TWEETSUM: Event-oriented Social Summarization Dataset

Ruifang He **Liangliang Zhao** **Huanyu Liu**
School of Computer Science and Technology
Tianjin University
Tianjin, China
{rfhe, liangliangzhao, huanyuliu}@tju.edu.cn

Abstract

With social media becoming popular, a vast of short and noisy messages are produced by millions of users when a hot event happens. Developing social summarization systems becomes more and more critical for people to quickly grasp core and essential information. However, the publicly available and high-quality large scale dataset under social media situation is rare. Constructing such corpus is not easy and very expensive since short texts have very complex social characteristics. Though there exist some datasets, they only consider the text on social media and ignore the potential user relations relevant signals on social network. In this paper, we construct TWEETSUM, a new event-oriented dataset for social summarization. The original data is collected from twitter and contains 12 real world hot events with a total of 44,034 tweets and 11,240 users. We create expert summaries for each event, and we also have the annotation quality evaluation. In addition, we collect additional social signals (i.e. user relations, hashtags and user profiles) and further establish user relation network for each event. To our knowledge, it is the first event-oriented social summarization dataset that contains social relationships. Besides the detailed dataset description, we show the performance of several typical extractive summarization methods on TWEETSUM to establish baselines. For further researches, we will release this dataset to the public.

1 Introduction

Social media has become an important real-time information source, especially during emergencies, natural disasters and other hot events. According to a new Pew Research Center survey, social media has surpassed traditional news platforms (such as TV and radio) as a news source for Americans: about two-thirds of American adults (68%) get news via social media. Among all major social media sites, Twitter is still the site Americans most commonly use for news, with 71% of Twitter's users get their news from Twitter. However, it can often be daunting to catch up with the most recent contents due to high volume and velocity of tweets. Hence, social summarization aiming to acquire the most representative and concise information from massive tweets when a hot event happens is particularly urgent.

In recent years, many large-scale summarization datasets have been proposed such as New York Times (Sandhaus, 2008), Gigaword (Napoles et al., 2012), NEWSROOM (Grusky et al., 2018) and CNN/DAILYMAIL (Nallapati et al., 2016). However, most of these datasets focus on formal document summarization. Actually, social media text has many different characteristics from formal documents: **1) Short.** The length of a tweet is limited to 140 characters, which is much shorter than formal document. **2) Informal.** Tweets usually contains informal expressions such as abbreviations, typos, special symbols and so on, which make tweets more difficult to deal with. **3) Social signal.** There are different kinds of social signals on social media such as hashtags, urls and emojis. **4) Potential relations.** Tweets are generated by users and hence have potential connections through user relationship. Because of these characteristics, traditional summarization methods often do not perform well on social media.

Proceedings of the 28th International Conference on Computational Linguistics, pages 5731–5736
Barcelona, Spain (Online), December 8-13, 2020

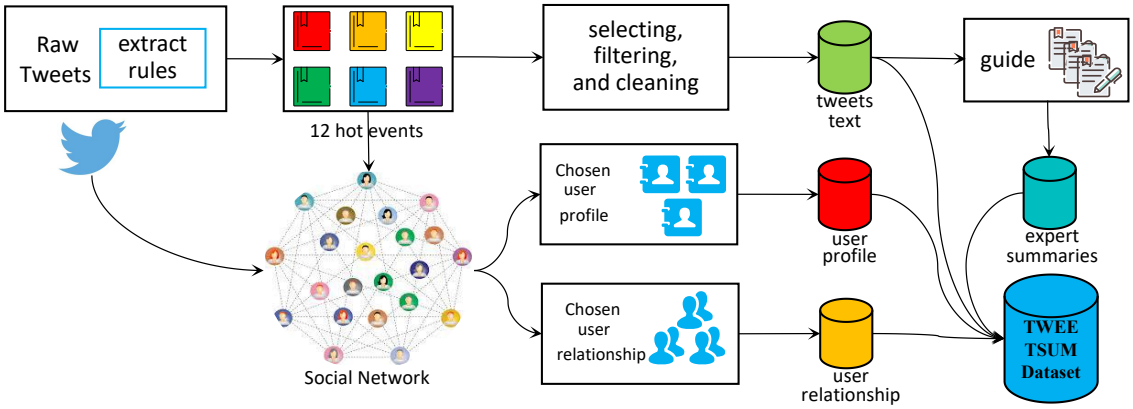


Figure 1: Diagram of the process for creating the TWEETSUM dataset.

Though there exists some social media summarization datasets(Hu et al., 2015; Li et al., 2016; P.V.S. et al., 2018; Duan et al., 2012; Cao et al., 2017; Nguyen et al., 2018). However, these datasets only consider the text on social media and ignore the potential social signals on social network. In a social context, the interactions between friends are obviously different from that between strangers. This phenomenon demonstrates that social relationship can affect user behavior patterns and consequently affect the tweets content they post. This inspires us to consider integrating social relations relevant signals when analyzing social information.

In this paper, we construct an event-oriented large-scale dataset with user relations for social summarization called TWEETSUM. It contains 12 real world hot events with a total of 44,034 tweets and 11,240 users. In summary, this paper provides the following contributions: (1) Construct an event-oriented social media summarization dataset, TWEETSUM, which contains social signals. To our knowledge, it is the first summarization dataset that contains user relationships relevant social signals, such as hashtags and user profiles and so on; (2) Create expert summaries for each social event and verified the existence of sociological theory in real data, including social consistency and contagion; (3) Evaluate the performance of typical extractive summarization models on our TWEETSUM dataset to provide benchmarks and validate the effectiveness of the dataset.

2 TWEETSUM DATASET

2.1 Task and Data Collection

Tweets summarization aims to find a group of representative tweets for a specific topic. Given a collection of tweets about an event $T = [t_1; t_2; .....; t_m]$, our goal is to extract a set of tweets $S = [s_1; s_2; ...; s_n]$ $(n << m)$, which contain as much important information and as little redundant information as possible at the same time (Rudrapal et al., 2018).

The dataset is created using the public Twitter data collected by University of Illinois[1] as the raw data and the overall creation process is shown as Figure 1. The detailed process of data collection is shown summarized as follows: (1) We first select twelve hot events happened in May, June and July 2011, including sports, technology and science, natural disasters, politics, terrorist attacks and so on. The events selected should satisfy the following conditions: (i) Widely spread on the Internet and cause a heated discussion on social media; (ii) Last longer than 30 days; (iii) Be impressive to news providers. (2) Since each hot event can have multiple hashtags, such as "#nba" and "#nbafinals", we then search the tweets which contain any of these hashtags or any of the keywords obtained by getting rid of "#" from hashtags. (3) After obtaining the event-oriented data, we carefully preprocess the data as follows: (i) Merge identical tweets; (ii) Remove tweets whose length are shorter than 3 other than hashtags, keywords, mentions, URLs and stop words; (iii) Delete tweets whose author has no connection with others. (4) For each event, we further collect user profiles and user relationships. We filter users whose

[1]https://wiki.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug20

degree is smaller than 1 and obtain 11,240 users with their relations. Finally, we collect user profiles including User ID, historical tweets records, Tweet timestamp and Retweet Count.

2.2 Expert Summaries Creation

To verify summarization performance, we create expert summaries for each event. Specifically, for each of the 12 events, we ask annotators to select most representative 25 tweets as expert summary. Since different annotators have different understandings of the same event, we ask 4 annotators to create expert summary individually for each event in order to eliminate the subjective factors of users. To evaluate the quality of all expert summaries, we further ask 3 other annotators to score all summaries in range $[1, 5]$ based on the coverage, diversity and readability. If only 0-6 tweets are satisfactory, the summary is scored as 1, 6-12 tweets as 2 scores, 12-18 tweets as 3 scores, 18-24 tweets as 4 scores. If all tweets are good, the score is 5. We remain the summaries with scores greater than or equal to 3 and require modifications to those low-quality summaries until they meet the criteria. To ensure the agreement of multiple expert summaries of each event, we conduct the mutual evaluation among them, and the results are shown in Figure 2.

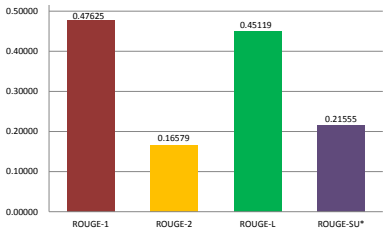


Figure 2: ROUGE scores of expert summaries.

3 Data Properties and Analysis

In this section, we introduce each part of the TWEETSUM dataset in detail. The dataset consists of 12 hot events, each of which contains four parts: tweets text, user relations, user profiles, and manually created expert summaries. The detailed statistics of each part are shown in Table 1. Due to the limited space, we only show the statistics of four events.

Events	Osama	Casey	Mavs	Oslo
# of Tweets	4,780	6,241	5,817	4,571
Avg.Length per Tweet	15.09	15.35	15.10	14.58
# of Hashtags	10,111	4,154	29,962	8,741
Avg.Hashtags per Tweet	2.11	0.66	5.15	1.91
# of Users	1,309	1,318	442	1,026
Avg.Degree of Users	6.20	7.82	6.53	10.10
Network Density(%)	0.47	0.59	0.36	0.98
p-value of contagion	1.258e-42	5.625e-172	2.433e-16	4.417e-32

Table 1: The detailed statistics for the TWEETSUM dataset.

Tweet text is the textual content of tweets, whose average length in all 12 events is 15.22 words. The number of tweets and average length per tweet in each event is shown in the first two rows in Table 1. In addition, hashtags in tweets contain important clues that can help understand the semantics of the tweets. Therefore, we also analyzed the distribution of hashtags in tweets, as shown in the third and fourth rows of Table 1.

User relations are the unique property of our dataset compared with other summarization datasets. We collect users and their corresponding relations in each event to construct social networks and further analyze the statistics of the generated network, which is shown in the second part of Table 1. Indicated by social theories, i.e. consistency (Abelson, 1983) and homophily (Mcpherson et al., 2001), social relations will affect the user behaviors and consequently influence the content. We visualize the structure of one social network as shown in Figure 3. Users and their relationships constitute an undirected graph $G(V, E)$, where V is user set and E is relation set.

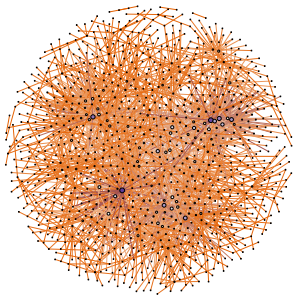
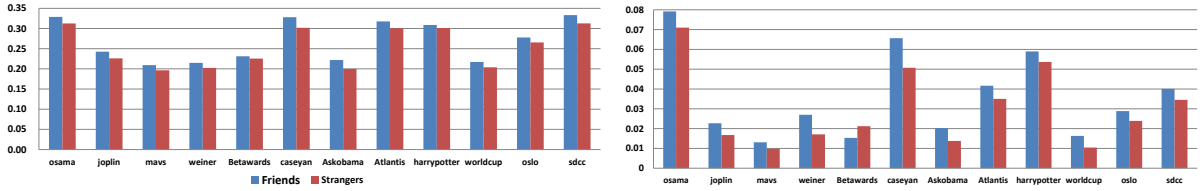


Figure 3: Visualization of one user social networks.



(a) 1-gram. (b) 2-gram.

Figure 4: 1-gram and 2-gram overlap between friends and strangers under 12 events.

We observe some homophily groups, which may indicate that users being friends tend to share similar opinions or topics. We further analyze the words overlap ratio between friends and strangers respectively. Figure 4 shows the 1-gram and 2-gram overlap ratio under all 12 events. The average 1-gram and 2-gram overlap ratio between friends (26.92% and 4.01%) are consistently higher than that between strangers (25.40% and 3.45%), which demonstrates the impact of social relations on user behavior. We further conduct two sample t-test where null hypothesis H_0 means there is no difference between tweets posted by friends and those randomly selected tweets, while alternative hypothesis H_1 means the distance between tweets posted by friends is smaller than that of randomly selected tweets. We define the distance of two tweets as : $D_{ij} = ||t_i - t_j||_2$, where t_i is the TFIDF representation of the i-th tweet. The p-value shown in Table 1 suggests to reject H_0, which proves the influence of social relation on the tweet content.

User profiles include the following information: 1) User ID. This is the unique identity of each user. 2) Historical tweet records. Tweets posted by users contain lots of user information, from which we can obtain abundant information, such as user interests and preference. 3) Tweet timestamp records the time of the creation of tweets. 4) Retweet Count is used to reflect the popularity of each tweet.

Expert summary has been described in section 2.2.

4 Experiments

4.1 Compared methods

To verify the effectiveness of our TWEETSUM dataset, we choose some typical extractive summarization methods as baseline methods. (1) **Expert**: denotes the average mutual assessment of expert summaries. (2) **Random**: selects tweets randomly from each hot events set to form summaries. (3) **LexRank**: adopts PageRank-like algorithm to rank and select tweets (Erkan and Radev, 2004). (4) **LSA**: exploits SVD (Gong and Liu, 2001) and selects the highest-ranked tweets from each singular vector. (5) **MDSS**: uses two-level sparse representation model (Liu et al., 2015). (6) **DSDR**: uses data reconstruction method (He et al., 2012). (7) **SNSR**: integrates the social relationship into a unified sparse coding optimization framework(He and Duan, 2018). (8) **Fine-Tuning BERT**: We fine tune the pre-trained BERT model (Devlin et al., 2019) and learn representations for tweets. Summaries are choosen according to cosine similarity.

4.2 Evaluation Methodologies

ROUGE is the most commonly used evaluation metric in summarization task, which counts number of overlapping units such as n-gram, word sequences and word pairs between the machine-generated summary and reference summaries. (Lin and Hovy, 2003) proposed different ROUGE matrices. Here, we use the F-measures of ROUGE-1, 2, ROUGE-L and ROUGE-SU* as our evaluation metric.

4.3 Results and Discussions

Table 2 shows the performance of different baselines on our dataset. As we can see, all of these models have improvement over the Random baseline, especially SNSR model, which achieves the best performance and outperforms the Random baseline with an absolute gain of 3.31% R-1, 4.45% R-2, 3.57% R-L and 3.39% R-SU*. The main reason is that SNSR captures social relations among tweets.

However, the improvement of other models are not as significant as SNSR. The reason is that most of these models are designed for formal documents such as news articles, thus not suitable for tweets. The neural network based model BERT has a strong ability in feature extraction. However, the BERT model still lags behind the best model. There are mainly three reasons: 1) Learning an efficient tweet representation still remains a big challenge since tweets are short and noisy. 2) It only considers text content and ignores relations among tweets. 3) The summary selection strategy is relatively simple. To further prove the effectiveness of social relations, we remove the relation component of SNSR (indicated by -social), which brings performance deterioration.

Methods	R-1	R-2	R-L	R-SU*
Expert	0.47625	0.16579	0.45119	0.21555
Random	0.41574	0.09440	0.39227	0.16596
LexRank	0.42132	0.13302	0.39965	0.18192
LSA	0.43524	0.13077	0.41347	0.18197
MDSS	0.42119	0.10059	0.40101	0.16686
DSDR	0.43335	0.13106	0.41055	0.17264
SNSR	**0.44886**	**0.13891**	**0.42800**	**0.19990**
SNSR-social	0.43578	0.10668	0.41056	0.18267
BERT	0.42577	0.10142	0.39953	0.17578

Table 2: The ROUGE scores of different baselines.

As we discussed above, there are multiple types of social signals in social media which can provide various kinds of additional information. These heterogeneous signals contain a large amount of information, which is conducive to generating summaries. This inspires us to further explore to integrate these additional signals to improve social summarization.

5 Conclusion and Future Work

In this paper, we construct an event-oriented social media summarization dataset, called TWEETSUM. To better explore how social signals help social summarization, we filter some outliers, keeping social network dense to some extent, and conduct experiments to verify the influence of social signals on user generated content. We further analyze the characteristics of this dataset in detail and validate the influence of social relations on tweets content selection. Both traditional summarization methods and neural network-based methods are tested on our dataset.

In the future, the dataset can be further expanded to include more events as well as more various social signals. In addition, manually annotating data is an expensive and labor-consuming task, therefore we will further try to explore approaches to construct social summarization dataset automatically. More research space can be extended based on this dataset and we hope the TWEETSUM dataset can foster the development of social summarization.

References

Robert P. Abelson. 1983. Whatever became of consistency theory? *Personality & Social Psychology Bulletin*, 9(1):37–64.

Ziqiang Cao, Chengyao Chen, Wenjie Li, Sujian Li, Furu Wei, and Ming Zhou, 2017. *TGSum: Build Tweet Guided Multi-Document Summarization Dataset: Natural Language Processing and Beyond*, pages 401–417. 11.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Yajuan Duan, Zhumin Chen, Furu Wei, Ming Zhou, and Heung-Yeung Shum. 2012. Twitter topic summarization by ranking tweets using social influence and content quality. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 763–780, 12.

G. Erkan and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 1925, New York, NY, USA. Association for Computing Machinery.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ruifang He and Xingyi Duan. 2018. Twitter summarization based on social network and sparse reconstruction. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 5787–5794. AAAI Press.

Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He. 2012. Document summarization based on data reconstruction. *Proceedings of the National Conference on Artificial Intelligence*, 1:620–626, 01.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal, September. Association for Computational Linguistics.

Chen Li, Zhongyu Wei, Yang Liu, Yang Jin, and Fei Huang. 2016. Using relevant public posts to enhance news article summarization. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 557–566, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 7178, USA. Association for Computational Linguistics.

He Liu, Hongliang Yu, and Zhi-Hong Deng. 2015. Multi-document summarization based on two-level sparse representation model. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 196–202. AAAI Press.

Miller Mcpherson, Lynn Smithlovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Review of Sociology*, 27(1).

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August. Association for Computational Linguistics.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100.

Minh-Tien Nguyen, Dac Viet Lai, Huy-Tien Nguyen, and Le-Minh Nguyen. 2018. TSix: A human-involved-creation dataset for tweet summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Avinesh P.V.S., Maxime Peyrard, and Christian M. Meyer. 2018. Live blog corpus for summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Dwijen Rudrapal, Amitava Das, and Baby Bhattacharya. 2018. A survey on automatic twitter event summarization. *Journal of Information Processing Systems*, 14, 03.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12).