Event-related Collections Understanding and Services

Liuqing Li

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Edward A. Fox, Chair
Zhiwu Xie
Andrea Kavanaugh
Chandan K. Reddy
Zhihong Deng

February 18, 2020
Blacksburg, Virginia 24061

Keywords: Event-related Collections, Collection Development, Curation, URL Analysis, User Classification, Tweeting Pattern Analysis, Webpages, Text Summarization

Event-related Collections Understanding and Services

Liuqing Li

(ACADEMIC ABSTRACT)

Event-related collections, including both tweets and webpages, have valuable information, and are worth exploring in interdisciplinary research and education. Unfortunately, such data is noisy, so this variety of information has not been adequately exploited. Further, for better understanding, more knowledge hidden behind events needs to be unearthed. Regarding these collections, different societies may have different requirements in particular scenarios. Some may need relatively clean datasets for data exploration and data mining. Social researchers require preprocessing of information, so they can conduct analyses. General societies are interested in the overall descriptions of events. However, few systems, tools, or methods exist to support the flexible use of event-related collections.

In this research, we propose a new, integrated system to process and analyze event-related collections at different levels (i.e., data, information, and knowledge). It also provides various services and covers the most important stages in a system pipeline, including collection development, curation, analysis, integration, and visualization. Firstly, we propose a query likelihood model with pre-query design and post-query expansion to rank a webpage corpus by query generation probability, and retrieve relevant webpages from event-related tweet collections. We further preserve webpage data into WARC files and enrich original tweets with webpages in JSON format. As an application of data management, we conduct an empirical study of the embedded URLs in tweets based on collection development and data curation techniques. Secondly, we develop TwiRole, an integrated model for 3-way user classification on Twitter, which detects brand-related, female-related, and male-related tweeters through multiple features with both machine learning (i.e., random forest classifier) and deep learning (i.e., an 18-layer ResNet) techniques. As guidance to user-centered social research at the information level, we combine TwiRole with a pre-trained recurrent neural network-based emotion detection model, and carry out tweeting pattern analyses on disaster-related collections. Finally, we propose a tweet-guided multi-document summarization (TMDS) model, which generates summaries of the event-related collections by using tweets associated with those events. The TMDS model also considers three aspects of named entities (i.e., importance, relatedness, and diversity) as well as topics, to score sentences in webpages, and then rank selected relevant sentences in proper order for summarization.

The entire system is realized using many technologies, such as collection development, natural language processing, machine learning, and deep learning. For each part, comprehensive evaluations are carried out, that confirm the effectiveness and accuracy of our proposed approaches. Regarding broader impact, the outcomes proposed in our study can be easily adopted or extended for further event analyses and service development.

Event-related Collections Understanding and Services

Liuqing Li

(GENERAL AUDIENCE ABSTRACT)

Event-related collections, including both tweets and webpages, have valuable information. They are worth exploring in interdisciplinary research and education. Unfortunately, such data is noisy. Many tweets and webpages are not relevant to the events. This leads to difficulties during data analysis of the datasets, as well as explanation of the results. Further, for better understanding, more knowledge hidden behind events needs to be unearthed. Regarding these collections, different groups of people may have different requirements. Some may need relatively clean datasets for data exploration. Some require preprocessing of information, so they can conduct analyses, e.g., based on tweeter type or content topic. General societies are interested in the overall descriptions of events. However, few systems, tools, or methods exist to support the flexible use of event-related collections.

Accordingly, we describe our new framework and integrated system to process and analyze event-related collections. It provides varied services and covers the most important stages in a system pipeline. It has sub-systems to clean, manage, analyze, integrate, and visualize event-related collections. It takes an event-related tweet collection as input and generates an event-related webpage corpus by leveraging Wikipedia and the URLs embedded in tweets. It also combines and enriches original tweets with webpages. As an application of data management, we conduct an empirical study of tweets and their embedded URLs. We developed TwiRole for 3-way user classification on Twitter. It detects brand-related, female-related, and male-related tweeters through their profiles, tweets, and images. To aid user-centered social research, we combine TwiRole with an existing emotion detection tool, and carry out tweeting pattern analyses on disaster-related collections. Finally, we propose a tweet-guided multi-document summarization (TMDS) model and service, which generates summaries of the event-related collections by using tweets associated with those events. It extracts important sentences across different topics from webpages, and organizes them in proper order.

The entire system is realized using many technologies, such as collection development, natural language processing, machine learning, and deep learning. For each part, comprehensive evaluations help confirm the effectiveness and accuracy of our proposed approaches. Regarding broader impact, our methods and system can be easily adopted or extended for further event analyses and service development.

Acknowledgments

First and foremost, I would like to show my sincere gratitude to my advisor, Dr. Edward A. Fox, an important person in my life. I genuinely appreciate his continuous motivation, encouragement, patience, rigor, and immense knowledge. His guidance and support helped me successfully go through my Ph.D. program. I cherish every moment we spent together on research discussion, project collaboration, paper and dissertation writing, and course preparation.

Besides my advisor, I would like to express deep appreciation for other professors in my committee: Dr. Zhiwu Xie, Dr. Andrea Kavanugh, Dr. Chandan K. Reddy, and Dr. Zhihong Deng, as well as earlier help from Dr. Jiepu Jiang. Special thanks go to Dr. Na Meng, Dr. Djavad Salehi-Isfahani, and Dr. Donald J. Shoemaker. It's fortunate to have a chance to collaborate with these professors, who gave me insightful suggestions and invaluable guidance in my research and courses.

I give my thanks to my colleagues in the Digital Library Research Laboratory (DLRL): Sunshin Lee, Yufeng Ma, Xuan Zhang, Xinyue Wang, Ziqian Song, Prashant Chandrasekar, Maanav Mehrotra, Yinlin Chen, Mohamed MG Farag, Matthew Bock, Yu Wang, Saurabh Chakravarty, Shuo Niu, Siyu Mi, Abigail Bartolome, etc. I thank other friends who helped me during my Ph.D. study: Liyan Li, Shuaicheng Zhang, Jack Geissinger, Tian Shi, Ashin Marin Thomas, Suraj Gupta, etc.

I give special thanks to Dr. Rao Shen for the intern opportunity and her tremendous help during my internship at Verizon Media, which was a valuable industry research experience in a U.S. company.

I must express heartfelt gratitude to my family – my father, Hong Li, my mother, Mi Liu, my grandparents, and other relatives. Without their continuous understanding, support, patience, and sacrifice, I could not have come back to school and gain a doctor degree after several years of work.

Thanks go to the Department of Computer Science, Advanced Research Computing (ARC), and DLRL for the funding and facilities which supported me through my Ph.D. research.

Thanks go to National Science Foundation for support through grant NSF IIS - 1319578 and IIS - 1619028. Thanks also go to Mayfair Group for their support through a grant.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In connection with our work on the Global Event and Trend Archive Research (GETAR) project, we have curated more than 1,700 different tweet collections about important events and topics since 2012. Both tweets and webpages (represented as short URLs) in the raw collections can be poured into a digital library system and interpreted by the 5S theory (i.e., Streams, Structures, Spaces, Scenarios and Societies) [41]. Though we have such "big data", it is still unclear how the Fourth Paradigm [49] can apply to help organize the data and support human access. Further, few systems, tools, or methods exist to support flexible use of event-related collections, which requires a pipeline approach, e.g., collection development, curation, analysis, integration, and visualization. Figure 1.1 depicts a proposed event ecosystem, showing some examples of *Scenarios* and *Societies* across the three levels (i.e., data, information, and knowledge) in the data pyramid.

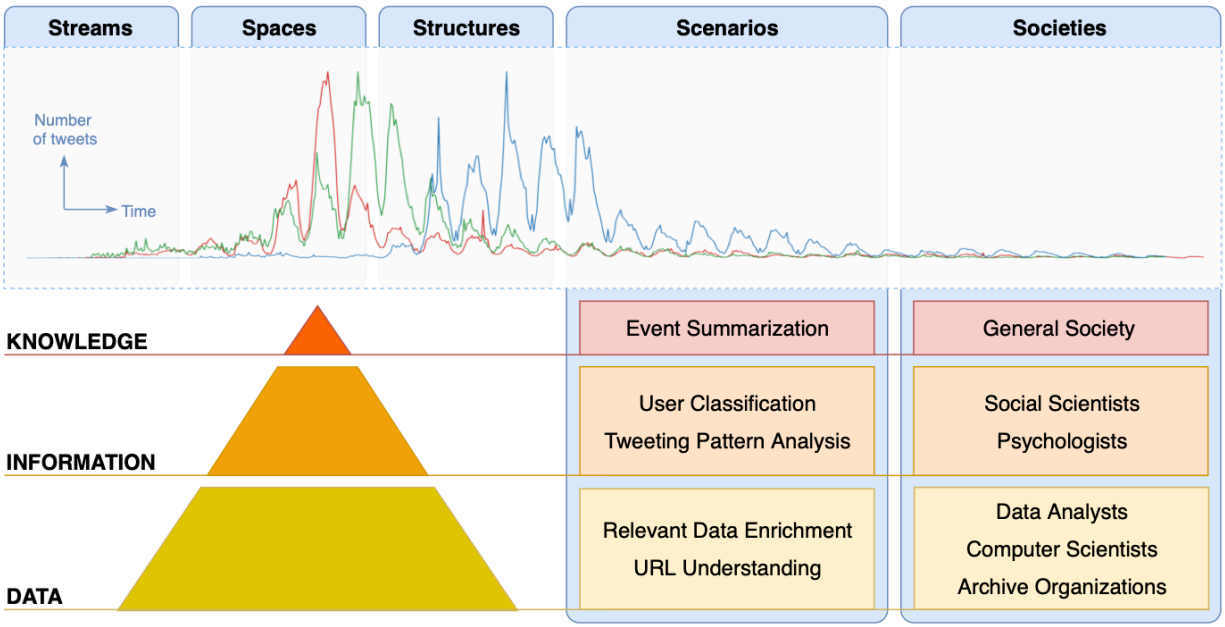


Figure 1.1: A proposed event ecosystem based on both the 5S theory and data pyramid

Our new, integrated system, guided by both the 5S theory and data pyramid, can aid interdisciplinary research and education. It connects most phases of the above pipeline into a digital library. Concerning the data pyramid, we aim to:

- generate rich data (i.e., tweets, webpages) and work on data curation, so that data-related societies, such as data analysts, computer scientists, and archive organizations, can benefit through data exploration;
- provide services at the information level, connecting the processed information and societies interested in social issues. Social scientists, psychologists, and other related communities can adopt such services in their research;
- share knowledge in the general society and help people summarize event-related collections.

1.2 Problem Statement

In this section, we concentrate on the specific problems that need to be investigated and resolved.

First, we work on identifying relevant webpages and tweets from event-related collections. Both webpage and tweet contents include various types of "noise" like spam or advertisements. Such non-relevant information needs to be spotted and tagged so that further interpretations can be connected with high-quality event-related collections, to support hypothesis testing and generalization. Further, we study the characteristics and utility of short URLs, which are a vital connection between webpages and tweets, and aid the web crawling process. We reduce this problem to that of investigating their lifespan and the scope of the Wayback Machine [53], a digital archive of Internet content, consisting of mementos (snapshots of webpages) across time.

Second, we move further to identify the roles of users on Twitter. Accurate classification of users as to role can be beneficial in both academic and industrial research. Social scientists can undertake more user-centered research, while consumer-oriented companies can provide targeted services. Unfortunately, due to privacy concerns, a user's role may not be explicitly revealed. For example, with Twitter, the gender value (i.e., female and male) can be set by a user or predicted by Twitter on the profile page, but that is unreadable to others. Brand users (i.e., newsmaker, organization, or institution) were not taken into account in most previous research. Moreover, since moods might vary with time among different roles of users, analyzing such changes will be helpful both academically and practically. We reduce this problem to that of studying the relevant tweets and labeled users, and carry out two empirical analyses of tweeting patterns (i.e., posting patterns, mood patterns) about disasters at both the event and role level.

Third, we investigate how to summarize the webpages for each event-related collection and improve the quality of summaries. Regarding our collections, webpage summarization requires a high level of integration and can be guided by tweets, some of which are firmly connected with the key elements in summaries. Currently, most researchers are still concentrating on single document summarization. Regarding multi-document summarization, some researchers just combine a set of documents into a long single document, with lack of consideration for contextual information. In our particular scenario, we reduce this problem to that of associating webpages with tweets, and developing an integrated ranking model for summarization at the sentence level.

1.3 Research Questions and Hypothesis

The primary research question for this dissertation is: *How can a digital library system with event-related content made of tweets and webpages be built so its key types of users can better analyze and utilize content of interest?* This question can be further decomposed into the following three research sub-questions.

- *How can study and processing of the short URLs in tweets, along with their webpages, be applied to improve the quality and relevance of collections?*
- *How can the roles of tweeters, along with tweeting patterns, be identified and analyzed to provide insights about tweeters and their behavior?*
- *How can multi-document summaries of the event-related collections be improved by using tweets associated with those events?*

Our research involves three topics, addressing the three problems stated, and answering the three questions.

Accordingly, the central hypothesis of this research is that *our solutions for these problems, questions, and topics will enhance our system's curation of event-related collections at the three levels in the data pyramid, and the services will more effectively support both Scenarios and Societies.*

1.4 Research Approach

Compared with existing content classification methods based on supervised learning, approaches based on collection development can better handle our event-related collections. We can use an unsupervised approach to identify relevant webpages; it is infeasible to manually label the webpages in hundreds of collections as is needed for supervised classification.

For each collection, its corresponding Wikipedia entry is a good query candidate. Further, a language model with query expansion can improve the query representation during collection development. Also, since the main Wikipedia page and external links are closely interrelated with a particular event, a pre-query can be designed with such prior knowledge to improve the retrieval results. By adding the event focused crawler (EFC) [33], our model should retrieve more relevant webpages from the open web. Regarding the short URL analysis, we assume that links to old webpages are likely to be broken, so the Wayback Machine can help us retrieve those webpages.

Twitter users assuming different roles (e.g., brand, female, male) behave differently. For example, preliminary work has shown that male users prefer technology and sports [10], female users are more emotional [97], and brand users act as broadcasters by spreading information or delivering messages. We believe that a hybrid role classification model performs better than a single model, while rich features (e.g., first-person words, profile images) can improve the classification results. Further, different types of events like disasters have distinct patterns. For example, natural disasters have dramatic effects and can bring huge loss, while man-made disasters often proceed rapidly, with long-lasting impact. These events might lead to different posting patterns as well as mood patterns. For instance, some users may be frightened by the aftershocks during an earthquake, some are angered by a bombing disaster, and others verbalize their sadness for the victims in a school shooting event.

The tweet-guided multi-document summarization approach proposed in this study can produce high-quality, understandable summaries for a set of event-related webpages. Different from most existing multi-document summarization (MDS) methods, we take tweets as an external source that can effectively support the summarization process. The reason is tweeters tend to refer to the key elements in webpages to state the facts or express their attitudes. These key elements, including persons, places, and dates, can also be considered as crucial factors in summaries. Further, the posted time of tweets is another good factor, leading to a summary with a suitable timeline of sub-events. Based on such assumptions, we extract relevant sentences from webpages through tweets during data generation and implement an integrated ranking model for summarization at the sentence level. We suppose that the summaries generated by our model could cover the major aspects of events in proper order.

1.5 Research Deliverables and Objectives

Figure 1.2 shows our research deliverables as part of the system data flow diagram.

At the data level, we propose a collection development model to generate collections of relevant webpages and tweets, organize event-related collections with rich metadata and appropriate formats (e.g., WARC, JSON), and carry out an empirical analysis on short URLs in tweets in large event-related collections.

At the information level, we develop a role-related user classification model that can detect

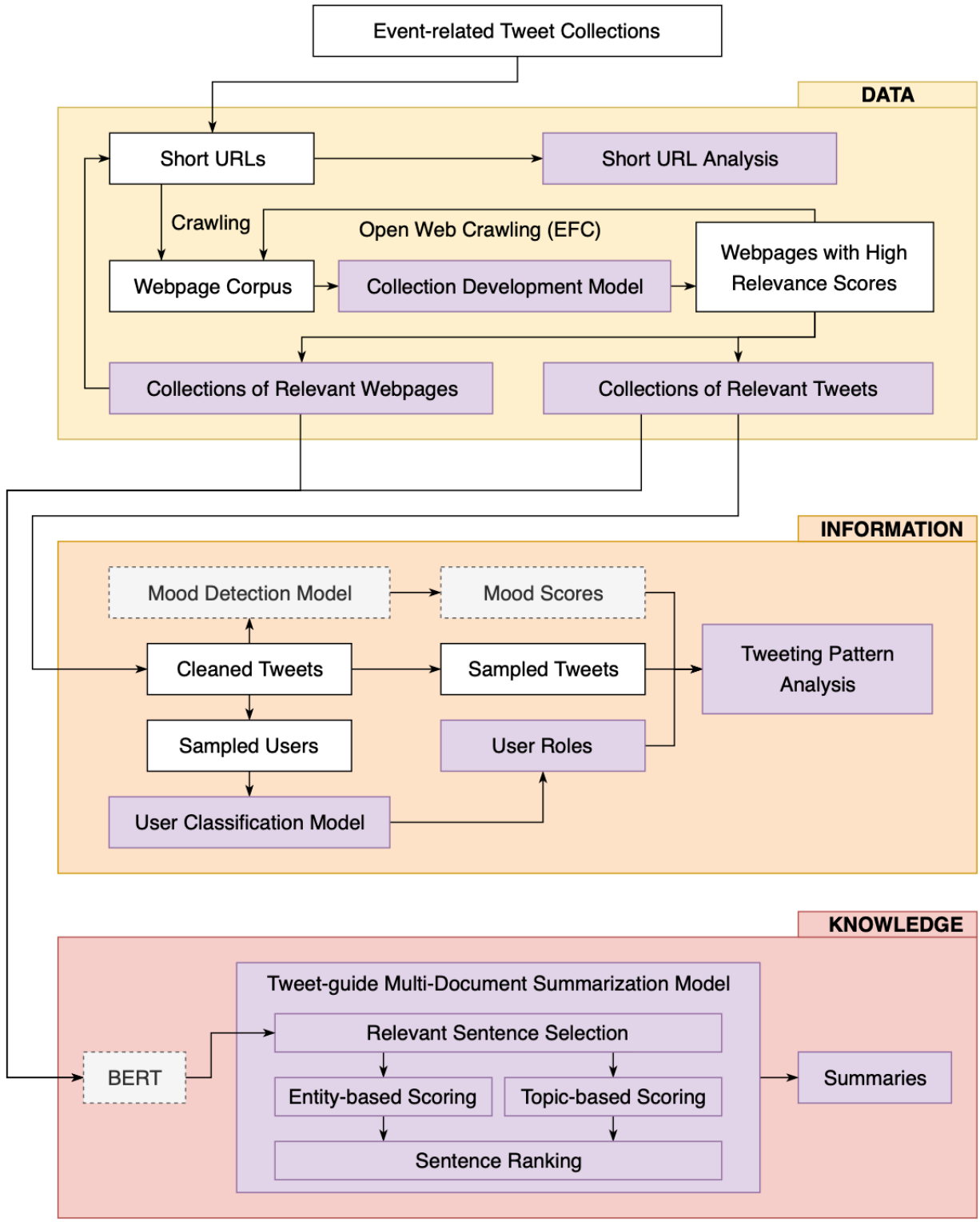


Figure 1.2: System data flow diagram, showing deliverables (light purple) and modules from other works (gray and dotted)

tweeters with different roles, and move further to study the tweeting patterns (i.e., posting patterns, mood patterns) in disasters.

At the knowledge level, we design and implement a tweet-guided multi-document summarization (TMDS) model which includes multiple components (i.e., relevant sentence selection, entity-based scoring, topic-based scoring, and sentence ranking) and is used to summarize a set of webpages for each event-related collection.

1.6 Dissertation Organization

Our study explores three research topics in establishing a system at different levels. The remaining part of this dissertation is organized as follows. Chapter 2 reviews the literature regarding the techniques used in our study. Chapter 3 introduces our collection development model, data curation, and the empirical analysis of short URLs. Chapter 4 shows how a hybrid model TwiRole identifies the roles of tweeters, followed by two tweeting pattern analyses on events, users, and moods, as applications. Chapter 5 explores how our proposed integrated ranking model can be applied to improve the quality of summaries of webpages with the guidance of tweets. Finally, the contributions and future work are summarized in Chapter 6.

Chapter 2

Literature Review

2.1 Document Retrieval

A critical problem in document retrieval or collection development is how to rank a collection of documents based on their relevance scores for a given query. Researchers have been focusing on the development of ranking methods for many years.

Traditionally, a small number of critical features (e.g., term frequency (TF), inverse document frequency (IDF), and document length) are applied to rank documents. The parameters in these methods can be fine-tuned to improve the ranking results. Okapi BM25 [98] and language models (e.g., query likelihood model) [62, 95] are such methods and are still being widely used. Automatic query expansion has long been suggested as a practical approach to deal with the fundamental issue of word mismatch in document retrieval. Global query expansion [120] can discover word relationships while local query expansion [64] extracts important words from documents retrieved by an initial query.

Additional structural features (e.g., title and URL) and query-independent features (e.g., PageRank and URL length) have proved useful to retrieve relevant documents. Supervised learning has drawn attention in document retrieval. Joachims [54] developed SVM^{rank} to optimize the ranking problem, and Burges et al. [15] proposed a neural network model for document retrieval. Recently, deep-learning models [27, 119] have become popular and perform better on many tasks.

Besides documents and queries, researchers also adopt other sources (e.g., Wikipedia, DBpedia) to improve ranking results by using prior knowledge [19, 28, 117, 118].

2.2 Short URL Analysis on Twitter

As a key link between tweets and webpages, short URLs are also worth exploring. Some researchers made use of short URLs to analyze Twitter users' activity and influence. Bakshy et al. [9] tracked the diffusion of short URLs instead of retweets to evaluate the influence score of a Twitter user. Ghosh et al. [44] also used these URLs as markers to trace the spread of information or content, and classify the retweeting activities. Other researchers [16, 22, 65, 82] detected suspicious/spam URLs through structural features (e.g., the number

of URLs posted or clicked, lifespan) and contextual information.

Currently, few researchers worked on analyzing short URLs themselves. Antoniades et al. [6] conducted a short URL analysis by focusing on two shorten URL websites and tracking the URLs. They analyzed the targeted webpages, their popularity, and activity over time. However, they were limited to two services and not concerned about broken links and archives. SalahEldeen and Nelson [101] carried out an empirical study on short URLs in a small number of collections over 2.5 years and discovered a relatively linear relationship between time of sharing of those URLs and the percentage of lost URLs, with a slight linear correlation between time of sharing of those URLs and coverage of resource archives. Compared with their work, we take broken links into account and analyze the URLs in a dozen collections during 5 years.

2.3 User Classification on Twitter

Researchers have long worked on Twitter user classification according to roles. Multiple features have been considered and extracted for gender classification. Liu and Ruths [79] investigated the relationship between first name and gender, and took the first name as an important feature for gender prediction. Some researchers [3, 4, 37] have found that female and male users may apply templates in different colors, and so utilized color-based features in gender inference. Information in user profiles, especially the descriptions of users, may also contain gender-based terms or phrases (e.g., man, woman, actor, mother) that are good indicators [26, 109]. Besides these features, a user's network and behavior [23, 63, 90], external sources (e.g., personal website, Facebook) [14], and tweets [8, 10, 43, 69, 92, 97], have been advocated for gender classification.

Additionally, deep neural networks have been leveraged in gender classification on social media in recent years. Various CNN models are designed to classify the gender in social media [68, 112, 113]. However, the dataset of [68] is produced from Flickr instead of Twitter, while [112, 113] only pays attention to profile images with faces. Geng et al. [43] proposed an ensemble approach by combining models for both tweet contents and profile images to improve the quality of bi-classification.

In addition to the traditional bi-classification, Purohit and Chan [96] categorized Twitter users into organization, organization-affiliated, and non-affiliated, while Pennacchiotti and Popescu [93] classified Twitter users as either Democrat or Republican. However, there are methodological concerns regarding the evaluation of such studies, because most approaches were evaluated on self-created datasets, and performance is still unclear on other datasets. In comparison, we develop a 3-way classification model and evaluate our model and baseline models on third-party datasets.

2.4 Tweeting Pattern Analysis

Regarding event-related collections, most researchers are focusing on single events or a single type of event, such as the 2011 Egyptian uprising [56], 2011 Japan Earthquake [115], 2012 Hurricane Sandy [17, 89], 2013 River Elbe Flood [48], 2016 Ghana Election [84], or 2016 Berlin Terrorist Attack [36]. Yang et al. [122] designed PhaseVis, a multi-view integrated visualization tool, and applied it to Hurricane Isaac. Lee et al. [66, 67] built a prototype of a digital library to detect and visualize water main break occurrences, and to geolocate tweets associated with other types of events. Alam et al. [1] analyzed the multimedia content from three hurricane collections. Novel theories, models, and techniques have been developed and applied, and researchers can make further discoveries or tell exciting stories about these events. However, the generality of these approaches is not clear; it is uncertain how well they can work on other events.

There has been growing interest in sentiment and emotion classification of tweets. Many studies have been published on basic sentiment analysis (i.e., positive, negative, and neutral) by applying traditional machine learning techniques (e.g., Naive Bayes, SVM, logistic regression). Schulz et al. [102] took unigram, part-of-speech, syntactic, and sentiment information as features for sentiment analysis. Nagy et al. [86] detected tweet sentiments during crises by using sentiment words, emoticons, and a list of out-of-vocabulary words. Caragea et al. [17] performed sentiment classification of tweets during Hurricane Sandy, and visualized these sentiments on a geographical map. Further, Wendland et al. [114] applied an appraisal system to account for the interpersonal assessment of speakers and associated attitudes, which evaluated the sentiments of tweets from several aspects such as appreciation, affect, and judgment. Instead of detecting the three basic sentiments, Colneriĉ and Demsar [24] proposed a recurrent neural network (RNN) model predicting tweet emotions, which can better describe the feelings of tweeters. Therefore, we carry out a study of different types of disasters, followed by a case study, based on our user classification model and the above-mentioned RNN-based emotion detection model.

2.5 Text Summarization

Text Summarization is condensing a source text into a shorter version while preserving its key information and overall meaning. Based on the input type, a summarization task can be divided into single document summarization (SDS) and multi-document summarization (MDS). Regarding summary type, it can be categorized into extractive or abstractive summarization. According to purpose, there are three types of approaches: generic, domain-specific, and query-based summarization.

In pioneering research, Luhn et al. [81] introduced a method to extract salient sentences from the text using features such as word and phrase frequency. Early researchers concen-

trated on extractive SDS. Basic features (e.g., TFIDF, location, length) have been applied to select key words or sentences to generate summaries. Graph-theoretic representation is another popular approach [60, 121]. Some hot sub-graphs with important nodes and more edges will be selected for summarization. Regarding MDS, topic modeling techniques have been extensively used. For example, Daume et al. [29] proposed BayeSum, a Bayesian summarization model for query-focused summarization. Wang et al. [110] introduced a Bayesian sentence-based topic model for summarization which used both term-document and term-sentence associations. Liu et al. [80] applied Restricted Boltzmann Machines (RBMs) for query-based MDS. Additionally, Banerjee [11] employed an integer linear programming (ILP) model to maximize the readability of the summary of multi-documents.

Recently, deep learning methods [20, 51, 87, 99, 103] have demonstrated impressive performance in term of SDS. Gao et al. [42] proposed a reader-aware summary generator (RASG) that incorporates reader comments to improve the performance of SDS. To the best of our knowledge, only a few papers [78, 111] have investigated how deep learning performs on abstractive MDS tasks. An issue with these papers is that the final summary may have invalid and useless contents because methods simply concatenate all the source documents from a collection into a long sequence. Baumel et al. [12] improved the query-based MDS on filtered passages with a relevance model and an attention model. Chu and Liu [21] presented a fully unsupervised abstractive MDS approach with an autoencoder and evaluated their model on a Yelp dataset. To date, however, little MDS research has led to effective solutions to the challenge of arranging sentences to improve their sequential order in final summaries.

Chapter 3

Relevant Data Enrichment and URL Understanding

The research question for this chapter is: *How can study and processing of the short URLs in tweets, along with their webpages, be applied to improve the quality and relevance of collections?* A sub-system for data enrichment and curation is proposed to answer this question. We mainly work on relevant data enrichment, along with the evaluation of our collection development (CD) model, and carry out an empirical analysis of short URLs from data curation.

3.1 Approach

Figure 3.1 shows the architecture of the two-part sub-system considered in this chapter. One part is mainly for relevant data enrichment (green box with slash lines), while the other is mainly for data curation (yellow box with backslash dotted lines). Since there is a gap between our raw event-related tweet collections and the CD model (orange box with gridlines), we first need to extract the short URLs and prepare a webpage corpus for the CD model. Regarding the relevant data enrichment, we further expand our webpage collection into the broader "open web" level. The data curation is flexible and can be easily extended with additional metadata.

3.1.1 Webpage Corpus Generation

Given an event-related tweet collection, we first designed and implemented a spam tweet detector with deep learning to filter out tweet spam. Then, as a "big data" solution, we uploaded the roughly-processed data into a Hadoop cluster [38] and used Spark [40] to distributedly tag retweets and extract short URLs. After processing, the metadata of each record has 4 fields: *tweet id*, *retweet status*, *posted date*, and *short URL(s)*. We expanded short URLs into long URLs through *wget*, retrieved the mementos of the URLs from the Internet Archive [7], and stored the response data including raw webpage contents into WARC files. A 24 virtual machine cluster has been deployed to speed up the cumbersome crawling process.

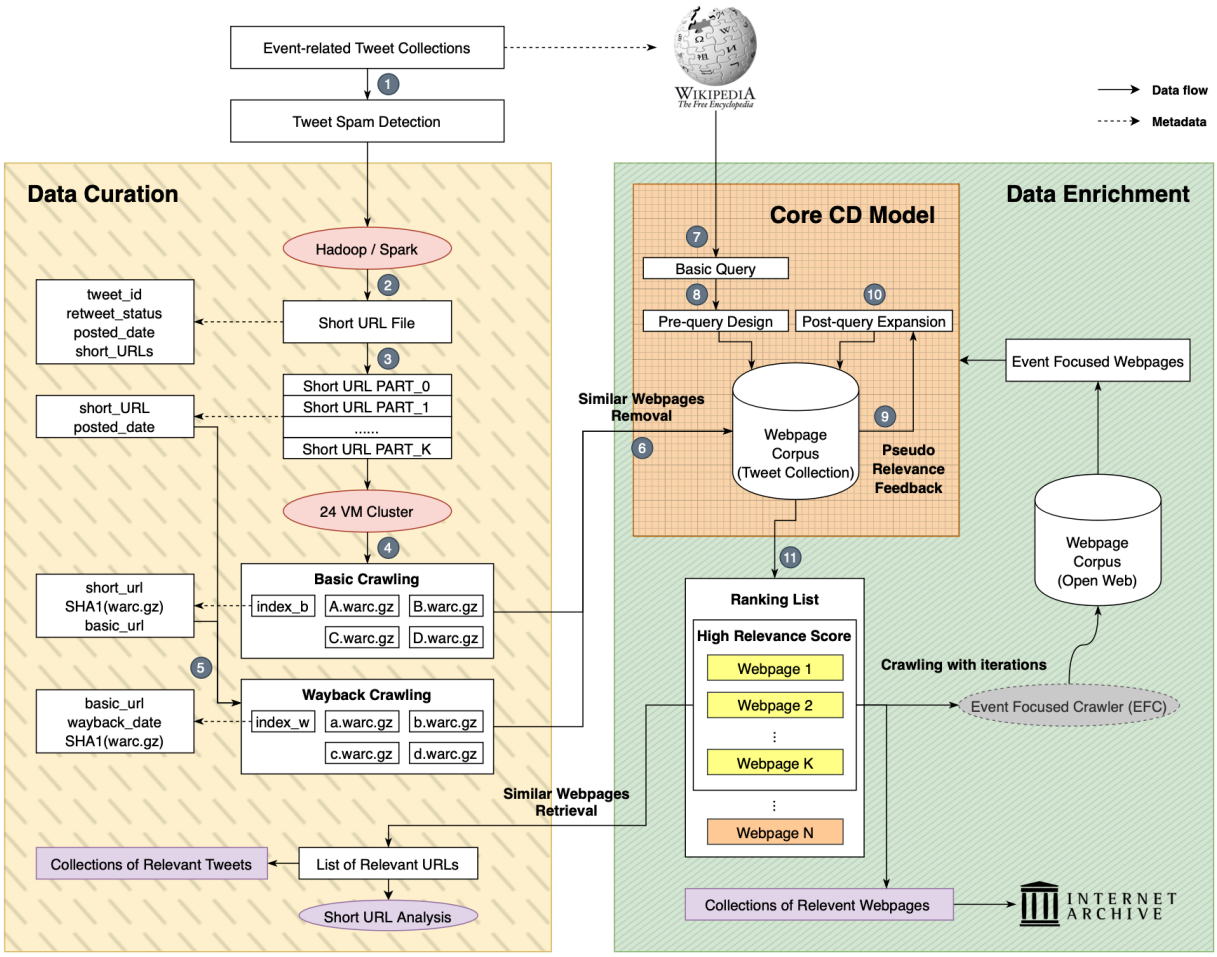


Figure 3.1: System architecture for relevant data enrichment and data curation

Regarding the webpages from both the World Wide Web (WWW) and Internet Archive, to improve page readability we applied jusText [94] to remove unnecessary content from webpages, such as HTML headers and footers, as well as navigation information. Based on our observation that a certain percentage of webpages are extremely similar to each other, we used TFIDF to represent webpages, created a similarity matrix, and eliminated webpages with similarity above a default threshold ($TH_{sim} = 0.9$). Thus, we have built a webpage corpus for the following CD task.

3.1.2 Collection Development Model Design

Our CD model mainly includes three components: pre-query design, query likelihood modeling, and post-query expansion. Theses are explained in the following subsections. The proposed model calculates the relevance score of each webpage and generates a ranked list

for each webpage corpus. As an extension, our model can also work with an event focused crawler (EFC) [32, 33] that has been applied to retrieve more relevant webpages.

3.1.2.1 Pre-query Design

Each event-related collection links with an entry in Wikipedia. For instance, the entry of the Connecticut school shooting in 2012 is "*Sandy Hook Elementary School shooting*". We first took each entry as a basic query and developed our CD model with pre-query design. During preprocessing, we lowercased and tokenized each query and webpage into terms with spaCy [50], and utilized the Krovetz stemmer [59] to reduce inflected terms into their base form/stem.

We assume that the main Wikipedia page and its external links are highly interrelated with each event, which should be helpful for our retrieval task. Based on this assumption, we crawled the web contents of these pages and calculated the TF scores of a set of words, such as nouns, proper nouns, pronouns, and numerals, with part-of-speech tagging. For example, Table 3.1 lists the top 10 frequent words along with their TF scores in the webpages related to the Connecticut school shooting. Besides the raw query terms, we added the most frequent word from Wikipedia pages and extended the raw query into "*sandy hook elementary school shoot lanza*". We noticed that "*lanza*" has high TF score, which is the shooter's name.

Table 3.1: Top 10 frequent words with TF scores in the webpages related to the Sandy Hook Elementary School shooting collection

No.	**Word**	**TF Score**		**No.**	**Word**	**TF Score**
1	school	2,608		6	gun	1,046
2	lanza	1,803		7	state	992
3	shoot	1,513		8	hook	973
4	newtown	1,440		9	children	704
5	police	1,047		10	connecticut	682

3.1.2.2 Query Likelihood Model

We applied a query likelihood (QL) model to rank our webpage corpus. To calculate the probability that the particular document is D given a query q, we converted the problem to an easier one with Bayes' rule and an independence assumption; see Equation 3.1.

$$P(D|q) = \frac{P(q|D) \cdot P(D)}{P(q)} \propto P(q|D)P(D) \propto P(q|D) \tag{3.1}$$

For each document D, we assume there is a unigram language model θ_D and compute the

likelihood of the query q generated by θ_D, using maximum likelihood estimation as the simplest way; see Equation 3.2.

$$\log P(q|\theta_D) = \sum_{t \in q} \log P(t|\theta_D) = \sum_{t \in q} \log \frac{c(t, D)}{|D|} \tag{3.2}$$

Here $c(t, D)$ is the TF score of term t in D and $|D|$ is the document length.

Further, we applied the Jelinek-Mercer smoothing method to avoid log 0 error; see Equation 3.3.

$$P(t|\theta_D) = (1 - \lambda) \cdot \frac{c(t, D)}{|D|} + \lambda \cdot \frac{c(t, C)}{|C|} \tag{3.3}$$

Here λ is the tuning parameter, $c(t, C)$ is the TF score of t in a sampled web corpus, and $|C|$ is the total number of terms in the corpus. We set $\lambda = 0.5$ by default and chose the ClueWeb09 Category B (ClueWeb09B) dataset as the corpus C. Since q, D, C, and λ are all known, we could rank our webpage corpus through Equations 3.2 and 3.3.

3.1.2.3 Post-query Expansion

The post-query expansion is based on pseudo-relevance feedback (PRF) [64]. We assume that the top 100 ranked webpages are relevant, and use Relevance Model 3 (RM3) [64] to calculate the term weights from those webpages; see Equation 3.4.

$$\begin{aligned} P_{RM3}(t|q, R) &= (1 - \lambda) \cdot \frac{c(t, q)}{|q|} + \lambda \cdot P_{RM1}(t|q, R) \\ P_{RM1}(t|q, R) &\propto \sum_{D \in D_R} P(t|D) \prod_{q_i \in q} P(q_i|D) \end{aligned} \tag{3.4}$$

We still focused on some types of words (i.e., nouns, proper nouns, pronouns, and numerals) for query expansion and assigned the ranking scores from the previous section to $\prod_{q_i \in q} P(q_i|D)$. We ranked all the selected words by weight and took the top 10 words and their weights for query expansion. For the Sandy Hook Elementary School shooting, the top 10 weighted words in the pseudo-relevance identified webpages are listed in Table 3.2.

Then, we leveraged the KL-divergence (KLD) model to rerank all the webpages with the terms in the extended query and their corresponding weights; see Equation 3.5.

Table 3.2: Top 10 weighted words in the pseudo-relevance webpages

No.	Word	Term Weight		No.	Word	Term Weight
1	lanza	0.0911		6	elementary	0.0890
2	school	0.0898		7	use	0.0076
3	shoot	0.0894		8	adam	0.0048
4	sandy	0.0893		9	weapon	0.0036
5	hook	0.0892		10	obama	0.0032

$$-\overline{KLD}(\theta_q||\theta_D) = \sum_t w_q^t \cdot \log P(t|\theta_D) \tag{3.5}$$

Here w_q^t is the term weight of t and $P(t|\theta_D)$ refers to Equation 3.3.

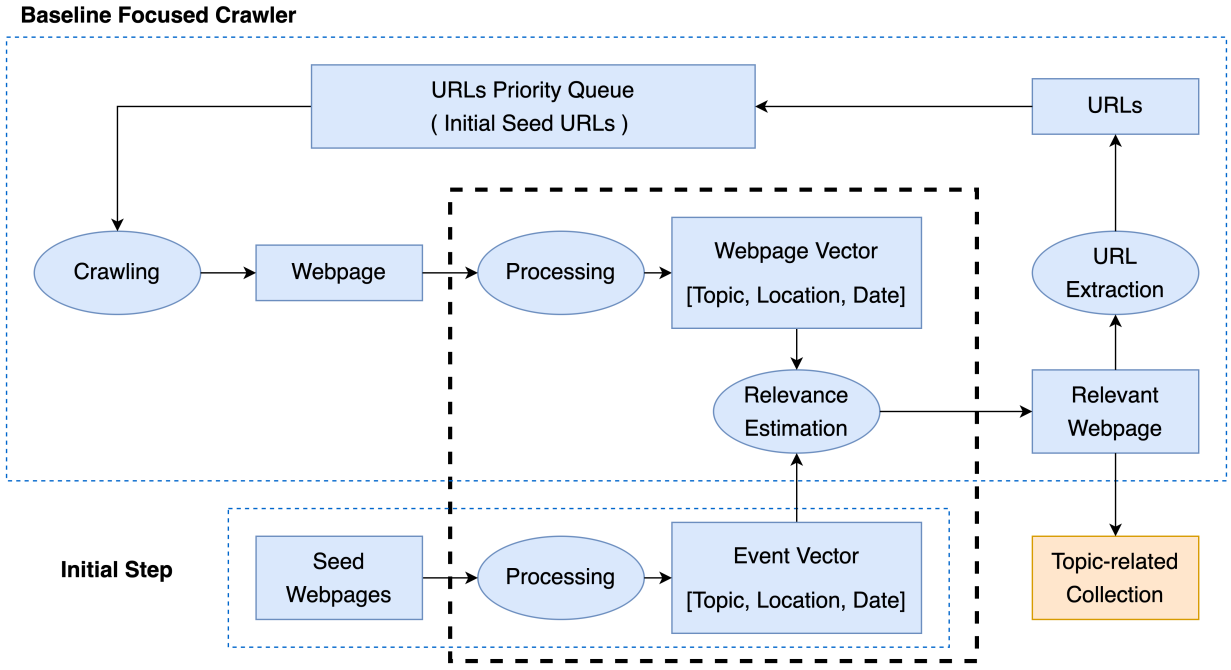


Figure 3.2: System architecture of EFC

3.1.2.4 More about Retrieval

In our specific scenario, we also hope to crawl more relevant webpages from the open web to enrich our collections.

Regarding our event-related collections, we selected the top 30 percent of the ranked webpages as seeds and applied an event focused crawler (EFC) [33] to crawl for additional

relevant webpages from the WWW. Figure 3.2 shows the system architecture of EFC, where the input is a list of webpage seeds, the output is a relevant webpage collection, and an event model is used to estimate the relevance between the seed webpapges and newly crawled webpages. We added the webpages crawled by EFC into our corpus and refined them with our CD model. This process could be executed with several iterations. As a result, the recall rate can efficiently increase during iteration while the precision can also be kept at the same level.

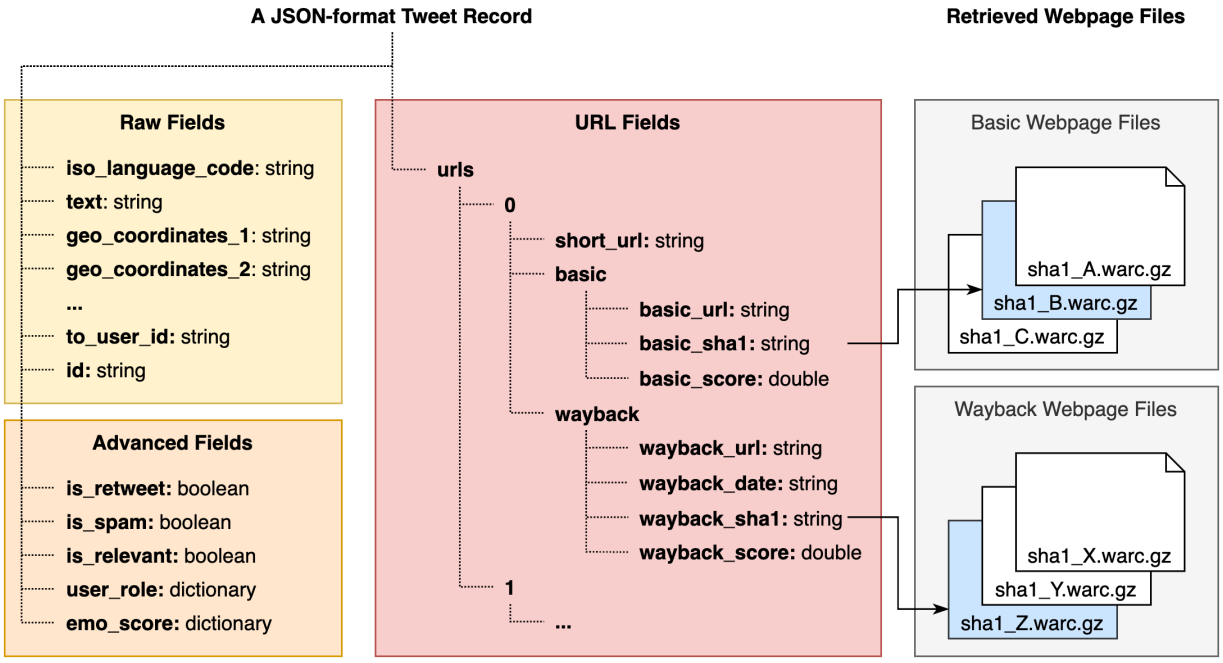


Figure 3.3: A JSON-format tweet record linked with retrieved webpage files

3.1.3 Data Curation

After the relevant data enrichment, we generated webpage collections in addition to their corresponding event-related tweet collections. We overwrote the original metadata of tweets and added more values for data curation. Figure 3.3 shows a JSON-format tweet record, linking with retrieved webpage files. Specifically, besides the raw fields, we have created multiple advanced fields that identify whether a tweet is a retweet or spam. The tweeter's role and tweet mood are two additional advanced fields. Chapter 4 describes how to extract such information in detail. Further, the URL fields include both the basic and Wayback Machine crawling results. The *basic_score* or *wayback_score* represents the relevance score of a webpage. We take tweets with highly relevant webpage links as relevant. By following the *basic_sha1* or *wayback_sha1* field, we can locate the raw webpage contents related to the

URLs in tweets, and also retrieve relevant tweets based on those relevant webpages, which will be used in Chapter 5. These extended fields can also support advanced search, webpage extraction, URL analysis, and other application scenarios.

As we mentioned in Section 3.1.1, *wget* has been utilized for URL expansion and WARC file generation. An advantage of *wget* is that it records the entire HTTP request and response data in WARC files, which can help people better understand which URLs can be accessed at present (with response code 2XX) and which have been moved (with response code 3XX) or broken (with response code 4XX and others). Figures 3.4a to 3.4c show the fragments of both request and response data in a WARC file. With this background, we carry out an empirical analysis of short URLs in a set of event-related tweet collections.

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://t.co/p9ZntCDrwq
Content-Type: application/http;msgtype=request
WARC-Date: 2018-10-16T21:22:36Z
WARC-Record-ID: <urn:uuid:f4b199a5-b5d5-460b-9731-16138474e71d>
WARC-IP-Address: 199.59.148.12
WARC-Warcinfo-ID: <urn:uuid:8ad7dbdc-71e3-4fb9-a587-9d90be50b5aa>
WARC-Block-Digest: sha1:VSC5KB4DIIAP2Y4BEZS5PWMLVXTY6BCX
Content-Length: 112

GET /p9ZntCDrwq HTTP/1.1
User-Agent: Wget/1.16 (linux-gnu)
Accept: */*
Host: t.co
Connection: Keep-Alive
```

(a) WARC request record

Figure 3.4: Fragments of both request and response data in a WARC file

3.2 Data

3.2.1 Data for Collection Development Model Evaluation

3.2.1.1 ClueWeb09B Dataset

Regarding the smoothing method in the QL model, a sampled web corpus is required. Therefore, we choose the ClueWeb09B dataset in our approach, since it is a widely used standard corpus of general Web documents. The original file size is about 155GB, which contains about 50 million English pages. We utilize Lucene [39] for indexing. Each page has been tokenized with a lowercase filter and a Krovetz stemmer filter. The indexed file size is 230GB, and the total number of terms is about 25.7 billion, represented as $|C|$ in Equation 3.3.

```
WARC/1.0
WARC-Type: response
WARC-Record-ID: <urn:uuid:58913a75-c7ec-4b17-ae4f-6b47fefb7ba7>
WARC-Warcinfo-ID: <urn:uuid:8ad7dbdc-71e3-4fb9-a587-9d90be50b5aa>
WARC-Concurrent-To: <urn:uuid:f4b199a5-b5d5-460b-9731-16138474e71d>
WARC-Target-URI: http://t.co/p9ZntCDrwq
WARC-Date: 2018-10-16T21:22:36Z
WARC-IP-Address: 199.59.148.12
WARC-Block-Digest: sha1:6Y5465U3NC7TA5UQL3O4YZNS7RMBGDMJ
WARC-Payload-Digest: sha1:3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ
Content-Type: application/http;msgtype=response
Content-Length: 213

HTTP/1.1 301 Moved Permanently
content-length: 0
date: Tue, 16 Oct 2018 21:22:36 GMT
location: https://t.co/p9ZntCDrwq
server: tsa_a
x-connection-hash: a79fca67760001f9ace28af33b339d55
x-response-time: 3
```

(b) WARC response record (code = 301)

```
WARC/1.0
WARC-Type: response
WARC-Record-ID: <urn:uuid:5a2a029f-c0ab-408b-908e-ddf7341b0679>
WARC-Warcinfo-ID: <urn:uuid:8ad7dbdc-71e3-4fb9-a587-9d90be50b5aa>
WARC-Concurrent-To: <urn:uuid:0cfdc291-356f-437a-b56c-5def8bf01fbf>
WARC-Target-URI: http://www.beliefnet.com/news/home-page-news-and-views/
how-to-respond-to-horror-of-school-shooting-in-connecticut.aspx
WARC-Date: 2018-10-16T21:22:37Z
WARC-IP-Address: 93.184.216.125
WARC-Block-Digest: sha1:UDXZXCJGPVWRUQBXZIOL5PZZVZ2YJZRC
WARC-Payload-Digest: sha1:2RHLHMKEOR3JAFTW7EN53FYCMIZ7GXTF
Content-Type: application/http;msgtype=response
Content-Length: 98114

HTTP/1.1 200 OK
Accept-Ranges: bytes
Cache-Control: max-age=10800
Content-Type: text/html; charset=utf-8
Date: Tue, 16 Oct 2018 21:22:37 GMT
Expires: Wed, 17 Oct 2018 00:22:37 GMT
Last-Modified: Tue, 16 Oct 2018 21:10:46 GMT
Server: ECS (dca/246B)
Vary: Accept-Encoding
X-Cache: HIT
X-Frame-Options: SAMEORIGIN
Content-Length: 97767
```

(c) WARC response record (code = 200)

Figure 3.4: Fragments of both request and response data in a WARC file (cont.)

3.2.1.2 Event-related Collections

We selected five event-related collections for model evaluation, including Sandy Hook Elementary School shooting, Hurricane Sandy, Chapel Hill shooting, Nepal Earthquake, and Hurricane Matthew. For each event, we first processed its corresponding collection to sample a smaller dataset and then built a golden standard dataset through Amazon Mechanical Turk (MTurk).

We took the Sandy Hook Elementary School shooting collection as an example to describe the subsampling process. The entire collection has 363,419 tweets posted from December 14, 2012 to October 12, 2018. According to the steps in Section 3.1.1, we first filtered out spam and unformatted tweets, leaving 360,621 tweets after processing. Then, we extracted all unique short URLs from tweets and crawled their basic webpages and Wayback mementos. We merged the two types of pages and kept all the readable webpages. After calculating the similarity scores among these webpages, all duplicates and similar ones were removed, and only 14,182 webpages were left. Due to the high cost of labor, we randomly sampled 1,000 webpages for MTurk labeling. Table 3.3 lists all the steps in webpage corpus generation on the Sandy Hook shooting collection, giving the number of tweets, URLs, or webpages at each step. Afterwards, we followed the same steps to process the other four collections and sample 1,000 webpages for each collection.

Table 3.3: Webpage corpus generation on the Sandy Hook Elementary School shooting collection

Step	Action	# of Tweets	# of URLs (Webpages)
0	—	363,419	—
1	Remove spam and unformatted tweets	360,621	—
2	Extract short URLs from tweets	—	150,477
3	Crawl basic webpages	—	150,477
4	Crawl Wayback mementos	—	84,332
5	Merge basic and Wayback webpages	—	234,809
6	Remove unreadable webpages	—	139,144
7	Remove duplicated webpages	—	15,772
8	Remove similar webpages	—	14,182
9	Subsampling	—	1,000

To conduct the MTurk labeling task, we submitted our MTurk study to the Institutional Review Board (IRB) and received an exemption approval; see Appendix A. Regarding the task design, each MTurk assignment has five single labeling tasks and is assigned to five MTurk workers, who are requested to check whether a webpage is relevant to a given event. Therefore, for each webpage, we gathered the labeling results from five MTurk workers and judged whether it is relevant (R) or non-relevant (NR) through majority vote.

Table 3.4 lists the basic statistical results from the MTurk labeling task. Regarding the

average time per assignment, the time cost of the first two collections is less than 30s, since we set the assignment time duration to 1 minute. Based on the feedback of some workers, we extended the time duration to 3 minutes, so workers could spent more time on labeling the other three collections. Further, among the five collections, the Sandy Hook Elementary School shooting collection seems noisier, where 52.3% of webpages are non-relevant to the event. Meanwhile, 81.4% of webpages in the Nepal Earthquake collection are related to the disaster.

Table 3.4: Basic statistical results across the MTurk labeling task

ID	Collection	Avg. Time/Assignment	R/NR webpages	Total
$\mathbf{D}_1$	Sandy Hook shooting	26s	477(47.7%) / 523(52.3%)	1,000
$\mathbf{D}_2$	Hurricane Sandy	23s	767(76.7%) / 233(23.3%)	1,000
$\mathbf{D}_3$	Chapel Hill shooting	51s	724(72.4%) / 276(27.6%)	1,000
$\mathbf{D}_4$	Nepal Earthquake	49s	814(81.4%) / 186(18.6%)	1,000
$\mathbf{D}_5$	Hurricane Matthew	46s	660(66.0%) / 340(34.0%)	1,000

We further examined the labeling results of five MTurk workers per assignment for each collection. For each webpage, there are six possible voting results (R vs. NR): 5 vs. 0, 4 vs. 1, 3 vs. 2, 2 vs. 3, 1 vs. 4, and 0 vs. 5, representing as Most/Very/Somewhat Relevant and Somewhat/Very/Most Non-relevant, respectively. Figure 3.5 depicts the distribution of different voting results across MTurk workers. From the figure we notice that only about 10% of webpages are somewhat relevant or non-relevant in each of our five collections, indicating most webpages are easy to distinguish.

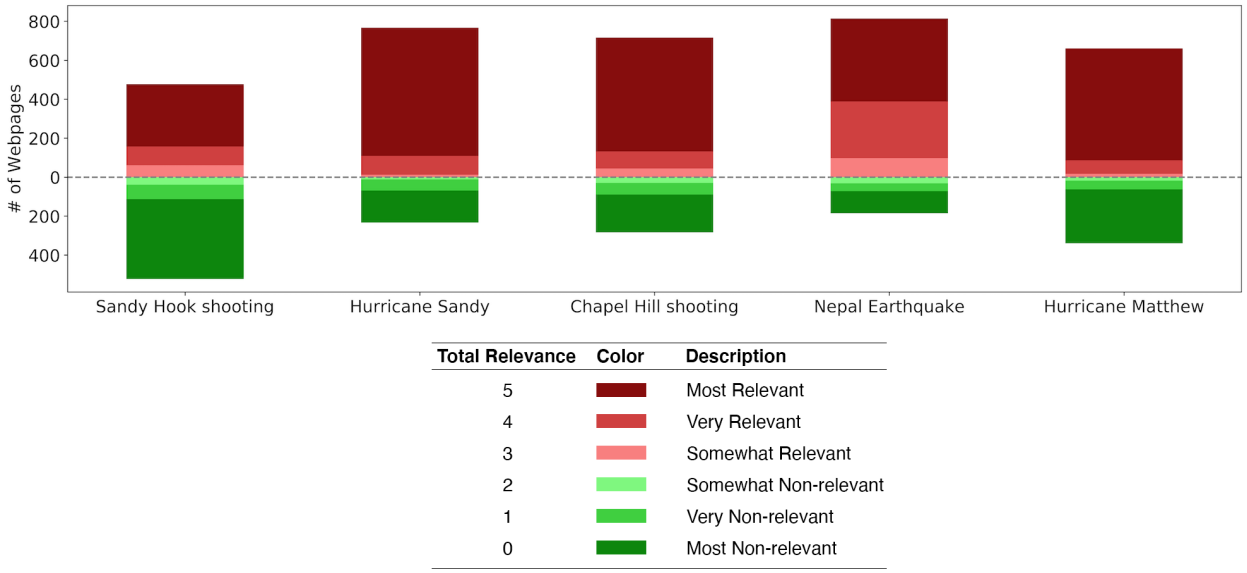


Figure 3.5: Distribution of different voting results across MTurk workers

3.2.2 Data for Short URL Analysis

We selected 12 event-related collections from January 1, 2013 to December 31, 2017 from our tweet archives, representing 4 categories: *Nature*, *Health*, *Man-made*, and *Particular Event*. The first three are general, while the fourth covers specific events. To reduce computation time, we randomly selected about 20% of the tweets as samples; see Table 3.5.

Table 3.5: Different categories of event-related collections

General Type	Keywords	# of Tweets
Nature	flood	2,201,160
	hurricane	2,103,014
	typhoon	1,158,824
Health	heart attack	3,659,421
	diabetes	2,135,363
	obesity	1,249,644
Man-made	terrorism	1,566,884
	gun control	1,206,863
	gun violence	783,040
Particular Event	hurricane sandy	385,337
	hurricane isaac	19,149
	connecticut school shooting	14,141

3.3 Evaluation and Analysis

3.3.1 Evaluation of Collection Development Model

We evaluate our CD model on the golden standard datasets of the five event-related collections in this section. In addition to our current model, we also propose several methods for comparison, including both basic methods and query expansion methods.

We calculate the average precision score (AP) and draw both the precision-recall (PR) curve and receiver operating characteristic (ROC) curve to evaluate our approach and the methods for comparison. AP summarizes a precision-recall curve as the weighted mean of precision achieved at each threshold; see Equation 3.6.

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{3.6}$$

where P_n and R_n are the precision and recall at the n^{th} threshold.

3.3.1.1 Baseline Methods

Random Method (*Rand*)

We shuffle the sampled webpages by their URLs, and take the top k records as relevant and the others as non-relevant.

Rule-based Method (*Rule*)

We build a filter rule to predict whether a URL points to an event-related webpage. Our rule considers two factors that we guess are likely too be associated with events. Based on the assumption that a URL linking to a news website is likely to be relevant to an event, we construct a list of top 100 USA news websites [34], and determine if the URL is associated with some table entry. Second, we assume that the length of a URL is correlated with the likelihood of it being event-related, and so experiment to set a length threshold to filter URLs. Table 3.6 lists the top 10 websites. We evaluate different length thresholds from 30 to 150 with step size equal to 10, and set the best threshold having the highest AP score during experiments; see Table B.1 in Appendix B. Then, if the domain name of a URL appears in the news site list and the URL length is greater than the best threshold, we predict that the URL points to a relevant webpage.

Table 3.6: Top 10 USA news websites

No.	News Site		No.	News Site
1	us.cnn.com/		6	www.reuters.com/
2	www.nytimes.com/		7	www.politico.com/
3	www.huffingtonpost.com/		8	www.yahoo.com/news/
4	www.foxnews.com/		9	www.npr.org/
5	www.usatoday.com/		10	www.latimes.com/

Okapi BM25 (*BM*25)

We use Okapi BM25 [98] as a basic content-based approach for comparison. The scoring function of BM25 is shown in Equation 3.7:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot |D|/avgdl)}$$
$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{3.7}$$

Here $f(q_i, D)$ is the TF score of q_i in the document D, $|D|$ is the number of words in the document D, and $avgdl$ is the average document length in the whole corpus. For IDF, N is

the total number of documents in the collection, and $n(q_i)$ is the number of documents that contain q_i.

In our case, the corpus is not huge, and many webpages are relevant to the given event, leading to negative IDF scores, so that the relevance scores may not precisely describe the relationship between query and documents. To solve this problem, since the original BM25 is for general web search, we use the ClueWeb09B dataset instead of our webpage corpus to update the IDF scores; see Equation 3.8.

$$IDF_w(q_i) = \log \frac{W - w(q_i) + 0.5}{w(q_i) + 0.5} \tag{3.8}$$

where W is the total number of documents in the ClueWeb09B corpus, and $w(q_i)$ is the number of ClueWeb09B documents containing q_i.

Query Likelihood Model (QL)

We compare the original QL model [62], excluding query expansion, with our approach.

Pre-query Design + Query Likelihood Model ($Pre + QL$)

To evaluate whether the relevance model can improve the ranking results, we drop the post-query expansion component and only use the first two parts for the retrieval task.

Query Likelihood Model + Post-query Expansion ($QL + RM3$)

Similar to the above, we remove the pre-query design component from our approach and use the most popular retrieval model for ranking [64]. In this way, we can evaluate whether the pre-query design can improve the ranking performance.

Since we do not have the prior knowledge (e.g., Wikipedia) about the five events when applying both BM25 and QL models, the initial queries for the two models consist of the following terms: year, typical words, and location; see Table 3.7.

Table 3.7: Initial queries for BM25 and QL models across five event-related collections

Collection	**Query**
Sandy Hook shooting	2012 Sandy Hook Shooting Connecticut
Hurricane Sandy	2012 Hurricane Sandy United States Atlantic October November
Chapel Hill shooting	2015 Chapel Hill Shooting North Carolina University
Nepal Earthquake	2015 Nepal Earthquake Kathmandu India Intensity
Hurricane Matthew	2016 Hurricane Matthew United States Atlantic September October

3.3.1.2 Results on Entire Collections

Tables 3.8 and 3.9 list the average precision (AP) scores and area under the curve (AUC) scores across all methods on the five entire collections. In general, our proposed model $Pre+QL+RM_3$ outperforms the other methods. We also notice that there are two exceptions. The best AP score is 0.982 ($QL+RM3$) in $\mathbf{D}_4$, while the best AUC score is 0.951 ($Pre+QL$) in $\mathbf{D}_1$. Afterward, we conduct a set of significance tests to evaluate the difference among methods and improve the interpretability of our results.

Table 3.8: Average precision (AP) scores across all methods on entire collections

		$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$	$\mathbf{D}_4$	$\mathbf{D}_5$
Relevance Ratio		0.477	0.767	0.724	0.814	0.660
Method/Model		**AP Score**				
Basic	*Rand*	0.480	0.775	0.718	0.808	0.651
	Rule	0.491	0.813	0.757	0.854	0.715
	$BM25$	0.927	0.979	0.976	0.979	0.955
	QL	0.943	0.984	0.978	0.980	0.960
Query Expansion	$QL+RM3$	0.948	0.985	0.983	**0.982**	0.962
	$Pre+QL$	0.960	0.986	0.985	0.980	0.965
	$Pre+QL+RM3$	**0.961**	**0.988**	**0.986**	0.981	**0.969**

Table 3.9: Area under the curve (AUC) scores across all methods on entire collections

		$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$	$\mathbf{D}_4$	$\mathbf{D}_5$
Relevance Ratio		0.477	0.767	0.724	0.814	0.660
Method/Model		**AP Score**				
Basic	*Rand*	0.506	0.523	0.500	0.478	0.480
	Rule	0.528	0.621	0.590	0.622	0.609
	$BM25$	0.917	0.945	0.944	0.928	0.923
	QL	0.932	0.956	0.950	0.932	0.932
Query Expansion	$QL+RM3$	0.938	0.959	0.956	0.934	0.936
	$Pre+QL$	**0.951**	0.961	0.968	0.938	0.941
	$Pre+QL+RM3$	0.950	**0.965**	**0.968**	**0.940**	**0.946**

Figures 3.6 through 3.10 show the PR curves (left [L]) and ROC curves (right [R]), which have already been summarized by the AP scores and AUC scores above. Because the random and rule-based methods only generate a binary value for each webpage, the curves are straight lines.

3.3.1.3 Significance Tests

For each collection, we create 15 random datasets and 15 balanced datasets, separately, for significance tests. For each random dataset, we randomly select 500 webpages at each time.

For each balanced dataset, we randomly select 150 relevant webpages and 150 non-relevant webpages from our imbalanced collections. For each subset created by either method, we calculate the AP and AUC scores across all baseline models and our proposed model. Then, we compute the average AP and AUC scores across all subsets for each model.

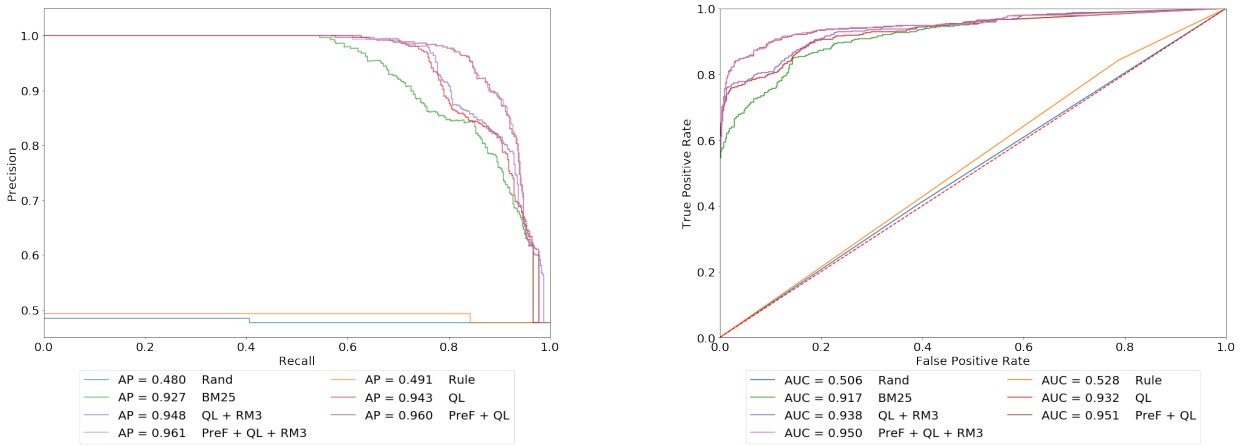


Figure 3.6: PR (L) and ROC (R) curves across all methods in the Sandy Hook shooting

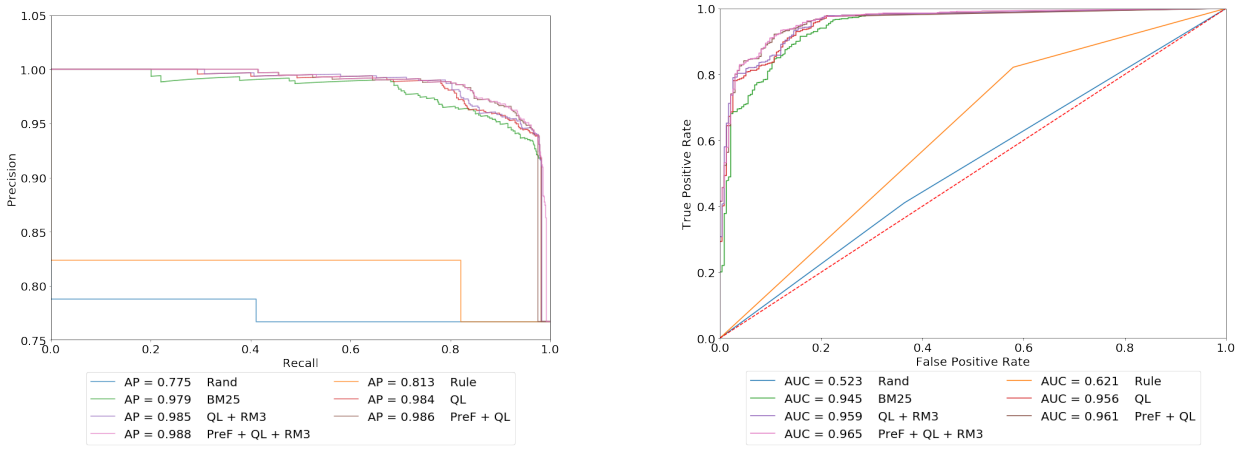


Figure 3.7: PR (L) and ROC (R) curves across all methods in Hurricane Sandy

We carry out paired t-tests between five pairs of models, including $BM25$ vs. QL, QL vs. $QL+RM3$, QL vs. $Pre+QL$, $QL+RM3$ vs. $Pre+QL+RM3$, and $Pre+QL$ vs. $Pre+QL+RM3$. We set the significance level α to 0.05, which is most commonly used. Tables 3.10 through 3.13 list the average AP and AUC scores, along with p-values in significance tests. If p-value is greater than 0.05, it means the average scores of a model pair have no significant difference. Otherwise, the average scores of two models are significantly different.

First, we examine the AP scores and p-values of $\mathbf{D}_4$ in Tables 3.10 and 3.12. Though the AP scores of $QL+RM3$ are still the highest in both random and balanced datasets, the two p-values of $QL+RM3$ vs. $Pre+QL+RM3$ are greater than 0.05, indicating that

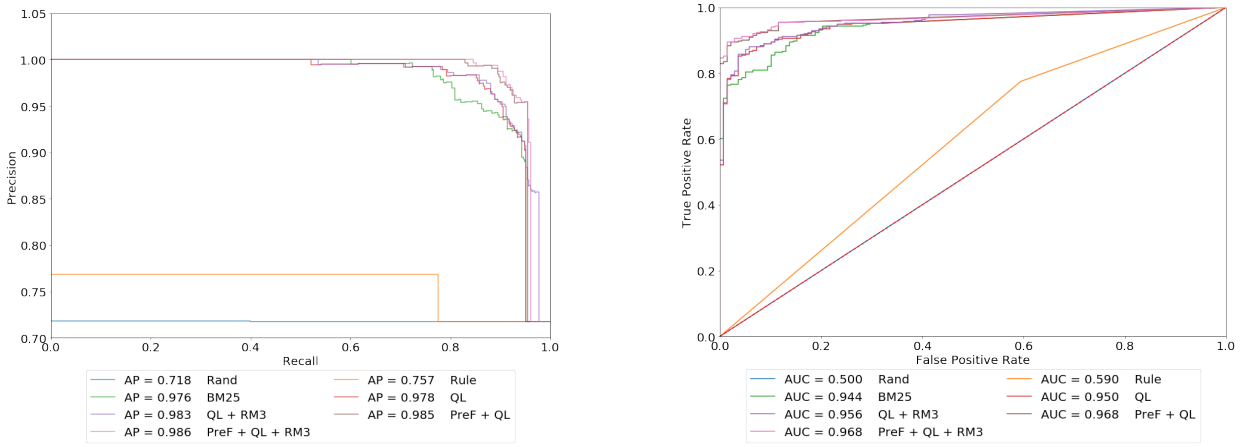


Figure 3.8: PR (L) and ROC (R) curves across all methods in the Chapel Hill shooting

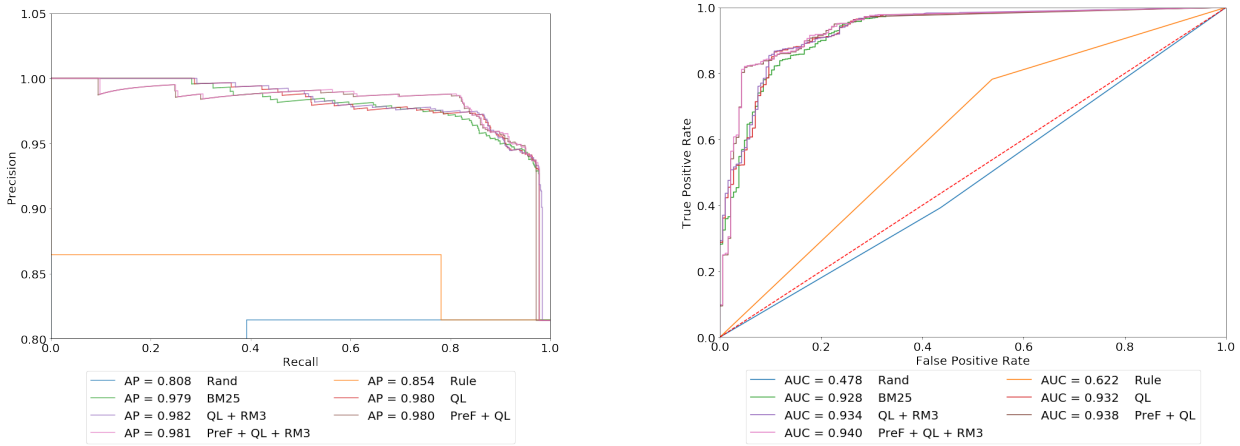


Figure 3.9: PR (L) and ROC (R) curves across all methods in Nepal Earthquake

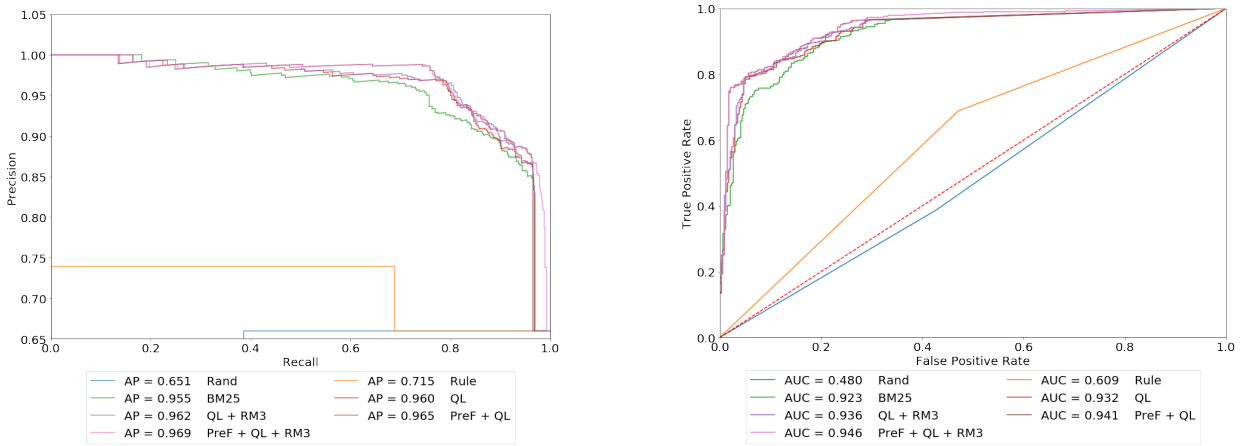


Figure 3.10: PR (L) and ROC (R) curves across all methods in Hurricane Matthew

there is no difference between the two models. Similarly, in Table 3.11, the p-value of $Pre+QL$ vs. $Pre+QL+RM3$ is 0.32, showing the two models have the same performance on the random dataset of $\mathbf{D}_1$. Specifically, regarding the balanced dataset of $\mathbf{D}_1$, $Pre+QL+RM3$ performs better than $Pre+QL$.

Second, we explain the average AP and AUC scores with the help of p-values in both random and balanced datasets. For each collection, the AP score of the random method is close to the relevance ratio, while the rule-based method has a higher AP score since it utilizes news sites and URL length as knowledge. Both methods perform worse, since no context information has been used. Regarding the other two basic methods, the QL model is always better than $BM25$ in both random and balanced datasets. Further, all methods with pre-query design perform better than or equally with their corresponding basic methods, showing that pre-query design can improve the ranking results. The AP and AUC scores of the $Pre+QL$ method are equal to or greater than the one of the QL method, while the AP scores of the $Pre+QL+RM3$ method are greater than for the $QL+RM3$ method in all collections. We also find the post-query expansion has a positive effect on ranking. The two methods with $RM3$ have higher AP and AUC scores than those without $RM3$, correspondingly.

From the results, we conclude that our proposed model $Pre+QL+RM3$ performs better than baseline methods; it can retrieve more relevant webpages from event-related collections.

Table 3.10: Average AP scores across all methods on random datasets

		$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$	$\mathbf{D}_4$	$\mathbf{D}_5$
Relevance Ratio		0.477	0.767	0.724	0.814	0.660
Method/Model		**AP Score**				
Basic	*Rand*	0.484	0.769	0.702	0.815	0.659
	Rule	0.502	0.823	0.767	0.863	0.721
	$BM25$	0.928	0.980	0.975	0.980	0.953
	QL	0.944	0.985	0.978	0.981	0.959
Query Expansion	$QL+RM3$	0.948	0.986	0.982	**0.982**	0.961
	$Pre+QL$	0.961	0.986	0.985	0.980	0.965
	$Pre+QL+RM3$	**0.962**	**0.988**	**0.988**	0.981	**0.969**

Pair	**D1**	**D2**	**D3**	**D4**	**D5**
$BM25$ vs. QL	6.01e-13	2.41e-10	2.00e-03	8.56e-04	9.70e-09
QL vs. $QL+RM3$	1.10e-06	3.68e-09	1.09e-06	3.58e-09	2.45e-09
QL vs. $Pre+QL$	5.39e-12	0.06	9.26e-07	0.1	1.51e-07
$QL+RM3$ vs. $Pre+QL+RM3$	1.33e-11	6.00e-04	2.18e-04	0.15	5.59e-10
$Pre+QL$ vs. $Pre+QL+RM3$	1.44e-05	2.47e-09	2.98e-05	4.18e-10	5.83e-11

Table 3.11: Average AUC scores across all methods on random datasets

		$\mathbf{D_1}$	$\mathbf{D_2}$	$\mathbf{D_3}$	$\mathbf{D_4}$	$\mathbf{D_5}$
Relevance Ratio		0.477	0.767	0.724	0.814	0.660
Method/Model		**AUC Score**				
Basic	*Rand*	0.512	0.498	0.492	0.503	0.498
	Rule	0.546	0.637	0.629	0.648	0.618
	$BM25$	0.918	0.946	0.943	0.931	0.921
	QL	0.933	0.957	0.950	0.934	0.930
Query Expansion	$QL + RM3$	0.937	0.959	0.958	0.936	0.934
	$Pre + QL$	0.951	0.961	0.969	0.938	0.941
	$Pre + QL + RM3$	**0.951**	**0.964**	**0.972**	**0.939**	**0.945**

Pair	**D1**	**D2**	**D3**	**D4**	**D5**
$BM25$ vs. QL	2.84e-13	1.04e-11	2.59e-03	7.48e-05	2.98e-10
QL vs. $QL + RM3$	1.73e-04	9.29e-07	6.59e-06	1.91e-06	4.30e-08
QL vs. $Pre + QL$	6.68e-13	1.20e-02	1.48e-08	0.079	2.35e-09
$QL + RM3$ vs. $Pre + QL + RM3$	1.66e-11	8.16e-04	1.39e-05	0.071	5.43e-11
$Pre + QL$ vs. $Pre + QL + RM3$	0.32	2.37e-07	2.29e-03	2.78e-06	8.86e-09

Table 3.12: Average AP scores across all methods on balanced datasets

		$\mathbf{D_1}$	$\mathbf{D_2}$	$\mathbf{D_3}$	$\mathbf{D_4}$	$\mathbf{D_5}$
Relevance Ratio		0.500	0.500	0.500	0.500	0.500
Method/Model		**AP Score**				
Basic	*Rand*	0.507	0.507	0.503	0.508	0.507
	Rule	0.523	0.585	0.578	0.601	0.569
	$BM25$	0.937	0.938	0.920	0.921	0.923
	QL	0.952	0.956	0.956	0.927	0.932
Query Expansion	$QL + RM3$	0.954	0.960	0.962	**0.932**	0.937
	$Pre + QL$	0.967	0.960	0.972	0.926	0.938
	$Pre + QL + RM3$	**0.968**	**0.964**	**0.979**	0.930	**0.944**

Pair	**D1**	**D2**	**D3**	**D4**	**D5**
$BM25$ vs. QL	2.25e-10	3.34e-12	3.77e-09	3.94e-06	1.19e-06
QL vs. $QL + RM3$	1.28e-05	7.93e-09	5.46e-05	1.85e-07	6.28e-06
QL vs. $Pre + QL$	3.12e-07	1.40e-02	1.23e-08	0.27	1.01e-02
$QL + RM3$ vs. $Pre + QL + RM3$	2.68e-07	9.20e-03	1.37e-09	0.11	2.35e-03
$Pre + QL$ vs. $Pre + QL + RM3$	1.50e-03	4.84e-08	3.95e-06	2.18e-06	3.21e-06

Table 3.13: Average AUC scores across all methods on balanced datasets

		$\mathbf{D}_1$	$\mathbf{D}_2$	$\mathbf{D}_3$	$\mathbf{D}_4$	$\mathbf{D}_5$
Relevance Ratio		0.500	0.500	0.500	0.500	0.500
Method/Model		**AUC Score**				
Basic	*Rand*	0.513	0.513	0.513	0.513	0.513
	Rule	0.542	0.636	0.614	0.652	0.610
	$BM25$	0.924	0.943	0.942	0.926	0.924
	QL	0.939	0.956	0.950	0.931	0.933
Query Expansion	$QL + RM3$	0.941	0.959	0.955	0.934	0.937
	$Pre + QL$	0.956	0.961	0.969	0.936	0.939
	$Pre + QL + RM3$	**0.956**	**0.964**	**0.974**	**0.938**	**0.944**

Pair	**D1**	**D2**	**D3**	**D4**	**D5**
$BM25$ vs. QL	1.96e-10	5.11e-13	1.49e-03	9.10e-08	2.79e-08
QL vs. $QL + RM3$	1.30e-03	1.79e-07	2.07e-04	1.02e-05	1.21e-04
QL vs. $Pre + QL$	1.96e-06	2.17e-03	7.43e-09	3.55e-04	1.32e-03
$QL + RM3$ vs. $Pre + QL + RM3$	3.34e-06	1.56e-03	3.46e-09	5.18e-03	4.09e-03
$Pre + QL$ vs. $Pre + QL + RM3$	0.17	3.06e-06	3.77e-04	3.60e-02	3.77e-04

3.3.2 Analysis of Short URLs in Event-related Collections

In this section, we carry out an empirical analysis of short URLs in event-related collections and present some interesting results. This study can be considered as an application of data curation, which utilizes the extended tweet records and response records in WARC files. It also helps researchers better understand the lifespan of short URLs and the role of the Wayback Machine.

3.3.2.1 Tweets with Short URLs

Basically, for each event-related collection, we are interested in how many URLs are embedded in a tweet; see Figure 3.11. Of all the tweets with URLs, it is clear that most (about 90%) tweets have only one URL. Meanwhile, 10% of tweets have two URLs on average, and less than 1% of the tweets have three or more URLs embedded.

For each year, we calculate the percentage of tweets with short URLs; see Figure 3.12. We notice that there is no significant difference between different categories of events. Percentages are lowest for "heart attack" and highest for "Connecticut school shooting"; for that there is a great reduction from 2016 to 2017. For most collections, the peak value appears in 2015 or 2016 instead of 2017. The reason might lie in Twitter's mid-2016 decision to exclude URLs from the tweet length limit.

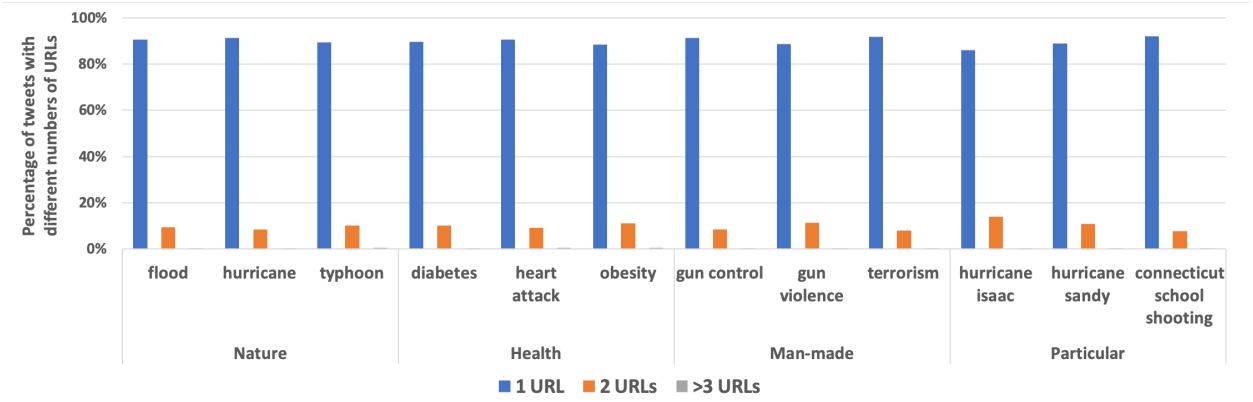


Figure 3.11: Percentage of tweets with different numbers of URLs

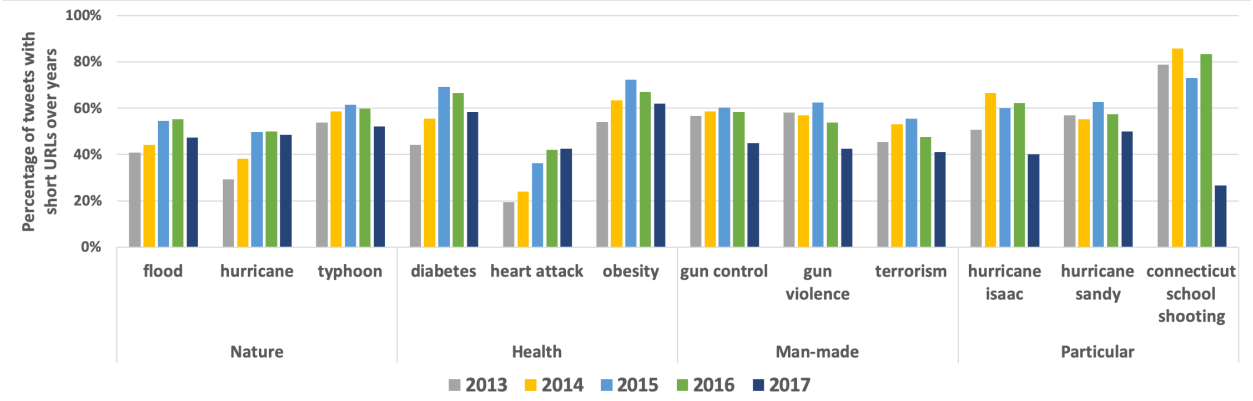


Figure 3.12: Percentage of tweets with short URLs over years (2013-2017)

3.3.2.2 Broken URLs over Years

Figure 3.13 shows that recent URLs are less likely to reflect broken links, but there is less difference for "hurricane isaac". Table 3.14 shows the average percentage values over the years. In general, the percentage of broken URLs dropped 3%-6% year by year. The average percentage of broken links over the past 5 years is about 33%.

Table 3.14: Average percentage of broken URLs over years (2013-2017)

	All Years	**2013**	**2014**	**2015**	**2016**	**2017**
Avg %	32.9	40.9	37.2	31.9	25.7	19.5

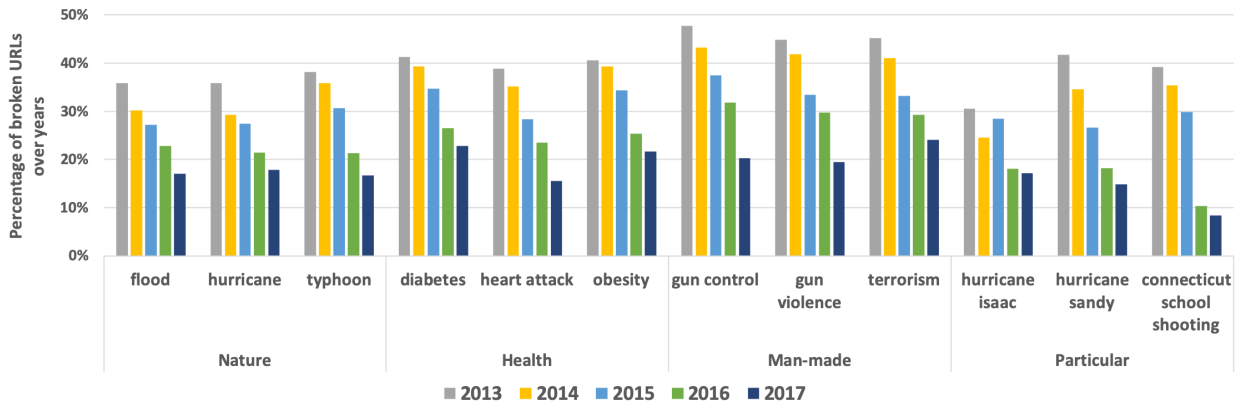


Figure 3.13: Percentage of broken URLs over years (2013-2017)

3.3.2.3 Retrievable URLs over Years

A memento in the Wayback Machine is an archived copy of a webpage. The Wayback Machine allows retrieval for many mementos in former years; see Figure 3.14. As shown in this figure, the Wayback Machine has mementos for a higher percentage of URLs for "man-made" events than for other types of events. Table 3.15, shows that the Wayback Machine provides webpages for 17.4% of the short URLs. We further split all URLs into two classes: broken and unbroken, and find older unbroken URLs are more likely to be saved.

Table 3.15: Average percentage of retrievable URLs over years (2013-2017)

	All Years	2013	2014	2015	2016	2017
Avg % - all	17.4	23.2	20.7	17.3	15.9	12.0
Avg % - broken	14.8	16.4	15.9	14.0	15.7	12.4
Avg % - unbroken	18.5	27.7	23.4	18.7	15.9	11.9

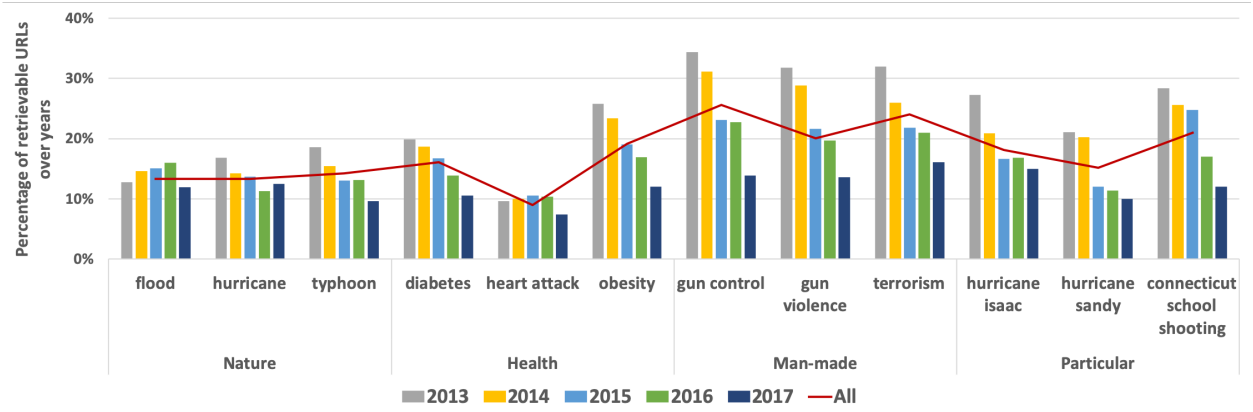


Figure 3.14: Percentage of retrievable URLs over years (2013-2017)

Chapter 4

User Classification and Tweeting Pattern Analysis

The research question for this chapter is: *How can the roles of tweeters, along with tweeting patterns, be identified and analyzed to provide insights about tweeters and their behavior?* A sub-system for user classification and tweeting pattern analysis is proposed to address this question. At the information level, we focus on identifying the roles of users (i.e., brand, female, and male), and analyzing the posting and mood patterns across event-related collections about disasters, through user classification and by considering other information.

4.1 Approach

We propose TwiRole, with an integrated model for 3-way role-related user classification on Twitter, which detects brand-related, female-related, and male-related users. Then we detect the moods expressed in tweets through an existing recurrent neural network (RNN) model [24]. By integrating the information of tweets, tweeters, and moods, we report on a tweeting pattern analysis of 12 large disaster-related collections and a case study of Hurricane Dorian.

4.1.1 Role-related User Classification

Using a hybrid model, TwiRole has a multi-classifier on basic features (BF), a multi-classifier on advanced features (AF), a CNN model, and a final multi-classifier, as shown in Fig 4.1. The BF multi-classifier takes the basic features from the user profiles and tweets as input. The AF multi-classifier focuses on the k-top words in user tweets. The CNN works on the user profile images. The final multi-classifier takes the output of the above three modules as input.

We focus on the following five types of features: *name*, *description*, *relationship*, *profile image*, and *tweet*. Each type has one or multiple features, as shown in Table 4.1; see below a brief discussion of each of the 7 features summarized in the table. Different classifiers are focusing on different features. For each feature, we compute a score to be used in classifiers; see details below. Adopting a supervised learning approach, we train and test TwiRole on third-party datasets with labels.

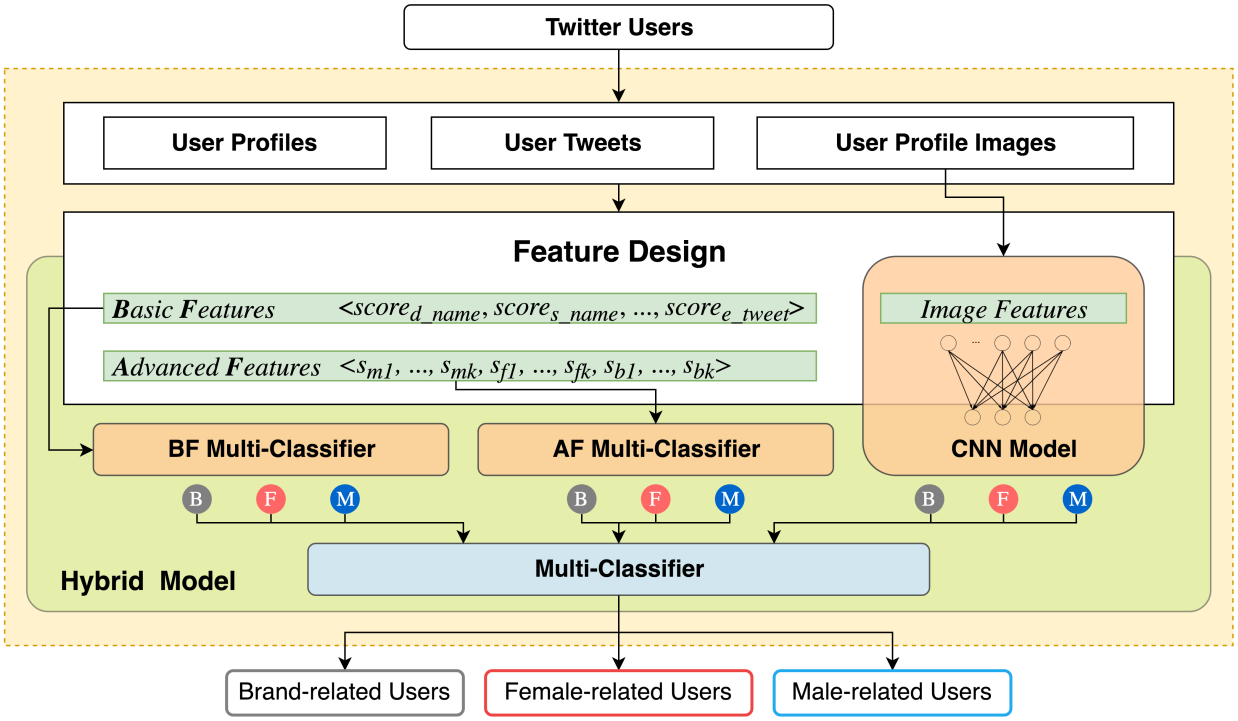


Figure 4.1: Architecture of our proposed TwiRole tool

Table 4.1: Feature types and details in TwiRole

No.	Feature Type	Feature Detail	No.	Feature Type	Feature Detail
BF1	*name*	display name screen name	BF3	*relationship*	TFF score
			BF4	*profile image*	brightness
BF2	*description*	first-person score number of tokens	BF5	*tweet*	first-person score interjection score emotion score
AF	*tweet*	k-top words	CNN	*profile image*	hidden in image

BF1 – name

To calculate the name score, we downloaded popular baby names from the US Social Security Administration [104] into a database, chose a subset covering the past 10 years, and then summarized the occurrences of names over years. Each name can be represented as a vector $< name, gender, frequency >$. We had a total of 71,299 records. It must be noted that for a given name, the gender could be either male or female. For instance, there are 406 female babies named "dallis" while 167 male babies have the same name. To expand the name dataset, we combined it with an Arabic name dictionary [105] that includes 979 female names and 898 male names. Because the Arabic dataset has no occurrence numbers, we simply set the field with the same value for each name. Duplicated names were removed

during combination. Finally, there were 72,134 rows in the name dictionary.

Given a display name d_name, TwiRole first splits it into tokens. It only takes the token that first appears in the name dictionary to calculate the display name score $score_{d\_name}$. If there is no token found, $score_{d\_name}$ is equal to 0, as shown in Equation 4.1:

$$score_{d\_name} = \begin{cases} \frac{tf_f - tf_m}{\max(tf_f, tf_m)} \in [-1, 1], & \text{token } t \text{ is found} \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

where tf_f and tf_m represent the female and male frequency of token t. For instance, given a display name "John Clemson", since "John" is the first token found in the name dictionary and there are 445 females and 256,166 males named "John", the display name score is calculated as $score_{d\_name} = (445 - 256166)/256166 = -0.998$.

For the screen name, TwiRole parses the entire string into tokens through an integrated method combined with four different methods: name-based, word-based, name-word-based, and wordninja [5], a popular word splitter. Based on the parsing results, TwiRole takes the result with the least number of tokens as the best candidate. Table 4.2 shows the different parsing results of two samples, where the candidates are shown in bold. Then, we reapply Equation 4.1 to calculate the screen name score $score_{s\_name}$.

Table 4.2: Results for different parsing methods

screen name	clemsonjohn	123tommy
Method	**Results**	
name-based	clem, son, john	tom, my
word-based	cl, ems, on, john	**tommy**
name-word-based	clem, son, john	**tommy**
wordninja	**clemson, john**	1, 2, 3, t, o, m, m, y

BF2 – description

Users on Twitter are likely to show their role information through the description, since it appears on the personal main page, and might give a brief introduction to the user. We proposed two word lists: first-person word list and brand word list, to calculate the first-person score in the description. The first list is represented as $list_{first}$, containing first-person words like $i, am, my, me, mine, i'm$, while the latter one is represented as $list_{brand}$, which has one word: *official*. While scanning the tokens in a user's description, we set the first-person score by:

$$score_{fp\_desc} = \begin{cases} 1, & \text{token } t \in list_{first} \text{ and } t \notin list_{brand} \\ -1, & \text{token } t \in list_{brand} \text{ and } t \notin list_{first} \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

Then, we removed hashtags, mentions, and URLs from the user's description, and stored the number of the remaining tokens for each user as $score_{tf\_desc}$.

BF3 – relationship

Twitter Follower-Friend (TFF) ratio is a widely used figure that represents the quantitative relation of a user [58]. Here, we make use of the number of followers $Num_{followers}$ and friends $Num_{friends}$ to calculate the TFF ratio, as shown in Equation 4.3:

$$score_{tff} = \log \frac{Num_{followers}{}^2 + 1}{Num_{friends} + 1} \tag{4.3}$$

Different from the basic TFF score, we add 1 to both denominator and numerator to avoid dividing by zero or $\log 0$ error. We also add an exponent on the numerator to strengthen the number of followers, and distinguish from cases where the follower number is proportional to the friend number.

BF4 – profile image

For the profile images, TwiRole focuses on the HSV format instead of RGB, because the latter format seems noisier. We selected brightness, also called "value" in HSV, as a basic feature. Given a profile image, we convert the original RGB format into HSV, accumulate the brightness score of each pixel, and compute the average brightness $score_{b\_image}$ of the entire image.

BF5 – tweet

To process the tweet contents of each user, TwiRole generates three scores: first-person score, interjection score, and emotion score, which are represented as $score_{fp\_tweet}$, $score_{i\_tweet}$, and $score_{e\_tweet}$, respectively. Besides the first-person word list above, we created two other word lists for the word matching: interjection and emotion. The interjection list includes words related to an expression of feeling, such as *oops*, *goodness*, *ah*, and *wow*. The emotion words can be categorized into seven types: *curiosity*, *urgency*, *confusion*, *anger*, *satisfied*, *inspired*, and *relaxed*. The three scores can be calculated through a linear scan of each tweet collection. Equation 4.4 shows how to calculate the first-person score; the other two scores can be computed in the same way.

$$score_{fp\_tweet} = \frac{\#\text{ of tweets that have words in } list_{first}}{\#\text{ of tweets in user tweet collection}} \tag{4.4}$$

AF1 – tweet

TwiRole follows the popular k-top words method [79] to further process each tweet collection. First, TweetNLP [91], a fast and robust tokenizer and part-of-speech tagger, has been leveraged to extract the important tags – nouns (N), verbs (V), adjectives (A), adverbs (R), emoticons (E), and hashtags (#) – from the raw texts. Then, in comparison

with the k-top words method, our tool applies a similar method to create a word list for each of the three roles and merge them into one vector, represented as $vector_{k\_top} =< w_{m1}, w_{m2}, ..., w_{mk}, w_{f1}, w_{f2}, ..., w_{fk}, w_{b1}, w_{b2}, ..., w_{bk} >$. For each word in the vector, TwiRole counts the number of tweets containing the given word, divides it by the size of the user tweet collection, and finally gets the k-top words score vector $score_{k\_top}$, shown in Equation 4.5. We assign different values to k during experiments and take the best value $k = 20$ as default; see Table B.2 in Appendix B.

$$score_{k\_top} =< s_{b1}, s_{b2}, ..., s_{bk}, s_{f1}, s_{f2}, ..., s_{fk}, s_{m1}, s_{m2}, ..., s_{mk} > \tag{4.5}$$

CNN1 – profile image

In addition to the basic and advanced features, our tool also applies a pre-trained ResNet-18 model [47] to extract the hidden features from the profile images. Because the original ResNet is designed for the ImageNet dataset [100] with 1,000 categories, we have changed the number of nodes in the output layer from 1,000 to 3 and employed a softmax function on the three nodes. There are 512 nodes in the fully-connected layer which produce the deep features in the network.

4.1.2 Tweeting Pattern Analysis

We analyze both the posting and mood patterns across different types of disasters [73]. Given event-related collections, we extract the screen names of users and use TwiRole to predict their roles. Afterwards, we apply a pre-trained RNN model [24] for predicting Ekman's emotions [31] of tweets, which include six basic emotions (i.e., joy, fear, surprise, sadness, anger, and disgust). Focusing on tweets and user roles, we uncover the posting patterns and user distribution among different disasters. Moreover, tweets, user roles, and moods help us analyze the tweeting patterns at both the event and user level; see details in Sections 4.3.2 and 4.3.3.

4.2 Data

4.2.1 Data for User Classification

To reduce bias resulting from the selection of data, we did not create any labeled dataset by ourselves. Instead, we reused two existing Twitter user classification datasets, each of which contains a huge number of users with labels.

4.2.1.1 Kaggle Dataset

The dataset on Kaggle [55] is a project of CrowdFlower [25], including the information of about 20,000 users. Project contributors manually labeled each user by checking the corresponding information, which contains part of the profile metadata, such as display name, screen name, description, link color, etc. There are three labels in the dataset: male, female, and brand. The contributors also provided a confidence score along with the role tag, which is a good indicator of labeling quality.

4.2.1.2 Gender-labeled Twitter Dataset

Liu and Ruths released a public gender-labeled dataset[1] to support the evaluation of different user detection approaches. Three Amazon Mechanical Turk workers manually labeled a user as male or female if all of them agreed on the same gender assignment. The dehydrated dataset only has two fields: user ID and gender; the gender field has two values: "M" and "F". In total, there are 4,449 male-related users and 8,232 female-related users.

4.2.1.3 Data Preprocessing

For the Kaggle dataset, we first remove duplicate users that have the same screen name. Since we focus on role-related users, those with blank labels or "unknown" labels have also been filtered out. We keep users with labels having confidence value 1, assuming their records are of high labeling quality. Regular expressions are utilized to detect and remove users with similar patterns (e.g., 02633gnlc, 02634gnlc, 02636gnlc). For the remaining users, we run our 24 virtual machine cluster to retrieve their tweets and eliminate any user whose tweet file size is less than 4KB. Next, we update the user profiles and get the "bigger" [2] profile image of each user through the Twitter API. Thus, we create a high quality Twitter user dataset that contains profile information, tweet contents, and profile images. The resulting high quality set of 8,625 users has unbalanced numbers among the different roles. Consequently, we randomly select 6,000 users (2,000 users in each category) to build a balanced subsampled dataset. Table 4.3 shows all of the steps, giving the number of users at each step.

For Liu and Ruths' dataset, we follow steps 5-7 to retrieve the profiles, tweets, and images of users, and complete the preprocessing task. Since the gender-labeled Twitter dataset only has two classes, we take 3,000 users as a subset of each class.

[1] Download link: http://www.networkdynamics.org/static/datasets/LiuRuthsMicrotext.zip. Accessed date: 03/01/2020

[2] Image sizes for user profile images: normal: 48×48; mini: 24×24; bigger: 73×73

Table 4.3: Data preprocessing on Kaggle dataset

Step	Action	# of Users Left
0	—	20,050
1	Remove duplicated users	18,795
2	Remove users with blank and "unknown" labels	17,660
3	Remove users with less confidence value	12,991
4	Remove users with similar patterns	12,889
5	Remove users with tweet file size < 4KB	8,714
6	Remove users with broken profile images	8,625 (B: 2,254, F: 3,176, M: 3,195)
7	Subsampling	6,000 (B: 2,000, F: 2,000, M: 2,000)

4.2.2 Data for Tweeting Pattern Analysis

We apply GetOldTweets3 [85] to create a dozen complete tweet collections with four different types of disasters (i.e., school shooting, bombing, earthquake, and hurricane), three event collections per type. The time range of each collection covers up to about one month after the corresponding disaster. For example, regarding hurricanes, we find the date they first formed, according to Wikipedia, and obtain tweets for four weeks starting on that date. Later, retweets (RTs) and non-English tweets are filtered out during cleaning. We also create a single collection related to Hurricane Dorian to conduct a case study of disaster response patterns across different user groups. Table 4.4 shows the details of our collections.

Table 4.4: An overview of tweet collections for two tweeting pattern analyses

Type	Collection	Starting (UTC)	# of Tweets
School Shooting	Sandy Hook Elementary School shooting	12/14/2012 14:40:00	97,283
	Stoneman Douglas High School shooting	02/14/2018 19:27:00	22,321
	Umpqua Community College shooting	10/01/2015 17:48:00	17,821
Bombing	Boston Marathon bombing	04/15/2013 18:49:00	211,142
	San Bernardino attack	12/02/2015 18:58:00	44,224
	Manchester Arena bombing	05/22/2017 21:31:00	6,105
Earthquake	Nepal earthquake	04/25/2015 06:11:00	671,323
	Japan earthquake	03/11/2011 05:46:00	566,627
	Taiwan earthquake	02/05/2016 19:57:00	48,894
Hurricane	Hurricane Sandy	10/22/2012 00:00:00	2,399,334
	Hurricane Matthew	09/28/2016 00:00:00	1,088,212
	Hurricane Florence	08/31/2018 00:00:00	636,281
	Hurricane Dorian	08/24/2019 00:00:00	565,911

4.3 Evaluation and Analysis

4.3.1 Evaluation of User Classification

We utilize 10-fold cross validation to evaluate TwiRole. In the training phase, we calculate all the feature scores from BF1 to BF5 for the users as input, and train the BF multi-classifier with the role-related labels. For each user, the output is a probability vector for three different roles. Then, we train the AF multi-classifier with the k-top words score vectors in the same way. Next, the profile images of all the training users and their labels are put into the ResNet-18 model to train the deep neural network, and we also get the probability vector for each user. At last, we concatenate the three probability vectors and train the final multi-classifier. Different types of classifiers (e.g., decision tree, Naive Bayes) can be applied on the AF, BF, and final multi-classifiers. To reduce the number of combinations, we set all the three multi-classifiers with the same type. In the testing phase, since all the modules have been trained and fixed, for each user, we follow the above steps to produce the prediction from the final multi-classifier and compare the result with the ground truth.

For the Kaggle dataset, we carry out an intra-comparison to verify our tool with different classifiers, features, and parameters, and also draw an inter-comparison between TwiRole and the method developed by Ferrari et al. [35]. Then, for the gender-labeled Twitter dataset, we slightly modify TwiRole into a bi-classification model TwiRole^{bi}, and compare it with Liu & Ruths' approach on their dataset.

We use a confusion matrix to calculate the recall (R), precision (P), and F1 score of each role. For a certain role r, the three values are computed as:

$$\begin{aligned} Recall_r &= \frac{\#\text{ of users correctly identified as } r}{\#\text{ of users labeled as } r}, \\ Precision_r &= \frac{\#\text{ of users correctly identified as } r}{\#\text{ of users predicted as } r}, \\ F1_r &= \frac{2 * Recall_r * Precision_r}{Recall_r + Precision_r} \end{aligned} \tag{4.6}$$

The performance of TwiRole is reflected in the overall accuracy; see Equation 4.7.

$$Accuracy = \frac{\sum_r (\#\text{ of users correctly identified as } r)}{\sum_r (\#\text{ of users labeled as } r)} \tag{4.7}$$

4.3.1.1 Kaggle Dataset

First, we evaluate TwiRole with different classifiers and also measure the performance of every single model as well as our hybrid model. Regarding a multi-classifier that considers the basic and advanced features, we experiment to compare classical individual classifiers like

decision tree and support vector machine (SVM), and ensemble classifiers such as AdaBoost, GradientBoosting, and random forest. For the CNN model, we use ResNet-18 as default.

Table 4.5 shows the accuracy of TwiRole's modules with different classifiers. The CNN model alone does well, but a combination is better. Among the five classifiers, GradientBoosting does best for both sets of features ($Acc_{BF} = 0.816$, $Acc_{AF} = 0.738$), but random forest has the highest accuracy ($Acc = 0.899$) regarding the entire model. Moreover, the ensemble classifiers perform better than the classical individual classifiers in every single model and the hybrid model. By comparing the performance of every single model and the hybrid model, we notice that the hybrid model is always better than each single model with different classifiers, except decision tree (with a tie). Accordingly, in further evaluation studies, the default is to use random forest, and a hybrid model is preferred.

Table 4.5: Accuracy of TwiRole's modules with different classifiers

Classifier Type	**Accuracy**			
	BF Multi-classifier	**AF Multi-classifier**	**CNN**	**Overall**
Decision Tree	0.721	0.618	0.790	0.721
SVM	0.739	0.685		0.800
AdaBoost	0.790	0.704		0.850
GradientBoosting	**0.816**	**0.738**		0.842
Random Forest	0.796	0.708		**0.899**

Then, we evaluate the feature sets of TwiRole that can help us find out which feature has a greater impact among the whole feature set. Specifically, we take the hybrid model with all features as our baseline method, and remove the features belonging to each feature type step by step to generate multiple feature subsets. Based on the remaining features, we retrain and reevaluate the entire model.

Table 4.6 shows TwiRole's performance with different feature sets. Using all features gives the best accuracy overall, as well as the best F1 score for each of the roles. The CNN features seem most important; omitting them leads to a 6.2% drop in accuracy. Similarly, as a basic feature, name (including display and screen name) also plays an important role among the features. On the other hand, some features, like description and relationship, may have a small impact; after removal, the overall accuracy only declined 0.2% and 0.3%, respectively. We also consider how TwiRole does with regard to each of the user roles. In most cases, we find that $F1_{female} > F1_{male} > F1_{brand}$. The exceptions mainly occur in sets 1 and 7, where the dropped features have a great impact on prediction results.

Focusing on the user tweet collections, we further investigate the parameters in TwiRole. We first choose the most recent 10, 30, 50, and all tweets posted by each user to calculate the three scores in BF5, then set the value k as 1, 5, 10, and 20 in k-top words. Table 4.7 shows the performance of TwiRole with different parameters. The accuracy achieves the best result in BF5 when we leverage the entire tweet collection, since it can integrally describe the users' behavior. For the k-top words, the best value of k is 10 or 20. It seems that a

Table 4.6: TwiRole's performance with different feature sets

Feature Set Description	Male	Female	Brand	Acc
	F1	F1	F1	
0. All Features	**0.903**	**0.908**	**0.885**	**0.899**
1. Without BF1 (name)	0.874	0.876	0.861	0.870
2. Without BF2 (description)	0.903	0.905	0.883	0.897
3. Without BF3 (relationship)	0.901	0.906	0.882	0.896
4. Without BF4 (profile image)	0.897	0.902	0.876	0.892
5. Without BF5 (tweet)	0.901	0.898	0.875	0.892
6. Without AF1 (tweet)	0.893	0.893	0.868	0.885
7. Without CNN1 (profile image)	0.814	0.843	0.854	0.837

small k is not helpful enough to differentiate the roles of users. When we set k to 30, 50, or 100, computation time increases, but with no significant performance increase.

Table 4.7: TwiRole's performance with different parameters in BF5 and AF1

Parameters in BF5 (tweet)	Acc
Recent 10 tweets	0.894
Recent 30 tweets	0.893
Recent 50 tweets	0.893
All user tweets	**0.899**

Parameters in AF1 (tweet)	Acc
1 top words	0.886
5 top words	0.890
10 top words	**0.899**
20 top words	**0.899**

We compare TwiRole with Ferrari et al.'s work on the same dataset. The classification results are shown in Table 4.8. Their model has an advantage in identifying the male-related users, where the F1 score is 0.947 and ours is 0.903. But TwiRole performs better in detecting both female-related ($F1_{female} = 0.908$) and brand-related users ($F1_{brand} = 0.885$), and the overall accuracy ($Acc = 0.899$) is higher than with Ferrari et al.'s approach ($Acc = 0.865$). Besides, the prediction results of our model are more balanced across different roles, because the difference in F1 score is only 0.023 in TwiRole while it is 0.136 for Ferrari et al.

Table 4.8: Performance of TwiRole and Ferrari et al.'s work

Model	Male			Female			Brand			Acc
	R	P	F1	R	P	F1	R	P	F1	
TwiRole	0.885	0.922	0.903	**0.920**	**0.897**	**0.908**	**0.891**	**0.879**	**0.885**	**0.899**
Ferrari et al.	**0.948**	**0.946**	**0.947**	0.806	0.857	0.831	0.837	0.786	0.811	0.865

4.3.1.2 Gender-labeled Twitter Dataset

Besides testing multi-classification, we test our hybrid model on the gender-labeled Twitter dataset. Because the dataset has only two classes – male and female – we slightly adjust

TwiRole to enable it for bi-classification; we name the variant model as TwiRole^{bi}. It makes use of the same features as TwiRole, but merges the basic and advanced features for bi-classification. Moreover, the output of each module has been changed into two classes to fit the data format.

We still apply 10-fold cross validation to train and evaluate TwiRole^{bi}. Table 4.9 shows the performance of TwiRole^{bi} and Liu & Ruth's method. Since there are no recall values in their paper, we are not able to compare the recall and F1 score. Based on precision and accuracy, we see that TwiRole^{bi} has better performance than their method in each role ($P_{male} = 0.901, P_{female} = 0.897$) and the overall evaluation ($Acc = 0.899$).

Table 4.9: Performance of TwiRole^{bi} and Liu & Ruths' work

Model	**Male**			**Female**			**Acc**
	R	**P**	**F1**	**R**	**P**	**F1**	
TwiRole^{bi}	0.896	**0.901**	0.898	0.901	**0.897**	0.898	**0.899**
Liu and Ruths, 2013	–	0.875	–	–	0.866	–	0.871

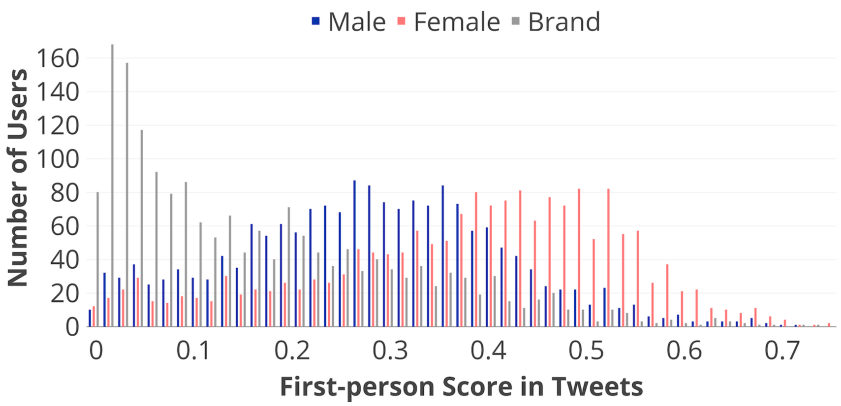
4.3.1.3 Real Twitter Environment

Besides experimental evaluation, since TwiRole is designed to be deployed for general use, as a production service, it is appropriate to evaluate its performance in settings other than with the two datasets discussed above. We applied TwiRole to find roles for the set of Twitter users who posted tweets found in multiple event-related collections (e.g., hurricanes, earthquakes, and shootings). Then, we randomly selected 100 users in each class from the predicted results, and manually checked their roles by browsing their Twitter pages. The precision score of brand users is 0.84, indicating that 84 out of 100 brand users are correctly identified by TwiRole. The precision scores of female and male users are 0.86 and 0.81, respectively.

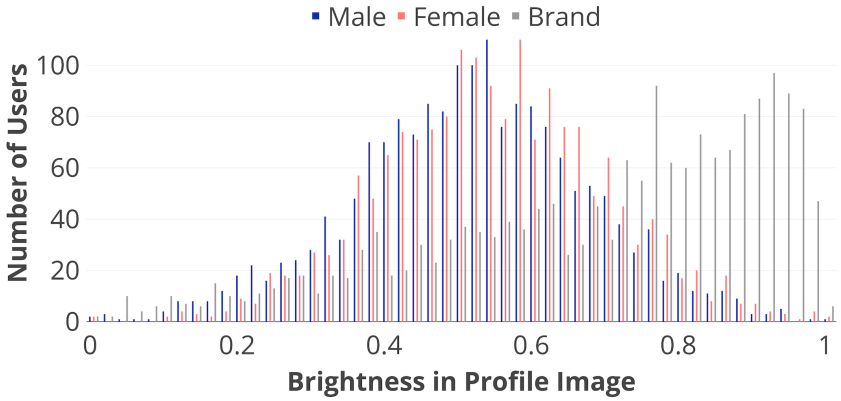
4.3.1.4 Relevant Features for User Classification

We take one fold as a sample during the training phase and draw the first-person score distribution, as shown in Figure 4.2a. Using an ANOVA test and requiring p-value < 0.05, we find that each role is statistically significantly different from any other role; see Table 4.10a. Thus, according to the sample data, brand-related users seldom use first-person words in their tweets, while female-related users are likely to mention themselves through tweets. Male-related users stay between the above two user groups.

Similarly, we investigate brightness; see the distribution in Figure 4.2b. We discover that there is a significant difference between each pair of the three roles; see Table 4.10b. Brand-related users have even brighter profile images; these may be more engaging.



(a) First-person score distribution



(b) Brightness score distribution

Figure 4.2: Relevant features discovered in TwiRole

Table 4.10: Significance tests between pairs of role-related users

(a) Significance test on first-person score

role	mean	variance		pair	p-value
brand	0.1697	0.0250		brand vs. female	5.532e-284
female	0.3788	0.0254		brand vs. male	3.287e-105
male	0.2816	0.0194		female vs. male	2.3463e-80

(b) Significance test on brightness score

role	mean	variance		pair	p-value
brand	0.6925	0.0539		brand vs. female	3.130e-102
female	0.5474	0.0232		brand vs. male	2.742e-136
male	0.5201	0.0255		female vs. male	1.6839e-7

4.3.2 Tweeting Patterns across Different Types of Disasters

4.3.2.1 Disaster Patterns

For each disaster collection, we calculate the number of tweets posted per hour during the time window of the event. Then we apply min-max normalization to convert the number of hourly tweets to a scale between 0 and 1 for inter-comparison. Different colors represent different collections in each disaster type. Figures 4.3 to 4.6 show the collection timelines across different disaster types.

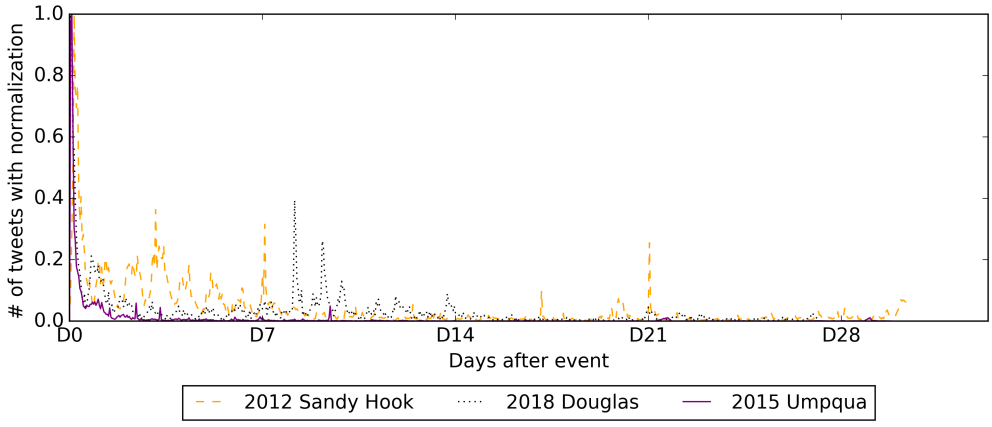


Figure 4.3: Number of tweets (scaled) per hour in school shootings

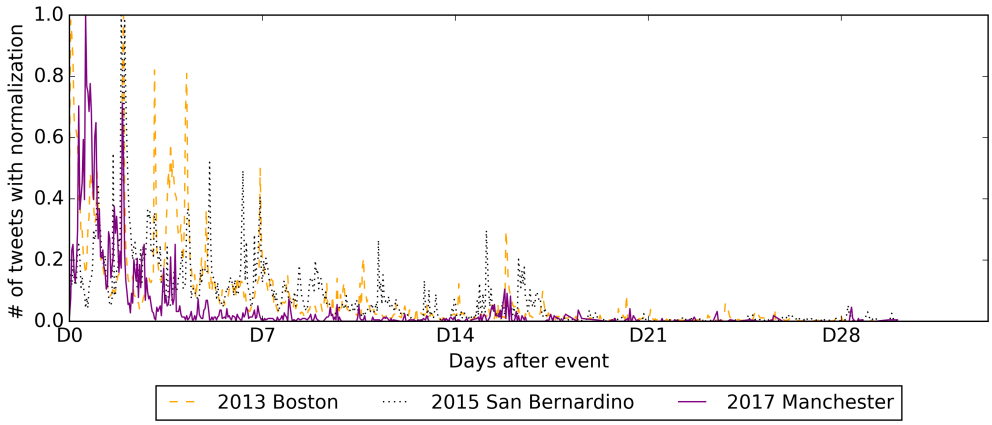


Figure 4.4: Number of tweets (scaled) per hour in bombings

The figures illustrate that each type of disaster has a pattern that differs from the others. These patterns can also help us better understand user mood changes.

- *School Shooting*: School shootings are not as complicated, among man-made disasters, as bombings. The number of hourly tweets dropped to a low level just one or two days after

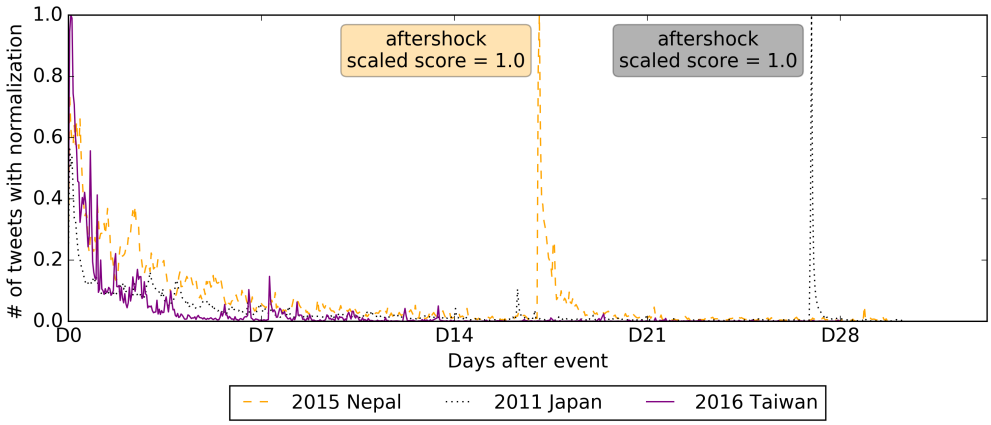


Figure 4.5: Number of tweets (scaled) per hour in earthquakes

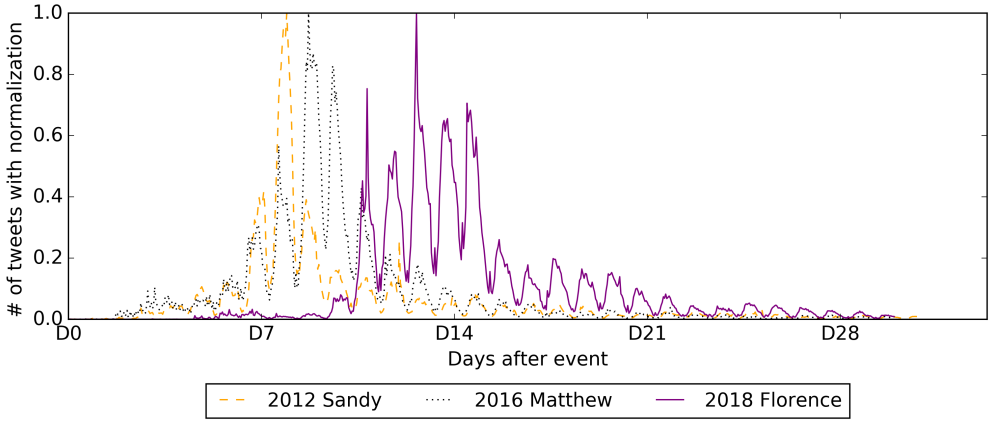


Figure 4.6: Number of tweets (scaled) per hour in hurricanes

a shooting. However, further along in the timeline, there are still some smaller peaks, e.g., corresponding to subsequent information releases or discoveries.

- *Bombing*: Each bombing timeline demonstrates highly variable numbers of tweets during the subsequent week. Tweets in significant numbers were posted even two or three days after the incident, which indicates that users paid close attention to information about the disaster.

- *Earthquake*: Twitter users posted many tweets, shown by a peak, when an earthquake occurred, and then the number of tweets gradually reduced over time. Later, when a massive aftershock occurred, another greater peak appeared, followed by a rapid reduction different from that observed after the original event.

- *Hurricane*: As a progressive natural disaster, each hurricane has a well-recognized timeline that approximates a normal distribution. We can roughly distinguish its stages and

estimate the time when it hit land and caused severe damage. Moreover, the day-night periodicity in tweet number variations is more significant than for other types of disasters.

4.3.2.2 User Distribution

We randomly select at most 10,000 users from each disaster collection; the total across all collections is 107,323 users. Some accounts have been deactivated or protected. Finally, we utilize TwiRole [76] to detect a user's role for 102,597 users (95.6% of 107,323 users). Afterward, we calculate the mean value and standard derivation of the percentages of each role across all collections. The results (Brand: 0.4737 ± 0.0093; Female: 0.2124 ± 0.0098; Male: 0.3139 ± 0.0085) indicate that the distribution of users is relatively consistent among all disasters, as shown in Figure 4.7. Brand users are the primary group who might publish information or deliver messages, while male users participated more actively than female users in our selected disasters.

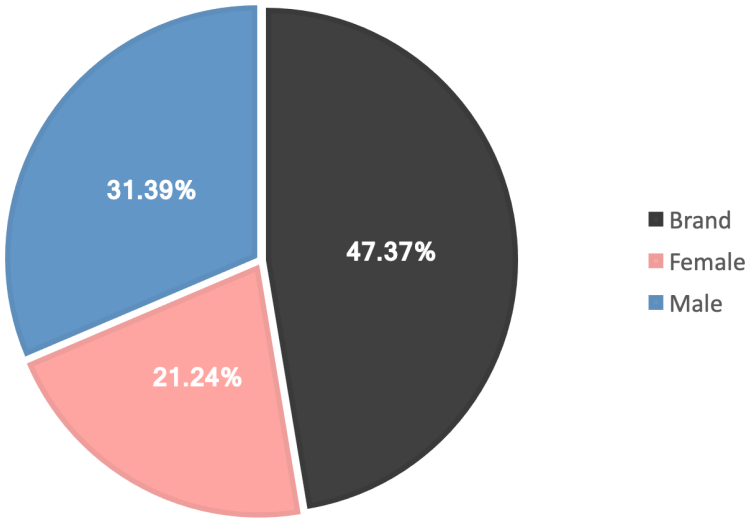


Figure 4.7: User distribution totals across all disasters

Bots are worth exploring, so we applied Botometer [107] for detection. Due to the rate limit, we detected 150 sampled users for each collection and calculated the percentage of bot users. We set the threshold to 0.5, which means a user is a bot user if the prediction score is higher than 0.5. The result (0.0694 ± 0.0270) indicates that only a small number of users are bots, which suggests they have little impact on our results.

4.3.2.3 Mood Changes in All Users

We randomly sample at most 100,000 tweets from each disaster collection and predict the mood scores (i.e., fear, sadness, and surprise) of each tweet with the pre-trained RNN clas-

sifier [24]. The scores of each mood are accumulated and divided by the total number of tweets in each collection. Figure 4.8 shows the average scores of the three moods in our collections.

Fear is the dominant feeling in eight out of the twelve disasters, which is consistent with our expectation [45]. After the Boston bombing in 2013, users posted more fear tweets and expressed their feelings with words: *fear*, *afraid*, *terror*, *deadly*, or *scared*. During Hurricane Sandy in 2012, the fear words include *fear*, *scary*, *terrifying*, *frightening*, and *threatening*.

Twitter users show more fear and surprise than sadness in ten collections. The two counter-examples are school shootings, where sadness is the significant mood. Users felt great sadness for the children and students who died or were injured in those massacres, and showed their feelings with sad words or phrases like *R.I.P*, *heart goes out*, *condolence*, *heart is broken*, and *depressed*.

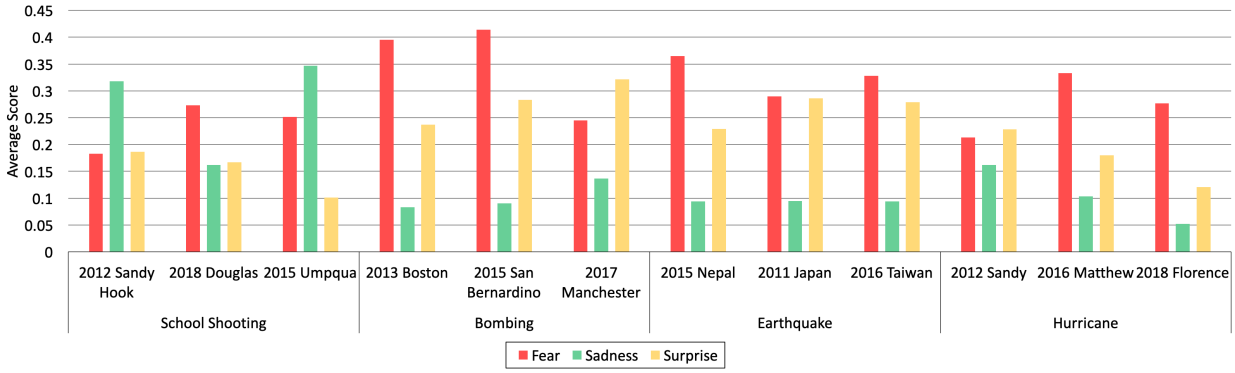
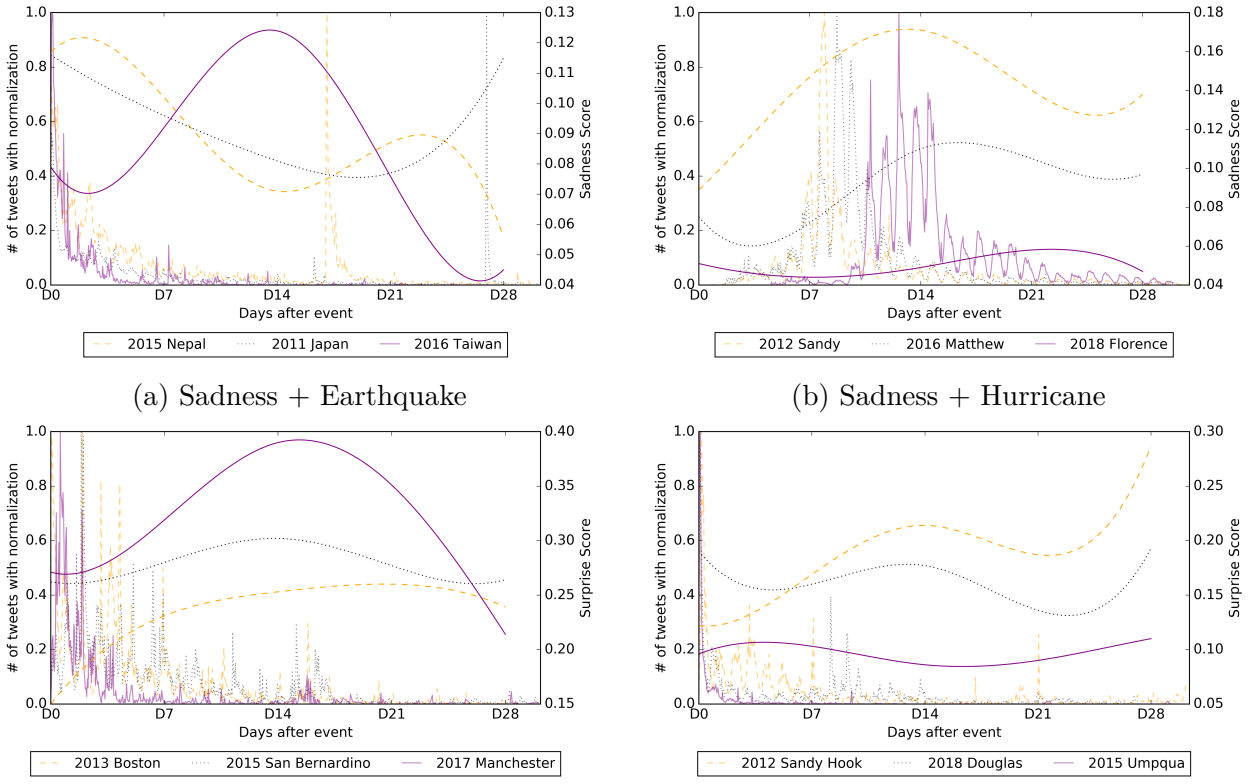


Figure 4.8: Average scores of fear, sadness, and surprise in different types of disasters

To explore further, we calculated the average score of each mood on the first day of each disaster and in every week after that. We selected two moods and two types of disaster for each mood, and showed changes over time with the corresponding tweet timelines, in Figures 4.9a through 4.9d. Figure 4.9a illustrates the sadness change in the earthquake disasters. The sadness score peaked two weeks after the main shock in Taiwan, while the aftershock played an important role in the Japan and Nepal earthquakes, leading to an increase in sadness. Figure 4.9b displays the sadness change in the selected three hurricanes. The sadness scores peaked about one week after the tweet count peaks of the three tweet timelines. Figure 4.9b also supports a quantitative comparison of the impacts of the three hurricanes. Users felt sadder in Hurricane Sandy, since it was the fourth-costliest hurricane in U.S. history, while Hurricane Florence made landfall as a weakened Category 1 hurricane, accompanied by a low sadness score.

The peaks in surprise were delayed by two or three weeks, for the bombing disasters shown in Figure 4.9c; we discuss more about mood delays later. As Figure 4.9d indicates, for school shootings, the scores of surprise were still increasing one month after the disasters, especially for the Sandy Hook Elementary School shooting (2012) and the Douglas High

School shooting (2018). After browsing through users' tweets, we noted that users posted tweets like "*The Sandy Hook shooting was a hoax?!*" and "*I find it very odd two of the sandy hook funds were created before the shooting*" after the former shooting event, while they tweeted "*Who's Behind the Real Scandal*" about the latter event.



(a) Sadness + Earthquake

(b) Sadness + Hurricane

(c) Surprise + Bombing

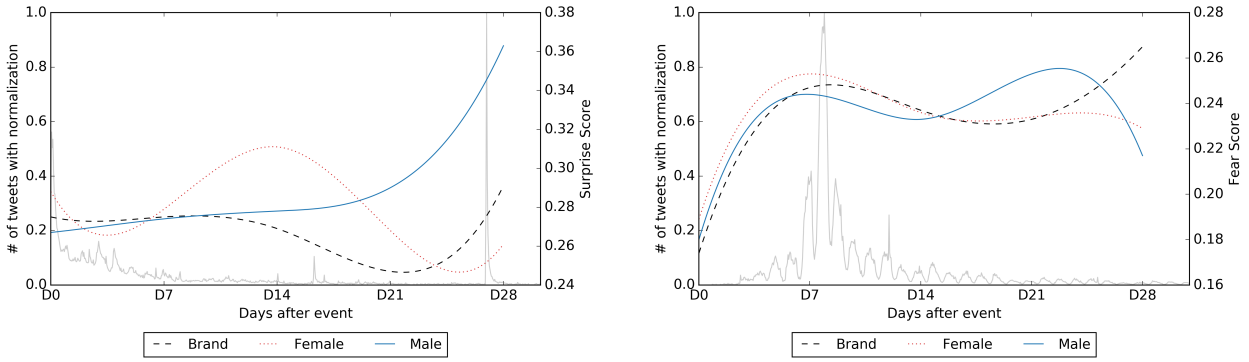
(d) Surprise + School Shooting

Figure 4.9: Mood changes (smoothed) for different types of disasters

4.3.2.4 Mood Changes for Different Roles of Users

Based on the user and mood classification results, we enriched each tweet collection with user roles and mood scores, analyzed the mood changes among different role-related users, and carried out two case studies. Each sub-figure in Figure 4.10 illustrates the mood change across brand, female, and male users in one specific disaster, with its corresponding tweet timeline. Then, for each case study, we sorted the tweets according to mood score, and presented the surprise value for the k-th (k=1, 10, 20, 30) rank tweets, for all three roles; see Table 4.11. Finally, we give the texts of some sampled tweets in Tables 4.12 and 4.13.

- *Case Study 1*: Surprise change in Japan Earthquake
  Figure 4.10a shows that the surprise score of male users had a significant increase after



(a) Surprise + 2011 Japan Earthquake (b) Fear + 2012 Hurricane Sandy

Figure 4.10: Mood changes (smoothed) among different roles of users in disasters

the aftershock, compared with the scores of brand and female users. The ranked k-th surprise scores also show male users were more surprised than brand users, while female users stayed relatively calm during that period.

Table 4.11: Ranked k-th surprise score comparison in two disasters

(a) Japan Earthquake

	Surprise Score		
k	**Brand**	**Female**	**Male**
1	0.993	0.938	0.996
10	0.899	0.775	0.959
20	0.866	0.664	0.925
30	0.818	0.570	0.880

(b) Hurricane Sandy

	Fear Score		
k	**Brand**	**Female**	**Male**
1	0.999	0.991	0.996
10	0.875	0.720	0.848
20	0.789	0.537	0.733
30	0.728	0.441	0.635

- *Case Study 2*: Fear change in Hurricane Sandy
  Figure 4.10b shows the divergence of fear scores among different roles that appeared during the last week within the one-month time window. Here, brand users expressed more fear, and the fear scores of female and male users decreased at the same time. From the sample tweets, we noticed that brand users posted fear-related tweets regarding the aftermath of the destructive hurricane.

4.3.3 Tweeting Patterns in Hurricane Dorian

4.3.3.1 Data Post-processing

We discovered that the timeline of tweets about a hurricane approximates a normal distribution [73], where there are few tweets posted at the start and end of the entire disaster.

Table 4.12: Sample tweets from male users in Japan Earthquake

Sample tweets from male users with high surprise scores
japan had another earthquake? #idontbelieveyou
Japan just had another earthquake. WHHAAAAAAAAA?!?
Mother nature #idontbelieveyou another earthquake Japan?
Another earthquake in Japan! #wow
There was another earthquake in Japan?!?! OMG #PRAYFORJAPAN

Table 4.13: Sample tweets from brand users in Hurricane Sandy

Sample tweets from brand users with high fear scores
Haiti fears food crisis in Hurricane #Sandy's aftermath
Hurricane Sandy 2012: What a nightmare!! Lack of power, gas rationing...
Gas Shortage In New York After Hurricane Sandy Caused By Poor Policy
Terrifying Note Left Behind By a New Jersey Man In Hurricane #Sandy...
#Forbes Obviously the destruction caused by Hurricane Sandy posed a risk...

Focusing on the principal part of the disaster, we chose a subset of the tweet collection for further analysis of disaster response across different users. We utilized a Gaussian distribution function to fit the hurricane timeline based on non-linear least squares. Figure 4.11 shows the collection timeline and its corresponding Gaussian fit curve. The mean μ and standard deviation σ of the curve are 226.48 (hours) and 80.73 (hours), respectively. We used $\mu \pm 2\sigma$ (pink box in Figure 4.11) for data selection. The dataset after filtering contains 521,886 tweets posted by 207,894 users, from 08/27/2019 to 09/09/2019.

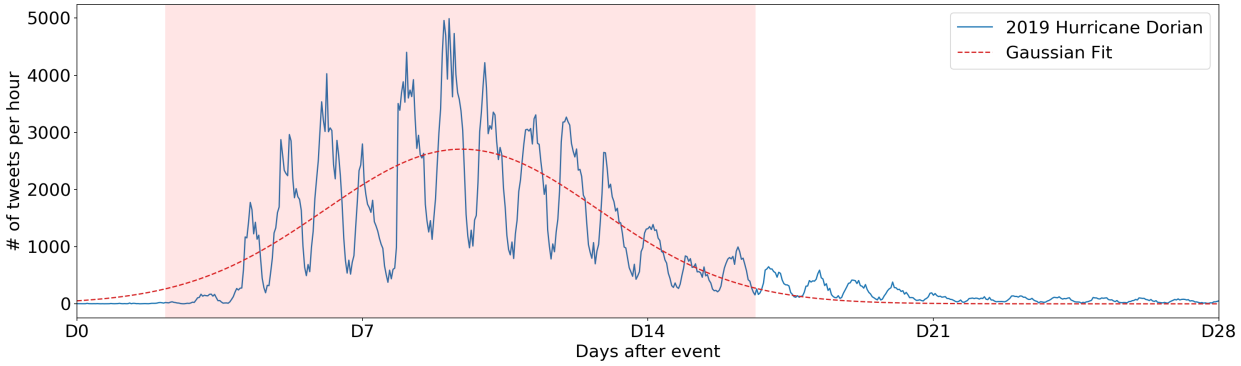


Figure 4.11: Number of tweets posted per hour during Hurricane Dorian

4.3.3.2 User Distribution

As was discussed in Section 4.3.2, we applied TwiRole [76] to classify users by role (i.e., brand, female, and male). Due to time constraints related to big data analysis, we randomly

selected 90,000 tweets, posted by 56,528 users, from the raw collection. TwiRole successfully analyzed 48,753 users (86.2% of 56,528 users). Considering the similar behavior of individual users (i.e., female and male), we further merged both female and male users as individual users for subsequent analysis.

Since we used TwiRole in our specific disaster scenario, we reevaluated its performance to make sure our analysis results can accurately describe the different patterns between brand and individual users. We randomly selected 150 users from both brand and individual groups for manual inspection. We examined the roles of users by browsing their Twitter pages, and calculated the precision score of each group; see Table 4.14. The result shows the precision score of TwiRole is 95.3% for brand users while the value is 93.3% for individual users.

Table 4.14: Precision of TwiRole for brand and individual users

		Total	**Manual Labeled**		
			Brand	**Individual**	**Precision**
Predicted	**Brand**	150	143	7	95.3%
	Individual	150	10	140	93.3%

We also employed Botometer [107] to detect the sampled users above (i.e., 150 users per group). We calculated the percentage of bot users for each group; see Table 4.15. The result indicates that brand users have relatively more bot users than individual users, but only a small percentage of users are bots, leading to little impact during our analysis. We extracted tweets with user labels during the above time range as our final dataset, including 267,842 tweets posted by 45,237 users.

Table 4.15: Percentage of bot users in brand and individual users

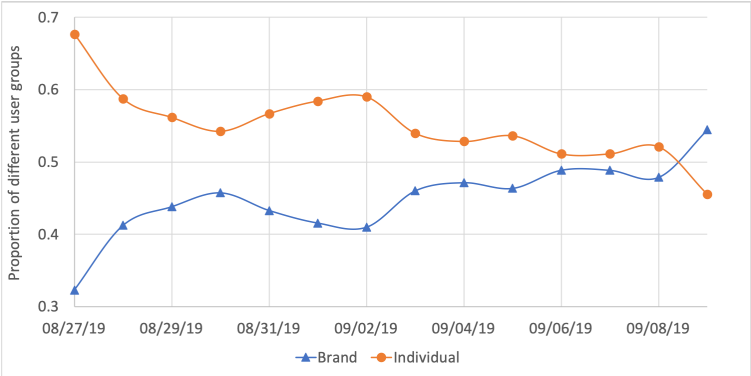
		Total	**Botometer**	
			Bot	**Percentage**
Predicted	**Brand**	150	10	6.67%
	Individual	150	3	2.00%

4.3.3.3 Basic Analysis

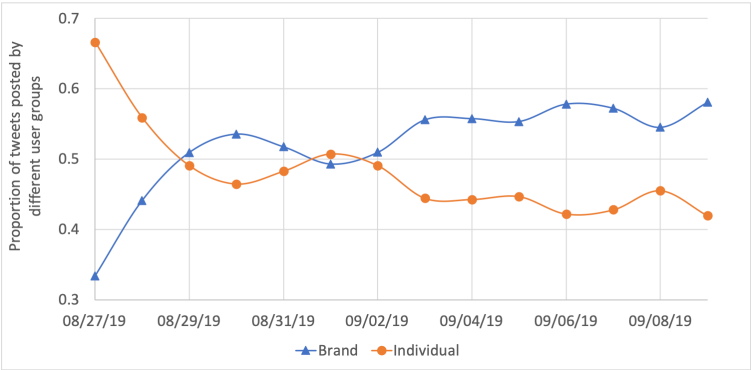
We conducted a basic analysis of the tweet posting patterns between brand and individual users during the disaster. Among the 267,842 tweets, there are 143,346 (53.5%) tweets posted by 16,835 (37.2%) brand users and 124,496 (46.5%) tweets posted by 28,402 (62.8%) individual users. We calculated the average number of tweets posted by different users within the entire time window. Each brand user posted 8.5 tweets on average, which is about two times greater than individual users (4.38 tweets per user).

Then we analyzed the user participation and proportion of tweets posted by different user groups at the day level. We counted the number of unique brand users and individual users

per day and calculated the proportion of each group with the development of the disaster; see Figure 4.12a. Similarly, we computed the proportion of tweets posted by different user groups every day; see Figure 4.12b. The two figures show that as the disaster unfolded, brand users participated more actively than individual users. When the disaster formed, brand users constituted only 30% of the entire users, but the proportion reached about 55% two weeks later. Meanwhile, the proportion of tweets posted by brand users also increased in general, having a similar trend as the proportion of brand users.



(a) Proportion of different user groups



(b) Proportion of tweets posted by different user groups

Figure 4.12: Proportion of different user groups (top) and tweets posted by different user groups (bottom) per day

Since both the proportion of brand users and tweets posted by them increased at the same time, we further investigated the average number of tweets posted by different user groups per day; see Figure 4.13. Both brand and individual users posted 1.2 tweets on average on August 27. Afterward, the daily average number increased for both user groups, but it is clear that brand users posted more tweets related to Hurricane Dorian than individual users. When the hurricane hit Florida on September 2, each brand user posted more than 3 tweets while each individual user posted about 2 tweets per day. The difference between the two groups kept up through when the hurricane dissipated after September 9. It seems

that brand users would most likely share information during the disaster. Table 4.16 lists the top 10 brand users with their descriptions and the total number of tweets posted. Based on the descriptions in their Twitter pages, we can consider them as falling into two groups, meteorological agencies and news sites – which is consistent with our expectations.

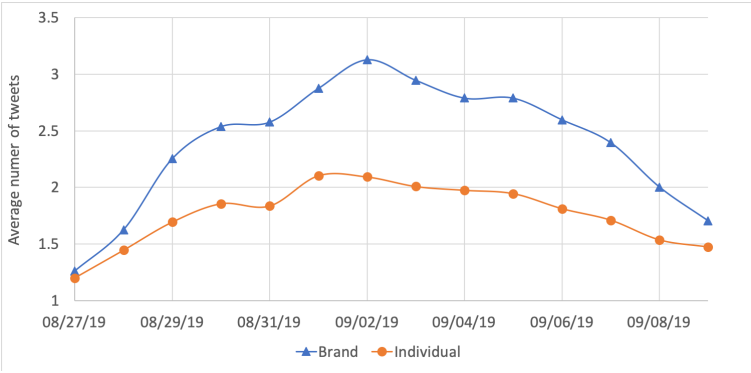


Figure 4.13: Average number of tweets posted by different user groups per day

Table 4.16: Top 10 brand users with their descriptions and the total number of tweets

Twitter Account	Description	# of Tweets
EcoInternetDrGB	Climate	887
TheChestnutPost	News	846
poandpo	News	840
TimMelino	Meteorologist	733
gridpointwx	Forecasting	733
PlanaWeather	Forecasting	496
wsbtv	News	464
NWS_LCH	Alerts	461
WFTV	News	409
NewsNetNews	News	375

4.3.3.4 Emotion Analysis

We reused the RNN-based emotion detection model [24] to predict Ekman's emotions [31].

First, for each user group, we applied the model to generate emotion scores of all tweets and calculated the average score of each mood. If users express a strong emotion (e.g., fear, surprise) through tweets, the average score of that mood should be high. Figure 4.14 depicts the score distribution of emotions between brand and individual users. For both groups, users posted more tweets with the joy emotion than with the other emotions, and the average mood score is about 0.40. Fear is the second dominant mood, followed by surprise, sadness, anger, and disgust. Because the average scores of anger and disgust are

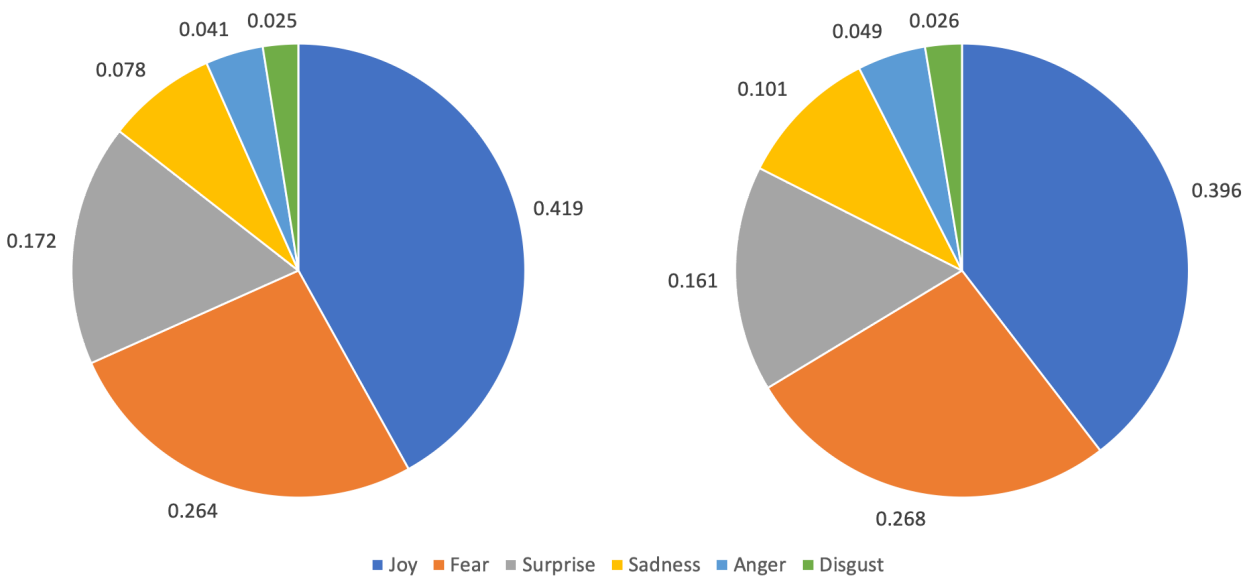


Figure 4.14: Average scores of emotions between brand (left) and individual users (right)

less than 0.05, indicating only a small number of tweets related to the two emotions, we filtered them out and chose the top four emotions for further analysis.

Second, to better understand the emotional tweets, we set a threshold $\theta = 0.7$ and considered tweets with scores greater than the threshold as good representatives for each emotion. Some typical words or phrases are extracted from those representative tweets and listed across different emotions and user groups in Table 4.17. We noticed that both brand and individual users shared some emotional words in their tweets. For example, they mentioned *wonderful* and *thank* in the joy emotion. *Frighten*, *fear*, and *terrifying* are used by both user groups to express the fear emotion while *sad*, *saddened*, and *r.i.p.* are common expressions related to sadness. Further, we also found that users were surprised by different events. Brand users were mainly surprised by the impact of the hurricane, such as *special track* and *100mph+ winds*. Meanwhile, individual users showed their surprise at personal events (e.g., *late birthday gift*, *hurricane party*) caused by the disaster.

Third, Figure 4.14 shows that brand users posted more joy-related and surprise-related tweets than individual users. In contrast, individual users seemed sadder than brand users. However, that is an overall description during the selected time range. In this case, we calculated the average scores of the four major emotions per day to verify whether there are some patterns of emotions across users. Figures 4.15a through 4.15d show the average scores of different emotions between brand and individual users per day. The overall description can also be applied at the day level. Later, we carried out a paired t-test (significance level $\alpha = 0.05$) to statistically examine the significant differences between the two user groups. Table 4.18 shows the results of the paired t-test, including the mean, variance, and p-values of emotion scores in each group. Based on the one-sided test, the p-values in the joy, surprise,

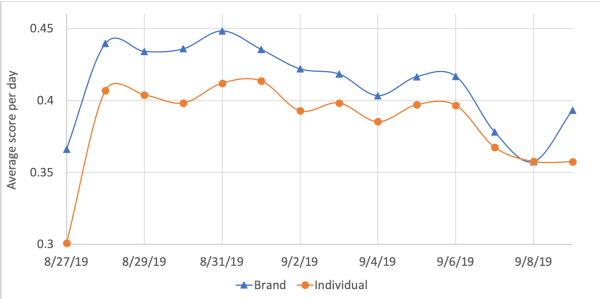
Table 4.17: Four major emotions and their corresponding typical words or phrases

Emotion	User Group	Typical Words / Phrases (lowercase)
Joy	Brand	*beautiful sunrise, nature's beauty, smiles evening universal, wonderful pilots, assist answering, thanks visiting, enjoy chatting, ocean rescue recovery, join charity, everyone donate*
	Individual	*great day, love leave, community safe, prayer, funny video, happy labor day, feeling better, prayed three rosaries, god bless you, amen, thank lord, stay safe, wonderful evening*
Fear	Brand	*feared bahamas, frightened, florida fears, terrifying, death toll, nightmare, tortured path, hyped fear mongering hurricanes, scared, life-threatening, deadly*
	Individual	*traumatized, terrified, drive fear, terrifying, big fear fear-mongering, scared, cape fear, threatened, nuclear coffin, frightening, terror, death toll*
Surprise	Brand	*what's coming, birth strong, special track, longer wait time, disney surprise, last minute evacuation, 100mph+ winds, protect launch infrastructure*
	Individual	*late brithday gift, i'm paper towels, surprise trip disney, people freaking out, get one free, mystery shipwreck, forgot anniversary, hurricane party, never know*
Sadness	Brand	*sad, r.i.p., saddened, queen left, heart broken, lost loved ones, dead, upsetting, miss, heartbreaking, empty disney, ghost town, sadness*
	Individual	*sad, queen left, saddened, precious children, r.i.p, sweet children, condolences, lost, dead, missing, deceased, tragedy, heart goes, saddening, helpless, drowned*

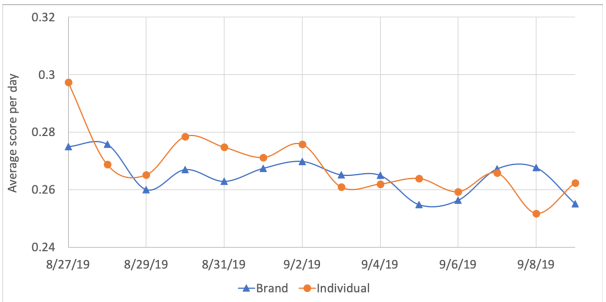
and sadness moods are much smaller than α, indicating the differences (i.e., greater or less) between both groups. Regarding the fear emotion, the p-value (0.097) is higher than 0.05, showing there is no significant difference between brand and individual users. Additionally, we also measured the Pearson's correlation across the four emotions and user groups. The r-values of the joy, surprise, and sadness emotions are above 0.85, which means the average scores in both brand and individual users are highly correlated. Also, we can consider the two user groups are moderately correlated in the fear mood since the r-value is about 0.5.

4.3.3.5 Text Analysis

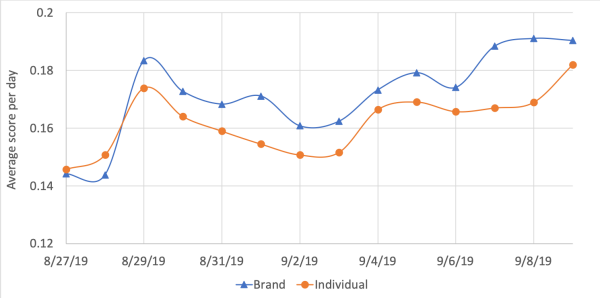
We first extracted both hashtags (e.g., *#hurricanedorian*) and words (e.g., *hurricane, dorian*) from the entire tweet corpus and counted their frequencies. Figure 4.16 shows the top 20 hashtags and words posted by both brand and individual users. The frequency of the top hashtag *#hurricanedorian* is 83,313, which is almost the sum of the frequencies of the other hashtags, is compared with the gradually decreased distribution of top words. The top hashtags include hurricane (e.g., *#hurricanedorian, #dorian*), weather (e.g., *#weather,*



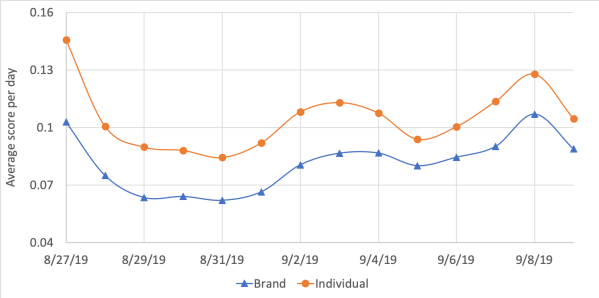
(a) Average score of the joy emotion



(b) Average score of the fear emotion



(c) Average score of the surprise emotion



(d) Average score of the sadness emotion

Figure 4.15: Average scores of emotions between brand and individual users per day

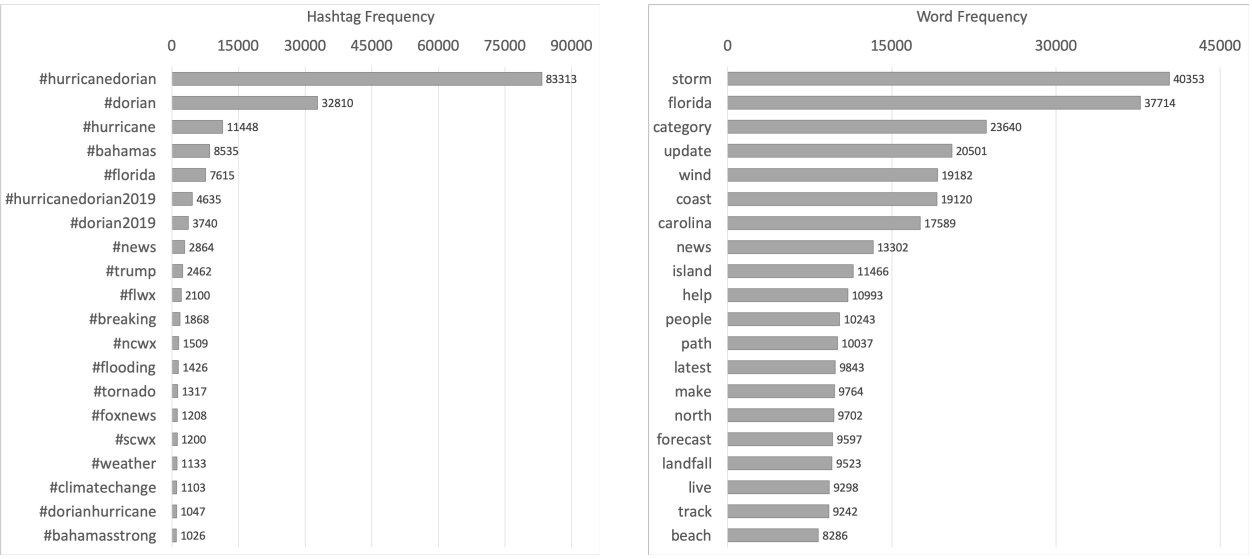
#flwx, #scwx), location (e.g., *#bahamas, #florida*), and prayers (e.g., *#bahamasstrong*). Compared with those hashtags, the top words mainly focus on the basic information (e.g., *category, path, landfall, track*) or impact (e.g., *storm, wind, coast, beach*) of the disaster.

Figure 4.16 shows the frequent hashtags and words posted by all users. Meanwhile, we also focused on the hashtags and words posted by different user groups. In other words, we aimed to seek for hashtags and words that have great difference of occurrence between brand and individual users. To this end, we counted the frequency of hashtags/words for both brand and individual users, separately. For each hashtag/word, we subtracted its frequency in one group from the one in the other group and sorted all hashtags/words by their subtraction values in each group. Then we selected the top hashtags and words from each group, as shown in Figure 4.17, which can describe the difference between brand and individual users.

Regarding hashtags, brand users likely posted simpler hashtags, which usually contain one specific word such as *#hurricane, #news, #dorian, #flooding*. Most hashtags are common and could be reused next time; they are easy and convenient for management. Individual users preferred to use compound words as hashtags like *#hurricanedorian* and *#hurricanedorian2019*, which seem to be used once and have rich and targeted information. Particularly, *#disasterassistteam*, *#dat*, and *#prayforthebahamas* are the popular hashtags from individual users, showing their concerns about hurricane rescue and the hurricane-affected country.

Table 4.18: Paired t-test and Pearson correlation coefficient across user groups and emotions

User Group	Joy				Fear			
	Mean	Variance	p-value	r-value	Mean	Variance	p-value	r-value
Brand	0.4117	8.12e-4	8.80e-6	0.8660	0.2650	4.38e-5	0.097	0.5202
Individual	0.3848	9.21e-4			0.2684	1.21e-4		
User Group	**Surprise**				**Sadness**			
	Mean	Variance	p-value	r-value	Mean	Variance	p-value	r-value
Brand	0.1716	2.27e-4	2.06e-4	0.8856	0.0813	1.98e-4	5.55e-9	0.9091
Individual	0.1621	1.07e-4			0.1051	2.77e-4		



(a) Hashtags and frequencies

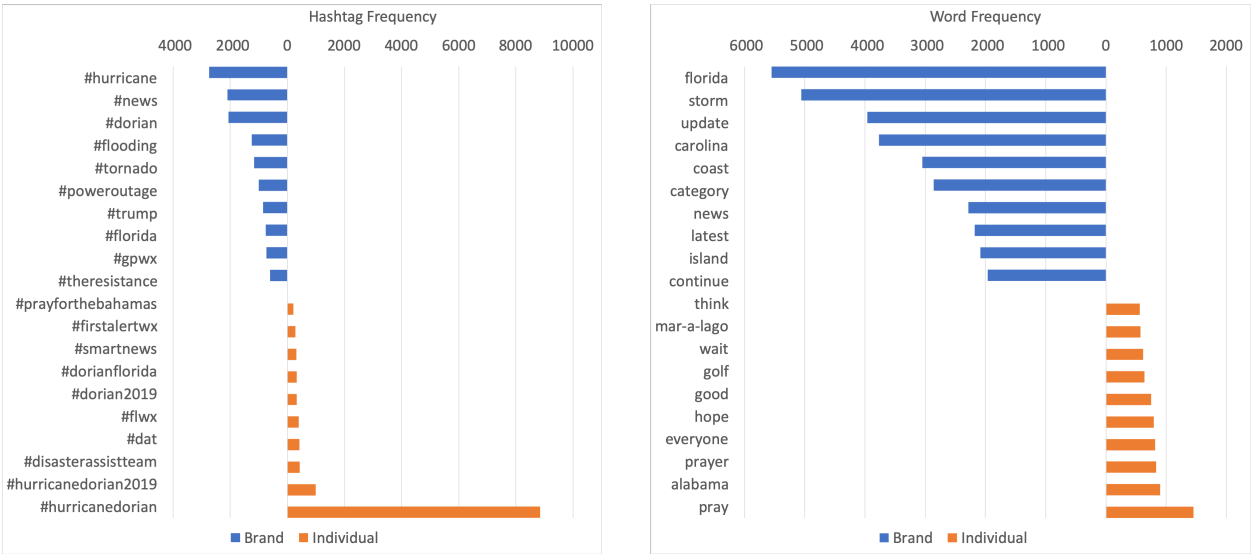
(b) Words and frequencies

Figure 4.16: Top 20 hashtags (left) and words (right) posted by all users

Regarding words, brand users applied hurricane-related words to publish the disaster information, such as *florida*, *storm*, *update*, and *carolina*. Different from brand users, individual users were more concentrated on personal feelings (e.g., *pray*, *hope*, *think*) and specific events (e.g., *alabama*, *golf*, and *mar-a-lago*).

For further analysis, we manually select three topics and their corresponding words during Hurricane Dorian; see Table 4.19. Specially, the Trump-related words describe two events during Hurricane Dorian. One is that President Trump went golfing during the disaster, and the other is he had erroneously stated that the hurricane threatened Alabama.

We counted the number of tweets containing those words in each topic for both brand and individual users at the day level, and divided it by the total number of tweets posted by the two types of users for normalization. Figure 4.18 shows the percentages of tweets related to the three topics posted by brand and individual users. From the figure we know that the percentages of casualty-related tweets for both groups increased sharply after the hurricane



(a) Top hashtags by frequency

(b) Top words by frequency

Figure 4.17: Top hashtags (left) and words (right) by frequency, showing |brand - individual| values

Table 4.19: Three selected topics and their corresponding words

Topic Type	**Words (lowercase)**
Casualty-related	*victim, dead, death, kill, die, loss, drown, missing*
Prayer-related	*pray, prayer, r.i.p, hope, god, wish, brace, crave, appeal*
Trump-related	*golf, mar-a-lago, alabama*

hit Florida on September 2. Brand users posted relatively more casualty-related tweets than individual users. The peak value appears on September 8, when over 10% of tweets posted by brand users contained words in the casualty topic. Regarding the prayer topic, the two peaks appeared on August 28 and August 31 for individual and brand users, separately. Different from the posting patterns about casualties, individual users likely posted prayer-related tweets (as opposed to brand users), and both trends gradually decreased after September 2. In addition, the percentages of tweets are highly correlated with the two events that mainly took place on September 1 and September 4 for the Alabama controversy, and September 2 for the golf course. Further, more tweets were posted by individual users regarding the Alabama controversy.

4.3.4 Code Release and Online Service

We shared our source code with pre-trained models on GitHub (https://github.com/liuqingli/TwiRole) [76] with developers so that they can fit them into other applications;

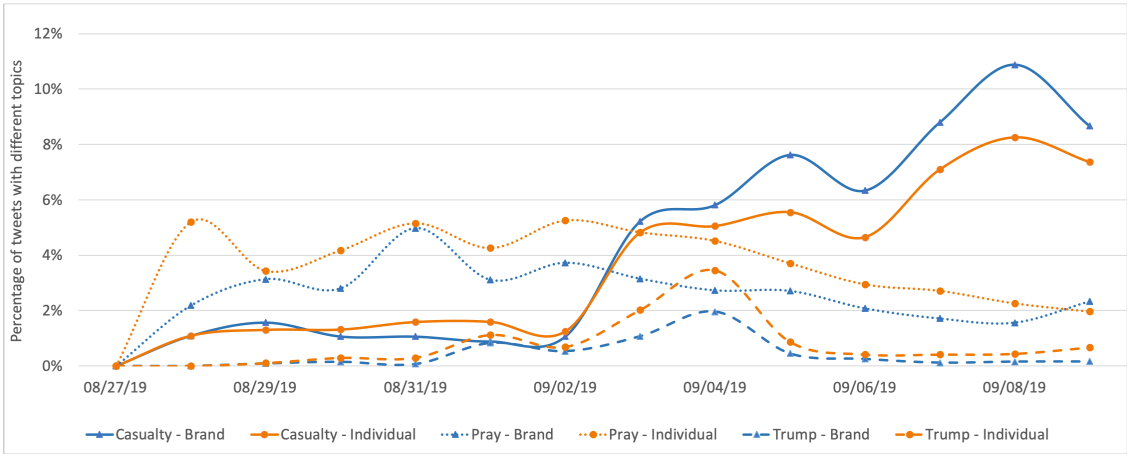


Figure 4.18: Percentages of tweets related to the three topics posted by brand and individual users

see Figure 4.19.

We also published a compute capsule (https://codeocean.com/capsule/9584745/tree/v4) on Code Ocean; see Figure 4.20. The capsule contains not only the code, but everything else that the code needs in order to run, namely 1) pre-trained classifiers, 2) three Twitter accounts' information for testing, and 3) a specification of the computational environment, including the operating system, packages, and dependent libraries. It also has been verified to be computationally reproducible, generating the same prediction results for the three Twitter users; see Figure 4.21.

As a visualization part of our system, we further developed a web application that hosts our TwiRole model, which is easy to use and navigate, providing an online user classification service for a wide range of users.

The website is made up of an HTML page (http://vis.dlib.vt.edu:3001) using React on the front-end. The back-end loads TwiRole and is made using Django to render the HTML page, host images, and other resources, and expose a GraphQL API. The front-end makes AJAX calls to the GraphQL API which obtains the information that will be displayed on the website. The web interface is aimed at being simple to manage and update, for those with experience in web development, specifically involving Django, React, and GraphQL.

The only input provided by users is the screen name of a Twitter account. After crawling relevant information, the application shows the account's screenname, along with her/his/its profile image. The role prediction results of the three sub-modules are shown through stacked bar charts while the final prediction result is presented through a donut chart. Figure 4.22 shows a screenshot of the prediction results for two Twitter accounts (i.e., CNN and edwardafox).

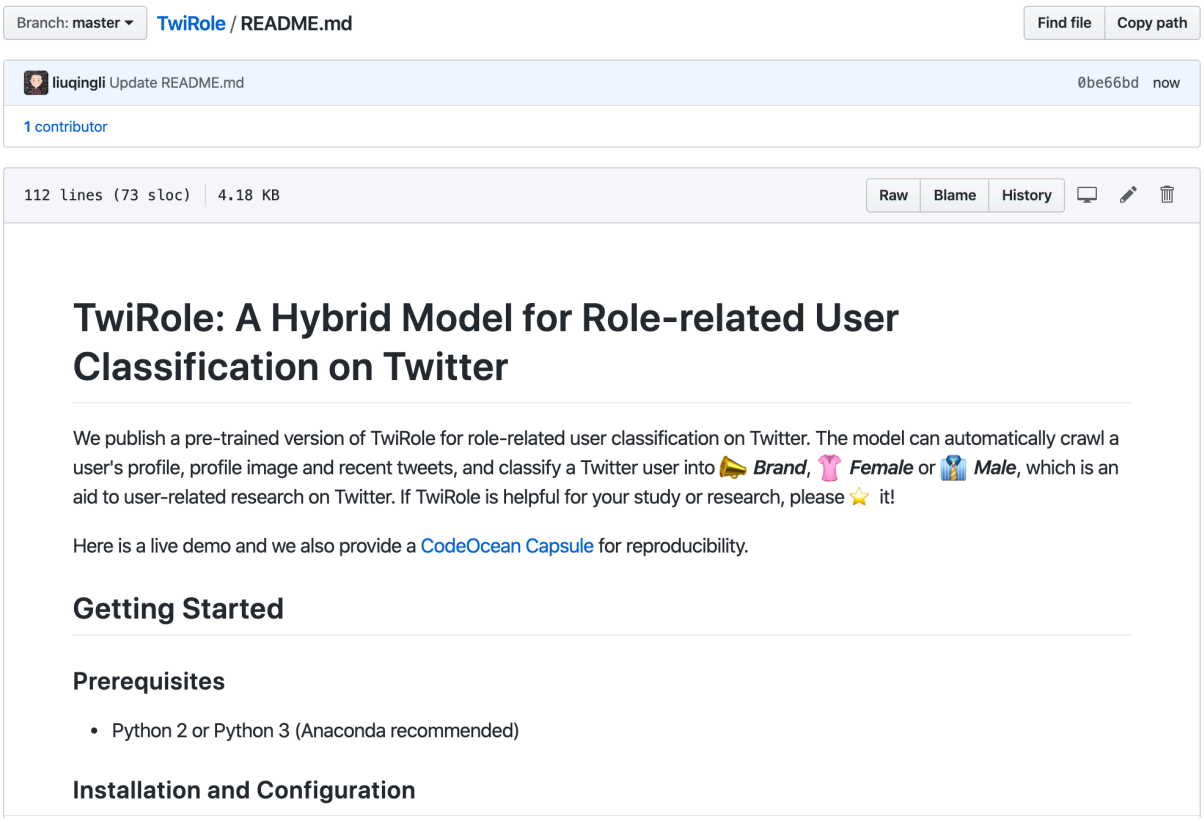


Figure 4.19: A screenshot of the GitHub page of TwiRole

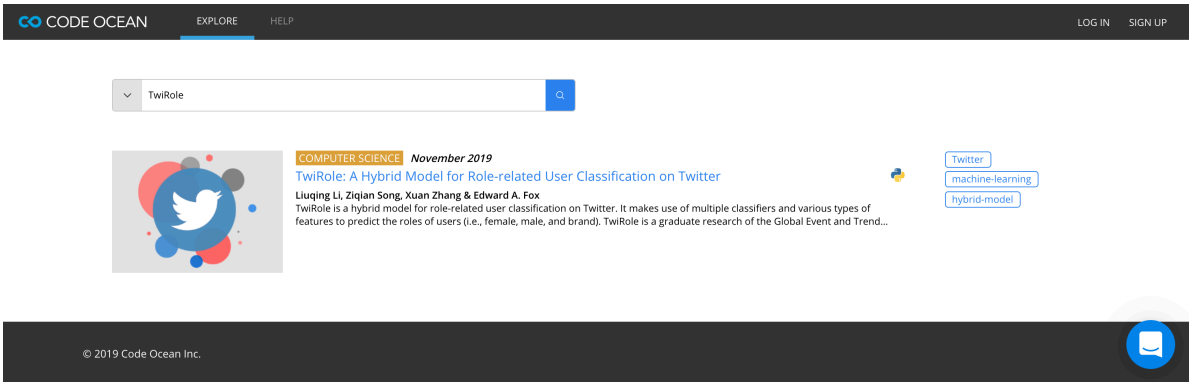


Figure 4.20: The capsule of TwiRole on Code Ocean

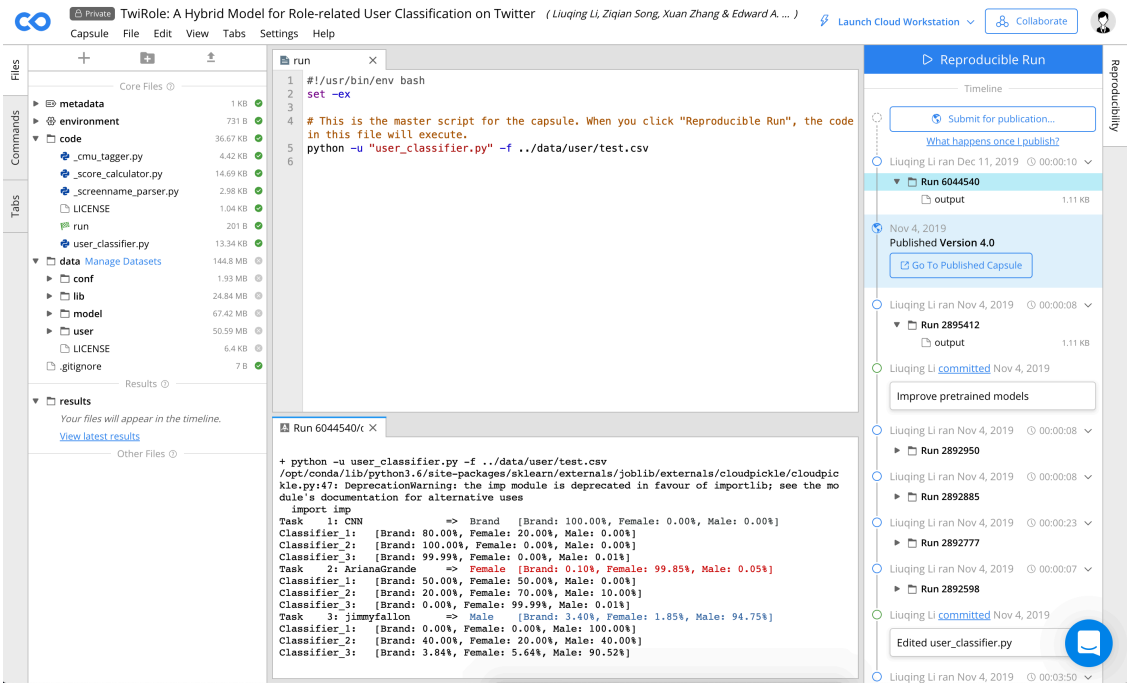


Figure 4.21: Three reproducible results predicted by TwiRole on Code Ocean

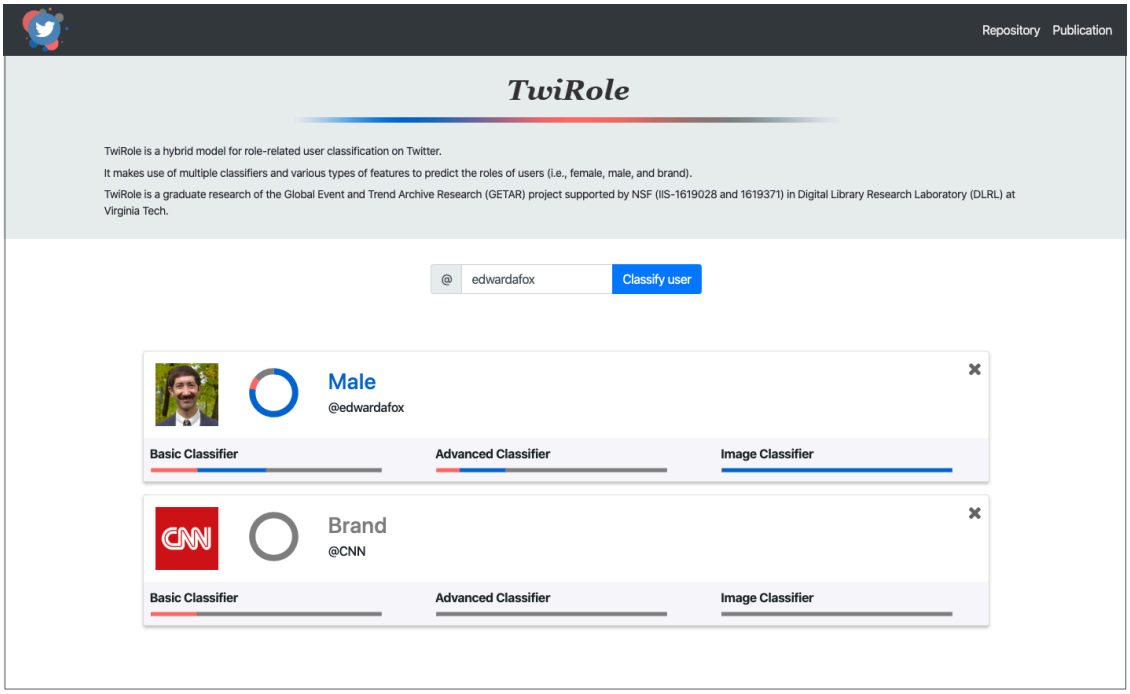


Figure 4.22: Online prediction results of two selected Twitter accounts

Chapter 5

Tweet-guided Multi-Document Summarization

The research question for this chapter is: *How can multi-document summaries of the event-related collections be improved by using tweets associated with those events?* A sub-system for tweet-guided multi-document summarization (TMDS) is proposed to solve this question. As part of the data, tweets play an essential role and can effectively guide the summarization task. The reason is that a webpage is closely interrelated with its corresponding tweets, and tweeters tend to mention details (e.g., persons, places, and numbers) of an event. Such information can be considered as prior knowledge to support the summarization task. Table 5.1 shows an example of the tweet-guided summarization. A good summary considers both the relevant webpages and concerns from tweets, as well as the timeline that can be inferred from the tweets, while a bad one only focuses on the webpages and ignores some key elements only found in the tweets.

Table 5.1: Example of tweet-guided summarization

<table>
<tr><th>Webpages</th><th>Tweets</th></tr>
<tr><td rowspan="2">mass shootings: there were 372 mass shootings in the us in 2015, killing 475 people and wounding 1,870, according to the mass shooting tracker, which catalogues such incidents. a mass shooting is defined as a single shooting incident which kills or ...</td><td>rt @xxx: connecticut elementary school shooting...</td></tr>
<tr><td>r.i.p to all the victims of the connecticut elementary school shooting</td></tr>
<tr><td>...</td><td>...</td></tr>
<tr><td rowspan="2">on december 14, 2012, the sandy hook elementary school in the quiet village of sandy hook in newton, connecticut was shattered when a gunman mindlessly open fired and took the lives of 20 children aged between 6 and 7 and 6 adults. the gunman, adam lanza, had access to the guns ...</td><td>watching about a school shooting in newton connecticut kills 20 students and 6 adults</td></tr>
<tr><td>rt @xxx: connecticut police released the names of 20 children and 6 staff members</td></tr>
<tr><th colspan="2">Good Summary</th></tr>
<tr><td colspan="2">the connecticut school shooting happened on december 14, 2012. adam lanza took the lives of 20 children and 6 adults ...</td></tr>
<tr><th colspan="2">Bad Summary</th></tr>
<tr><td colspan="2">there were 372 mass shootings in the us in 2015. a mass shooting is defined as a single shooting ...</td></tr>
</table>

At the knowledge level, with the guidance of tweets, we focus on extracting relevant sentences from webpages and reorganizing those sentences through entity-based and topic-based

scoring methods for summary generation.

5.1 Approach

Figure 5.1 shows the architecture for our proposed TMDS model, which consists of four main components:

- a relevant sentence selection module that selects sentences from webpages, which have high contextual similarity with their corresponding tweets;
- an entity-based scoring module that calculates the entity-based score of each sentence from three aspects;
- a topic-based scoring module that learns the different topics of relevant sentences and computes the topic-score of each sentence through topics; and
- an integrated ranking method to rank all sentences based on multiple factors, and generate summaries for event-related collections.

5.1.1 Relevant Sentence Selection

As discussed in Chapter 3, we can retrieve relevant webpages and tweets, when given an event-related tweet collection. In most cases, tweeters may post various tweets when focusing on a particular webpage. There is a one-to-many mapping between webpage and tweets. Though most webpages and tweets are relevant to a specific event, we find that tweets are closer to the event itself, and webpages may describe other information. Also, a webpage being relevant does not mean all its paragraphs or sentences are relevant. Therefore, we use tweets to guide the relevant sentence selection and filter out those significantly non-relevant sentences that are not helpful for our summarization task.

Figure 5.2 gives a data flow diagram of the relevant sentence selection process. Given a set of webpages and their corresponding tweet lists, we first preprocess each pair (i.e., a webpage and its tweet list) with splitting and cleaning. Specifically, we apply jusText [94] to split each webpage into paragraphs, and further divide each paragraph into multiple sentences. Since the length of a tweet is usually short, we consider each tweet as a sentence to guide the selection process. Regarding cleaning, at the beginning we refine sentences in webpages with a rule-based filter to improve the readability of our final summary. Table 5.2 lists the rules in our customized filter, and some examples. For tweets, we remove all mentions (e.g., @CNN), hashtags (e.g., #HurricaneSandy), URLs, and retweets tags (RT). Because most retweets have the same tweet content, we also eliminate duplicated tweets to ensure that each tweet list only contains unique tweets.

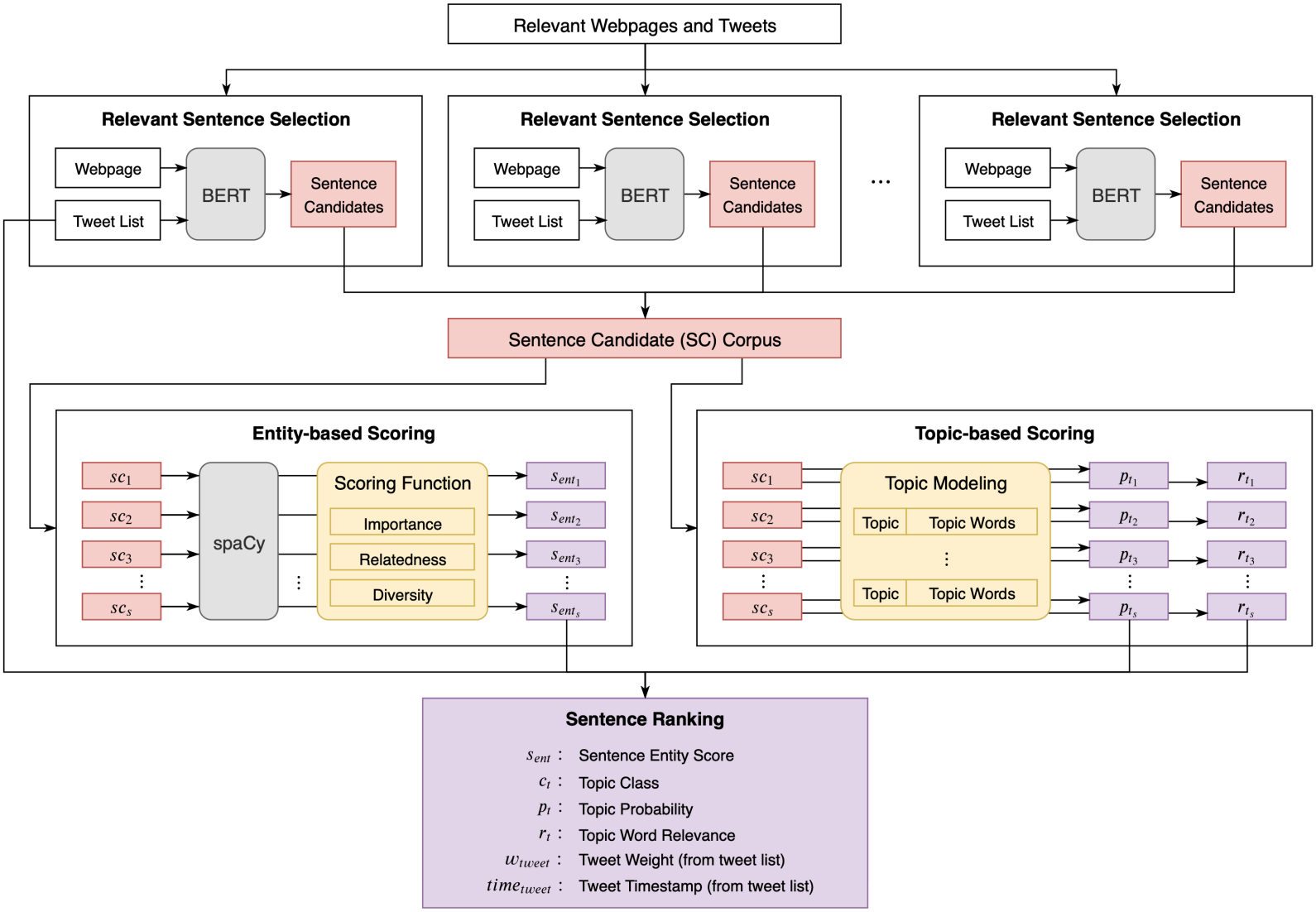


Figure 5.1: System architecture for the TMDS model

Table 5.2: Customized rule-based filter and examples

Rule	Example
news header	(reuters) -; (cnn) -; (nbc) -
URL-like string	telegraph.co.uk; members.pic.twitter.com/kyal0rradw
non-alphabet start	[] you think you've seen the worst that can happen
webpage menu	latest news >newtown students ... >nra breaks silence ... >...
special tags	you can add location information; all rights reserved

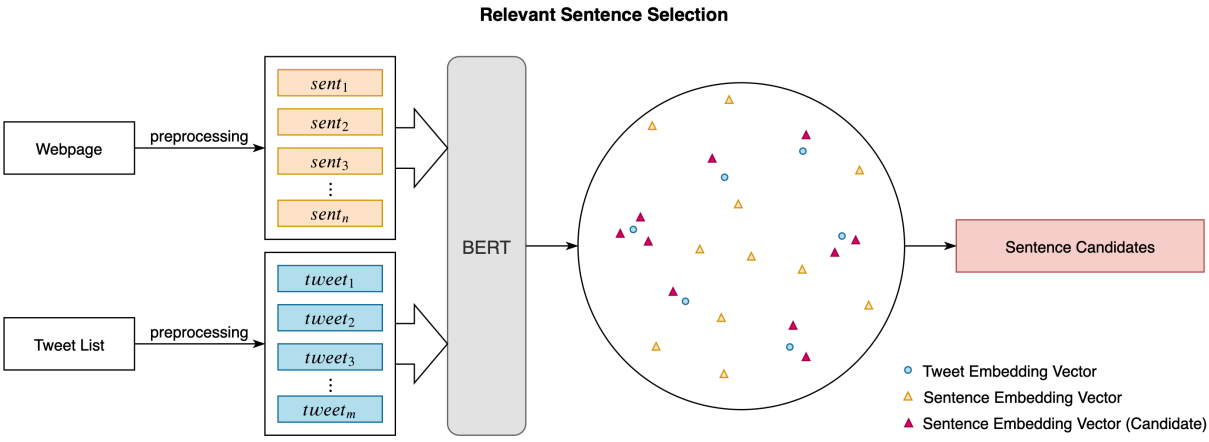


Figure 5.2: Data flow diagram of relevant sentence selection

Then, we introduce BERT [30], a pre-trained contextual representation model, to generate embeddings for both sentences and tweets. The model has been trained on a large text corpus, which is well adapted for our scenario. Figure 5.3 shows the architecture of the BERT model. The embedding of each token in the input layer has three parts: token embedding, sentence embedding, and positional embedding. A transformer [108] encoder is used to generate hidden states, and a fully-connected layer is applied to predict the same words. Therefore, the word embeddings could be learned during training. We apply "bert-as-service" [116] as a sentence encoder in our approach since it is easy for installation and deployment. We further use a pre-trained model "BERT-Large" released by Google for embedding. Each sentence or tweet was converted into a 768-dimensional vector by taking the average of the hidden states of words, shown in Equation 5.1:

$$\begin{aligned} sent_j &\Longrightarrow e_s^j = [e_{s_1}^j, e_{s_2}^j, \cdots, e_{s_{|E|}}^j] \\ tweet_k &\Longrightarrow e_t^k = [e_{t_1}^k, e_{t_2}^k, \cdots, e_{t_{|E|}}^k] \end{aligned} \tag{5.1}$$

Here $sent_j$ represents the j^{th} sentence in a given webpage while $tweet_k$ represents the k^{th} tweet in its tweet list. e_s^j and e_t^k are the contextual embedding vectors of the above sentence and tweet, respectively, and $|E|$ is the embedding length.

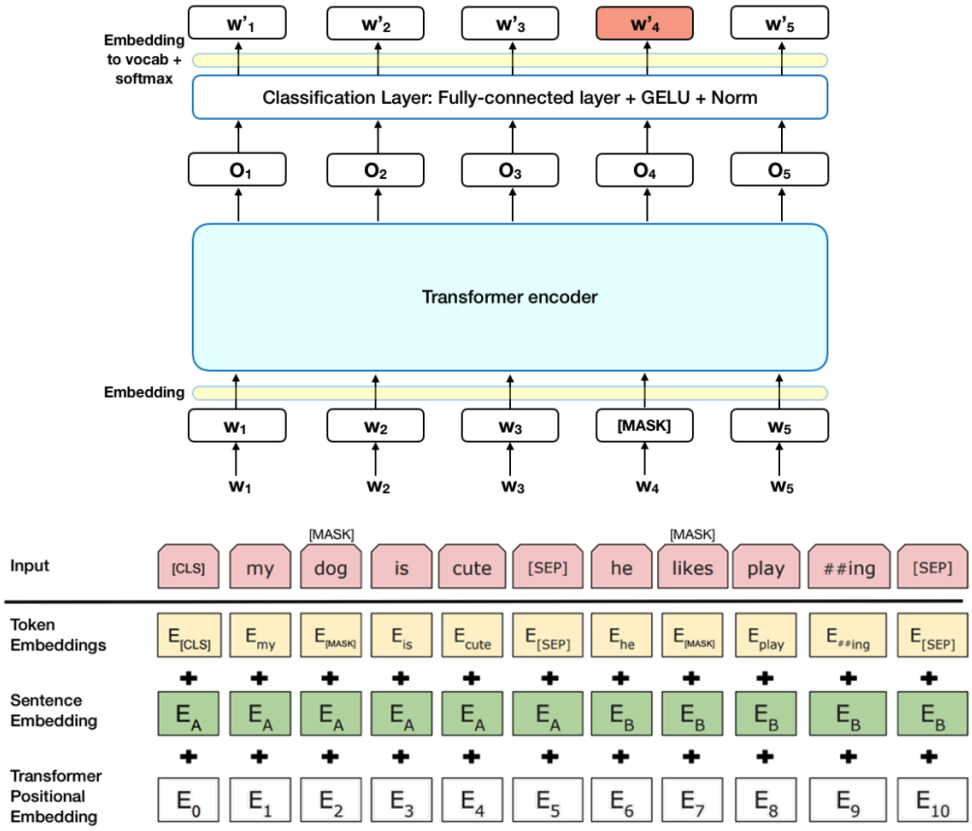


Figure 5.3: Architecture of the BERT model [30]

$$D = \{d(e_s^j, e_t^k) : j = 1, \cdots, n; k = 1, \cdots, m\}$$

$$d(e_s^j, e_t^k) = \sqrt{\sum_{i=1}^{|E|} (e_{s_i}^j - e_{t_i}^k)^2} \tag{5.2}$$

Therefore, all sentences and tweets are represented as points in the high dimensional space. We indicate the two types of data sources as orange triangle markers and blue circles in a two-dimensional circle diagram in Figure 5.2.

Next, we calculate the similarity scores between each tweet and all sentences through Euclidean distance; see Equation 5.2. We rank all the scores, choose the top 10 scores, and retrieve their corresponding webpage sentences as candidates, represented as red triangle markers in Figure 5.2. Then we choose the top 10 sentences from a webpage as sentence candidates. Last, we merge every 10 sentences candidates from each pair of a webpage and its tweet list to build a large sentence candidate corpus for further processing.

5.1.2 Entity-based Scoring

Our summarization task can significantly benefit from different types of entities, including people's names, locations, datetimes, and numbers, since these entities contain a wealth of information. Nenkova [88] discovered that entity-driven noun phrase rewriting for multi-document summarization of news leads to 20% to 50% different content in summaries, having higher linguistic quality in comparison to other baseline methods. Therefore, we took entities as an important factor in our TMDS model.

We apply spaCy [50] to identify named entities from our sentence candidate corpus, choose 9 out of 18 named entities as important ones, and categorize these named entities into five groups. Table 5.3 lists the selected groups of named entities, along with descriptions and examples.

Table 5.3: Selected groups of named entities, along with descriptions and examples

Group	Type	Description	Example
Person	PERSON	People, including fictional.	Hillary Clinton
Organization	ORG	Companies, agencies, institutions, etc.	Apple
Location	GPE	Countries, cities, states.	U.K.
Datetime	DATE	Absolute or relative dates or periods.	20 years
	TIME	Times smaller than a day.	11:45:00
Numeric	PERCENT	Percentage, including "%".	60%
	MONEY	Monetary values, including unit.	$1 billion
	QUANTITY	Measurements, as of weight or distance.	5 miles
	CARDINAL	Numerals that do not fall under another type.	13

Further, we take the Sandy Hook Elementary School shooting as an example event to illustrate the importance of named entities. By following the groups in Table 5.3, we count named entities in the sentence candidate corpus of the given event and visualize the word frequency of the top 10 words in each group; see Figures 5.4 through Figures 5.8.

We notice that most key elements of the event could be discovered with high frequencies. Such basic information includes the shooter's name *Adam Lanza*, the school's name *Sandy Hook Elementary School*, the event location *Connecticut* and *Newtown*, the occurrence date *Friday*, and the number of victims (i.e., 20 children and six adults). Other information can also be uncovered. Regarding persons, Barack Obama was mentioned many times since the president participated in multiple activities after the tragedy, while Nancy Lanza, Adam Lanza's mother, was also reportedly among the victims. Regarding organizations, the Congress and National Rifle Association (NRA) were popular entities, which were strongly related to gun control and gun laws.

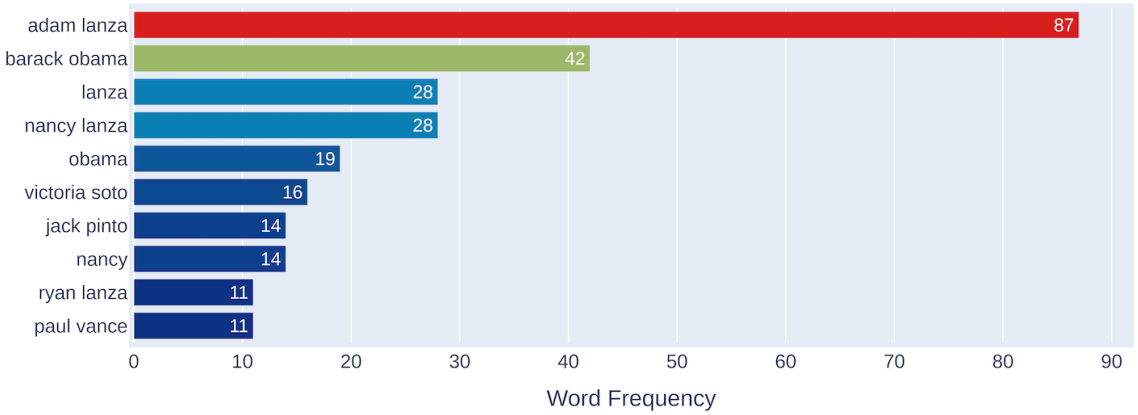


Figure 5.4: Named entity (person) distribution in the Sandy Hook shooting

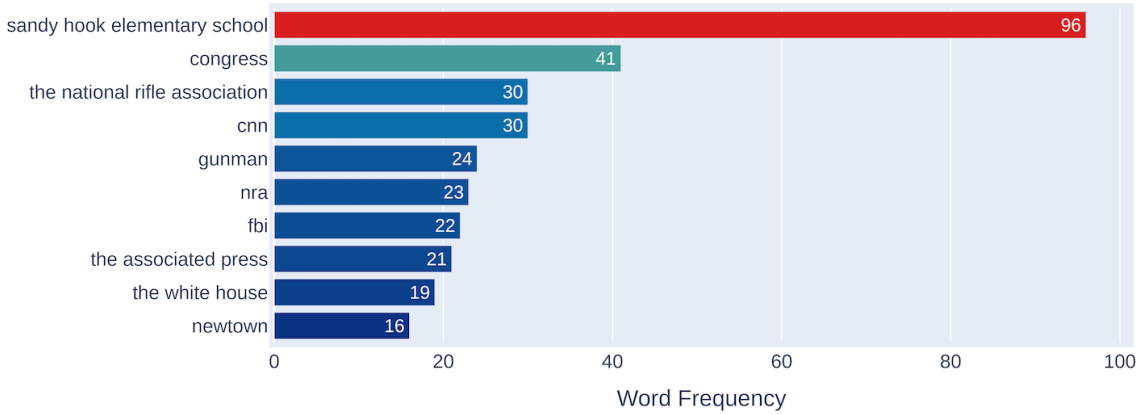


Figure 5.5: Named entity (organization) distribution in the Sandy Hook shooting

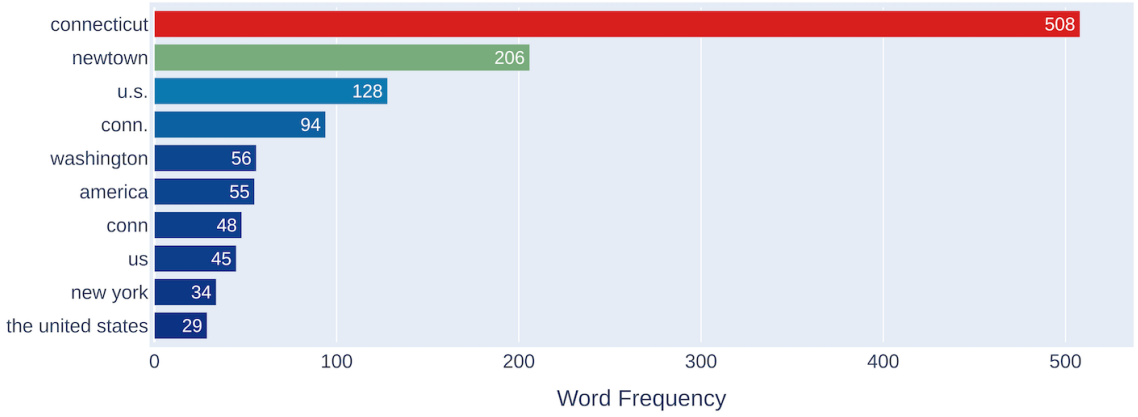


Figure 5.6: Named entity (location) distribution in the Sandy Hook shooting

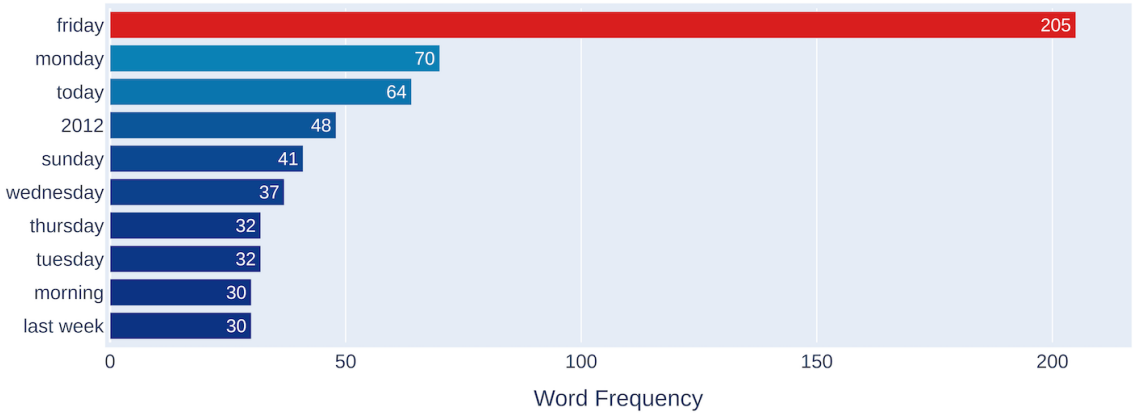


Figure 5.7: Named entity (datetime) distribution in the Sandy Hook shooting

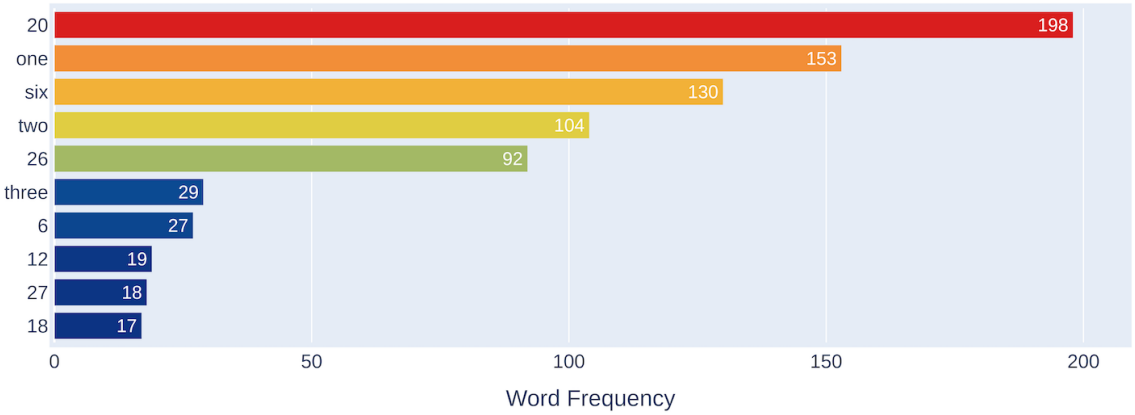


Figure 5.8: Named entity (numeric) distribution in the Sandy Hook shooting

Given a sentence candidate corpus $SC = \{sc_1, sc_2, \cdots, sc_s\}$, we extract the five types of entities from each sentence candidate. Our entity-based scoring module calculates the entity-based score of the i^{th} sentence candidate sc_i from three aspects: importance, relatedness, and diversity.

Entity Importance

If a named entity occurs many times in its corresponding group, it is likely to be more important and appear in the final summary. Based on this assumption, we extract named entities from each sentence, calculate the probability of each entity based on the corpus, and apply it to describe the importance of that sentence. The importance of a person-related named entity is calculated by Equation 5.3:

$$Imp_{per} = \sum_{i=1}^{|ENT_{per}|} p(ent^i_{per}) = \frac{f(ent^i_{per})}{|SC_{per}|} \in (0, 1] \tag{5.3}$$

Here $|ENT_{per}|$ is the number of person-related named entities in a sentence, $f(ent^i_{per})$ is the TF score of ent^i_{per} in the corpus, and $|SC_{per}|$ is the total number of person-related named entities in the corpus. Additionally, if a sentence has no person-related named entity, we simply set Imp_{per} to 0. Similarly, we applied the same equation to calculate the importance scores of other groups of named entities. Afterward, we took the average score of the five scores as the overall importance score of a sentence, shown in Equation 5.4:

$$Importance = \frac{Imp_{per} + Imp_{org} + Imp_{loc} + Imp_{dat} + Imp_{num}}{5} \in [0, 1] \tag{5.4}$$

Entity Relatedness

Besides single named entities, we also consider a co-occurrence based relatedness of entity pairs. We assume that if an entity pair occurs in a number of sentences, there is a close correspondence between the two entities. In other words, the sentence having that entity pair might have more information that can support the final summary. We calculate the relatedness of an entity pair by the following measure. First, we list all the combinations of entity pairs given a sentence. Then, for each entity pair, we scan the entire corpus, count the number of sentence candidates containing the entity pair, and divide it by the total number of sentences in the corpus. The probabilities of all entity pairs are summed up later; see Equation 5.5:

$$rel_{pair} = \sum_{i=1}^{|ENT_{pair}|} p(ent^i_{pair}) = \frac{c(ent^i_{pair}, SC_{sent})}{|SC_{sent}|}; \quad p(ent^i_{pair}) \in (0, 1] \tag{5.5}$$

Here $|ENT_{pair}|$ is the total number of entity pairs in one sentence, $c(ent^i_{pair}, SC_{sent})$ is the number of sentence candidates containing ent^i_{pair}, and $|SC_{sent}|$ is the total number of sentences in the corpus.

We also divide the above score by $|ENT_{pair}|$ for normalization; see Equation 5.6. For those sentences with no entity pairs, we simply set the entity relatedness score to 0. The final score should also be between 0 and 1.

$$Relatedness = \frac{rel_{pair}}{|ENT_{pair}|} \in [0, 1] \tag{5.6}$$

Entity Diversity

Different from previous research [46, 106], we propose entity diversity as another factor to evaluate sentence candidates. We introduce Simpson's diversity index as a measure of diversity, which is often used to quantify the biodiversity of habitats. Equation 5.7 shows the original Simpson's index of diversity:

$$Diversity = 1 - \sum_{i=1}^{R} (\frac{n_i}{N})^2 \in [0, 1] \tag{5.7}$$

Here n_i is the number of individuals of each species, N is the total number of individuals of all species, and R is the number of species. With this index, 1 represents infinite diversity, and 0 represents no diversity.

In our task, we not only expect each sentence in the final summary has more frequent entities and entity pairs, but also hope some sentences have different groups of entities to uncover the general information of a specific event. Therefore, we update Equation 5.7 to match our summarization scenario. Specifically, R is the number of groups of named entities, which is equal to 5. n_i is the number of named entities in each group, while N is the total number of named entities in each sentence. We still set the diversity score to 0 if there is no named entity in one sentence.

Finally, we measure the entity-based score of each sentence with the average value of the three scores above, shown in Equation 5.8:

$$s_{sent} = \frac{Importance + Relatedness + Diversity}{3} \tag{5.8}$$

5.1.3 Topic-based Scoring

Most previous work on multi-document summarization simply concatenates multiple documents into a single one and implements single document summarization approaches. Such a process is not following the cognitive habits of humans. Often, given multiple documents, human beings first classify them into different aspects/topics, and then summarize each topic with several sentences. Then they reorganize these sentences to create a summary. Regarding our event-related collections, those aspects may include preparedness, response, impact, recovery, etc. [52]. We have discovered that tweeters posted tweets with different topics in school shooting-related collections [70]; see Table 5.4.

Based on the assumption that sentence candidates also describe different aspects of a given event, we apply topic modeling on the sentence candidate corpus and implement a topic-based scoring module.

We take each sentence in our corpus as one document for preprocessing. We tokenize each sentence into words through spaCy [50] and filter out English stopwords and words with

Table 5.4: Major topics and typical words in school shooting-related tweets

Category	Typical Words
Information	news, shooter, accidental, gunman, arrest, victim, dead, report, kill, suspect, breaking, injured, custody, wounded, police, fatality, update, fire, campus, student, confirm
Gun Control	gun, control, nra, congress, senate, legislation, safety, violence, illegal, antigun, debate, governor, laws
Mental Health	mental, health, illness
Emotion	absurd, anger, angry, awful, crazy, disgust, frighten, heart, pray, rip, sad, tragic, dedicated, deep, silence

length less than 5 characters. We then use NLTK's Wordnet [13] for word lemmatization.

We use Gensim to create a dictionary from our corpus, then convert it into a bag-of-words corpus. Later, we employ the Latent Dirichlet Allocation (LDA) model for topic modeling. To evaluate topic models and identify the number of topics, we exploit a topic coherence measure. It calculates pairwise word similarity scores of the words in a topic and considers topic models with high topic coherence scores as good candidates. Here, we set the number of topics from 2 to 80 with step = 2 and suppose the best number of topics is t.

As a result, a sentence candidate sc_i has a probability distribution across all topics, represented as $\{p^1_{t_i}, p^2_{t_i}, \cdots, p^t_{t_i}\}$. We choose the topic with the maximum probability as the topic of that sentence, notated as c_{t_i} with its probability p_{t_i}. Further, all words in the dictionary have their probabilities in each topic, and the top words can be more helpful in describing the meaning of each topic. We select the top 10 words from each topic and calculate the relevance score between the top words and each sentence through the Jaccard index, shown in Equation 5.9:

$$r_{t_i} = \frac{|p(sc_i) \cap q(t_z)|}{|p(sc_i) \cup q(t_z)|} = \frac{|p(sc_i) \cap q(t_z)|}{|p(sc_i) + q(t_z) - p(sc_i) \cap q(t_z)|} \tag{5.9}$$

Here r_{t_i} is the relevance score of the i^{th} sentence candidate, $p(sc_i)$ is its word list, and $q(t_z)$ is the top word list of its corresponding topic. As the output of our topic-based scoring module, both topic probability and relevance score are used for sentence ranking.

5.1.4 Sentence Ranking

Besides the above measures, we also consider the potential influence of tweets. First, if many tweeters cite a webpage in their tweets, the webpage likely comes from a good news source, and its sentences are of high quality. We assign a tweet weight w^i_{tweet} to each sentence candidate sc_i, and the value is the number of tweets bonded to the sentence's original webpage.

For each input pair (i.e., a webpage and its tweet list), the selected sentence candidates have the same tweet weight. Second, a great challenge in multi-document summarization is how to arrange sentences to improve their sequential order in the final summary. The current evaluation metrics, such as ROUGE and BLEU, are still focusing on word matching and ignore the potential relationship among sentences. More concretely, if we concatenate all documents into a long one, we neglect their potential order. Though we may get a high evaluation score of the system summary, it is also possible that similar sentences appear in different positions of the summary, and the temporal sequence of sub-events is confusing. As prior knowledge, tweets are much more time-sensitive than webpages. Each tweet has a posted timestamp, indicating the date and time when a tweeter posts it. We consider tweet timestamps as an important factor to further create a proper sequence of sentence candidates. Table 5.5 lists the variables used for sentence ranking.

Table 5.5: Multiple variables for sentence ranking

Type	**Variable**		**Notation**
Entity	Sentence Entity Score	Importance	s_{sent}
		Relatedness	
		Diversity	
Topic	Topic Class		c_t
	Topic Probability		p_t
	Topic Word Relevance		r_t
Tweet	Tweet Weight		w_{tweet}
	Tweet Timestamp		$time_{tweet}$

As mentioned in Section 5.1.3, each sentence candidate has a topic probability score p_t and a topic word relevance score r_t. Accordingly, we set two thresholds to refine sentence candidates. We filter out sentence candidates with $p_t < 0.9$, assuming they are not closely related to their topics. Similarly, we remove sentence candidates with $r_t < 0.05$, preferring sentence candidates containing more topical words. The value of β is small because the length of a sentence is usually longer than ten words in most cases. Then, we select the sentence with the highest sentence entity score and tweet weight from each topic and sort all sentences by tweet timestamp to generate the summary with a proper sequence order.

5.2 A Summarization Example

As a simple start, we apply our proposed TMDS model to process the Sandy Hook Elementary School shooting collection and present intermediate results at major stages. Regarding data preprocessing, we randomly select 3,000 unique URLs that exclude Twitter pages containing `twitter.com`. We utilize our collection development model to calculate the relevance score of each webpage and select the top 1,000 webpages from the ranking list. We assume that they should be highly relevant to the given event. Then, we use either *basic_sha1* or

wayback_sha1 index to retrieve their corresponding tweets. We also add the timestamp of each tweet for further use. Table 5.6 shows an example of a webpage and its tweet list from the Sandy Hook shooting collection, where PARA_SEP is the separator of paragraphs, and the 10 digits string in each tweet is the Unix timestamp, which can be easily converted into a human-readable datetime.

Table 5.6: A webpage and its tweet list from the Sandy Hook shooting collection

docid	8ee15f64e878496f79d84517e351c893a57355fa
webpage	the national rifle association says armed police officers should be present in every school in america, as protests interrupt a news conference following the connecticut shootings. PARA_SEP nra chief executive wayne lapierre claimed that the 26 innocent victims at sandy hook elementary school could still be alive if adam lanza had been confronted by gunmen. PARA_SEP he also blamed the media for promoting violence through video games and said it demonised lawful gun owners. PARA_SEP speaking today for the first time since the shootings, he said the only thing that stops a bad guy with a gun is a good guy with a gun. PARA_SEP the nra is going to bring all its knowledge, all its dedication and all its resources to develop a model national school shield emergency response programme for every school in america that wants it. PARA_SEP he speculated that another school gunman was waiting in the wings and said every single school in america should immediately deploy a protection programme, including armed security. PARA_SEP mr lapierre was interupted on two occasions by protesters holding up signs saying: nra killing our kids and nra blood on your hands. PARA_SEP it came after connecticut governor dannel malloy called for residents of his state to observe a moment of silence at 9:30 a.m. (2.30pm gmt) today, a week after the shootings, and his fellow governors from maine to kansas followed suit. PARA_SEP the national cathedral in washington rang its bell 28 times as part of an interfaith memorial. PARA_SEP we have the moral obligation to stand for and with the victims of gun violence and to work to end it, said reverend gary hall, dean of washington national cathedral, who called on americans to pray that we may have courage to act, so that the murderous violence done on friday may never be repeated. PARA_SEP the company that operates the nasdaq stock exchange observed a moment of silence at 9:30 a.m., although markets opened trading at that time as usual. PARA_SEP the rampage, in which 28 people died, including 20 children and the gunman, has sparked new discussion on tightening gun laws, a thorny political issue in the united states, which has a strong culture of individual gun ownership. PARA_SEP it comes as vice president joe biden convened a white house task force to search for ways to quell gun violence. PARA_SEP with funerals for a half-dozen victims on thursday, services have now been held for more than half of the 27 people shot and killed last friday by 20-year-old man adam lanza, who carried out his attack armed with an assault rifle ...
tweets	1356092827 RT @Channel4News: America prepares to remember victims of Connecticut shootings: http://t.co/KARL4bn3
	1356092693 RT @Channel4News: America prepares to remember victims of Connecticut shootings: http://t.co/KARL4bn3
	1356092666 America prepares to remember victims of Connecticut shootings: http://t.co/KARL4bn3

Table 5.7 lists 10 sentence candidates selected from a webpage and its tweet list. Though the webpage was mentioned in 46 tweets, all these tweets are the same after preprocessing, describing a senator who used the victims in the shooting to raise campaign cash. The 10

sentences are sorted by their distance to the given tweet. We notice that the top sentence is semantically close to the tweet, which states both the terrible shooting event and the senator's re-election campaign. The last sentence is just a general description of the election campaign.

Table 5.7: 10 sentence candidates selected by their corresponding tweet

Tweet	***RT @PatDollard: Connecticut Senator Uses Newtown's Dead Kids 'To Raise Campaign Cash' http://t.co/JGlGF418Bq #nra #tcot #lnyhbt***
sc_1	in the wake of the horror of the december 14, 2012, massacre of 20 beautiful children and 6 dedicated educators, blumenthal is asking supporters to send money to his 2016 re-election campaign!
sc_2	it's an important part of the discussion around restricting the kind of mass murder weaponry used to kill 20 children and six educators on dec. 14.
sc_3	malloy went so far as to send public twitter messages to u.s. senators who were blocking a vote on gun legislation, asking that they return phone calls from the daughter of murdered sandy hook elementary school principal dawn hochsprung.
sc_4	using the "horror" of the "massacre of 20 beautiful children" at a time when critical legislation honoring their memory is at stake to beg for $5 for your next political campaign is as tasteless as it gets.
sc_5	members of congress need to know that dylan hockley "loved jumping on trampolines and watching movies" and that he "died in his teacher's arms."
sc_6	there are lots of ways and lots of time for sen. blumenthal to raise money for his re-election campaign before 2016.
sc_7	we're glad that u.s. sen. chris murphy and gov. dannel malloy are talking in detail about the victims of sandy hook as part of a push for federal gun control legislation.
sc_8	please contribute $5 now as the senate debate continues on common-sense gun reform legislation this week.
sc_9	the issue took a disgusting political turn on thursday, though, when u.s. sen. richard blumenthal, d-conn., used sandy hook to raise money.
sc_{10}	but the campaign – and its drive for money to buy elections – is constant

Regarding our entity-based scoring module, we compute the three scores (i.e., importance, relatedness, and diversity) of each sentence candidate, and list the top 10 sentences for each score, along with the top 10 sentences with high entity-based scores; see Tables 5.8 through 5.11. From the results, we discover that the top sentences vary a lot across the three aspects, and only one sentence candidate appears in both the importance list (ID = 2) and the diversity list (ID = 9). Sentences in Table 5.9 have shorter lengths than those in Table 5.10 since sentences with high diversity scores cover more entity types that are likely to appear in long sentences. Focusing on the top 10 sentences with high entity scores, we find that only half of them occur in the three sub-lists, and most of them give an overall description of the shooting event.

Regarding our topic-based scoring module, as can be seen in Figure 5.9, the coherence score peaks for 44 topics. Following our sentence ranking strategy, Table 5.12 lists the sentences in our final summary with datetime.

Table 5.8: Top 10 sentences with high entity importance scores

ID	Sentence
1	connecticut state police friday released thousands of pages of police documents from the investigation into the dec. 14, 2012 slaying of 20 children and six educators at sandy hook elementary school in newtown.
2	connecticut shooter adam lanza used a weapon in the bushmaster ar-15 family to shoot all of his victims at a school in a rampage that killed 20 young children and six staff members on friday in newtown, connecticut, police said.
3	on this day in 2012, at sandy hook elementary school in newtown, connecticut, adam lanza killed 20 first graders and six school employees before turning a gun on himself.
4	a heavily armed gunman killed 26 people, including 20 children from 5 to 10 years old, in a rampage at a connecticut elementary school on friday, one of the worst mass shootings in u.s. history.
5	a heavily armed gunman killed 26 people, including 20 children from 5 to 10 years old, in a rampage at a connecticut elementary school on friday, one of the worst mass shootings in u.s. history.
6	connecticut's highest court on tuesday said it would consider whether families of nine victims, plus one survivor, can recover damages from remington outdoor co and others over the gunman's use of a bushmaster ar-15 rifle in the attack in newtown, connecticut, which killed 20 students and six staffers.
7	world leaders expressed shock and horror after a gunman massacred 20 small children and six teachers on friday in the us state of connecticut, in one of the worst school shootings in history.
8	world leaders expressed shock and horror after a gunman massacred 20 small children and six teachers on friday in the us state of connecticut, in one of the worst school shootings in history.
9	the families of two of the 20 students killed in a 2012 massacre at sandy hook elementary school, are suing the town of newtown, connecticut and the local school board over alleged lax security, media reported on monday.
10	a gunman opened fire at the sandy hook elementary school in newtown, connecticut, on friday morning, killing 26 people — including 20 children.

5.3 Data

We select 32 event-related collections for summarization, including four categories: shooting, hurricane, earthquake, and others. Table 5.13 shows the detailed descriptions of these collections. Similar to Section 5.2, we sample webpages from the entire tweet collections and select the top 1000 webpages for summarization, along with their corresponding tweets and timestamps.

Further, we take the event-related Wikipedia pages as gold summaries so that we can carry out the evaluation on our model and other baselines. Each event-related collection links with an entry in Wikipedia. For instance, the Wikipedia entry of the El Paso shooting is "2019 El Paso shooting". Later, we use the Wikipedia API to directly extract readable content from each Wikipedia page.

Table 5.9: Top 10 sentences with high entity relatedness scores

ID	Sentence
1	newtown, conn.
2	the boy, whose name has not been released because he is a juvenile, indicated that he wanted to defend himself if there was an incident similar to what happened in newtown, conn.
3	a gunman walked into sandy hook elementary school in newtown, conn.
4	a gunman walked into sandy hook elementary school in newtown, conn.
5	a woman waits to hear about her sister, a teacher, following a mass shooting at the sandy hook elementary school in newtown, conn.
6	teachers and parents across the country were wrestling with how best to quell children's fears about returning to school for the first time since the killings at sandy hook elementary school in newtown, conn.
7	bells from a nearby church chimed for each victim during a somber gathering at edmond town hall in newtown, conn.
8	using news audio, we've highlighted the song's relevance to the horrific shootings in newtown, conn.
9	a 27-year-old teacher of puerto rican descent has emerged as a hero in the tragic shooting at an elementary school in newtown, connecticut.
10	names of victims are displayed on a flag in newtown, connecticut.

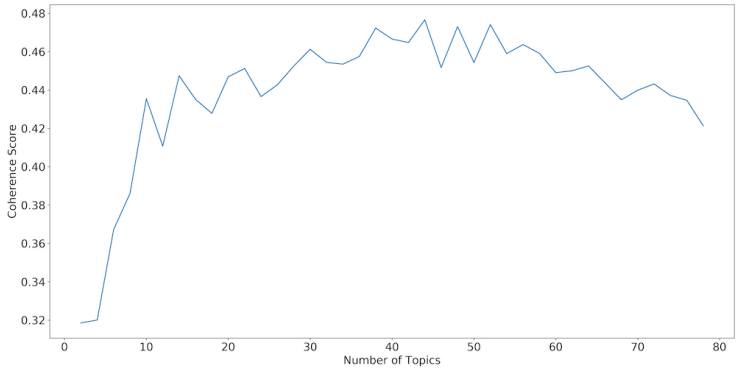


Figure 5.9: Coherence scores among different numbers of topics

5.4 Evaluation and Visualization

5.4.1 Evaluation of TMDS Model

In this section, we evaluate our TMDS model on the golden standard datasets of the 32 event-related collections. In addition to our current model, we also propose several methods for comparison, including both basic methods and extractive methods.

For quantitative evaluation, we report ROUGE [77] scores, which measure the overlap of unigrams (ROUGE-1), bigrams (ROUGE-2), and skip bigrams with a max distance of four words (R-SU). We apply *pythonrouge*, a Python wrapper for evaluating summarization quality by ROUGE package, to calculate the ROUGE scores between system summaries and gold

Table 5.10: Top 10 sentences with high entity diversity scores

ID	Sentence
1	manchester was the lead agency in the hartford distributors shooting in which a fired employee, omar thornton, killed eight people before killing himself in august 2010.
2	dennis carlson, superintendent of anoka-hennepin school district in minnesota, said a mental health consultant will meet with school officials monday, and there will be three associates — one to work with the elementary, middle and high schools, respectively.
3	that move came days after president barack obama called on congress to impose a similar prohibition nationwide following the fatal shooting of 14 people in california by a married couple inspired by islamic state militants.
4	l.a. times national correspondent tina susman talks with video reporter ann simmons about the demolition of sandy hook elementary school in newtown, conn., the site of last december's massacre of 20 first-graders and six adults.
5	newtown shooter adam lanza used a remington-made, ar-15-style rifle to kill 20 children and six educators.
6	one of those filings, on thursday, showed that the washington, d.c.-based entertainment software association — which represents producers of computer and video games, and whose website denies any "link between computer and video games and violence" — has agreed to pay $36,000 by june 30 to brown rudnick government relations strategies, the lobbying firm of former democratic state house speaker thomas d. ritter.
7	the national rifle association has broken its silence on the mass school shooting in newtown, conn., saying it is ready to "offer meaningful contributions" to the effort to make sure there are no more incidents like the one in which 20-year-old adam lanza used an assault rifle to kill 27 people before killing himself.
8	debora seifert, a spokeswoman for the federal bureau of alcohol, tobacco, firearms and explosives, told cnn on friday the riverview gun sales shop in east windsor, connecticut, lost its federal firearms license december 20 - six days after the massacre about 65 miles southwest in newtown.
9	connecticut shooter adam lanza used a weapon in the bushmaster ar-15 family to shoot all of his victims at a school in a rampage that killed 20 young children and six staff members on friday in newtown, connecticut, police said.
10	one of the proposed post-newtown reform bills — which a key leader last week called "low-hanging fruit" — is one being co-sponsored by democratic state sen. gary lebeau of east hartford and senate republican leader john mckinney of fairfield to ban ammunition magazines for rifles that contain more than 10 bullets lanza used numerous 30-round magazines during his onslaught.

summaries. We set the parameter *length_limit* to *False*, considering the entire summaries from the two sources above. Finally, we compute the average ROUGE scores across all 32 events for evaluation.

For qualitative evaluation, we list some system summaries generated by different models and measure them from various aspects

5.4.1.1 Baseline Methods

Basic Method

First (First-k)

Table 5.11: Top 10 sentences with high entity scores

ID	Sentence
1	connecticut shooter adam lanza used a weapon in the bushmaster ar-15 family to shoot all of his victims at a school in a rampage that killed 20 young children and six staff members on friday in newtown, connecticut, police said.
2	on this day in 2012, at sandy hook elementary school in newtown, connecticut, adam lanza killed 20 first graders and six school employees before turning a gun on himself.
3	connecticut's highest court on tuesday said it would consider whether families of nine victims, plus one survivor, can recover damages from remington outdoor co and others over the gunman's use of a bushmaster ar-15 rifle in the attack in newtown, connecticut, which killed 20 students and six staffers.
4	connecticut state police friday released thousands of pages of police documents from the investigation into the dec. 14, 2012 slaying of 20 children and six educators at sandy hook elementary school in newtown.
5	james dobson suggests friday's shooting of 20 children in newtown, connecticut was a consequence of gay marriage.
6	the sandy hook elementary school shooting occurred on december 14, 2012, in newtown, connecticut, when 20-year-old adam lanza fatally shot 20 children between six and seven years old, as well as six adult staff members.
7	their effort follows president barack obama's remarks in connecticut on monday night on gun control, an issue catapulted into the national arena by december's gruesome slaying of 20 young children and six educators at the sandy hook elementary school in newtown, connecticut.
8	the sandy hook elementary school shooting occurred on december 14, 2012, in newtown, connecticut, united states, when 20-year-old adam lanza fatally shot 20 children between six and seven years old, as well as six adult staff members.
9	the sandy hook elementary school shooting occurred on december 14, 2012, in newtown, connecticut, united states, when 20-year-old adam lanza fatally shot 20 children between six and seven years old, as well as six adult staff
10	the families of two of the 20 students killed in a 2012 massacre at sandy hook elementary school, are suing the town of newtown, connecticut and the local school board over alleged lax security, media reported on monday.

We concatenate the first sentence of each webpage in each event-related collection as the system summary. For our dataset, *First-k* means the first k sentences from each webpage will be concatenated as the summary. Comparing with other multi-document summarization datasets (e.g., DUC2004, Multi-News), our event-related collections have much more webpages, so it is infeasible to do the sentence concatenation across all webpages. Therefore, we develop two subsets for evaluation. One is to randomly select 10 webpages from each webpage corpus and the other is to choose the top 10 webpages since they are highly relevant to each event based on our collection development model.

Extractive Method

TextRank (TextRank)

TextRank is a graph-based ranking model, developed by Mihalcea and Tarau [83]. Sentence importance scores are computed based on eigenvector centrality within a global graph from

Table 5.12: Summary sentences with datetimes in the Sandy Hook shooting collection

ID	Datetime	Sentence
1	2012-12-14 13:43:00	a gunman opened fire at the sandy hook elementary school in newtown, connecticut, on friday morning, killing 26 people — including 20 children.
2	2012-12-14 17:38:40	two law enforcement sources briefed on the investigation confirmed to reuters the shooter had been identified as adam lanza, 20.
3	2012-12-15 10:29:48	with the death toll at 26, the massacre in newtown is the second-deadliest school shooting in u.s. history, behind the 2007 virginia tech mass shooting that left 32 dead.
4	2012-12-15 10:29:48	hartford, connecticut, mayor padro segarra speaks emotionally about the students and teachers who died earlier in the day at sandy hook elementary school in nearby newtown at a candlelight vigil at bushnell park in hartford on friday.
5	2012-12-18 06:03:01	president obama is scheduled to deliver a statement on today's shooting at the sandy hook elementary school in newtown, connecticut at 3:15 pm et from the white house.
6	2012-12-21 21:11:28	the nra proposal would take one of every seven u.s. police officers off the streets during school days, based on a reuters analysis of u.s. government data.
7	2012-12-22 18:49:26	paul simon performed his classic track the sound of silence at the funeral of a teacher who died in the school shooting in connecticut on 14 december.
8	2013-12-27 20:33:08	however, a prosecutor's report released last month concluded that lanza acted alone and took the motive for the bloodbath to his grave.
9	2015-10-12 13:50:18	in this photo provided by the newtown bee, connecticut state police lead children from the sandy hook elementary school in newtown, conn., following a reported shooting there friday, dec. 14, 2012.
10	2016-04-14 18:53:25	a connecticut judge ruled thursday that a wrongful-death lawsuit filed by families of victims killed at sandy hook elementary school against the manufacturer of the rifle used in the 2012 shooting in newtown, conn., can proceed.

the corpus.

Maximal Marginal Relevance (MMR)

Maximal Marginal Relevance (MMR) is an iterative method for content selection from multiple documents, introduced by Carbonell and Goldstein [18]. It greedily chooses the best sentence to expand the current summary and stops when the desired length is reached. The best sentence should be maximally relevant to a query and minimally redundant with the current summary.

BERT-based Extractive Summrization Model (BERT-based)

It is a Python-based RESTful service that utilizes the BERT model for text embeddings and K-Means clustering to identify sentences closest to the centroid for summary selection. The difference between the *BERT-based* model and our TMDS model is that the *BERT-based* model does not leverage guided-knowledge, named entity, and topic modeling during

summarization.

We evaluate the above extractive methods on both most relevant webpages (top 10 webpages) and the entire dataset (1,000 webpages). For both datasets, we concatenate all webpages into a single document for summarization. For *TextRank* and *MMR*, we set the length of system summaries to 1,000 words. Since *BERT-based* generates a list of sentence candidates, we sequentially add each sentence into the system summary and truncate it to no more than 1,000 words. We further remove the evaluation of *MMR* on the entire dataset due to the overly long run-time of the script.

Table 5.13: Selected event-related collections for summarization

Collection Type	Collection	# of Tweets	Starting
Shooting	Sandy Hook Elementary School shooting	540,701	12/14/2012
	Los Angeles International Airport shooting	119,445	11/01/2013
	Charleston church shooting	505,172	06/17/2015
	Orlando nightclub shooting	186,549	06/12/2016
	Las Vegas shooting	187,910	10/01/2017
	Marshall County High School shooting	330,805	01/23/2018
	Stoneman Douglas High School shooting	1,313,128	02/14/2018
	Santa Fe High School shooting	1,300,311	03/18/2018
	Noblesville West Middle School shooting	173,474	03/25/2018
	Christchurch mosque shootings	867,725	03/15/2019
	El Paso shooting	1,302,810	08/03/2019
	Dayton shooting	832,935	08/04/2019
Hurricane	Hurricane Sandy	1,500,440	10/22/2012
	Hurricane Matthew	3,576,764	09/28/2016
	Hurricane Harvey	2,706,818	08/17/2017
	Hurricane Irma	1,085,654	08/30/2017
	Hurricane Florence	5,114,477	08/31/2018
	Hurricane Michael	1,336,660	10/07/2018
	Hurricane Barry	109,964	07/11/2019
	Hurrcane Dorian	3,444,573	08/24/2019
Earthquake	Guatemala earthquake	124,690	11/07/2012
	Nepal earthquake	600,267	04/25/2015
	Kaikoura earthquake	470,711	11/14/2016
	Iran–Iraq earthquake	222,475	11/12/2017
	Anchorage earthquake	290,343	11/30/2018
	Ridgecrest earthquakes	458,595	07/04/2019
	Puerto Rico earthquakes	388,626	01/07/2020
Others	Ohio State University attack	250,539	11/28/2016
	Berlin truck attack	214,849	12/19/2016
	Westminster attack	806,040	03/22/2017
	Manchester Arena bombing	149,791	05/22/2017
	California wildfires	504,911	07/15/2018

5.4.1.2 Evaluation Results

Quantitative Evaluation

Table 5.14 shows the ROUGE F1 results on our event-related collections. Our approach is *TMDS*, which has the highest R-1, R-2, and R-SU scores. We also analyze the performance of baseline methods. Focusing on both random and top datasets, as the number of first sentences increases, all types of ROUGE scores increase, since more relevant words could be added into the system summaries. By comparing the same *First-k* between the two datasets, we notice that the summaries from top webpages are better than those from random datasets. Top webpages are more relevant to the given event, so their corresponding first sentences likely overlap with the sentences in gold summaries. Regarding the *TextRank* and *MMR* methods on the top dataset, they outperform all *First-k* methods. Further, the ROUGE scores of the *BERT-based* method are lower than the two methods above and the *First-3* method. On the other hand, the *BERT-based* method has much better performance on the entire dataset than other baseline methods. The reason is that the top dataset only has 10 webpages, leading to more non-relevant sentence centroids that reduce the quality of system summaries.

Table 5.14: ROUGE F1 results on our event-related collections

Dataset	Method	R-1	R-2	R-SU
Random	*First-1*	4.74	1.01	1.49
	First-2	8.45	1.83	2.65
	First-3	12.13	2.49	3.81
Top	*First-1*	5.63	1.56	2.02
	First-2	9.71	2.44	3.32
	First-3	12.51	2.97	4.24
	TextRank	16.13	3.40	5.11
	MMR	16.40	3.38	5.15
	BERT-based	11.77	2.54	3.78
All webpages	*TextRank*	15.58	3.20	4.81
	BERT-based	18.00	3.79	5.80
	TMDS	**21.05**	**4.33**	**6.56**

Qualitative Evaluation

For qualitative evaluation, we present part of the system summaries generated by *TextRank*, *BERT-based*, and our TMDS model on the Sandy Hook shooting collection. Since all models yield extractive summaries, the readability of each sentence in the system summaries is acceptable. Focusing on these summaries, we find that the entire summary generated by our TMDS model performs well in readability, logicality, and fluency.

The *TextRank* summary has a good start but turns to describe some detailed subevents of

the shooting event. There are also two long repeated contents, which reduce the fluency of the summary.

TextRank Summary

newtown, connecticut (reuters) - a heavily armed gunman opened fire at a connecticut elementary school on friday, killing 26 people including 20 children in the latest in a series of shooting rampages across the united states this year, u.s. media reported. a mother runs with her children as police above canvass homes in the area following **a shooting at the sandy hook elementary school in newtown, conn., about 60 miles (96 kilometers) northeast of new york city, friday, dec**. in this photo provided by the newtown bee, connecticut state police lead children from the sandy hook elementary school in newtown, conn., following a reported shooting there friday, dec. people embrace at a firehouse staging area for family around near the scene of **a shooting at the sandy hook elementary school in newtown, conn., about 60 miles (96 kilometers) northeast of new york city, friday, dec**. hartford, connecticut (reuters) - members of a connecticut panel charged with recommending ways to prevent gun violence in schools after last year's massacre at sandy hook elementary school on friday said a state attorney's report failed to address the role of the shooter's mental health in the attack. abc news has learned that investigators have seized computers belonging to adam lanza from the home he shared with his mother nancy, the same place he killed her before going to sandy hook elementary school, where he fatally shot students in two first-grade classes along with teachers and staff. noah's twin sister, arielle, who was assigned to a different classroom, survived the killing frenzy by 20-year-old adam lanza that left 20 children and six adults dead last week at sandy hook elementary ···

BERT-based Summary

it has been nearly four months since the massacre at sandy hook elementary school and now legislative leaders are finally planning to pass what they call the most far-reaching gun-legislation package in the country. "in a time of universal deceit, telling the truth becomes a revolutionary act" - george orwell, author ... "all truth passes through three stages. yes indeed the facts do not add up.. i have my own comments and thoughts to follow:. i for one am seriously now smelling a rat... this has all the earmarks of being an operation for the sole purpose of indeed scaring the public into accepting gun control and a repeal of the second amendment of the us constitution!. again, as i stated before, i do feel for the families of this massacre. the fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors. canadian male, age 58.(hard to believe i am that old.) i am in this for the truth and where ever the truth leads me! ortherntruthseeker is fully responsible for all content in this blog. if any errors or omissions are deemed necessary for correction of material, it will be placed into the articles affected, and notice will be put up for all readers. the information articles in this blog are free for everyone to take and spread to others. in these times of uncertainty, people deserve the truth no matter how harsh it may be! connecticut state police officers walk out of st. paul vance addresses the press on december 15. j. paul vance, center, briefs the media on the elementary school shootings during a press conference at treadwell memorial park on december 14 in newtown. people weep and embrace near sandy hook elementary school on friday, december 14. ···

The *BERT-based* summary has some problem with logicality. It mainly introduces something

related to gun control instead of the shooting event. Moreover, there are many first-person words (e.g., i, my, us) in sentences, which should not occur frequently in a summary.

Based on the tweet posted timestamps, our TMDS summary is more logical and fluent. It gives an overview of the entire event and points out the shooter. Then, it describes the impact of the massacre, followed by detailed subevents.

TMDS Summary

a gunman opened fire at the sandy hook elementary school in newtown, connecticut, on friday morning, killing 26 people — including 20 children. two law enforcement sources briefed on the investigation confirmed to reuters the shooter had been identified as adam lanza, 20. with the death toll at 26, the massacre in newtown is the second-deadliest school shooting in u.s. history, behind the 2007 virginia tech mass shooting that left 32 dead. hartford, connecticut, mayor padro segarra speaks emotionally about the students and teachers who died earlier in the day at sandy hook elementary school in nearby newtown at a candlelight vigil at bushnell park in hartford on friday. president obama is scheduled to deliver a statement on today's shooting at the sandy hook elementary school in newtown, connecticut at 3:15 pm et from the white house. the nra proposal would take one of every seven u.s. police officers off the streets during school days, based on a reuters analysis of u.s. government data. paul simon performed his classic track the sound of silence at the funeral of a teacher who died in the school shooting in connecticut on 14 december. however, a prosecutor's report released last month concluded that lanza acted alone and took the motive for the bloodbath to his grave. in this photo provided by the newtown bee, connecticut state police lead children from the sandy hook elementary school in newtown, conn., following a reported shooting there friday, dec. 14, 2012. a connecticut judge ruled thursday that a wrongful-death law suit filedby families of victims killed at sandy hook elementary school against the manufacturer of the rifle used in the 2012 shooting in newtown, conn., can proceed. · · ·

Moreover, we have submitted an MTurk evaluation task entitled "Human Evaluation on Summary Quality" to the IRB and received an exemption approval; see Appendix A. We plan to evaluate the summaries generated by our TMDS model and baseline models from four aspects: readability, correctness, completeness, and compactness. A ranking of summaries will also be provided by the MTurk workers to judge the overall quality of summaries during human evaluation.

5.4.2 Visualization: Timeline Service

We employ TimelineJS [61] to build a storytelling timeline [2] of each summary. TimelineJS is an open-source tool that is developed by the Knight Lab at Northwestern University and built in JavaScript. We convert our summary into a suitable JSON format for visualization.

For the timeline service, we first add a link (`https://cdn.knightlab.com/libs/timeline3/latest/css/timeline.css`) tag loading the Timeline CSS, and a script tag (`https://cdn.knightlab.com/libs/timeline3/latest/js/timeline.js`) loading the Timeline JavaScript.

Then we pass our JSON data into a TimelineJS object and configure other options to optimize the web interface. Figure 5.10 shows part of the summary timeline of the Sandy Hook shooting.

Figure 5.10: Event summary visualization through TimelineJS

Regarding user interaction, users can drag the timeline bar to select a time window. The bar can also be zoomed in or out, to show less or more information about the event. By clicking the items in the timeline bar, users can get detailed information from the content box above.

Given an event-related collection, the timeline service can help the general society and specific groups (e.g., historians, journalists) with a lack of background knowledge have an overview of an event and uncover some of the five Ws (i.e., who, what, when, where, why), with the help of different types of entities.

Chapter 6

Contributions and Future Work

6.1 Contributions

This research introduces a new, integrated system that supports the flexible use of event-related collections for interdisciplinary research and education at different levels (i.e., data, information, and knowledge). Enhanced modules in sub-systems have been designed in terms of our event-related scenarios, targeting specific applications.

At the data level, our sub-system takes each event-related tweet collection as input and produces an event-related webpage corpus. The data curation strategy establishes the mappings between both tweets and webpages. Researchers in the data-related societies can explore such data for further data mining work, since the processed data is relatively clean and highly related to an event.

At the information level, TwiRole predicts relatively accurate results for role-related user classification. Our tweeting pattern analysis demonstrates how to apply TwiRole for user-centered research. Moreover, features in TwiRole can be modified or extended for the detection of other types of tweeters, and different modules (e.g., topic modeling, social network) could be combined with TwiRole, for various types of user analyses in social societies. With the help of the service of TwiRole, people in the general society can test their interested Twitter accounts, while researchers are able to detect a set of Twitter users to support their research.

At the knowledge level, with the guidance of tweets, our TMDS model takes advantage of named entities and sentence topics, to summarize a set of webpages related to an event. By considering data stream (i.e., tweet timestamps), we also provide a timeline service to illustrate important sub-events at different time points, providing an overview of each event to the general society.

These sub-systems cover some of the most important stages in a digital library, including collection development, curation, analysis, integration, and visualization. They can not only be applied independently to solve problems in a specific scenario but also be integrated as a whole based on the 5S theory. The outcomes and services proposed in our study can also be easily adopted or extended by different societies for further event analyses and service developments.

The detailed contributions are shown below:

- We designed and implemented a collection development model with both pre-query design and post-query expansion, which retrieves relevant webpages with both high precision and recall, from a given event-related tweet collection. If results are fed into an event focused crawler (EFC), the model can provide relevant URL seeds to EFC for open web crawling and thereby further expand the collection size.
- We designed and implemented a data curation strategy for our event-related collections. The metadata of both tweets and webpages are represented in JSON and WARC files, respectively. Both types of files are linked by SHA-1 hashes. As an application, we carried out an empirical study of short URLs to help people better understand the lifespan of URLs on Twitter and the coverage of the Wayback Machine.
- We developed TwiRole, a hybrid model for role-related user classification on Twitter. TwiRole outperforms existing methods and obtains more balanced results over several roles. We also found that first-person words and profile images are good indicators for user classification.
- We deployed TwiRole and an existing mood detection model to discover tweeting patterns of different events like disasters. We found differences in the daily posting patterns, between the different types of events. The mood patterns also vary across events and tweeters.
- We shared the source code of TwiRole, published a capsule on Code Ocean for code reproducibility, and provided a web application for online role-related user classification. The web application has a simple input interface and the output layout can show the prediction results of different classifiers.
- We implemented TMDS, a tweet-guided multi-document summarization model, to generate summaries from event-related collections with high quality. The proposed model considers tweets, named entities, and sentence topics, and the sentences in summaries are organized with proper order and cover different aspects of events.

6.2 Publications

A number of papers have been published or accepted during my Ph.D. program. A selected set of publications is shown in the list below.

- **Liuqing Li**, Jack Geissinger, William A. Ingram, Edward A. Fox. Teaching Natural Language Processing through Big Data Text Summarization with Problem-Based Learning. Data and Information Management, ISSN:2543-9251, Vol. 4, 2020, in press [75].

- **Liuqing Li** and Edward A. Fox. Disaster Response Patterns across Different User Groups on Twitter: A Case Study during Hurricane Dorian. Accepted for the 17th Information Systems for Crisis Response and Management Conference (ISCRAM), Virgnia, USA, May 2020 [74].
- **Liuqing Li**, Rishabh Anand, and Edward A. Fox. Users, User Roles, and Topics in School Shooting Collections of Tweets. Web Archiving and Digital Libraries (WADL), a workshop held in conjunction with ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Illinois, USA, June 2019 [70].
- Andrea Kavanaugh, Ziqian Song, **Liuqing Li** and Edward A. Fox. Communication Behavior in an Emerging Democracy: Political Expression via Tweets during the 2014 Tunisian Elections. In Proceedings of the 20th Annual International Conference on Digital Government Research, Dubai, United Arab Emirates, June 2019 [57].
- **Liuqing Li** and Edward A. Fox. Understanding Patterns and Mood Changes through Tweets about Disasters. In Proceedings of the 16th Information Systems for Crisis Response and Management Conference (ISCRAM), Valencia, Spain, May 2019 [73].
- **Liuqing Li**, Ziqian Song, Xuan Zhang and Edward A. Fox. TwiRole: A Hybrid Model for Role-related User Classification on Twitter, published on Arxiv: https://arxiv.org/abs/1811.10202 (GitHub link: https://github.com/liuqingli/TwiRole) [76].
- **Liuqing Li** and Edward A. Fox. A Study of Historical Short URLs in Event Collections of Tweets. Web Archiving and Digital Libraries (WADL), a workshop held in conjunction with ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), Texas, USA, June 2018 [72].
- **Liuqing Li**, He Feng, Wenjie Zhuang, Na Meng and Barbara Ryder. CCLearner: A Deep Learning-Based Clone Detection Approach. In Proceedings of the 33rd IEEE International Conference on Software Maintenance and Evolution (ICSME), Shanghai, China, September 2017 [71].

6.3 Future Work

We will continue the work on the three research projects, optimize the system architecture and work flow, and report results, as follows.

- For relevant data enrichment, we will further optimize the interfaces of each part and eventually integrate them into an automatic sub-system.
- For user classification, we will employ TwiRole in various types of application scenarios to uncover more role-related results.

- For tweet-guided multi-document summarization, we will carry out a human evaluation on our TMDS model and baseline models soon. We will conduct an intra-comparison among different factors used in our TMDS model for optimization. We will also improve the timeline service for visualization and further evaluation.

Bibliography

[1] Firoj Alam, Ferda Ofli, and Muhammad Imran. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. In *Proceedings of the 2018 International ISCRAM Conference*, 2018.

[2] Yasmin AlNoamany. *Using web archives to enrich the live web experience through storytelling*. Old Dominion University, 2016.

[3] Jalal S Alowibdi, Ugo A Buy, and Philip Yu. Empirical Evaluation of Profile Characteristics for Gender Classification on Twitter. In *Proceedings of the 12th International Conference on Machine Learning and Applications (ICMLA)*, volume 1, pages 365–369. IEEE, 2013.

[4] Jalal S. Alowibdi, Ugo A. Buy, and Philip Yu. Language Independent Gender Classification on Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 739–743. ACM, 2013.

[5] Derek Anderson. Wordninja. https://pypi.python.org/pypi/wordninja/0.1.3, 2017. Accessed: 2019-12-01.

[6] Demetris Antoniades, Iasonas Polakis, Georgios Kontaxis, Elias Athanasopoulos, Sotiris Ioannidis, Evangelos P Markatos, and Thomas Karagiannis. We.B: The Web of Short URLs. In *Proceedings of the 20th International Conference on World Wide Web*, pages 715–724. ACM, 2011.

[7] Internet Archive. Internet Archive. https://archive.org/, 2018. Accessed: 2019-12-01.

[8] Claudette G Artwick. News Sourcing and Gender on Twitter. *Journalism*, 15(8):1111–1127, 2014.

[9] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 65–74. ACM, 2011.

[10] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[11] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. Multi-Document Abstractive Summarization Using ILP Based Multi-Sentence Compression. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1208–1214, 2015.

[12] Tal Baumel, Matan Eyal, and Michael Elhadad. Query Focused Abstractive Summarization: Incorporating Query Relevance, Multi-Document Coverage, and Summary Length Constraints into seq2seq Models. *arXiv preprint arXiv:1801.07704*, 2018.

[13] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[14] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1301–1309. Association for Computational Linguistics, 2011.

[15] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to Rank using Gradient Descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96. ACM, 2005.

[16] Cheng Cao and James Caverlee. Detecting Spam URLs in Social Media via Behavioral Analysis. In *European Conference on Information Retrieval*, pages 703–714. Springer, 2015.

[17] Cornelia Caragea, Anna Cinzia Squicciarini, Sam Stehle, Kishore Neppalli, and Andrea H. Tapia. Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy. In *Proceedings of the 2014 International ISCRAM Conference*, 2014.

[18] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

[19] Jing Chen, Chenyan Xiong, and Jamie Callan. An Empirical Study of Learning to Rank for Entity Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 737–740. ACM, 2016.

[20] Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, 2016.

[21] Eric Chu and Peter J Liu. Unsupervised Neural Multi-document Abstractive Summarization. *arXiv preprint arXiv:1810.05739*, 2018.

[22] Zi Chu, Indra Widjaja, and Haining Wang. Detecting Social Spam Campaigns on Twitter. In *International Conference on Applied Cryptography and Network Security*, pages 455–472. Springer, 2012.

[23] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1136–1145. Association for Computational Linguistics, 2013.

[24] Niko Colneriĉ and Janez Demsar. Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. *IEEE Transactions on Affective Computing*, 2018.

[25] CrowdFlower. Using Machine Learning to Predict Gender. https://www.crowdflower.com/using-machine-learning-to-predict-gender, 2015. Accessed: 2019-12-01.

[26] PJ Daas, Joep Burger, Quan Le, Olav ten Bosch, and MJ Puts. Profiling of Twitter users: a big data selectivity study. Technical report, Discussion paper 201606, Statistics Netherlands, 2016.

[27] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 126–134. ACM, 2018.

[28] Jeffrey Dalton, Laura Dietz, and James Allan. Entity Query Feature Expansion using Knowledge Base Links. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 365–374. ACM, 2014.

[29] Hal Daumé III and Daniel Marcu. Bayesian Query-focused Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 305–312. Association for Computational Linguistics, 2006.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[31] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[32] Mohamed Magdy Gharib Farag. *Intelligent Event Focused Crawling*. PhD dissertation, Virginia Polytechnic Institute and State University, 23 September 2016. http://hdl.handle.net/10919/73035.

[33] Mohamed Magdy Gharib Farag, Sunshin Lee, and Edward A Fox. Focused Crawler for Events. *International Journal on Digital Libraries*, 19(1):3–19, 2018.

[34] Feedspot. Top 100 USA News Websites on the Web. https://blog.feedspot.com/usa_news_websites/, 2018. Accessed: 2019-12-01.

[35] Francesco Ferrari, Ayush Jasuja, Manu Seth, and Ranti Dev Sharma. Gender Inference Based On Twitter Profiles. https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a119.pdf, 2017. Accessed: 2019-12-01.

[36] Diana Fischer, Carsten Schwemmer, and Kai Fischbach. Terror Management and Twitter: The Case of the 2016 Berlin Terrorist Attack. In *Proceedings of the 2018 International ISCRAM Conference*, 2018.

[37] Scott Fortmann-Roe. Effects of hue, saturation, and brightness on color preference in social networks: Gender-based color preference on the social networking site Twitter. *Color Research & Application*, 38(3):196–202, 2013.

[38] The Apache Software Foundation. Apache Hadoop. https://hadoop.apache.org/, 2018. Accessed: 2019-12-01.

[39] The Apache Software Foundation. Apache Lucene. http://lucene.apache.org/, 2018. Accessed: 2019-12-01.

[40] The Apache Software Foundation. Apache Spark. https://spark.apache.org/, 2018. Accessed: 2019-12-01.

[41] Edward Fox, Marcos André Gonçalves, and Rao Shen. *Theoretical Foundations for Digital Libraries: the 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach*. Morgan & Claypool Publishers, July 2012.

[42] Shen Gao, Xiuying Chen, Piji Li, Zhaochun Ren, Lidong Bing, Dongyan Zhao, and Rui Yan. Abstractive Text Summarization by Incorporating Reader Comments. *arXiv preprint arXiv:1812.05407*, 2018.

[43] Li Geng, Ke Zhang, Xinzhou Wei, and Xin Feng. Soft Biometrics in Online Social Networks: A Case Study on Twitter User Gender Recognition. In *Applications of Computer Vision Workshops (WACVW), 2017 IEEE Winter*, pages 1–8. IEEE, 2017.

[44] Rumi Ghosh, Tawan Surachawala, and Kristina Lerman. Entropy-based Classification of 'Retweeting' Activity on Twitter. *CoRR*, abs/1106.0346, 2011.

[45] Anna Grimm, Lynn Hulse, and Silke Schmidt. Human responses to disasters: A pilot study on peritraumatic emotional and cognitive processing. *Europe's Journal of Psychology*, 8(1):112–138, Mar. 2012.

[46] Kalpa Gunaratna, Krishnaprasad Thirunarayan, and Amit Sheth. Faces: Diversity-aware entity summarization using incremental hierarchical conceptual clustering. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, AAAI'15, pages 116–122. AAAI Press, 2015.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[48] Benjamin Herfort, Svend-Jonas Schelhorn, João Porto de Albuquerque, Alexander Zipf, et al. Does the spatiotemporal distribution of tweets match the spatiotemporal distribution of flood phenomena? A study about the River Elbe Flood in June 2013. In *Proceedings of the 2014 International ISCRAM Conference*, 2014.

[49] Tony Hey, Stewart Tansley, Kristin M Tolle, et al. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, volume 1. Microsoft Research, Redmond, WA, 2009.

[50] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/, 2019. Accessed: 2019-12-01.

[51] Baotian Hu, Qingcai Chen, and Fangze Zhu. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. pages 1967–1972, 2015.

[52] Qunying Huang and Yu Xiao. Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery. *ISPRS International Journal of Geo-Information*, 4(3):1549–1568, Aug 2015.

[53] InternetArchive. Wayback Machine. https://archive.org/web/, 2017. Accessed: 2019-12-01.

[54] Thorsten Joachims. Optimizing Search Engines using Click through Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142. ACM, 2002.

[55] Kaggle. Twitter User Gender Classification. https://www.kaggle.com/crowdflower/-twitter-user-gender-classification, 2016. Accessed: 2019-12-01.

[56] Andrea Kavanaugh, Steven D Sheetz, Riham Hassan, Seungwon Yang, Hicham G Elmongui, Edward A Fox, Mohamed Magdy, and Donald J Shoemaker. Between a Rock and a Cell Phone: Communication and Information Technology Use during the 2011 Uprisings in Tunisia and Egypt. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 5(1):1–21, 2013.

[57] Andrea Kavanaugh, Ziqian Song, Liuqing Li, and Edward Fox. Communication behavior in an emerging democracy. In *Proceedings of the 20th Annual International Conference on Digital Government Research*, pages 445–455, 2019.

[58] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A Few Chirps About Twitter. In *Proceedings of the 1st Workshop on Online Social Networks*, pages 19–24. ACM, 2008.

[59] Robert Krovetz. Viewing Morphology as an Inference Process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202. ACM, 1993.

[60] Canasai Kruengkrai and Chuleerat Jaruskulchai. Generic Text Summarization Using Local and Global Properties of Sentences. In *Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, pages 201–206. IEEE, 2003.

[61] The Knight Lab. TimelineJS: Easy-to-make, beautiful timelines. https://timeline.knightlab.com/, 2019. Accessed: 2019-12-01.

[62] John Lafferty and Chengxiang Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of the 24st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 111–119. ACM, 2001.

[63] Dominic Lasorsa. Transparency and Other Journalistic Norms on Twitter. *Journalism Studies*, 13(3):402–417, 2012.

[64] Victor Lavrenko and W Bruce Croft. Relevance-based Language Models. In *ACM SIGIR Forum*, volume 51, pages 260–267. ACM, 2017.

[65] Sangho Lee and Jong Kim. WarningBird: Detecting Suspicious URLs in Twitter Stream. In *NDSS*, volume 12, pages 1–13, 2012.

[66] Sunshin Lee. *Geo-Locating Tweets with Latent Location Information*. PhD dissertation, Virginia Polytechnic Institute and State University, 13 February 2017. http://hdl.handle.net/10919/75022.

[67] Sunshin Lee, Noha Elsherbiny, and Edward A Fox. A Digital Library for Water Main Break Identification and Visualization. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 335–336. ACM, 2012.

[68] Gil Levi and Tal Hassner. Age and Gender Classification using Convolutional Neural Networks. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[69] Jiwei Li, Alan Ritter, and Eduard Hovy. Weakly Supervised User Profile Extraction from Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 165–174, 2014.

[70] Liuqing Li, Rishabh Anand, and Edward A. Fox. Users, User Roles, and Topics in School Shooting Collections of Tweets. In *Web Archiving and Digital Libraries Workshop*, 2019.

[71] Liuqing Li, He Feng, Wenjie Zhuang, Na Meng, and Barbara Ryder. Cclearner: A deep learning-based clone detection approach. In *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 249–260. IEEE, 2017.

[72] Liuqing Li and Edward A Fox. A study of historical short urls in event collections of tweets. In *Web Archiving and Digital Libraries (WADL 2018), a workshop held in conjunction with ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2018)*, 2018.

[73] Liuqing Li and Edward A. Fox. Understanding Patterns and Mood Changes through Tweets about Disasters. In *Proceedings of the 2019 International ISCRAM Conference*, pages 756–767, 2019.

[74] Liuqing Li and Edward A. Fox. Disaster Response Patterns across Different User Groups on Twitter: A Case Study during Hurricane Dorian. In *Proceedings of the 2020 International ISCRAM Conference*, 2020. In press.

[75] Liuqing Li, Jack Geissinger, William A. Ingram, and Edward A Fox. Teaching Natural Language Processing through Big Data Text Summarization with Problem-Based Learning. *Data and Information Management*, 4(1), 2020. In press.

[76] Liuqing Li, Ziqian Song, Xuan Zhang, and Edward A Fox. A Hybrid Model for Role-related User Classification on Twitter. *arXiv preprint arXiv:1811.10202*, 2018.

[77] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[78] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating Wikipedia by Summarizing Long Sequences. *arXiv preprint arXiv:1801.10198*, 2018.

[79] Wendy Liu and Derek Ruths. What's in a Name? Using First Names as Features for Gender Inference in Twitter. In *AAAI Spring Symposium: Analyzing Microtext*, volume 13, pages 10–16, 2013.

[80] Yan Liu, Sheng-hua Zhong, and Wenjie Li. Query-Oriented Multi-Document Summarization via Unsupervised Deep Learning. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

[81] Hans Peter Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[82] Federico Maggi, Alessandro Frossi, Stefano Zanero, Gianluca Stringhini, Brett Stone-Gross, Christopher Kruegel, and Giovanni Vigna. Two Years of Short URLs Internet Measurement: Security Threats and Countermeasures. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 861–872. ACM, 2013.

[83] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[84] Andrés Moreno, Philip Garrison, and Karthik Bhat. WhatsApp for Monitoring and Response during Critical Events: Aggie in the Ghana 2016 Election. In *Proceedings of the 2017 International ISCRAM Conference*, 2017.

[85] Dmitry Mottl. GetOldTweets3. https://github.com/Mottl/GetOldTweets3, 2018. Accessed: 2019-12-01.

[86] Ahmed Nagy and Jeannie Stamberger. Crowd Sentiment Detection during Disasters and Crises. In *Proceedings of the 2012 International ISCRAM Conference*, pages 1–9, 2012.

[87] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[88] Ani Nenkova. Entity-driven rewrite for multi-document summarization. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[89] Venkata Kishore Neppalli, Murilo Cerqueira Medeiros, Cornelia Caragea, Doina Caragea, Andrea H. Tapia, and Shane E. Halse. Retweetability Analysis and Prediction during Hurricane Sandy. In *Proceedings of the 2016 International ISCRAM Conference*, 2016.

[90] Shirin Nilizadeh, Anne Groggel, Peter Lista, Srijita Das, Yong-Yeol Ahn, Apu Kapadia, and Fabio Rojas. Twitter's Glass Ceiling: The Effect of Perceived Gender on Online Visibility. In *Proceedings of the 10th International Conference on Web and Social Media (ICWSM)*, pages 289–298, 2016.

[91] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2013.

[92] Marco Pennacchiotti and Ana-Maria Popescu. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the 5th International Conference on Web and Social Media (ICWSM)*, pages 281–288, 2011.

[93] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. In *Proceedings of the 17th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, pages 430–438. ACM, 2011.

[94] Jan Pomikálek. *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis, Masarykova Univerzita, Fakulta Informatiky, 2011.

[95] Jay M Ponte and W Bruce Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281. ACM, 1998.

[96] Hemant Purohit and Jennifer Chan. Classifying User Types on Social Media to inform Who-What-Where Coordination during Crisis Response. In *Proceedings of the 2017 International ISCRAM Conference*, 2017.

[97] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying Latent User Attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, pages 37–44. ACM, 2010.

[98] Stephen E Robertson and Steve Walker. Some Simple Effective Approximations to the 2–Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241. Springer-Verlag New York, Inc., 1994.

[99] Alexander M Rush, Sumit Chopra, and Jason Weston. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[100] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[101] Hany M SalahEldeen and Michael L Nelson. Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost? In *International Conference on Theory and Practice of Digital Libraries*, pages 125–137. Springer, 2012.

[102] Axel Schulz, Tung Dang Thanh, Heiko Paulheim, and Immanuel Schweizer. A Fine-Grained Sentiment Analysis Approach for Detecting Crisis Related Microposts. In *Proceedings of the 2013 International ISCRAM Conference*, 2013.

[103] Abigail See, Peter J Liu, and Christopher D Manning. Get To The Point: Summarization with Pointer-Generator Networks. *arXiv preprint arXiv:1704.04368*, 2017.

[104] SocialSecurity. Beyond the Top 1000 Names. https://www.ssa.gov/oact/babynames-/limits.html, 2015. Accessed: 2019-12-01.

[105] Packet Storm. Arabic Names Dictionary. https://packetstormsecurity.com/files/101267-/Arabic-Names-Dictionary.html, 2011. Accessed: 2019-12-01.

[106] Andreas Thalhammer, Nelia Lasierra, and Achim Rettinger. LinkSum: Using link analysis to summarize entity data. pages 244–261, 2016.

[107] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online Human-Bot Interactions: Detection, Estimation, and Characterization. *arXiv preprint arXiv:1703.03107*, 2017.

[108] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, USA, 2017. Curran Associates Inc.

[109] Marco Vicente, Fernando Batista, and Joao Paulo Carvalho. Twitter Gender Classification Using User Unstructured Information. In *Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE, 2015.

[110] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong. Multi-Document Summarization using Sentence-based Topic Models. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 297–300. Association for Computational Linguistics, 2009.

[111] Lu Wang and Wang Ling. Neural Network-Based Abstract Generation for Opinions and Arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, 2016.

[112] Yu Wang, Yang Feng, Jiebo Luo, and Xiyang Zhang. Voting with feet: who are leaving Hillary Clinton and Donald Trump. In *Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM)*, pages 71–76. IEEE, 2016.

[113] Yu Wang, Yuncheng Li, and Jiebo Luo. Deciphering the 2016 US Presidential Campaign in the Twitter Sphere: A Comparison of the Trumpists and Clintonists. In *Proceedings of the 10th International Conference on Web and Social Media (ICWSM)*, pages 723–726, 2016.

[114] Jan Wendland, Christian Ehnis, Rodney Clarke, and Deborah Bunker. Sydney Siege, December 2014: A Visualisation of a Semantic Social Media Sentiment Analysis. In *Proceedings of the 2018 International ISCRAM Conference*, 2018.

[115] Hiroko Wilensky. Twitter as a Navigator for Stranded Commuters during the Great East Japan Earthquake. In *Proceedings of the 2014 International ISCRAM Conference*, 2014.

[116] Han Xiao. bert-as-service. https://github.com/hanxiao/bert-as-service, 2018. Accessed: 2019-12-01.

[117] Chenyan Xiong and Jamie Callan. EsdRank: Connecting Query and Documents through External Semi-Structured Data. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 951–960. ACM, 2015.

[118] Chenyan Xiong and Jamie Callan. Query Expansion with Freebase. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*, pages 111–120. ACM, 2015.

[119] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–64. ACM, 2017.

[120] Jinxi Xu and W Bruce Croft. Query Expansion Using Local and Global Document Analysis. In *ACM SIGIR Forum*, volume 51, pages 168–175. ACM, 2017.

[121] Wei Xu, Ralph Grishman, Adam Meyers, and Alan Ritter. A Preliminary Study of Tweet Summarization using Information Extraction. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics*, page 20, 2013.

[122] Seungwon Yang, Haeyong Chung, Xiao Lin, Sunshin Lee, Liangzhe Chen, Andrew Wood, Andrea L. Kavanaugh, Steven D. Sheetz, Donald J. Shoemaker, and Edward A. Fox. PhaseVis: What, When, Where, and Who in Visualizing the Four Phases of Emergency Management Through the Lens of Social Media. In *Proceedings of the 2013 International ISCRAM Conference*, 2013.

Appendix A

IRB Approval and Supporting Files

A.1 Event-related Webpage Relevance Judgement

A.1.1 WIRB Exemption Approval

February 4, 2019

Edward Alan Fox, PhD, BS, MS
Virginia Tech
Department of Computer Science
114 McBryde Hall M/C 0106
Blacksburg, VA 24061

Dear Dr. Fox:

SUBJECT: IRB EXEMPTION—REGULATORY OPINION
Investigator/Sponsor Contact: Edward Alan Fox, PhD, BS, MS
Sponsor/Institution Protocol #:18-1008
Protocol Title: Event-related Webpage Relevance Judgement

This is in response to your request for an exempt status determination for the above-referenced protocol. Western Institutional Review Board's (WIRB's) IRB Affairs Department reviewed the study under the Common Rule and applicable guidance.

We believe the study is exempt under 45 CFR § 46.104(d)(3), because the research involves only benign behavioral interventions, and no identifiers will be recorded.

This exemption determination can apply to multiple sites, but it does not apply to any institution that has an institutional policy of requiring an entity other than WIRB (such as an internal IRB) to make exemption determinations. WIRB cannot provide an exemption that overrides the jurisdiction of a local IRB or other institutional mechanism for determining exemptions. You are responsible for ensuring that each site to which this exemption applies can and will accept WIRB's exemption decision.

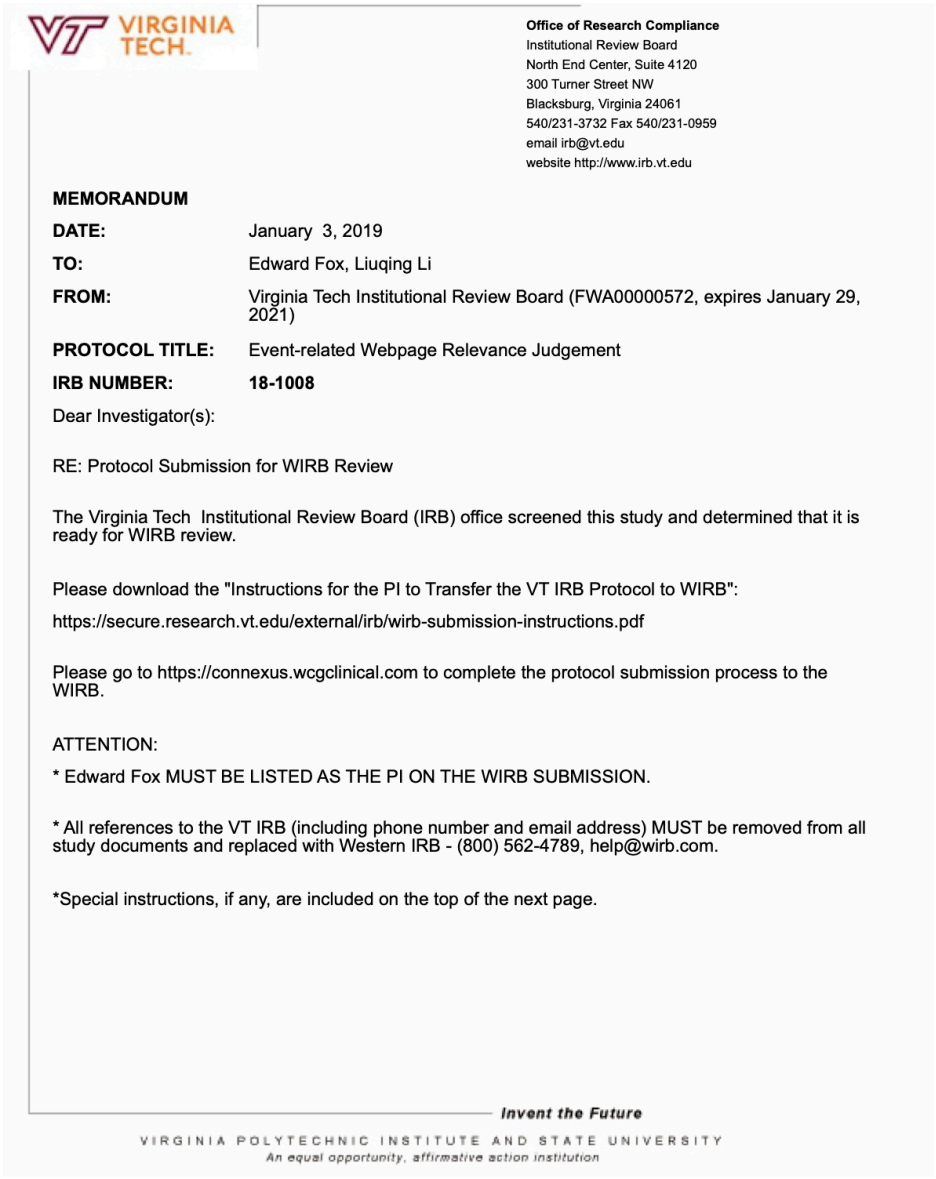
Please note that any future changes to the project may affect its exempt status, and you may want to contact WIRB about the effect these changes may have on the exemption status before implementing them. WIRB does not impose an expiration date on its IRB exemption determinations.

If you have any questions, or if we can be of further assistance, please contact Kelly FitzGerald, PhD, at 360-252-2578, or e-mail RegulatoryAffairs@wirb.com.

KAF:dao
D3 Exemption-Fox (02-04-2019)
cc: WIRB VA Tech
WIRB Accounting
WIRB Work Order #1-1153244-1

Western Institutional Review Board®
1019 39th Avenue SE Suite 120 | Puyallup, WA 98374-2115
Office: (360) 252-2500 | **Fax**: (360) 252-2498 | **www.wirb.com**

A.1.2 VT IRB Authorization Letter



VIRGINIA TECH

Office of Research Compliance
Institutional Review Board
North End Center, Suite 4120
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: January 3, 2019

TO: Edward Fox, Liuqing Li

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires January 29, 2021)

PROTOCOL TITLE: Event-related Webpage Relevance Judgement

IRB NUMBER: **18-1008**

Dear Investigator(s):

RE: Protocol Submission for WIRB Review

The Virginia Tech Institutional Review Board (IRB) office screened this study and determined that it is ready for WIRB review.

Please download the "Instructions for the PI to Transfer the VT IRB Protocol to WIRB":

https://secure.research.vt.edu/external/irb/wirb-submission-instructions.pdf

Please go to https://connexus.wcgclinical.com to complete the protocol submission process to the WIRB.

ATTENTION:

* Edward Fox MUST BE LISTED AS THE PI ON THE WIRB SUBMISSION.

* All references to the VT IRB (including phone number and email address) MUST be removed from all study documents and replaced with Western IRB - (800) 562-4789, help@wirb.com.

*Special instructions, if any, are included on the top of the next page.

IRB SPECIAL INSTRUCTIONS:

*** This study has received funding from the National Science Foundation. ***

Date*	OSP Number	Sponsor	Grant Comparison Conducted?
01/02/2019	PAOWS5UF	National Science Foundation (Title: III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR))	Not required

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irb@vt.edu) immediately.

A.1.3 Online Recruitment

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Online Recruitment
in Research Projects Involving Human Subjects

Title of Project: Event-related Webpage Relevance Judgement

Investigator(s): Principal Investigator:
Dr. Edward A. Fox (Professor of Computer Science)
Email: fox@vt.edu Phone: 540-231-5113

Co-Principal Investigator:
Liuqing Li
Email: liuqing@vt.edu
(Ph.D. Student, Virginia Tech)

I. Introduction

The task is to manually label a large number of webpages as "relevant" or "non-relevant". These manually labeled webpages will help us to improve an information retrieval model so that it is able to retrieve more relevant webpages from a given event collection. Participant will need to complete one or more assignments. In each assignment, there are five webpages that need to be labeled.

II. Requirement

The following eligibility requirement must be satisfied:

1. Participants should be at least 18-years-old.

III. Compensation

Each assignment can be completed within 1 minute. Each participant will be compensated $0.12 per assignment (about $7.25 per hour).

If you have questions, concerns, or complaints, or think this research has hurt you, talk to the research team at the phone number listed above. This research is being overseen by an Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may talk to them at (800) 562-4789, help@wirb.com if you have questions, concerns, or complaints that are not being answered by the research team or you have questions about your rights as a research subject.

A.1.4 Sample Task

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Sample Task
in Research Projects Involving Human Subjects

Title of Project: Event-related Webpage Relevance Judgement

Investigator(s): Principal Investigator:
Dr. Edward A. Fox (Professor of Computer Science)
Email: fox@vt.edu Phone: 540-231-5113

Co-Principal Investigator:
Liuqing Li
Email: liuqing@vt.edu
(Ph.D. Student, Virginia Tech)

If you have questions, concerns, or complaints, or think this research has hurt you, talk to the research team at the phone number listed above. This research is being overseen by an Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may talk to them at (800) 562-4789, help@wirb.com if you have questions, concerns, or complaints that are not being answered by the research team or you have questions about your rights as a research subject.

Please find appended below a sample labeling assignment, which contains five single labeling tasks.

Sample Labeling Assignment

Instructions

In this task, you are requested to read the webpage document excerpts to check whether they are related to the Sandy Hook Elementary School shooting.

- The Sandy Hook Elementary School shooting occurred on December 14, 2012, in Newtown, Connecticut, United States, when 20-year-old Adam Lanza fatally shot 20 children between six and seven years old, as well as six adult staff members.
- When you decide whether a document is related to the specific event, many factors can be taken into account, including but not limited to: shooters, victims, parents, investigations, anniversaries, recoveries.
- Overall reading is NOT always required, which means you can stop reading the document if you are sure of its relevance.
- All DOCIDs are for data management, you can just ignore them.

1. DOCID: 18379981006c07f907de9e8cb3433e4a892335a7

former us ambassador to saudi arabia robert jordan warns not to "read too much into" the relaxed images coming in from secretary of state mike pompeo's meeting with the saudi crown prince over the missing journalist. former us ambassador to saudi arabia robert jordan warns not to "read too much into" the relaxed images coming in from secretary of state mike pompeo's meeting with the saudi crown prince over the missing journalist.

- ○ Relevant
- ○ Non-relevant

2. DOCID: 796ecb91ea9d6d71c09fec51f1d697bb66f07d8a

hartford, conn. (ap) — lawyers are set to ask the connecticut supreme court to reinstate a wrongful death lawsuit against the maker of the rifle used in the 2012 newtown school massacre. justices are scheduled to hear arguments tuesday in an appeal by a survivor and relatives of nine people killed in the shooting. they're trying to sue remington arms, the north carolina company that made the bushmaster ar-15-style rifle used to kill 20 first-graders and six educators at sandy hook elementary school. gunman adam lanza's mother legally purchased the rifle. a lower court judge dismissed the lawsuit, saying federal law shields gun makers from most lawsuits over criminal use of their products. the company denies the lawsuit's allegations that it violated state law by selling such a dangerous weapon to the public.

- ○ Relevant
- ○ Non-relevant

3. DOCID: 3e623180d2f89daea31025290bd3fa5193427d38

miami - miami heat coach erik spoelstra walked off the court following his team's practice saturday morning, took a seat behind a microphone and started speaking his mind. friday's school shooting in newtown, conn., was the primary topic for the heat, even though they were gathered to finish preparing for a game against washington. spoelstra called the deaths of so many innocent victims "despicable." lebron james used the word "devastating." udonis haslem said he tears up whenever he thinks of what happened. said james, the league's reigning mvp: "basketball, this is nothing. these games are nothing compared to when you have a tragedy like that." yahoo can sensationalize anything....."athletes around

the globe....." and it is some nba people? hardly global!! sad that they have to have so many stories about the same thing and none of them are worth reading.....looked at this because i actually forgot yahoo's moronic headlines are usually #$%$!!! who cares what the atheletes, or hollywood folks do. this isn't about them. how is it that we always look for something to talk about other than the real story. a lot of babies died for nothing. that's the only thing that matters. right now everything else can go to hell! in a situation like this, we need accurate reporting, from relevant sources. nobody cares what athletes "around the globe??" think. secondly, media outlets that continue to milk this tragedy for ratings are giving potential school shooters the idea that they will be famous, not infamous. athletes comment on how despicable the act was and are truly broken up over these children and this horrible act. conversely, the hollywood celebrity idiots comment on how we need stricter gun control. jay leno and jimmy fallon had a sober moment at the beginning of their programs last night. they both said that they were going to cancel their programs, but decided to remain on their scheduled program.... i thought it was touching... in the aftermath of friday's newtown school shooting, we've heard tales mostly horrifying and occasionally heroic, from surviving witnesses and mourning citizens alike, but this one lies somewhere in between, all the more unshakeable. one six-year-old sandy hook student played dead in her first-grade classroom, her family pastor said late sunday, with the kind of quick thinking that ended up saving her life but now leaves her with the unshakeable memories of watching all her classmates being shot and killed. ... arguably, next to recess, lunchtime is one part of the day that schoolchildren look forward to the most. it's a time to grab some grub and socialize with friends. dianne brame, the cafeteria manager at hudson elementary school in webster groves, missouri, said she knows each one of [...] rome (reuters) - the egyptian pharaoh ramses iii, whose death has puzzled historians for centuries, had his throat slit in a succession plot concocted by his wife and son, a new analysis suggests. new ct scans have revealed a deep and wide cut that was hidden by the bandages covering the throat of the mummified king, which could not be removed in the interests of preservation, researchers said on tuesday. ... we realize there's only so much time one can spend in a day watching new trailers, viral video clips, and shaky cell phone footage of people arguing on live television. this is why every day the atlantic wire highlights the videos that truly earn your five minutes (or less) of attention. today: the nurse at sandy hook elementary school says she and the school secretary stayed hidden in a supply closet for almost four hours after the connecticut school massacre had ended, leaving 20 children and six adults dead.

- ○ Relevant
- ○ Non-relevant

4. DOCID: 047e1be75e21da96b028b524db8fcec28950b063

police stand before grieving residents following a shooting at sandy hook elementary school on december 14, 2012 in newtown, connecticut. tragedies such as the connecticut school shootings can raise unique issues for parents of teenagers, psychologists say. adolescence is often a turbulent time, when youngsters worry “about controlling their anger or frustrations, or even have fantasies about doing weird or crazy things,” says rich chaifetz, chief executive of compsych, chicago, a provider of employee-assistance programs. “even the healthiest kids at that age are struggling to establish their identity and trying to fit into peer groups,” he says. for troubled kids, news of shootings or other violence can deepen the turmoil. in the wake of a widely publicized tragedy, parents of teens should watch their children for signs of social withdrawal, agitation or anger, or changes in sleep or study habits, dr. chaifetz says. a teenager who is struggling to control his or her feelings may “become more aggressive at home or get into more fights,” dr. chaifetz says. some may isolate themselves as they wrestle with inner conflicts. parents also should avoid projecting their own sadness or anxiety about the tragedy onto their kids, dr. chaifetz says. it’s important, however, to listen closely to their questions or remarks about the incident and draw them out, encouraging them to express feelings in words and helping them gain perspective. if behavioral changes persist, parents should seek help from a therapist or counselor, he adds. the juggle examines the choices and tradeoffs people make as they juggle work and family. the site provides readers with news, insight and tips on parenting, workplace issues, commuting, caregiving and other issues busy readers with

families face. it is also a place for readers to share and compare their own work-and-family experiences and to seek advice and recommendations. the juggle is includes regular contributions from other staffers at the journal. contact the juggle with ideas or suggestions at thejuggle@wsj.com

- ◯ Relevant
- ◯ Non-relevant

5. DOCID: 233830aa363bdbf6bdae17223b1c1ff06dc8f4c0

this makes think sometimes americans have no idea what's going on in these middle eastern countries and americas role in it. i wish they would see that there are numerous children dying everyday and people live in fear all the time. 20 children die here and the whole country is mourning, don't they see what is happening in other countries? seriously. they need to open their eyes. guys, you miss the point. this speech is not about the "fake" tears. this speech is for the parents, families, neighbors, teachers, classmates, sisters, brothers the ones who loved the victims. this great man above shows respect and that it touches him. as a parent you all want to see your kids growing and you want to protect them from the bad, unfortunately some people can't anymore. we should all pray for them and listen to this speech again maybe you started think different. lets not talk take heart. stay calm. ignorance is always weeded out by war. soon today's "liberal" will only be found in history books & balance will be restored to justice & freedom. i pray your success & peace for your family. thanks a lot liberals...you guys have created these protection free zones with your brain dead ideas... you idiots put these stupid signs up at schools advertizing to nuts and murderers that our schools are totally unprotected and then you make sure that there is absolutely, positively no protection for our children in the schools so that when the kids die you can blame the nra and have some talking points for your twisted world view....thanks a lot liberals for leaving our kids unprotected... 'scuse me? that's like saying america doesn't give a shit. true fact: we do care. that's like your best friend's grandpa dying and you don't shed a tear, yet you feel pity for her and try to help her, then some one comes up and says, "oh, you don't care! you're just happy it didn't happem to you!" smh. just shut up, please. every death should be treated with sadness and absolute grief, whether if be here or in other countries. we can see just fine, thank you. no one ever said we didn't care. you should have also mentioned to junegrey, that a japanese general who studied here in a us college warned the emperor of japan, never to invade america "because there will be a gun behind every blade of grass!"

- ◯ Relevant
- ◯ Non-relevant

You must ACCEPT the HIT before you can submit the results.

A.1.5 Waiver of Documentation of Consent

RESEARCH SUBJECT CONSENT FORM

Title:	Event-related Webpage Relevance Judgement
Protocol No.:	Grant #: 479619
Sponsor:	National Science Foundation
Investigator:	Dr. Edward A. Fox
	114 McBryde Hall, Dept. of CS, M/C 0106, Virginia Tech
	Blacksburg, VA 24061, USA
Daytime Phone Number:	540-231-5113
Co-Investigator:	Liuqing Li
	2030 Torgersen Hall, Virginia Tech
	Blacksburg, VA 24061, USA

You are being invited to take part in a research study. Participation is voluntary. You can choose not to take part, or agree to take part and later change your mind. There will be no penalty or loss of benefits to which you are otherwise entitled.

The purpose of this research is to improve our models and software. We will ask you questions and determine your feedback. Your participation in this research will last until you have completed the questionnaire. The only risk is effort involved in the questionnaire. There are no direct benefits to you from your taking part in this research, but the general public may benefit from the information gained during this research. Your alternative is to not take part in the research. We may publish the results of this research. As we are not collecting any identifiable information, your information will be confidential.

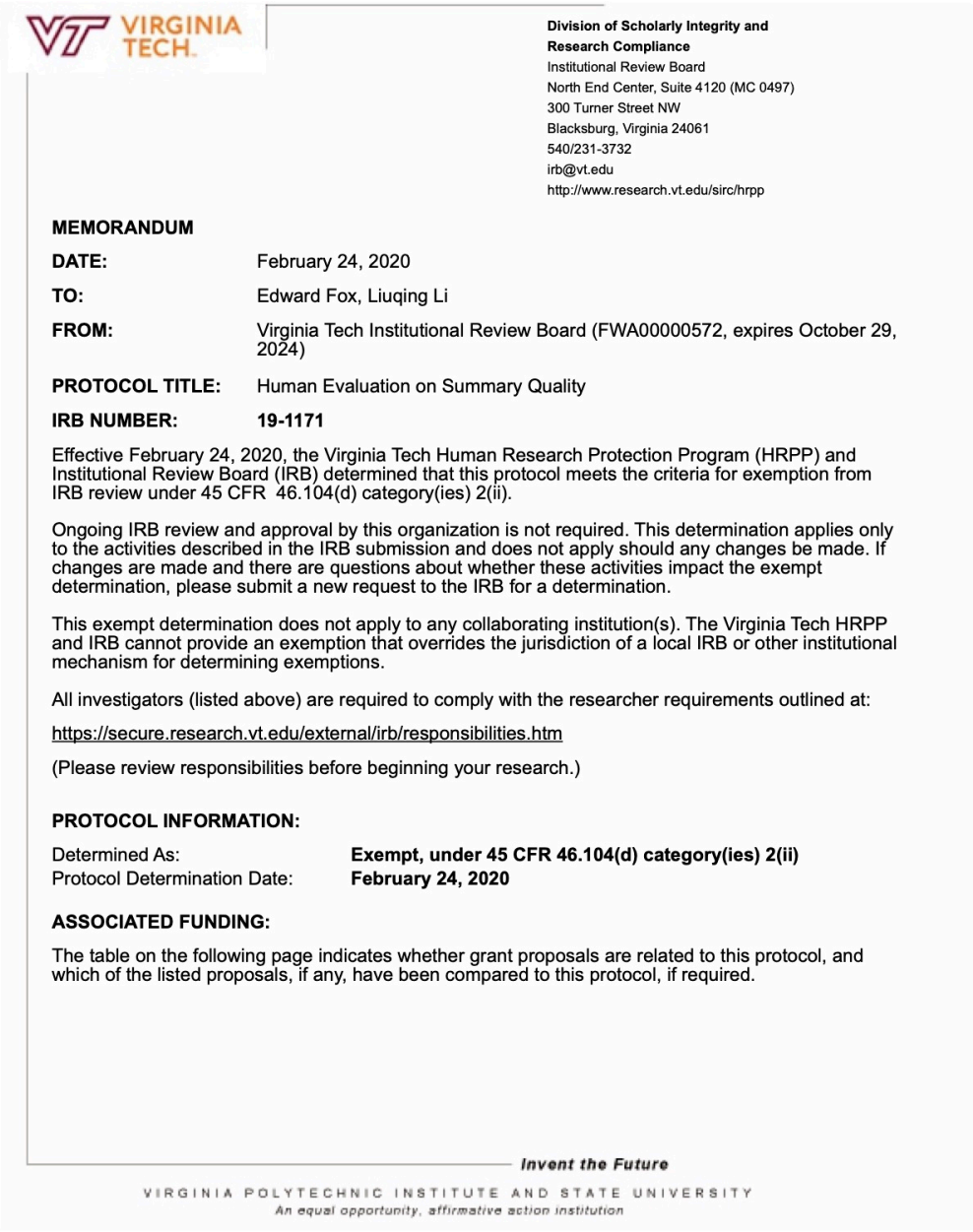
If you have questions, concerns, or complaints, or think this research has hurt you, talk to the research team at the phone number listed above. This research is being overseen by an Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may talk to them at (800) 562-4789, help@wirb.com if you have questions, concerns, or complaints that are not being answered by the research team or you have questions about your rights as a research subject.

For taking part in this research, you may be paid up to a total of $0.12 per assignment.

By continuing in the survey, you are consenting to continue.

A.2 Human Evaluation on Summary Quality

A.2.1 VT IRB Authorization Letter



VIRGINIA TECH

Division of Scholarly Integrity and Research Compliance
Institutional Review Board
North End Center, Suite 4120 (MC 0497)
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732
irb@vt.edu
http://www.research.vt.edu/sirc/hrpp

MEMORANDUM

DATE: February 24, 2020

TO: Edward Fox, Liuqing Li

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires October 29, 2024)

PROTOCOL TITLE: Human Evaluation on Summary Quality

IRB NUMBER: **19-1171**

Effective February 24, 2020, the Virginia Tech Human Research Protection Program (HRPP) and Institutional Review Board (IRB) determined that this protocol meets the criteria for exemption from IRB review under 45 CFR 46.104(d) category(ies) 2(ii).

Ongoing IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities impact the exempt determination, please submit a new request to the IRB for a determination.

This exempt determination does not apply to any collaborating institution(s). The Virginia Tech HRPP and IRB cannot provide an exemption that overrides the jurisdiction of a local IRB or other institutional mechanism for determining exemptions.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

https://secure.research.vt.edu/external/irb/responsibilities.htm

(Please review responsibilities before beginning your research.)

PROTOCOL INFORMATION:

Determined As: **Exempt, under 45 CFR 46.104(d) category(ies) 2(ii)**
Protocol Determination Date: **February 24, 2020**

ASSOCIATED FUNDING:

The table on the following page indicates whether grant proposals are related to this protocol, and which of the listed proposals, if any, have been compared to this protocol, if required.

Date*	OSP Number	Sponsor	Grant Comparison Conducted?
02/11/2020	PAOWS5UF	National Science Foundation (Title: III: Small: Collaborative Research: Global Event and Trend Archive Research (GETAR))	Compared on 02/11/2020

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this protocol is to cover any other grant proposals, please contact the HRPP office (irb@vt.edu) immediately.

A.2.2 Online Recruitment

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Online Recruitment
in Research Projects Involving Human Subjects

Title of Project: Human Evaluation on Summary Quality

Protocol No.: IRB # 19-1171

Investigator(s):
Principal Investigator:
Dr. Edward A. Fox (Professor of Computer Science)
Email: fox@vt.edu Phone: 540-231-5113

Co-Principal Investigator:
Liuqing Li
Email: liuqing@vt.edu
(Ph.D. Candidate, Virginia Tech)

I. Introduction

The task is to manually score a set of machine generated summaries. These manually labeled results will help us to evaluate our proposed multi-document summarization model with other baseline methods. Participants will need to complete one or more assignments. In each assignment, participants are requested to score eight summaries from four aspects and rank all these summaries according to their overall quality.

II. Requirement

The following eligibility requirement must be satisfied:

1. Participants should be at least 18-years-old;
2. Participants should have been granted the Mechanical Turk Masters Qualification.

III. Compensation

Each assignment can be completed within 5 minutes. Each participant will be compensated no more than $0.25 per assignment.

If you have questions, concerns, or complaints, or think this task has hurt you, talk to the research team using the contact information listed above. This task is being overseen by an Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may also contact IRB at Virginia Tech at irb@vt.edu if you have questions, concerns, or complaints that are not being answered by the research team or you have questions about your rights as a research subject.

A.2.3 Sample Task

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Sample Task
in Research Projects Involving Human Subjects

Title of Project: Human Evaluation on Summary Quality

Protocol No.: IRB # 19-1171

Investigator(s): Principal Investigator:
Dr. Edward A. Fox (Professor of Computer Science)
Email: fox@vt.edu Phone: 540-231-5113

Co-Principal Investigator:
Liuqing Li
Email: liuqing@vt.edu
(Ph.D. Candidate, Virginia Tech)

If you have questions, concerns, or complaints, or think this task has hurt you, talk to the research team using the contact information listed above. This task is being overseen by an Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may also contact IRB at Virginia Tech at irb@vt.edu if you have questions, concerns, or complaints that are not being answered by the research team or you have questions about your rights as a research subject.

Please find appended below a sample human evaluation assignment.

Sample Labeling Assignment

Instructions

In this task, you first need to read 8 summaries that are related to the Sandy Hook Elementary School shooting. Then, you are requested to evaluate each summary by rating based on the following four summary criteria:

- Readability: a summary is readable if it is easy to read and understand
- Correctness: a summary is correct if it expresses the opinions in the reviews
- Completeness: a summary is complete if it captures the whole range of opinions in the reviews
- Compactness: a summary is compact if it does not repeat information

At last, you need to rank all summaries based on their overall quality.

Summary 1

it has been nearly four months since the massacre at sandy hook elementary school and now legislative leaders are finally planning to pass what they call the most far-reaching gun-legislation package in the country. in connecticut, the new gun laws will require new state-issued eligibility certificates for the purchase of any rifle, shotgun or ammunition. laws will also mandate that offenders convicted of any of more than 40 weapons offenses register with the state. "in a time of universal deceit, telling the truth becomes a revolutionary act" - george orwell, author ... "all truth passes through three stages. first: it is ridiculed, second: it is violently opposed, and third: it is accepted as self-evident" – arthur schopenhauer, philosopher. connecticut state police officers search outside st. rose of lima roman catholic church in newtown, connecticut, on sunday, december 16, after a threat prompted authorities to evacuate the building. investigators found nothing to substantiate the reported threat, a police official said, declining to provide additional details. you can add location information to your tweets, such as your city or precise location, from the web and via third-party applications. you always have the option to delete your tweet location history. learn more. you think you've seen the worst that can happen, and then a day like this comes and breaks your heart all over again. just last week in our backyard (portland in neighboring state, oregon), a gunman killed two people and then himself at a shopping mall. today, the violence is targeted at children, while they were in school, and all these just a few weeks, days to christmas. …

Readability

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Correctness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Completeness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Compactness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Summary 2

it has been nearly four months since the massacre at sandy hook elementary school and now legislative leaders are finally planning to pass what they call the most far-reaching gun-legislation package in the country. in fact, i am now smelling a rat, and am seriously looking at this as an operation conducted for the sole purpose of scaring the crap out of the american people so that they will blindly accept new gun control legislation and surrender their right to bear arms. it is entitled: "the sandy hook school shooting in newtown connecticut..2nd shooter, guns, etc., the story is not adding up. the church held sunday services following last week's mass shooting at sandy hook elementary school in newtown. firefighters attach black bunting to a fire truck as a memorial at the fire station down the street from the sandy hook elementary school in newtown, connecticut, on saturday, december 15. police officers keep guard at the entrance to the street leading to the sandy hook elementary school on saturday, december 15. police officers stand at the entrance to the street leading to the sandy hook elementary school on december 15. corinne mclaughlin, a student at the university of hartford, bows her head during a candlelight vigil at hartford, connecticut's bushnell park on friday, december 14, honoring the students and teachers who died at sandy hook elementary school in nearby newtown earlier in the day. a child and her mother leave a staging area

outside sandy hook elementary school in newtown, connecticut, on december 14. faisal ali, right, of colorado springs, colorado, joins other people outside the white house on december 14 to participate in a candlelight vigil to remember the victims of the sandy hook elementary school shooting. j. paul vance, center, briefs the media on the elementary school shootings during a press conference at treadwell memorial park on december 14 in newtown. people weep and embrace near sandy hook elementary school on friday, december 14. a woman leans on a man as she weeps near sandy hook elementary school on december 14. president barack obama wipes a tear as he speaks about the shooting at sandy hook elementary school during a press briefing at the white house on december 14. …

Readability

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Correctness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Completeness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Compactness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Summary 3

a woman weeps near the site of a shooting at sandy hook elementary school on december 14. the information articles in this blog are free for everyone to take and spread to others. the singer is now speaking out on the decision — and she's being incredibly empathetic and sensitive. supporters of gun control often cite australia's dramatic response to a 1996 shooting spree in the southern state of tasmania that killed 35 people. nts notes: sherrie has done her homework quite well.. #neveragain#parkland#schoolshooting amen (ap photo/alex brandon) did it work? poll results suggest most americans wouldn't agree. connecticut chief medical examiner h. frontman nate ruess, dr. "all truth passes through three stages. § 107. america stands armed. learn more. source. massacre. the track faced an enormous radio airplay decline following the horrific school shooting on dec. yes indeed the facts do not add up.. the threat came in… . the associated press. by the next year, the ban had become law. the total included the shooter, who media said was a 24-year-old man. the suspected gunman entered the school as children were gathered in their classrooms for morning meeting. the bill is expected to go to both houses of the general assembly on wednesday; passage seemed assured. 15 when programmers and djs pulled it from rotation. you always have the option to delete your tweet location history. below are photos of the scene at the school immediately after the shooting. it's worth noting that 'die young' isn't actually about dying young, but about living life to the fullest, though the title is repeated in the chorus: "i hear your heart beat to the beat of the drums / oh what a shame that you came here with someone / so while you're here in my arms / let's make the most of the night like we're gonna die young / we're gonna die young / we're gonna die young / let's make the most of the night like we're gonna die young." …

Readability

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Correctness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Completeness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Compactness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Summary 4

it has been nearly four months since the massacre at sandy hook elementary school and now legislative leaders are finally planning to pass what they call the most far-reaching gun-legislation package in the country. when you take all the elements and compare it, i think you could judiciously say this is the strongest bill in the nation.". i came across a fascinating article from a fellow researcher named sherrie, who writes the great blog: sherrie questioning all, at www.sherryquestioningall.blogspot.com. yes indeed the facts do not add up.. i have my own

comments and thoughts to follow:. the fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors. canadian male, age 58.(hard to believe i am that old.) i am in this for the truth and where ever the truth leads me! ********** northerntruthseeker is fully responsible for all content in this blog. if any errors or omissions are deemed necessary for correction of material, it will be placed into the articles affected, and notice will be put up for all readers. connecticut state police officers walk out of st. zulma sein is hugged by a family member outside of the entrance to the sandy hook school on saturday. members of the media converge on december 14 in front of an apartment at 1313 grand street in hoboken, new jersey. faisal ali, right, of colorado springs, colorado, joins other people outside the white house on december 14 to participate in a candlelight vigil to remember the victims of the sandy hook elementary school shooting. people weep and embrace near sandy hook elementary school on friday, december 14. a woman weeps near sandy hook elementary school on december 14. a man takes in the scene near sandy hook elementary school on december 14….

Readability

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Correctness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Completeness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Compactness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Summary 5

newtown, connecticut (reuters) - a heavily armed gunman opened fire at a connecticut elementary school on friday, killing 26 people including 20 children in the latest in a series of shooting rampages across the united states this year, u.s. media reported. a mother runs with her children as police above canvass homes in the area following a shooting at the sandy hook elementary school in newtown, conn., about 60 miles (96 kilometers) northeast of new york city, friday, dec. in this photo provided by the newtown bee, connecticut state police lead children from the sandy hook elementary school in newtown, conn., following a reported shooting there friday, dec. people embrace at a firehouse staging area for family around near the scene of a shooting at the sandy hook elementary school in newtown, conn., about 60 miles (96 kilometers) northeast of new york city, friday, dec. hartford, connecticut (reuters) - members of a connecticut panel charged with recommending ways to prevent gun violence in schools after last year's massacre at sandy hook elementary school on friday said a state attorney's report failed to address the role of the shooter's mental health in the attack. abc news has learned that investigators have seized computers belonging to adam lanza from the home he shared with his mother nancy, the same place he killed her before going to sandy hook elementary school, where he fatally shot students in two first-grade classes along with teachers and staff. noah's twin sister, arielle, who was assigned to a different classroom, survived the killing frenzy by 20-year-old adam lanza that left 20 children and six adults dead last week at sandy hook elementary in an attack so horrifying that authorities could not say whether the school would ever reopen….

Readability

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Correctness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Completeness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Compactness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Summary 6

it has been nearly four months since the massacre at sandy hook elementary school and now legislative leaders are finally planning to pass what they call the most far-reaching gun-legislation

package in the country. "in a time of universal deceit, telling the truth becomes a revolutionary act" - george orwell, author ... "all truth passes through three stages. yes indeed the facts do not add up.. i have my own comments and thoughts to follow:. i for one am seriously now smelling a rat... this has all the earmarks of being an operation for the sole purpose of indeed scaring the public into accepting gun control and a repeal of the second amendment of the us constitution!. again, as i stated before, i do feel for the families of this massacre. the fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors. canadian male, age 58.(hard to believe i am that old.) i am in this for the truth and where ever the truth leads me! ********** northerntruthseeker is fully responsible for all content in this blog. if any errors or omissions are deemed necessary for correction of material, it will be placed into the articles affected, and notice will be put up for all readers. the information articles in this blog are free for everyone to take and spread to others. in these times of uncertainty, people deserve the truth no matter how harsh it may be! connecticut state police officers walk out of st. paul vance addresses the press on december 15. j. paul vance, center, briefs the media on the elementary school shootings during a press conference at treadwell memorial park on december 14 in newtown. people weep and embrace near sandy hook elementary school on friday, december 14. a woman leans on a man as she weeps near sandy hook elementary school on december 14….

Readability

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Correctness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Completeness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Compactness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Summary 7

a gunman opened fire at the sandy hook elementary school in newtown, connecticut, on friday morning, killing 26 people — including 20 children. two law enforcement sources briefed on the investigation confirmed to reuters the shooter had been identified as adam lanza, 20. with the death toll at 26, the massacre in newtown is the second-deadliest school shooting in u.s. history, behind the 2007 virginia tech mass shooting that left 32 dead. hartford, connecticut, mayor padro segarra speaks emotionally about the students and teachers who died earlier in the day at sandy hook elementary school in nearby newtown at a candlelight vigil at bushnell park in hartford on friday. president obama is scheduled to deliver a statement on today's shooting at the sandy hook elementary school in newtown, connecticut at 3:15 pm et from the white house. the nra proposal would take one of every seven u.s. police officers off the streets during school days, based on a reuters analysis of u.s. government data. paul simon performed his classic track the sound of silence at the funeral of a teacher who died in the school shooting in connecticut on 14 december. however, a prosecutor's report released last month concluded that lanza acted alone and took the motive for the bloodbath to his grave. in this photo provided by the newtown bee, connecticut state police lead children from the sandy hook elementary school in newtown, conn., following a reported shooting there friday, dec. 14, 2012. a connecticut judge ruled thursday that a wrongful-death law suit filedby families of victims killed at sandy hook elementary school against the manufacturer of the rifle used in the 2012 shooting in newtown, conn., can proceed….

Readability

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Correctness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Completeness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Compactness

☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Summary 8
the sandy hook elementary school shooting occurred on december 14, 2012, in newtown, connecticut, united states, when 20-year-old adam lanza shot and killed 26 people, including 20 children between six and seven years old, and six adult staff members. a november 2013 report issued by the connecticut state attorney's office concluded that lanza acted alone and planned his actions, but provided no indication why he did so, or why he targeted the school. a report issued by the office of the child advocate in november 2014 said that lanza had asperger syndrome and as a teenager suffered from depression, anxiety and obsessive-compulsive disorder, but concluded that they had "neither caused nor led to his murderous acts." the report went on to say, "his severe and deteriorating internalized mental health problems ... violent crime had been rare in the town of 28,000 residents; there was only one homicide in the town in the ten years before the school shooting.under connecticut law at the time, the 20-year-old lanza was old enough to carry a long gun, such as a rifle or shotgun, but too young to own or carry handguns. she was later treated at danbury hospital.a nine-year-old boy stated that he heard the shooter say: "put your hands up!" and someone else say "don't shoot!" he also heard many people yelling and many gunshots over the intercom while he, his classmates, and his teacher took refuge in a closet in the gymnasium. when she reached her mother, she said, "mommy, i'm okay, but all my friends are dead." the child described the shooter as "a very angry man." a girl hiding in a bathroom with two teachers told police that she heard a boy in the classroom screaming, "help me! police also investigated whether lanza was the person who had been in an altercation with four staff members at sandy hook school the day before the massacre. connecticut state police indicated their concern about misinformation being posted on social media sites and threatened prosecution of anyone involved with such activities.a large quantity of unused ammunition was recovered inside the school along with three semi-automatic firearms found with lanza: a .223-caliber bushmaster xm15-e2s rifle, a 10mm glock 20sf handgun, and a 9mm sig sauer p226 handgun. police found that lanza had downloaded videos relating to the columbine high school massacre, other shootings and two videos of suicide by gunshot.details of the investigation were reported by law enforcement officials at a meeting of the international association of police chiefs and colonels held during the week of march 11, 2013….
Readability ☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely **Correctness** ☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely **Completeness** ☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely **Compactness** ☐ Not at all ☐ Not very ☐ Somewhat ☐ Very ☐ Absolutely

Overall Ranking based on Overall Quality	
Summary	Ranking Position
Summary 1	
Summary 2	
Summary 3	
Summary 4	
Summary 5	
Summary 6	
Summary 7	
Summary 8	

A.2.4 Waiver of Documentation of Consent

RESEARCH SUBJECT CONSENT FORM

Title:	Human Evaluation on Summary Quality
Protocol No.:	IRB # 19-1171
Sponsor:	National Science Foundation
Investigator:	Dr. Edward A. Fox
	114 McBryde Hall, Dept. of CS, M/C 0106, Virginia Tech
	Blacksburg, VA 24061, USA
Daytime Phone Number:	540-231-5113
Co-Investigator:	Liuqing Li
	2030 Torgersen Hall, Virginia Tech
	Blacksburg, VA 24061, USA

You are being invited to take part in a research study. Participation is voluntary. You can choose not to take part, or agree to take part and later change your mind. There will be no penalty or loss of benefits to which you are otherwise entitled.

The purpose of this task is to manually evaluate summaries generated by our proposed model and other baseline methods. We will ask you questions and determine your feedback. Your participation in this task will last until you have completed all questions in this assignment. There are no direct benefits to you from your taking part in this task, but the general public may benefit from the information gained during this task. Your alternative is to not take part in the task. We may publish the results based on your assessments. As we are not collecting any identifiable information, no information about you will be recorded or shared.

If you have questions, concerns, or complaints, or think this task has hurt you, talk to the research team using the contact information listed above. This task is being overseen by an Institutional Review Board ("IRB"). An IRB is a group of people who perform independent review of research studies. You may also contact IRB at Virginia Tech at irb@vt.edu if you have questions, concerns, or complaints that are not being answered by the research team or you have questions about your rights as a research subject.

For taking part in this task, you may be paid up to a total of $0.25 per assignment.

By clicking the "Next" button, you agree to participate in our study.

Appendix B

Additional Results of Experiments

Table B.1: AP scores with different URL length thresholds across all collections

	D1	**D2**	**D3**	**D4**	**D5**
Relevance Ratio	0.477	0.767	0.724	0.814	0.660
URL Length Threshold	**AP Score**				
30	0.479	0.780	0.758	0.841	0.693
40	0.480	0.787	0.763	0.847	0.702
50	0.481	0.797	0.762	0.848	0.703
60	0.482	0.801	0.760	0.847	0.706
70	0.491	0.813	0.757	0.854	0.715
80	0.498	**0.824**	0.775	0.863	0.721
90	0.503	0.819	**0.777**	**0.864**	0.722
100	0.504	0.810	0.762	0.860	**0.726**
110	0.505	0.813	0.752	0.852	0.718
120	0.498	0.805	0.751	0.847	0.712
130	0.506	0.805	0.737	0.838	0.700
140	**0.512**	0.801	0.736	0.832	0.688
150	0.498	0.789	0.734	0.827	0.673

Table B.2: Accuracy scores with different K top words in AF1 (tweet)

Parameters in AF1 (tweet)	**Accuracy**
1 top words	0.886
5 top words	0.890
10 top words	**0.899**
20 top words	**0.899**

The best values of k are 10 and 20. It seems that with a smaller k, the roles of users are not differentiated sufficiently. When we set k to 30, 50, or 100, however, the computation time is much longer, and there is no significant increase in effectiveness.