click on the tweet time (highlighted as a red box in Figure 5b) and the related summaries will be displayed in descending order of their relevance scores as shown in Figure 5c.

RESULTS, EVALUATION, AND DISCUSSION

Evaluation is one of the critical steps in DSR to demonstrate the usefulness of the developed artifacts (Gregor and Hevner, 2013), which involves comparing the results against the objectives, and analyzing the effect of the artifact based on functionality and performance (Peffers et al., 2007). We evaluate the utility of the overall protocol based on the topics and summaries generated from the artifacts we developed. We consider validity, utility, quality, efficiency and applicability as the key dimensions in our evaluation. Validity is measured by comparing the
extracted topics with those retrieved from mainstream news media and see if the majority of the topics reported in media were also identified by our tools. Utility is determined by whether our analytic tools can provide additional value beyond what an individual user can obtain using a conventional approach (e.g. by showing that our tools derive topics not reported by search engine, online news websites, etc.), which is especially important when the public emergency system (e.g. 911) is overwhelmed and could not provide such alerts. Quality is measured for both generated topics and summaries by benchmarking the generated summaries/topics against well-established evaluation metrics such as Recall and Precision. Efficiency is assessed by whether our tools can summarize messages on time, which is measured by the time needed to generate the topics and summaries using our artifacts after the relevant tweets have been posted. Finally, to assess the applicability of our protocol, we demonstrate the results using two case studies on hurricanes occurred during 2017 and 2018. In particular, in the second study we generated the topics as we collected the data on an hourly basis, which makes the artifacts a near real-time solution to achieve SA.

**Validity and Utility – Evaluation of extracted topics against online news media topics**

To some extent, validity and utility serve as the necessary and sufficient conditions for justifying the value of our protocol and tools. Essentially, we want to ensure that the major topics reported by other information sources are not left out in the topics we generated, yet at the same time our tool can uncovers topics not reported by other conventional approaches. These two aspects together establish the artifact’s legitimacy and help us assess whether the intended purpose of the developed artifact is fulfilled (Sonnenberg and Vom Brocke, 2011, Gregor and Hevner, 2013). Table 4 presents all the widely-discussed topics overlapping over multiple hours during the three hurricanes in our sample time frame, along with the critical topics reported on popular mainstream news media.
<table>
<thead>
<tr>
<th>Hurricane</th>
<th>Date</th>
<th>Derived Topic</th>
<th>Date</th>
<th>News Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gert</td>
<td>8/15/17</td>
<td>Gert impact to UK via North Atlantic</td>
<td>8/15/17</td>
<td>Atlantic storm upgraded to Hurricane Gert</td>
</tr>
<tr>
<td>Gert</td>
<td>8/16/17</td>
<td>Food supply impact. Excessive shopping</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Irma</td>
<td>8/19/17</td>
<td>Category 2 Hurricane Irma possible</td>
<td>8/19/17</td>
<td>Hurricane Center Watching Three Possible Systems - Florida Storms</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/19/17</td>
<td>Possible Florida Storm</td>
<td>8/18/17</td>
<td>Tropical depression, coastal areas of Florida and Texas on alert</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/21/17</td>
<td>Austin pride festival</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Irma</td>
<td>8/21/17</td>
<td>Join helping hands to rescue stranded vehicles</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Irma</td>
<td>8/24/17</td>
<td>Alligators concern Houston residents</td>
<td>8/25/17</td>
<td>Hurricane Harvey Warning: Beware of Alligators</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/24/17</td>
<td>Gas prices at risk of hike</td>
<td>8/25/17</td>
<td>Hurricane Harvey fuels rise in gas prices as refineries shut down. Harvey could hurt oil production long after it hits the Houston area</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/25/17</td>
<td>Rockport residents act quickly to evacuate</td>
<td>8/26/17</td>
<td>‘Catastrophic’ flooding expected near Rockport</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/25/17</td>
<td>Too soon to flee Rockport, too late to stay</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/25/17</td>
<td>Immigration checkpoint continue operations</td>
<td>8/25/17</td>
<td>Border Patrol says Texas checkpoints to remain open during Hurricane Harvey</td>
</tr>
<tr>
<td>Harvey/Irma</td>
<td>8/28/17</td>
<td>Possible Hurricane Harvey donation scams</td>
<td>8/28/17</td>
<td>Watch out for these charity scams after Hurricane Harvey Where to Donate to Harvey Victims</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/28/17</td>
<td>Lakewood church not open for shelter</td>
<td>8/28/17</td>
<td>Joel Osteen responds to accusations for not opening the church</td>
</tr>
<tr>
<td>Harvey</td>
<td>8/29/17</td>
<td>Houston doctors helping victims</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Irma</td>
<td>8/30/17</td>
<td>Florida evacuation plan</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Irma</td>
<td>9/2/17</td>
<td>Increase in Gas prices</td>
<td>9/1/17</td>
<td>Gas prices spike in Texas &amp; Atlanta</td>
</tr>
<tr>
<td>Harvey</td>
<td>9/4/17</td>
<td>Volunteers helping with food trucks</td>
<td>9/4/17</td>
<td>A food truck offers a cooked meal to residents with no gas or power</td>
</tr>
<tr>
<td>Harvey</td>
<td>9/5/17</td>
<td>Tesla cars battery issues and don’t start</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Irma</td>
<td>9/7/17</td>
<td>Hurricane Irma pummeling Turks and Caico Islands</td>
<td>9/8/17</td>
<td>Turks and Caicos Islands hit by Hurricane Irma</td>
</tr>
<tr>
<td>Harvey/Irma</td>
<td>9/8/17</td>
<td>Post Harvey ruins brings more mosquitoes and diseases</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Harvey</td>
<td>9/10/17 05:00</td>
<td>Tesla extends range on cars to help owners avoid Hurricane Irma</td>
<td>9/10/17 12:00</td>
<td>Tesla gives battery boost to those fleeing Hurricane Irma</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>Harvey</td>
<td>9/11/17 11:00</td>
<td>communities partner to provide basic supplies for Harris County</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Florence</td>
<td>9/9/18 17:00</td>
<td>Available Sand Bags for protection</td>
<td>9/9/18 09:00</td>
<td>Sandbags available in Charleston County as Hurricane Florence nears</td>
</tr>
<tr>
<td>Florence</td>
<td>9/10/18 11:00</td>
<td>Walmart in Georgetown close at midnight</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Florence</td>
<td>9/10/18 05:00</td>
<td>Mobile App for Hurricane Planning</td>
<td>9/11/18 14:00</td>
<td>Hurricane Florence prep: apps to help you stay safe</td>
</tr>
<tr>
<td>Florence</td>
<td>9/10/18 22:00</td>
<td>Farm animals trapped in cages</td>
<td>9/11/18 09:00</td>
<td>As hurricane nears, U.S. farmers rush to clear crops but animals stay in storm's path</td>
</tr>
<tr>
<td>Florence</td>
<td>9/10/18 23:00</td>
<td>Long lines for Gas</td>
<td>9/11/18 13:00</td>
<td>Hurricane Florence gas shortages pop up in North Carolina</td>
</tr>
<tr>
<td>Florence</td>
<td>9/10/18 23:00</td>
<td>Hospitals across the Lowcountry preparing for Hurricane Florence</td>
<td>9/11/18 05:00</td>
<td>Local hospitals evacuate patients ahead of Hurricane Florence</td>
</tr>
<tr>
<td>Florence</td>
<td>9/11/18 03:00</td>
<td>Prison inmates not evacuated</td>
<td>9/11/18 10:00</td>
<td>Inmates at South Carolina prison in evacuate zone not being moved ahead of Florence</td>
</tr>
<tr>
<td>Florence</td>
<td>9/11/18 23:00</td>
<td>A caravan of ambulances helped transport patients to safety</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Florence</td>
<td>9/12/18 10:00</td>
<td>Rumor Control Page by FEMA</td>
<td>9/14/18</td>
<td>FEMA creates Hurricane Florence rumor control page</td>
</tr>
</tbody>
</table>

For comparison purpose, the time of the initial appearance of media news article for any given topic is recorded and verified against the time when the same topic is extracted by our artifact. On average, more than 80% of the topics covered by mainstream news media were detected earlier by our tool. For example, crucial and important topics that impact normal life, such as “gas shortage”, “evacuation path”, “shelter occupancy”, “congestion on highway”, and “donation scams” are all extracted by our protocol earlier than by the news websites. More critically, some of the highly-ranked topics extracted by our tool are not reported on the news websites (e.g. cells with “-” in Table 4), which shows that such critical information for victims, first respondents, and policy makers can be inadvertently ignored by the mainstream media as
these are not their usual focuses of coverage during the time of the crisis. In general, the favorable results can be attributed to the reverse boosting technique in our algorithm, which effectively ignores the previously generated trending topics that tend to be the focus of the mainstream media due to the high frequency of appearance. In addition, our algorithm is very sensitive to the emergence of the newly trending topics. For example, the reminder of Houston doctors providing help to hurricane victims with specific location information and the warning of potential Tesla cars being stranded due to battery issue immediately caught the attention of our analytic tools due to the surge of discussions in a short amount of time. Such critical information was, however, never reported in mainstream news media.

**Quality – Comparing Generated Topics/Summaries against Reference Topics/Summaries**

The quality of artifacts developed depends on how successful the algorithm detects all the relevant topics and whether the summary messages effectively reveal the key information revolving around each topic (Lin, 2004). We use the widely-adopted Recall and Precision metrics (Aiello et al., 2013, Gong and Liu, 2001, Chakrabarti and Punera, 2011) to assess the quality of our outputs. Recall measures the percentage of relevant instances that have been retrieved by the analytic tool over the total number of possible relevant instances (i.e. the completeness of results), while Precision measures the percentage of the relevant instances among all the retrieved instances (i.e. the accuracy of the results). Given the context of our study, we specifically examine the Recall measures for both topics and keywords, and the Precision measure for keywords only due to the same challenge faced in (Aiello et al., 2013).\(^5\)

To capture Recall and Precision, we need the reference topic keywords and the reference summaries to serve as benchmarks. These references were generated by two independent

\(^5\) Measuring topic precision requires us to analyze all the topics that emerged in our sample (to compute the denominator of the Precision measure), including the topics that are not hurricane related, which is practically impossible and adds very little marginal value to this research. Hence it is not conducted as done in other studies (e.g. [2]).
researchers who do not know the purposes of this study. They were provided a number of tweets in multiple time intervals from both datasets, and were asked to generate topic keywords similar to how they are generated using a BN Gram model. For reference summaries, they ranked the tweets in the descending order of importance based on the keywords provided to them. After each of them independently completed the process, the results are compared and discussed between the two researchers until mutual agreements are reached. These references are then used in producing the Recall and Precision measures are explained below.

**The recall measure for topics**

In our study, a topic is considered successfully detected when the automatically produced set of keywords contained all the relevant keywords for this given topic. For comparison, we also compute the topics Recall generated using the LDA model, the original BN Gram model, the BN Gram with aggressive filtering and contrast them with that measure generated from our model. As shown in Figure 7, our topics Recall measure (the blue bar of the rightmost group) clearly outperforms those of other three methods.

**Figure 7: Precision and Recall Measures for Topics & Keywords**

![Comparison of Topic Extraction Methods](image)

**Keyword “Recall” and keyword “Precision”**

Since the primary step to generate the summaries is to pass the topic keywords that represent the topics for which the summary needs to be generated. Hence, we focus on the Recall and Precision measures for keywords within each hourly window. Keyword Recall is the percentage
of correctly detected keywords over the total number of keywords from all the relevant topics in a given interval from the reference summaries by the independent evaluators (samples shown in Table 5), and keyword Precision is the percentage of correctly detected keywords over the total number of keywords in the summaries generated by our analytic tool (Chakrabarti and Punera, 2011, Gong and Liu, 2001, Aiello et al., 2013). For both metrics, we consider the fact that the same keyword may be related to multiple topics/events. Hence, the computation is conducted within each topic. As shown in Figure 7, both measures outperform their counterparts generated by the LDA algorithm by more than 50%, and are both comparably better than the same measures generated by the two BN Gram models.

**Table 5:** Samples of reference summaries vs. generated summaries

<table>
<thead>
<tr>
<th>Topic</th>
<th>Date</th>
<th>Reference Summaries</th>
<th>Generated Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandbags</td>
<td>9/9/18</td>
<td>Helping hands: Hanahan High School student athletes volunteered to fill sandbags for the community in advance of Florence. @SCDCNews: In advance of Hurricane Florence @SCDCNews sandbagging operations have commenced.</td>
<td>In advance of Hurricane Florence sandbagging operations have commenced. People in North Carolina have been filling sandbags and stocking up on wooden panels to reinforce their homes ahead of the arrival of the potentially devastating. Helping hands Hanahan High School students volunteered to fill sandbags.</td>
</tr>
<tr>
<td>Walmart</td>
<td>9/10/18</td>
<td>Walmart in Georgetown will close at midnight tonight (Monday) due to Hurricane Florence. The store will remain closed until further notice. Can confirm there is no water to be found at food lion, Walmart, Harris teeter, and surrounding gas stations in Morrisville. I saw people stealing things at Walmart. Police are here now. People are losing their dang mind…</td>
<td>Walmart in Georgetown will close at midnight tonight Monday due to Hurricane Florence. The store will remain closed until further notice. Can confirm there is no water to be found at food lion, Walmart, Harris teeter, and surrounding gas stations in Mor. The shelves at Walmart are bare as Hurricane Florence approaches New Bern.</td>
</tr>
<tr>
<td>FEMA Mobile App</td>
<td>9/10/18</td>
<td>@fema: Hurricane #Florence is forecast to cause inland flooding, life-threatening storm surge, and damaging winds along the Carolinas.</td>
<td>FEMA website, A mobile app is also available, which might come in handy in the coming week</td>
</tr>
<tr>
<td>Farm Animals</td>
<td>9/10/18</td>
<td>Hurricane Florence threatens to kill thousands of farm animals and trigger catastrophic spills of waste as it bears</td>
<td>Hurricane Florence threatens to kill thousands of farm animals and trigger catastrophic spills of waste as it bears</td>
</tr>
</tbody>
</table>
down on Carolina coastal area dotted with sewage treatment plants, hog waste lagoons, poultry farms and coal ash ponds.

Great to see evacuating pets for #HurricaneFlorence. But what about farm animals NC factory farms left millions to drown.

Text Summarization Results

With the list of topics, the associated topic words, and the SA dashboards, users will be able to determine what topics have higher priority and need to be further explored (step 5c). Once the topics are identified, users can selectively perform text summarization using artifact 3 of the protocol to gain a more comprehensive understanding of the evolution of the emerging events. Situationally or contextually related tweets are given higher priority based on the modified text summarization ranking proposed earlier and summaries will be generated for the topic selected by the user. Figure 8 shows a variety of sample summaries through which we can gain some important insights and enhance SA. For example, there is an important notification that residents should not solely rely on social media for help, and another important alarm that cyber criminals may try to take advantage of the sympathy for hurricane victims and set up various donation scams to defraud the community. It also provides valuable help information for people stranded in certain areas, or provide information for people who want to volunteer to assist these victims.

Figures 8 and 9: SA through extracted summaries and Sample summaries for 4 different periods
Figure 8 provides an automated form of summary report that consists of multiple top-ranked tweets displayed in the visualization tool related to a selective set of topic keywords (e.g. “donations”, “food”, “gas”, “victim”, and “evacuees”). In Figure 9, the size of the individual boxes that contain the summary tweets varies based on the number of tweets within the interval are summarized. The upper left summary alerts citizens (or the community) to two different atrocities during hurricane donations, while the upper right summary provides very important information about the various shelter locations for Harvey evacuees. The lower left summary offers timely update about gas shortage and availability, and the lower right summary helps victims locate the food donation information. These are various types of information can enhance situational awareness and help hurricane victims stay informed during such disastrous moments.

Figure 9 also shows the complete hourly summaries related to the topic words “evacuate, road, inundate, avoid, route, enroute, gas, rescue, map, plan, Harvey, help, donate” derived from more than 18,000 tweets in this interval. Through these summaries, we observe various different topics being discussed related to roads being inundated with water, the map of regions affected by hurricane, victims requesting for help, and people lending support through donations. Such a single screen representation saves the user from spending a lot of time and effort browsing through the vast number of tweets, thus enabling timely decisions and responses to these emergencies.
In addition to the rich information provided by our tools, our improved algorithm also leads to a much-reduced dataset for crisis-related tweets after performing initial filtering, and the adoption of the updated crisis lexicon. In Figure 11, it can be observed that over 40% of the tweets are filtered using the updated CrisisLex lexicon that includes the 36 new keywords, thus reducing the time required in subsequent processing of the tweets using the clustering and text summarization algorithms. In the actual execution of these algorithms to generate topics and summaries, the reverse boosting technique we develop not only improves the accuracy of the results, but also substantially shortens the processing time. For example, Figure 12 shows considerable differences in processing time when reverse boosting is adopted versus when it is not used, suggesting a significant efficiency gain.
LIMITATIONS AND FUTURE RESEARCH

Even though we have demonstrated that our proposed protocol can effectively recover emerging topics from a large number of social media messages (tweets), produce comprehensible summaries and outperform several contemporary algorithms, as any information management system, our tool is not without limitations. First, our approach focuses on capturing topics or emerging events in crisis situations and does not explicitly incorporate a sentiment analysis component to capture the development / convergence / divergence of specific sentiments or opinions. With natural disasters such as hurricanes, there is a relatively small chance that the messages contain a lot of sarcastic comments or disputes over a particular topic. However, such obscure data characteristics may be present in political events such as elections, protests, or political rallies. Even though our research is not constrained to a particular domain, the context of the application may affect the effectiveness of the tool. Thus, to develop a more versatile data analytic tool, future research may need to consider the inclusion of sentiment analysis and/or other deep learning components to offer an integrated solution for a broad range of contexts.

Second, our summarization results are based on the segregation of tweets posted in multiple one-hour windows. Even though the division of data based on this time interval yields good result in our study, such an hourly interval specification may need to be adjusted or optimized for other
datasets. Ideally the tool should allow users to easily adjust the time interval on a real-time basis, or optimize the time interval based on the characteristic of the social media data (e.g. volume, length, matching keywords, data noise, etc.). Third, when generating new keywords, lexicons such as CrisisLex require classified text marked in binary coding as crisis related or not, which is a manual process. Therefore, the implementation of our protocol can be slowed down if no lexicon exists (e.g. when applied to a new type of disaster or emerging event) and a significant portion of the lexicon needs to be updated. Although this problem can be overcome using content analysis or by classifying the new keywords via an unsupervised classification algorithm like Naïve Bayes algorithm, we did not explore the possibility of integrating these tools into our framework as the main focus of the study is to demonstrate the application of our proposed protocol in some frequently occurring crisis situations. We encourage other researchers to consider one of these extensions in future studies.

CONCLUSIONS

SA calls for a multi-faceted approach due to its complexity and unpredictable nature, particularly when dealing with social media data. With the ever-growing amount of data generated on social media and the increasing utilization of these platforms during emergency situation for information retrieval and dissemination, the capability to process the large volume of data has become a major bottleneck for the society to maximize the values of these platforms. Aiming to enhance situational awareness and near real-time data analytics capability, we have developed a very practical protocol that encompasses the ability to filter noise, classify the data into well-defined topics, visualize the development of events, and produce highly relevant and comprehensible summaries. To test the applicability of these tools, we applied the protocol to 2.5 million tweets collected during a six-week period for hurricane in 2017, 3 million tweets for hurricane in 2018. The utility of our protocol is demonstrated through a much-improved lexicon
specific to crisis situation, a trending topics extraction algorithm (through the use of Hierarchical clustering), a SA dashboard with a set of visualization options to offer the decision makers the perception, comprehension and near future prediction of emerging events, and finally, short and precise summaries generation through Social Media Messages Summarization (SMMS).

Moreover, by examining the performance of some of the most widely used text summarization techniques, we have shown that these conventional techniques do not work well for unstructured social media data and how they could be improved with the inclusion of tweet characteristics such as the number of retweets, favorites and the presence of the media information as new parameters in our algorithm. Also, generating timely alerts is a crucial task in crisis and emergency situations. We have shown that the protocol and its associated artifacts can significantly reduce the amount of time and efforts relative to other existing methods, and generate effective SA information more efficiently than the conventional online news media. Given the similarity among the various leading social media platforms, we believe our protocol and tools can be easily applied to other social media data with minimal changes and customization.

To increase the usability of our protocol, the tools developed in our study all build on publicly available data and software tools. No advanced computing resources are needed to implement the protocol. Users of our tools include but are not limited to general citizens who want to learn about the development of an event, victims seeking rescue or evacuation information, first respondents and volunteers, public safety officials who need to make real time decisions, and government leaders who need to assess the scope of the impact and degree of severity. In addition to provide almost real time decision support, our tools can also be used to analyze historical data to generate statistics and information for future policy making. With the
possible extensions discussed in the previous sections, we are also confident that our protocol and algorithms can be leveraged to be applied to a wide range of contexts.

References


