EnDSUM: Entropy and Diversity based Disaster Tweet Summarization

Piyush Kumar Garg  
IIT Patna  
India  
piyush_2021cs05@iitp.ac.in

Roshni Chakraborty  
Aalborg University  
Denmark  
roshnic@cs.aau.dk

Sourav Kumar Dandapat  
IIT Patna  
India  
sourav@iitp.ac.in

ABSTRACT

The huge amount of information shared in Twitter during disaster events is utilized by government agencies and humanitarian organizations to ensure quick crisis response and provide situational updates. However, the huge number of tweets posted makes manual identification of the relevant tweets impossible. To address the information overload, there is a need to automatically generate summary of all the tweets which can highlight the important aspects of the disaster. In this paper, we propose an entropy and diversity based summarizer, termed as EnDSUM, specifically for disaster tweet summarization. Our comprehensive analysis on 6 datasets indicates the effectiveness of EnDSUM and additionally, highlights the scope of improvement of EnDSUM.

KEYWORDS

Entropy, Disaster tweets, Social media, Summarization

1 INTRODUCTION

Social media platforms, like Twitter, are highly important mediums of information during disasters. For example, humanitarian organizations and government agencies rely on Twitter to identify relevant information on different categories, such as affected population, urgent need of resources, infrastructure damage, etc [14]. However, the huge number of tweets posted and the high vocabulary diversity [4, 7] make it a challenging task to manually find the relevant information [15, 26]. In order to address this issue, several research works [9, 23] have proposed specific tweet summarization approaches for disaster events.

Existing disaster tweet summarization approaches could be segregated into content based [23], graph based [9], deep learning based [8], and ontology based [12] approaches on the basis of the mechanism they follow. While content based approaches [22, 23] rely on only the importance of the words present in a tweet to determine its selection to the summary, deep learning based approaches [8] consider both content and contextual importance of the tweet. However, none of these approaches consider the vocabulary diversity and therefore, fails to always ensure diversity in summary and coverage of all the important categories present in the tweets. In order to address these, graph based approaches [9, 10] initially group similar tweets together such that each group represents a category by community detection algorithms, thereby handling the vocabulary diversity followed by selecting representative tweets from each group to create the summary to ensure coverage. However, automatic community detection algorithms fail to automatically segregate the tweets into different categories due to the vocabulary overlap among tweets of different categories. Therefore, Garg et al. [12] initially identify the category of each tweet by an ontology based approach and then, select tweets from each category to generate the summary. However, none of these approaches try to handle the vocabulary diversity simultaneously while selecting the tweets into the summary. For example, these existing approaches are dependent on identifying the categories initially which lead to bad summaries, such as reduced diversity in summary, if the categories are not identified correctly.

In order to resolve this, we propose EnDSUM, an entropy and diversity based disaster summarizer where we automatically select that tweet into summary which provides the best information coverage of all the tweets, i.e., entropy and most novel information, i.e., diversity. Therefore, EnDSUM can generate the summary automatically without explicitly identifying the category of a tweet. Although there are few single and multiple document summarization approaches [1, 11, 17, 19, 21] that have highlighted the relevance of entropy based selection of sentences into summary, these approaches are not directly applicable to disaster tweets. The reason being the informal structure of tweets, absence of storyline in tweets and the high vocabulary diversity in user generated tweets. Our evaluation of EnDSUM with existing state-of-the-art disaster tweet summarization approaches on 6 different disasters shows its high effectiveness on 5 datasets. However, we observe that the performance of EnDSUM degrades when there is considerable vocabulary overlap among the tweets which belong to different categories of the same disaster event. The reason being we consider only content based information for calculation of entropy and diversity. The organization of the paper is as follows. We discuss problem definition and proposed approach in Section 2 followed by the experiment details in Section 3 and conclusions in Section 4.

2 PROPOSED APPROACH

Given a disaster event, $E$, that consists of $m$ tweets, $T = \{T_1, T_2, ..., T_m\}$, we aim to prepare a summary, $S$, by selecting $L$ tweets from $T$ such that it provides the maximum information coverage from $T$ with minimum redundant information in the final summary. Therefore, we propose EnDSUM where we iteratively selecting the tweet that can ensure the maximum entropy of all the tweets and maximum diversity in summary. While selection of the tweet with maximum entropy ensures information coverage of a category, selection of the tweet with the maximum diversity ensures not multiple tweets from the same category are selected [5, 6]. Therefore, at every iteration, we select the tweet ($T^*$), which has the maximum score by Equation 1.

$$T^* = \arg \max_{T_i \in \mathcal{T}} \alpha \cdot E(T_i) + \beta \cdot D(T_i, S)$$  \hspace{1cm} (1)
where, $E(T_i)$ represents the entropy of tweet, $T_i$, and $D(T_i, S')$ represents the information diversity provided by $T_i$ with respect to the already selected tweets in summary, $S'$. $\alpha$ and $\beta$ are the tunable parameters which represent the importance of $E(T_i)$ and $D(T_i, S')$ respectively. We consider $\alpha$ and $\beta$ as 0.5 to provide equal importance to both entropy and diversity. Although there are several available mechanisms to calculate $E(T_i)$, we rely on Karci Entropy [13] for EnDSUM. Karci Entropy can resolve the inherent vocabulary diversity in disaster tweets as it calculates the entropy of a tweet, $E(T_i)$, by considering the similarity of $T_i$ with the other tweets as shown in Equation 2.

$$E(T_i, K) = \sum_{j=1}^{K} [1 - P_{ij}^\gamma \log P_{ij}], \ 0 < \gamma \ (2)$$

where, $\gamma$ represents the importance of similarity. We consider $\gamma$ as 0.5 as highlighted by Hark et al. [13]. $K$ is the list of similar tweets of $T_i$, where a tweet is said to be similar to $T_i$ if the content based cosine similarity, i.e. $P_{ij}$ between them is higher than 0 (as shown in [13]) and $P_{ij}$ is the normalized number of overlapping between $T_i$ and $T_j$ normalized by the total number of overlapping keywords of $T_i$ with any tweet. We calculate $D(T_i, S')$ as $(1 - \text{Sim}(T_i, S'))$ where $\text{Sim}(T_i, S')$ represents the overlap in keywords between $T_i$ and $S'$ by

$$\text{Sim}(T_i, S') = \sum_{k \in S} \frac{\text{Overlap}(T_i, T_k)}{\text{Length}(T_i)} \ (3)$$

where, $\text{Length}(T_i)$ is the number of keywords of $T_i$. We follow Khan et al. [16] to identify the keywords of $T_i$ as the nouns, verbs, adjectives present in $T_i$ and similarly, for $S'$, we consider the distinct set of nouns, verbs, adjectives present in all the tweets of $S'$. Therefore, a lower $\text{Sim}(T_i, S')$ ensures $T_i$ has minimum redundant content information with respect to already generated summary, $S'$, and a higher $E(T_i)$ ensures $T_i$ has higher information coverage of the category.

### 3 EXPERIMENTS AND RESULTS

In this Section, we provide details of the experiment and results. For the datasets, we consider Los Angeles International Airport Shooting 1 ($D_1$) provided by Olteanu et al. [20], Hurricane Matthew 2 ($D_2$), Puebla Mexico Earthquake 3 ($D_3$), Pakistan Earthquake 4 ($D_4$) and Midwestern U.S. Floods 5 ($D_5$) provided by Alam et al. [2] and Sandy Hook Elementary School Shooting 6 ($D_6$) provided by Dutta et al. [9]. We perform lemmatization, convert to lower case and remove of Twitter specific keywords [3] and retweets as pre-processing. We consider the ground truth summary provided by Garg et al. [12] for $D_1$-$D_4$ and by Dutta et al. [9] for $D_5$. We compare EnDSUM with content based [24] ($B_4$), graph based [9] ($B_2$), sub-event based [25] ($B_4$) and ontology based [12] ($B_4$) disaster summarization approaches.  

**Results and Discussion:** We evaluate the performance of EnDSUM and the existing research with the ground truth summary using ROUGE-N [18] F1-score score when N=1, 2, and L. Our observations from Table 1 indicate that EnDSUM ensures better ROUGE-N F1-score over all baselines for $D_2$-$D_6$. The improvement is highest over $B_1$ baseline and lowest over $B_4$ baseline. EnDSUM performs worse than $B_4$ for Rouge-N scores and worse than $B_1$ for Rouge-2 and Rouge-L scores on $D_1$. Therefore, although EnDSUM has highly effective performance in most scenarios, it sometimes fails to resolve the vocabulary overlap across different categories in a disaster, as seen for $D_1$. Therefore, to resolve this, we are working towards making EnDSUM resilient irrespective of the vocabulary diversity by considering semantic and contextual similarity along

---

1. https://en.wikipedia.org/wiki/2013_Los_Angeles_International_Airport_shooting
2. https://en.wikipedia.org/wiki/Hurricane_Matthew
3. https://en.wikipedia.org/wiki/2017_Puebla_earthquake
4. https://en.wikipedia.org/wiki/2019_Kashmir_earthquake
5. https://en.wikipedia.org/wiki/2019_Midwestern_U.S._floods
6. https://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting
with the already considered content similarity for entropy and diversity calculation.

4 CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a novel entropy and diversity based tweet summarizer, EnDSUM for disaster events. Our experimental analysis on 6 disaster datasets indicates both the effectiveness of EnDSUM and its scope of improvement. For example, to handle the the high vocabulary overlap among categories, we are working to both include semantic and contextual similarity while calculating entropy and diversity in EnDSUM. Furthermore, while most summarization algorithms generate a predefined length summary, we intend to extend EnDSUM such that it provides complete information coverage of the disaster event. For example, intuitively, the summary length varies on the basis of the information diversity in a disaster event, therefore a summary of length less than the required length leads to less information coverage whereas a summary with more number of tweets than the required will reduce the information diversity. We believe by incorporating these changes in EnDSUM, it would provide an effective performance irrespective of the disaster.

REFERENCES

[1] Subhanpurno Aji and Ramachandra Kaimal. 2012. Document summarization using positive pointwise mutual information. ABRUC’s International Journal of Computer Science and Information Technology 4, 2 (2012), 47–55.
[2] Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Oflı. 2021. HumAID: Human-annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks. arXiv preprint arXiv:2104.03090 (2021).
[3] Chulhem Arachie, Munas Gaur, Sam Anzaroot, William Groves, Ke Zhang, and Alejandro Jaimez. 2020. Unsupervised detection of sub-events in large scale disasters. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34: 354–361.
[4] Carlos Castillo. 2016. Big crisis data: social media in disasters and time-critical situations. Cambridge University Press.
[5] Roshni Chakraborty, Maitry Bhavsar, Sourav Dandapat, and Joydeep Chandra. 2017. A network based stratification approach for summarizing relevant comment tweets of news articles. In International Conference on Web Information Systems Engineering. Springer, 33–48.
[6] Roshni Chakraborty, Maitry Bhavsar, Sourav Kumar Dandapat, and Joydeep Chandra. 2019. Tweet summarization of news articles: An objective ordering based perspective. IEEE Transactions on Computational Social Systems 6, 4 (2019), 761–777.
[7] Roshni Chakraborty, Abhijeet Khara, Apalak Khatua, Sourav Kumar Dandapat, and Joydeep Chandra. 2018. Predicting tomorrow’s headline using Twitter Debates. In CRK Workshops.
[8] Alexis Dusart, Karen Pinel-Sauvagnat, and Gilles Hubert. 2021. TTSuBERT: Tweet Stream Summarization Using BERT. arXiv preprint arXiv:2106.08770 (2021).
[9] Soumi Dutta, Vibhash Chandar, Kanav Mehra, Asit Kumar Das, Tannoy Chakraborty, and Saptarshi Ghosh. 2018. Ensemble algorithms for microblog summarization. IEEE Intelligent Systems 33, 3 (2018), 4–14.
[10] Soumi Dutta, Soujat Ghatak, Moumita Roy, Saptarshi Ghosh, and Asit Kumar Das. 2015. A graph based clustering technique for tweet summarization. In 2015 4th international conference on reliability, infocom technologies and optimization (ICRITO) (trends and future directions) IEEE, 1–6.
[11] Guy Pergenblatt, Haggai Rotman, Odelia Beni, and David Konopnicki. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In Proceedings of the 4th International ACM SIGIR Conference on research and development in information retrieval. 961–964.
[12] Priyush Kumar Garg, Roshni Chakraborty, and Sourav Kumar Dandapat. 2022. OnToRealSumm: Ontology based Real-Time Tweet Summarization. arXiv preprint arXiv:2201.06545 (2022).
[13] Cengiz Hark and Ali Karcu. 2020. Karci summarization: A simple and effective approach for automatic text summarization using Karci entropy. Information Processing & Management 57, 3 (2020), 102187.
[14] Muhammad Imran and Carlos Castillo. 2015. Towards a data-driven approach to identify crisis-related topics in social media streams. In Proceedings of the 24th International Conference on World Wide Web. 1205–1210.