 OntoDSumm: Ontology-Based Tweet Summarization for Disaster Events

Piyush Kumar Garg, Roshni Chakraborty, and Sourav Kumar Dandapat, Member, IEEE

Abstract—The huge popularity of social media platforms, such as Twitter, attracts a large fraction of users to share real-time information and short situational messages during disasters. A summary of these tweets is required by the government organizations, agencies, and volunteers for efficient and quick disaster response. However, the huge influx of tweets makes it difficult to manually get a precise overview of ongoing events. To handle this challenge, several tweet summarization approaches have been proposed. In most of the existing literature, tweet summarization is broken into a two-step process where, in the first step, it categorizes tweets, and in the second step, it chooses representative tweets from each category. There are both supervised and unsupervised approaches found in the literature to solve the problem of first step. Supervised approaches require a huge amount of labeled data, which incurs cost as well as time. On the other hand, unsupervised approaches could not cluster tweet properly due to the overlapping keywords, vocabulary size, lack of understanding of semantic meaning, and so on, while, for the second step of summarization, existing approaches applied different ranking methods where those ranking methods are very generic, which fail to compute proper importance of a tweet with respect to a disaster. Both problems can be handled far better with proper domain knowledge. In this article, we exploited already existing domain knowledge by the means of ontology in both steps and proposed a novel disaster summarization method OntoDSumm. We evaluate this proposed method with six state-of-the-art methods using 12 disaster datasets. Evaluation results reveal that OntoDSumm outperforms the existing methods by approximately 2%–66% in terms of ROUGE-1 F1-score.

Index Terms—Crisis scenario, disaster events, maximum marginal relevance, ontology, social media, tweet summarization.

I. INTRODUCTION

SOCIAL media platforms, such as Twitter, have become extensively popular in the last decade. A large fraction of users use Twitter for a variety of purposes, such as sharing opinions and updates, promoting products, and spreading awareness. For example, a good number of users share real-time updates during disaster events [1], [2], [3] on social media platforms. Government and volunteer organizations utilize these updates to ensure effective disaster response. However, the high volume of tweets generated continuously makes it challenging for government and volunteer organizations to manually identify the relevant information. Therefore, an automated summary of these tweets could immensely help these organizations to decide their immediate plan of action. Although there is a plethora of tweet summarization approaches [4], [5], [6], these approaches are not directly applicable for tweet summarization related to disaster events. This is mainly due to the difference in inherent characteristics and features, which vary across domains. Each domain consists of several prominent categories from which representative tweets need to be selected for the summary. The set of categories, as well as the importance of each category, varies across domains. Hence, representative tweets from those categories also vary across domains. Prior studies [7], [8], [9] show that tweets related to a disaster comprise

\[ \sum_{i=1}^{n} C_i \]

To handle this challenge, several tweet summarization approaches have been proposed. In most of the existing literature, tweet summarization is broken into a two-step process where, in the first step, it categorizes tweets, and in the second step, it chooses representative tweets from each category. There are both supervised and unsupervised approaches found in the literature to solve the problem of first step. Supervised approaches require a huge amount of labeled data, which incurs cost as well as time. On the other hand, unsupervised approaches could not cluster tweet properly due to the overlapping keywords, vocabulary size, lack of understanding of semantic meaning, and so on, while, for the second step of summarization, existing approaches applied different ranking methods where those ranking methods are very generic, which fail to compute proper importance of a tweet with respect to a disaster. Both problems can be handled far better with proper domain knowledge. In this article, we exploited already existing domain knowledge by the means of ontology in both steps and proposed a novel disaster summarization method OntoDSumm. We evaluate this proposed method with six state-of-the-art methods using 12 disaster datasets. Evaluation results reveal that OntoDSumm outperforms the existing methods by approximately 2%–66% in terms of ROUGE-1 F1-score.

Index Terms—Crisis scenario, disaster events, maximum marginal relevance, ontology, social media, tweet summarization.

I. INTRODUCTION

SOCIAL media platforms, such as Twitter, have become extensively popular in the last decade. A large fraction of users use Twitter for a variety of purposes, such as sharing opinions and updates, promoting products, and spreading awareness. For example, a good number of users share real-time updates during disaster events [1], [2], [3] on social media platforms. Government and volunteer organizations utilize these updates to ensure effective disaster response. However, the high volume of tweets generated continuously makes it challenging for government and volunteer organizations to manually identify the relevant information. Therefore, an automated summary of these tweets could immensely help these organizations to decide their immediate plan of action. Although there is a plethora of tweet summarization approaches [4], [5], [6], these approaches are not directly applicable for tweet summarization related to disaster events. This is mainly due to the difference in inherent characteristics and features, which vary across domains. Each domain consists of several prominent categories from which representative tweets need to be selected for the summary. The set of categories, as well as the importance of each category, varies across domains. Hence, representative tweets from those categories also vary across domains. Prior studies [7], [8], [9] show that tweets related to a disaster comprise

\[ \sum_{i=1}^{n} C_i \]

To handle this challenge, several tweet summarization approaches have been proposed. In most of the existing literature, tweet summarization is broken into a two-step process where, in the first step, it categorizes tweets, and in the second step, it chooses representative tweets from each category. There are both supervised and unsupervised approaches found in the literature to solve the problem of first step. Supervised approaches require a huge amount of labeled data, which incurs cost as well as time. On the other hand, unsupervised approaches could not cluster tweet properly due to the overlapping keywords, vocabulary size, lack of understanding of semantic meaning, and so on, while, for the second step of summarization, existing approaches applied different ranking methods where those ranking methods are very generic, which fail to compute proper importance of a tweet with respect to a disaster. Both problems can be handled far better with proper domain knowledge. In this article, we exploited already existing domain knowledge by the means of ontology in both steps and proposed a novel disaster summarization method OntoDSumm. We evaluate this proposed method with six state-of-the-art methods using 12 disaster datasets. Evaluation results reveal that OntoDSumm outperforms the existing methods by approximately 2%–66% in terms of ROUGE-1 F1-score.

Index Terms—Crisis scenario, disaster events, maximum marginal relevance, ontology, social media, tweet summarization.

I. INTRODUCTION

SOCIAL media platforms, such as Twitter, have become extensively popular in the last decade. A large fraction of users use Twitter for a variety of purposes, such as sharing opinions and updates, promoting products, and spreading awareness. For example, a good number of users share real-time updates during disaster events [1], [2], [3] on social media platforms. Government and volunteer organizations utilize these updates to ensure effective disaster response. However, the high volume of tweets generated continuously makes it challenging for government and volunteer organizations to manually identify the relevant information. Therefore, an automated summary of these tweets could immensely help these organizations to decide their immediate plan of action. Although there is a plethora of tweet summarization approaches [4], [5], [6], these approaches are not directly applicable for tweet summarization related to disaster events. This is mainly due to the difference in inherent characteristics and features, which vary across domains. Each domain consists of several prominent categories from which representative tweets need to be selected for the summary. The set of categories, as well as the importance of each category, varies across domains. Hence, representative tweets from those categories also vary across domains. Prior studies [7], [8], [9] show that tweets related to a disaster comprise

\[ \sum_{i=1}^{n} C_i \]

To handle this challenge, several tweet summarization approaches have been proposed. In most of the existing literature, tweet summarization is broken into a two-step process where, in the first step, it categorizes tweets, and in the second step, it chooses representative tweets from each category. There are both supervised and unsupervised approaches found in the literature to solve the problem of first step. Supervised approaches require a huge amount of labeled data, which incurs cost as well as time. On the other hand, unsupervised approaches could not cluster tweet properly due to the overlapping keywords, vocabulary size, lack of understanding of semantic meaning, and so on, while, for the second step of summarization, existing approaches applied different ranking methods where those ranking methods are very generic, which fail to compute proper importance of a tweet with respect to a disaster. Both problems can be handled far better with proper domain knowledge. In this article, we exploited already existing domain knowledge by the means of ontology in both steps and proposed a novel disaster summarization method OntoDSumm. We evaluate this proposed method with six state-of-the-art methods using 12 disaster datasets. Evaluation results reveal that OntoDSumm outperforms the existing methods by approximately 2%–66% in terms of ROUGE-1 F1-score.

Index Terms—Crisis scenario, disaster events, maximum marginal relevance, ontology, social media, tweet summarization.

I. INTRODUCTION

SOCIAL media platforms, such as Twitter, have become extensively popular in the last decade. A large fraction of users use Twitter for a variety of purposes, such as sharing opinions and updates, promoting products, and spreading awareness. For example, a good number of users share real-time updates during disaster events [1], [2], [3] on social media platforms. Government and volunteer organizations utilize these updates to ensure effective disaster response. However, the high volume of tweets generated continuously makes it challenging for government and volunteer organizations to manually identify the relevant information. Therefore, an automated summary of these tweets could immensely help these organizations to decide their immediate plan of action. Although there is a plethora of tweet summarization approaches [4], [5], [6], these approaches are not directly applicable for tweet summarization related to disaster events. This is mainly due to the difference in inherent characteristics and features, which vary across domains. Each domain consists of several prominent categories from which representative tweets need to be selected for the summary. The set of categories, as well as the importance of each category, varies across domains. Hence, representative tweets from those categories also vary across domains. Prior studies [7], [8], [9] show that tweets related to a disaster comprise

\[ \sum_{i=1}^{n} C_i \]
several categories\(^1\) [7], such as infrastructure damage [10], victim needs [12], [13], volunteer operations [14], [15], emotional response [16], and affected population [17]. Our experimental result (which will be discussed shortly) reveals that the importance of these categories mentioned above varies. Therefore, it is required to have the disaster-specific domain knowledge to generate an effective disaster summary.

Existing literature on disaster event-based summarization tries to achieve an effective summary by following two major steps: 1) segregation of the tweets into categories and 2) selection of the representative tweets from each category to form a summary [8], [18], [19], [20]. We found both unsupervised and supervised approaches in the literature to achieve the first step of summarization. Existing unsupervised approaches, such as graph-based approaches [21], [22], [23] and topic-based approaches [20], utilize the content similarity of the tweets to automatically categorize the tweets into categories by community detection algorithms [24], [25], [26] or latent Dirichlet allocation (LDA)-based topic modeling approaches [27]. These approaches do not use domain knowledge and, therefore, fail to achieve the required performance. At the same time, another set of literature uses a supervised approach for the segregation of tweets into categories. Imran et al. [28], Rudra et al. [29], [30], and Nguyen and Rudra [31] proposed a supervised approach to achieve this objective. However, there is an inherent issue with the supervised approach, which needs a labeled dataset that is costly and time-consuming to obtain.

For the second step of disaster event summarization, i.e., selection of representative tweets from each category to generate the summary, existing approaches assign similar importance to all the categories [22], [23], [32]. For example, Rudra et al. [32] considered all the categories as equally important and, therefore, select two tweets from each category into the final summary for a disaster, whereas Dutta et al. [23] iteratively selected a tweet from each category until it reaches the desired summary length. Other than event summarization, there are a few summarization approaches that aim to come up with category-specific summaries [29], [30], [31]. Hence, these kinds of summarization approaches do not need to choose representative tweets across categories.

Our initial experiments and ground-truth summary of a few disaster events reveal that the importance of categories varies within a disaster as well as across disasters, as shown in Fig. 1. Furthermore, we observe that while some categories, such as affected population\(^2\) and infrastructure damage\(^3\), are always present irrespective of the disaster, the presence of other categories, such as international aid\(^4\) and aftermath,\(^5\) depends on the disaster, as shown in Table I.

In this article, we propose an ontology-based disaster summarization approach, OntoDSumm, which resolves each of these challenges sequentially in three phases. In Phase-I, we propose an ontology-based category identification approach, which utilizes the domain knowledge from ontology to map each tweet into a category. Phase-I of OntoDSumm is an unsupervised approach; however, it performs far better than other unsupervised approaches as it utilizes the domain knowledge of disaster through ontology. As previously discussed, the importance of a category varies across disasters, so we propose a method to automatically predict the importance of every category for the given disaster event in Phase-II. We propose disaster similarity index (we discuss in detail in Section V), which can effectively determine the importance of each category related to an event given the information of a similar disaster event. Finally, in Phase-III, we propose a modified version of maximal marginal relevance (MMR) [33] specifically designed for disaster events, which we refer to as disaster-specific maximal marginal relevance (DMMR). The novelty of DMMR is that it utilizes ontology knowledge with respect to each category and, therefore, can ensure maximum information coverage of each category in summary. Therefore, by systematic resolution of each of the specific objectives of disaster summarization, OntoDSumm can ensure better performance than the existing state-of-the-art summarization approaches by 2%–66% in terms of ROUGE-1 F1-score, 4%–77% in terms of ROUGE-2 F1-score, and 3%–43% in terms of ROUGE-L F1-score [34]. We also exhaustively analyze the performance of each phase of OntoDSumm with nine different variants of OntoDSumm to validate the requirement and role of each phase of OntoDSumm.

The rest of this article is organized as follows. We discuss related works in Section II and the dataset details in Section III. In Section IV, we present the problem definition and discuss details of OntoDSumm in Section V. We discuss the experiment details in Section VI and results in Section VI-B, and finally, we draw conclusions of OntoDSumm in Section VII.

### II. RELATED WORKS

During ongoing important events, a huge surge of tweets is found on Twitter. As a result, many such events are often being

---

\(^1\)A category consists of information related to the same topic/subevent of a disaster.

\(^2\)Information on the number of people missing, displaced, injured, or died.

\(^3\)Damage in buildings, roads, bridges, and so on.

\(^4\)Assistance provided by a country or multilateral institutions to the affected country.

\(^5\)Consequences of the disaster.

---

| SNo | Category                  | D\(_1\) | D\(_2\) | D\(_3\) | D\(_4\) | D\(_5\) | D\(_6\) |
|-----|---------------------------|---------|---------|---------|---------|---------|---------|
| 1   | Affected Population       | Yes     | Yes     | Yes     | Yes     | Yes     | Yes     |
| 2   | Infrastructure Damage     | Yes     | Yes     | Yes     | Yes     | Yes     | Yes     |
| 3   | Aftermath                 | Yes     | No      | No      | Yes     | Yes     | Yes     |
| 4   | Donations                 | No      | Yes     | No      | Yes     | Yes     | Yes     |
| 5   | International Aid         | No      | No      | Yes     | No      | Yes     | Yes     |

---

\(D\(_i\)\) denotes the disaster in which the summary is generated.
reported on social media platforms such as Twitter, even before any mainstream media [35] and being considered an important source of information. However, due to noise, duplicate tweets, and a humongous amount of tweets, it becomes really difficult to get a precise understanding about ongoing events. Hence, tweet summarization for events got huge attention from the research community. Tweet summarization approaches have been proposed for various domains such as sports events [3], [36], [37], political events [38], [39], social events [6], [40], disasters [41], [42], and news events [4], [43], [44].

Broadly, a disaster summarization approach, such as any other summarization approach can be categorized as either abstractive [31], [45], [46], [47] or extractive summarization [20], [42], [48], [49] approaches. As in this article, our focus is to propose an extractive tweet summarization approach, and we restrict ourselves to discuss related literature on the extractive disaster tweet summarization approach.

Existing extractive disaster tweet summarization approaches can further be segregated based on their proposed methodologies, such as graph-based [23], content-based [49], deep learning-based approaches [50], [51], or hybrid of multiple approaches [52]. For example, existing content-based disaster summarization approaches propose different mechanisms to select the tweets into the summary, such as the presence of important words in the tweet [18], [53] and coverage of relevant concepts by the tweets. While the presence of important words is measured by the frequency and information content of the words [18], [53], relevant concepts are determined by segregating tweets into either relevant or irrelevant by semisupervised learning [54] or supervised learning [8], [19], [55], [56]. In addition, deep learning-based approaches have proposed different neural network architectures, such as the disaster-specific bidirectional encoder representations from transformers (BERT) model [51] or graph convolutional neural network-based model [50], to identify the tweets to be selected in summary. However, both content- and deep learning-based techniques require a huge number of labeled tweets for training, which is very difficult to obtain for disaster events. Furthermore, these approaches select a tweet into the summary based on the relevance of that tweet to the disaster and, therefore, do not consider the presence of categories and, correspondingly, the category importance for summary selection and, therefore, fail to ensure the information coverage of each category in summary [57].

In order to incorporate the category-specific information, several research works have proposed graph-based tweet summarization approaches [22], [58], which initially create the tweet similarity graph with tweets as nodes and an edge as the similarity between a pair of tweets and then group similar tweets together by identifying communities that represent categories. Finally, these approaches select representative tweets from each category based on the length, degree, or centrality-based measures to generate the summary [59]. Therefore, these approaches ensure the integration of similar information together by the edge relationships in the graph, implicit identification of categories, and further ensure information coverage and reduction of redundancy by selecting representative tweets from each category. For example, Dutta et al. [21] proposed a community-based approach to identify the different subgroups as categories from the tweet similarity graph and finally select representative tweets by centrality-based measures to create a summary. However, these approaches rely on community-based measures to inherently identify the categories of the disaster, which is very challenging due to the high vocabulary overlap across categories in a disaster. In addition, these approaches consider only content-based similarity to identify the category, which cannot ensure handling of the inherent issues of tweet summarization. These approaches also do not consider the difference in importance of categories and their information content across different disasters.

Therefore, Rudra et al. [29], [32] used an existing category identification classifier, i.e., artificial intelligence for digital response (AIDR) [28], to identify the categories and then select representative tweets from each category based on information coverage of a tweet [29] or the presence of disaster information [32]. However, AIDR requires human
TABLE II
DETAILS OF 12 DISASTER DATASETS, INCLUDING DATASET NUMBER, YEAR, NUMBER OF TWEETS, TYPE OF DISASTER, AND CONTINENT

| SNo | Dataset | Year | Number of tweets | Type of disaster | Continent |
|-----|---------|------|------------------|------------------|-----------|
| 1   | $D_1$   | 2012 | 2080             | Man-made         | USA       |
| 2   | $D_2$   | 2013 | 2069             | Natural          | Asia      |
| 3   | $D_3$   | 2014 | 1461             | Natural          | Asia      |
| 4   | $D_4$   | 2013 | 1413             | Man-made         | Asia      |
| 5   | $D_5$   | 2015 | 1676             | Man-made         | Asia      |
| 6   | $D_6$   | 2013 | 1409             | Natural          | USA       |
| 7   | $D_7$   | 2016 | 1654             | Natural          | USA       |
| 8   | $D_8$   | 2017 | 2015             | Natural          | USA       |
| 9   | $D_9$   | 2019 | 1958             | Natural          | Asia      |
| 10  | $D_{10}$| 2019 | 1880             | Natural          | USA       |
| 11  | $D_{11}$| 2016 | 2195             | Natural          | Oceania   |
| 12  | $D_{12}$| 2015 | 2004             | Natural          | Oceania   |

intervention for each new disaster event and is applicable only to real-time disaster events. Furthermore, none of these approaches considers the difference in category vocabulary and importance across disasters. Therefore, there is a need to develop a system that can automatically identify the categories of the disaster with minimum human intervention, capture the importance of each category given a disaster, and ensure the representation of each category based on their specific importance in summary. In this article, we propose OntoDSumm that utilizes disaster-specific knowledge in the form of ontology to identify the category of a tweet that requires no human intervention, followed by importance prediction of a category based on an existing similar disaster, and finally, select tweets from each category by a disaster-specific selection mechanism to ensure information coverage of each category. We discuss datasets details next.

III. DATASET

In this section, we discuss the datasets, preprocessing details, and gold standard summary.

A. Dataset Details and Preprocessing

We evaluate the performance of OntoDSumm on 12 disaster datasets that are given as follows. We show an overview of these datasets in Table II.

1) $D_1$: This dataset is prepared based on the Sandy Hook Elementary School Shooting\(^6\) in which around 26 people, including 20 children and six adults, were killed in December 2012. This dataset is taken from [23].

2) $D_2$: This dataset is prepared based on the Uttarakhand Flood, India,\(^7\) which caused dreadful floods and landslides in June 2013. This dataset is also taken from [23].

3) $D_3$: This dataset is prepared based on the devastating impact of the strong cyclone, Hagupit Typhoon, Philippines,\(^8\) in December 2014, which led to the death of around 18 people and evacuation of 916 people. This dataset is also taken from [23].

4) $D_4$: This dataset is prepared based on the Hyderabad Blast, India,\(^9\) in which two consecutive bomb blasts killed 17 people and injured 119 people in February 2013. This dataset is also taken from [23].

5) $D_5$: This dataset is prepared based on the Harda Twin Train Derailment, India,\(^10\) in which 31 people died and 100 got injured. The incident happened in August 2015. This dataset is taken from [53].

6) $D_6$: This dataset is prepared based on the Los Angeles International Airport Shooting\(^11\) in which around 15 people were injured and one person was killed. The incident happened in November 2013. This dataset is taken from [60].

7) $D_7$: This dataset is prepared based on the devastating impact of the terrible hurricane, Hurricane Matthew, Haiti,\(^12\) in October 2016, which led to the death of 603 people and evacuation of 1.5 million people. This dataset is taken from [61].

8) $D_8$: This dataset is prepared based on the Puebla Mexico Earthquake\(^13\) in which 370 people died and 6011 got injured. The incident happened in September 2017. This dataset is taken from [61].

9) $D_9$: This dataset is prepared based on the Pakistan Earthquake\(^14\) in which 40 people died and 850 got injured. The incident happened in September 2019. This dataset is taken from [61].

10) $D_{10}$: This dataset is prepared based on the Midwestern U.S. Floods,\(^15\) which caused dreadful floods and massive damages in Midwestern United States from March 2019 to December 2019. This dataset is taken from [61].

---

\(^6\)https://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting
\(^7\)https://en.wikipedia.org/wiki/2013_North_India_floods
\(^8\)https://en.wikipedia.org/wiki/Typhoon_Hagupit_(2014)
\(^9\)https://en.wikipedia.org/wiki/2013_Hyderabad_blasts
\(^10\)https://en.wikipedia.org/wiki/Harda_twin_train_derailment
\(^11\)https://en.wikipedia.org/wiki/2013_Los_Angeles_International_Airport_shooting
\(^12\)https://en.wikipedia.org/wiki/Hurricane_Matthew
\(^13\)https://en.wikipedia.org/wiki/2017_Puebla_earthquake
\(^14\)https://en.wikipedia.org/wiki/2019_Kashmir_earthquake
\(^15\)https://en.wikipedia.org/wiki/2019_Midwestern_U.S._floods
11) \( D_{11} \): This dataset is prepared based on the Kaikoura Earthquake\(^{16}\) in which two people died and 57 got injured. The incident happened in November 2016. This dataset is taken from [61].

12) \( D_{12} \): This dataset is prepared based on the devastating impact of the strong cyclone, Cyclone Pam, Vanuatu,\(^{17}\) in March 2015, which led to the death of around 15 people and displacement of 3300 people. This dataset is taken from [9].

**Preprocessing and Gold Standard Summary:** As we consider only the tweet text, we perform preprocessing to remove URLs, usernames, emoticons, punctuation marks, and stop words. In addition, as shown by Alam et al. [62], we observe that the words with length less than three characters do not provide any relevant information specific to disasters, so we remove these words as noise from tweet text. We use the gold standard summary provided by Dutta et al. [23] for \( D_1-D_4 \) and by Rudra et al. [53] for \( D_5 \). For \( D_6-D_{12} \), we ask three annotators to prepare a summary of 40 tweets for each dataset. We follow the procedure by Dutta et al. [23] to combine the individual summaries to prepare the final gold standard summary.

**IV. PROBLEM STATEMENT**

Given a disaster event, \( D \), that comprises \( n \) tweets, \( T = \{T_1, T_2, \ldots, T_n\} \), we aim to create a summary, \( S \) of \( T \). As in most summarization applications, we assume that the length of the summary, \( m \), is provided. We previously discussed in Section I that a disaster tweet summarization approach must ensure information coverage of all the categories present in \( T \) where information coverage of a category refers to the representation of all the important aspects of that category in \( S \) [63]. As there are many different mechanisms, such as topics, keywords, or a combination of both content and context-based information, to represent aspects [47], we do not provide any specific method to calculate information coverage and only provide an intuition of information coverage next. We refer \( \text{In}(C_i) \) as a measure of aspect/information covered by the \( i \)th category \( C_i \), while \( \text{ICov}(T_j, \text{In}(C_i)) \) measures the number of aspects that are present in \( C_i \), which are covered by a tweet \( T_j \) that belongs to category \( C_i \).

In addition, apart from information coverage of all the categories present in \( T \), the tweets selected in summary, \( S \), must be diverse among each other, i.e., no two tweets selected in \( S \) convey the same information [33], [63]. Therefore, it is required to select a tweet into \( S \), which maximizes the diversity in \( S \). Existing research works measure diversity by the presence of keywords, aspects, and content or contextual information that has not been covered by the tweets already selected in summary [18], [32]. We use \( \text{Div}(T_j, S) \) to measure the diversity provided by selecting \( T_j \) with respect to the already selected tweets of \( S \). Therefore, to create \( S \), we need to select the tweet, \( T^* \), that can maximize both \( \text{ICov}(T_j, \text{In}(C_i)) \) and \( \text{Div}(T_j, S) \) simultaneously as shown in the following equation:

\[
T^* = \bigcup_{i=1}^{K} \bigcup_{T_j \in C_i} \max(\alpha \cdot \text{ICov}(T_j, \text{In}(C_i)) + \beta \cdot \text{Div}(T_j, S))
\]

subject to \( \sum_{i=1}^{K} I_i = m \) \hspace{1cm} (1)

where \( I_i \) represents the importance of the \( i \)th category (which is the same as the number of tweets that need to be selected from that category), \( \alpha \) and \( \beta \) are tunable parameters of information coverage and diversity, respectively [33], and \( m \) refers to summary length. We refer to the list of categories as \( C = \{C_1, C_2, \ldots, C^K\} \) such that there are \( K \) categories present in \( T \). Thus, we intend to select \( m \) tweets from \( T \) such that it maximizes the information coverage present in each category, \( \text{In}(C'_i) \) in \( C \) and diversity in \( S \) based on \( I_i \).

We show all the used notations and corresponding descriptions for OntoDSumm in Nomenclature.

**V. PROPOSED APPROACH**

In this section, we discuss the phases of OntoDSumm briefly followed by the details. An overview of OntoDSumm is shown in Fig. 2.

1) **Identification of the Category of a Tweet, Phase-I:** We propose an unsupervised approach that utilizes the disaster-specific domain knowledge of an existing ontology, Empathi [64], to identify the category of a tweet.

2) **Determination of Importance of Each Category, Phase-II:** We propose a novel score to automatically determine the importance of a category with respect to a given disaster event.

3) **Representative Tweets Selection From Each category, Phase-III:** We propose DMMR to select representative tweets from each category to generate the summary.

**A. Phase-I**

In Phase-I, we propose an ontology-based pseudo-relevance feedback approach to identify the category of a tweet, which does not require any human intervention. Given an existing ontology, \( O \), with \( K \) categories and their corresponding vocabulary, we propose a semantic similarity score, \( \text{SemSIM}(T_j, C_i) \), of \( T_j \) with each category, say \( C_i \), as shown in the following equation:

\[
\text{SemSIM}(T_j, C_i) = \text{Kw}(T_j) \cap \text{Kw}(C_i)
\]

where \( \text{Kw}(T_j) \) comprises the keywords (nouns, verbs, and adjectives) of \( T_j \) proposed by Khan et al. [65], while \( \text{Kw}(C_i) \) comprises the keywords associated with category \( C_i \) provided by ontology. On the basis of \( \text{SemSIM}(T_j, C_i) \), we assign the category of \( T_j \) as that with which \( T_j \) has the highest semantic similarity score, \( \text{MaxSIM}(T_j) \), as shown in the following equation:

\[
\text{MaxSIM}(T_j) = \arg \max_{i \in K} (\text{SemSIM}(T_j, C_i))
\]

We do not propose an ontology for disasters in this article as it requires a considerable amount of human intervention and

\(^{16}\)https://en.wikipedia.org/wiki/2016_Kaikoura_earthquake

\(^{17}\)https://en.wikipedia.org/wiki/Cyclone_Pam
uses the existing available ontology, Empathi [64]. Although several different disaster-specific ontologies are available [66], [67], [68], [69], we choose Empathi [64] as it provides the maximum information related to different types of categories compared to others. However, the ontology requires preprocessing. For example, we merge some categories that have similar information, such as three categories, infrastructure damage, broken bridge, and blocked road that are merged as broken bridge and blocked road, which is comprised of information that is already covered by definition in infrastructure damage. In addition, we observe that the existing keywords of the ontology are not sufficient enough to identify the category of every tweet due to the high vocabulary diversity in tweets of a disaster. For example, we observe that we could not classify 10%–36% of the total tweets with respect to a disaster when we consider only the existing vocabulary of the ontology (as shown in Table III). In order to handle this challenge, we use some disaster-related Wikipedia pages to extend the vocabulary of each category of the ontology. We identify more relevant keywords of each category from the Wikipedia pages based on similarity between the existing keywords of the ontology category and Wikipedia, thus using the existing keywords as a query to identify more relevant keywords. We discuss the detailed procedure next.

We, initially, select 20 Wikipedia pages that belong to different locations and types of disasters. We follow existing research works [9], [70] to identify the relevant Wikipedia pages. To extend the vocabulary of an ontology category, we first identify all the sentences from Wikipedia pages, which consists of at least one keyword of the ontology category. Then, we form a potential keyword list corresponding to this ontology category by adding nouns, verbs, and adjectives [18] from those sentences. Then, we further filter out this potential keyword list based on the frequency and ignore those keywords having a frequency of less than three. Before the final inclusion of these keywords as extended vocabulary, it is given to a human annotator. A keyword is finally added to the vocabulary of the ontology category if the annotator finds it
relevant. This procedure makes sure that extended vocabulary does not have any irrelevant keywords. We follow the same procedure for all ontology categories. Therefore, we now use this extended vocabulary to identify the category of a tweet. On using this extended vocabulary, we observe that around 68%–94% of the tweets could be classified with respect to a disaster, which is more by around 3%–9% than without using the extended vocabulary. However, we could not classify around 6%–32% of the tweets across disasters even after using the extended vocabulary. We show in Table III the fraction of tweets classified using the vocabulary of the ontology and extended vocabulary. We observe that around 3%–30% of the tweets, which are unclassified across disasters, are irrelevant. This procedure makes sure that extended vocabulary does not have any irrelevant keywords. We follow the same procedure for all ontology categories. Therefore, we now use this extended vocabulary to identify the category of a tweet.

In order to understand the performance of Phase-I, we validate it by three manual annotators who manually determine the category of all the classified tweets based on their understanding of the tweet text and category definition. We consider the ground-truth category of a tweet as that category, which is selected by the majority of the annotators. We, then, calculate the F1-score based on the ground-truth category and automated category as given by OntoDSumm. We show our results in Table IV, which indicates almost perfect classification performance by Phase-I of OntoDSumm irrespective of the dataset.

### B. Phase-II

Given a disaster, $D_x$, we propose an approach to automatically determine the importance of each category with respect to $D_x$ in this phase. We need the importance of a category, $I_i$, to determine the number of tweets to be selected from each category in the final summary. Therefore, $I_i$ of $C_i$ represents the total number of tweets to be selected from $C_i$ given $D_x$. The content of the tweets across the categories and the distributions of the tweets across categories vary with disasters. We use a linear regression model to predict the number of tweets to be included in the final summary. We train this regression model with a disaster, $D_x$, similar to $D_x$, on eight features of each category, where six features are based on the highest semantic similarity score, which is obtained in Phase-I: min, max, average, median, variance, and range, and interquartile range (IQR), and two features are based on the category size and category importance, such as the fractional size of the category [i.e., (number of tweets in a category)/(total number of tweets in $D_x$)] and fractional importance of the category [i.e., (number of tweets in summary form the category)/(total number of tweets in summary)]. Then, we predict the fractional importance of each category in $D_x$, which represents the number of tweets to be selected from each category in summary. To identify $D_x$ for a given $D_x$, we propose a metric disaster similarity index, which we discuss next.

**Disaster Similarity Index:** We propose a disaster similarity index, $\text{DisSIM}(D_x, D_y)$, to compute the similarity between any pair of disasters, $D_x$ and $D_y$. We define $\text{DisSIM}(D_x, D_y)$ as the weighted score of the information content of the categories, $\text{Cat}_x(D_x, D_y)$ and similarity in probability distribution between categories $\text{Cat}_y(D_x, D_y)$. We compute $\text{Cat}_p(D_x, D_y)$ as $(1 - \text{Cat}_a(D_x, D_y))$, where $\text{Cat}_a(D_x, D_y)$ is the Jensen–Shannon divergence of two events. $\text{DisSIM}(D_x, D_y)$ is calculated as

$$\text{DisSIM}(D_x, D_y) = w_1 \ast \text{Cat}_x(D_x, D_y)$$

$$+ w_2 \ast \text{Cat}_p(D_x, D_y)$$  \hspace{1cm} (4)

s.t. $w_1 + w_2 = 1$  \hspace{1cm} (5)

$$w_1, \ w_2 \in (0, 1)$$  \hspace{1cm} (6)

where $w_1$ and $w_2$ are the weights of $\text{Cat}_x(D_x, D_y)$ and $\text{Cat}_p(D_x, D_y)$, respectively, and we consider equal weighted of $\text{Cat}_x(D_x, D_y)$ and $\text{Cat}_p(D_x, D_y)$ as $w_1 = w_2 = 0.5$. We calculate $\text{Cat}_x(D_x, D_y)$ as the cosine similarity [71] of the most frequently occurring keywords between $D_x$ and $D_y$ for a category, $i$ and $\text{Cat}_x(D_x, D_y)$ as the average cosine similarity overall categories, as shown in the following equation:

$$\text{Cat}_x(D_x, D_y) = \frac{1}{K} \sum_{i=1}^{K} \text{Cat}_x^i(D_x, D_y)$$  \hspace{1cm} (7)

where $K$ is the total number of categories. We calculate $\text{Cat}_p(D_x, D_y)$ by the Jensen–Shannon divergence score of a disaster pair $D_x$ and $D_y$. $\text{Cat}_p(D_x, D_y)$ measures the similarity of the probability distributions of the categories between $D_x$ and $D_y$. We select $D_y$ as the disaster, which has the maximum similarity with $D_x$ based on $\text{DisSIM}(D_x, D_y)$, where $y \in 0, 1, \ldots, Q$ and $Q$ is the total number of disasters we have. We discuss this experiment and our observations in detail in Section VI-E.

### C. Phase-III

Several existing tweet summarization approaches [44], [58] have proposed MMR to select tweets iteratively such that
it selects that tweet into the summary, which provides the maximum relevance with respect to the query and can maintain maximum diversity among the already selected tweets into the summary. Therefore, MMR inherently does not consider the category relevance, which is necessary for disaster summarization. Therefore, we propose DMMR for tweet selection from each category. For $C^i$, we iteratively select the tweet, which is most relevant to the category, and ensure the maximum diversity among the already selected tweets into summary until $I$, number of tweets are selected. We select the tweet, $T^{max}$, that has the maximum score by (8) as follows:

$$T^{max} = \arg \max_{T_j \in C^i} \left[ \lambda \ast \text{Sim}_1(T_j, \text{OV}(C^i)) - (1 - \lambda) \left(\max_{T_i \in S} \text{Sim}_2(T_j, T_i)\right) \right]$$

where $S$ is the set of tweets already selected in summary and $\lambda$ is a hyperparameter in the interval $[0, 1]$, which measures the importance of relevance provided by $T_j$, i.e., $\text{Sim}_1(T_j, \text{OV}(C^i))$, and diversity of $T_j$ with respect to $S$, i.e., $\text{Sim}_2(T_j, T_i)$. We set the $\lambda$ value as 0.5 by giving equal importance to relevance and diversity. We explain how we calculate $\text{Sim}_1(T_j, \text{OV}(C^i))$ and $\text{Sim}_2(T_j, T_i)$ next. In order to capture the relevance of $T_j$ with respect to a category, $\text{Sim}_1(T_j, \text{OV}(C^i))$, we calculate the contextual similarity of $T_j$ with the ontology vocabulary of $C^i$, i.e., $\text{OV}(C^i)$. Therefore, $\text{Sim}_1(T_j, \text{OV}(C^i))$ ensures consideration of disaster-specific category information with respect to $C^i$ as we consider ontology vocabulary. In addition, the contextual similarity-based calculation can handle the high vocabulary diversity inherent in tweets, as in the following equation:

$$\text{Sim}_1(T_j, \text{OV}(C^i)) = \sum_{v \in \text{Kw}(T_j)} \text{CSim}(v, \text{OV}(C^i))$$

$$\text{CSim}(v, \text{OV}(C^i)) = \max_{k \in \text{Kw}(\text{OV}(C^i))} (\text{VSim}(v, \text{OV}(C^i)_k))$$

$$\text{VSim}(v, \text{OV}(C^i)_k) = \frac{\tilde{v} \cdot \text{OV}(C^i)_k}{|\tilde{v}| |\text{OV}(C^i)_k|}$$

where $\text{Kw}(T_j)$ and $\text{Kw}(\text{OV}(C^i))$ are the set of keywords of $T_j$ and $\text{OV}(C^i)$, respectively, while $\tilde{v}$ and $\text{OV}(C^i)_k$ are the word embedding of $v$ and $\text{OV}(C^i)_k$ generated by Word2Vec [9], respectively. We follow Rudra et al. [18] and only consider nouns, verbs, and adjectives of $T_j$ as the keywords ($v$) of $T_j$. We calculate $\text{Sim}_2(T_j, T_i)$ as the cosine similarity [71] between the keywords of $T_j$ and $T_i$, where $T_i$ is the set of tweets that are in $S$ from $C^i$. We calculate $\text{Sim}_2(T_j, T_i)$ as

$$\text{Sim}_2(T_j, T_i) = \frac{|\text{Kw}(T_j) \cap \text{Kw}(T_i)|}{\sqrt{|\text{Kw}(T_j)| |\text{Kw}(T_i)|}}$$

where $\text{Kw}(T_j)$ and $\text{Kw}(T_i)$ are the keywords of $T_j$ and $T_i$, respectively.

VI. EXPERIMENTS

In this section, we initially discuss the details of the existing research works related to disaster summarization which we use as baselines. Then, we provide a performance comparison of the baselines and OntoDSumm.

A. Baselines

We compare OntoDSumm with the following state-of-the-art summarization approaches.

1) $B_1$: Rudra et al. [30] proposed a summarization framework where they initially create a graph where the nodes are the most important disaster-specific keywords, and the edges represent the bigram relationship between a pair of keywords. Finally, they select the tweets in summary, which can ensure maximum information coverage of the graph.

2) $B_2$: Dutta et al. [23] proposed an ensemble graph-based summarization approach, which initially generates summary by nine existing text summarization algorithms. The authors create a tweet similarity graph where the nodes represent the tweets present in the summary of any of the nine existing text summarization algorithms, and the edges represent their content and context similarity. Finally, the authors follow a community detection algorithm to automatically identify the categories and then select representative tweets from each category based on length, informativeness, and centrality scores to create the summary.

3) $B_3$: Rudra et al. [32] proposed a subevent-based summarization approach that initially identifies the subevents and then generates a summary by selecting representative tweets by integer linear programming-based selection.

4) $B_4$: Nguyen and Rudra [31] proposed an abstractive summarization approach for disaster tweet summarization. Nguyen and Rudra [31] utilized a pretrained BERT model to identify key phrases from tweets and further select those phrases that provide the maximum information to generate the summary. For our experiments, we select those tweets in the summary, which provide maximum coverage of key phrases in the final summary.

5) $B_5$: Dusart et al. [51] proposed a summarization framework, which integrates the context of the tweets using the whole vocabulary related to the event (i.e., frequencies of the terms appearing in the event) and a pretrained language model as BERT model [72] to determine the tweet importance. Finally, they generate a summary by iteratively selecting the higher important tweets in the summary.

6) $B_6$: Li and Zhang [50] proposed a summarization framework that initially creates a tweet relation graph where the nodes represent the tweets and edge represent their content similarity. Then, they apply a graph convolutional network (GCN) on a tweet relation graph to generate tweet hidden features for each tweet. They calculate each tweet’s salience score by considering both the tweet’s hidden state and the event embedding that represents the whole event embeddings. Finally, they generate a summary by iteratively selecting the most salient tweets.
### B. Comparison With Existing Research Works

We compare the summary generated by the OntoDSumm and the existing research works with the ground-truth summary based on ROUGE-N score [34]. ROUGE-N score computes the overlapping words in the generated summary with a set of ground-truth summaries. We calculate precision, recall, and F1-score for three different variants of ROUGE-N score, i.e., $N = 1$, 2, and L, respectively. Our observations, as shown in Tables V and VI, indicate that OntoDSumm ensures better ROUGE-N precision, recall and F1-score in comparison with baselines. The improvement in summary scores of ROUGE-1 F1-score ranges from 1.85% to 65.46%, ROUGE-2 F1-score ranges from 4.16% to 76.47%, and ROUGE-L F1-score ranges from 2.63% to 43.33%. The improvement is highest over the $B_1$ baseline and lowest with the $B_2$ baseline. The performance of the OntoDSumm is the best for $D_6$ with 0.61–0.57 and worst for $D_2$ with 0.47–0.18 in ROUGE-1, ROUGE-2, and ROUGE-L F1-score.

#### C. Ablation Experiments

To understand the effectiveness of each phase of OntoDSumm, we compare and validate the performance of...
OntoDSumm with its different variants and study the role of ontology-based summarization with the identification of category importance.

1) Phase-I: In this experiment, we compare the proposed Phase-I with the following variant of the Phase-I of OntoDSumm.

1) In OntoDSumm-1A, we use the original keywords provided in Empathi and do not integrate our proposed extended vocabulary for each category to identify the category of a tweet.

2) Phase-II: To understand the different aspects of Phase-II and its role in the performance of OntoDSumm, we generate three different variants of Phase-II and compare their performance with OntoDSumm, which are given as follows.

1) In OntoDSumm-2A, we assume that each category is given equal importance, i.e., an equal number of tweets are selected from each category.

2) In OntoDSumm-2B, we use the ridge regression model to predict the number of tweets to be selected from a category.

In the name of a variant of OntoDSumm, i.e., OntoDSumm-1A, the first number represents the phase and the second letter represents the sequence of the ablation study in that phase.
3) In OntoDSumm-2C, we use the Bayesian regression model to predict the number of tweets to be selected from a category.

3) Phase-III: To validate the performance and specificity of DMMR for summarization with respect to disasters, we compare the proposed DMMR with its different variants, which are given as follows.

1) In OntoDSumm-3A, we select that tweet in each iteration, which provides the maximum information coverage of a category, has the maximum semantic similarity score (as discussed in Section V-A) with the category keywords into the summary, and therefore, we do not consider diversity while selecting a tweet.

2) In OntoDSumm-3B, we use the K-means clustering algorithm [73] to select the K tweets from a category into the summary. This K is chosen depending on the importance of a category. We select the centroid of each K-cluster as the representative tweets of the category.

3) In OntoDSumm-3C, we select that tweet in each iteration from a category, which has the maximum eigenvector centrality score [74] into the summary.

4) In OntoDSumm-3D, we select that tweet in each iteration from a category, which has the maximum PageRank score [75] into the summary.

5) In OntoDSumm-3E, we select that tweet in each iteration from a category, which has the maximum MMR-based score [33] into the summary.

4) Results and Discussion: We now compare and validate the performance of OntoDSumm with the discussed different variants on seven different disasters, such as D_2, D_3, D_6, D_8, D_{10}, D_{11}, and D_{12} based on ROUGE-1 F1-score. Our observations, as shown in Fig. 3, indicate that OntoDSumm is better by around 2%–20% in ROUGE-1 F1-score across the disasters.

1) Phase-I: We observe that OntoDSumm outperforms OntoDSumm-1A, which uses only the original keywords provided in Empathi by 5.35%–17.39% in ROUGE-1 F1-score, which highlights the effectiveness of keyword extension using Wikipedia pages.

2) Phase-II: We observe that OntoDSumm outperforms OntoDSumm-2A, which gives equal importance to each category by 3.57%–18.87% in ROUGE-1 F1-score, which highlights the need to consider the specific importance of each category corresponding to a disaster. In addition, we validate the decision to use linear regression over the ridge regression or Bayesian regression as OntoDSumm outperforms OntoDSumm-2B and OntoDSumm-2C by 1.78%–9.09% in ROUGE-1 F1-score.

3) Phase-III: Finally, we observe that the inclusion of DMMR in OntoDSumm provides an increase of 1.78%–19.56% in ROUGE-1 F1-score when compared to different variants of Phase-III, i.e., OntoDSumm-3A, OntoDSumm-3B, OntoDSumm-3C, OntoDSumm-3D, and OntoDSumm-3E. This justifies that domain knowledge is immensely helpful even in finding out important tweets from each category. In addition, we observe that the performance gain of OntoDSumm is highest over the OntoDSumm-3D, which uses PageRank as Pagerank does not consider the diversity criteria, while it is lowest against OntoDSumm-3E, which uses the MMR for selecting the tweets from each category. We also observe that the inclusion of the novel DMMR for Phase-III provides the maximum increase in the performance of OntoDSumm.

D. Identification of Category of a Tweet

In this section, we evaluate the effectiveness of Phase-I, i.e., the proposed unsupervised tweet category identification with an existing unsupervised approach used in [22]. We select the approach followed by Dutta et al. [22], where they utilize the Leuven community detection algorithm [76] to identify communities that inherently represent categories related to the disaster. For our experiments, we select 20% of the total tweets randomly for a disaster event. We then asked three annotators to manually determine the category of selected tweets. We assign the final category level based on majority voting. We then compare the identified category by OntoDSumm and the approach used in [22]. We repeat this for six disaster events. We found that F1-score by OntoDSumm and the approach used in [22] are in the range of 97.8%–99.7% and 53.3%–74.9%, respectively, as shown in Table VII. Therefore, based on our experiments, we can confirm that utilization of the domain knowledge of ontology ensures high efficiency in tweet categorization.

E. Understanding Disaster Similarity Index

On comparing DisSIM(D_x, D_y) for every disaster pair, we observe that any disaster (say D_x) having maximum similarity score with another disaster (say D_y) actually belongs to the same continent (such as USA or Asia or Oceania) and is of the same type (i.e., man-made or natural). For example, Uttarakhand Flood (D_2) has the maximum similarity

| Dataset | Approach | F1-score |
|---------|----------|----------|
| D_2     | Phase-I of OntoDSumm | 0.978    |
|         | Approach used in [22] | 0.653    |
| D_6     | Phase-I of OntoDSumm | 0.995    |
|         | Approach used in [22] | 0.533    |
| D_8     | Phase-I of OntoDSumm | 0.984    |
|         | Approach used in [22] | 0.571    |
| D_{10}  | Phase-I of OntoDSumm | 0.997    |
|         | Approach used in [22] | 0.749    |
| D_{11}  | Phase-I of OntoDSumm | 0.997    |
|         | Approach used in [22] | 0.550    |
| D_{12}  | Phase-I of OntoDSumm | 0.991    |
|         | Approach used in [22] | 0.584    |
score with disaster Pakistan Earthquake (D9) among all the disasters (both the disasters belong to Asia and type is natural). Similarly, Sandy Hook Elementary School Shooting (D1) has the maximum similarity score with Los Angeles International Airport Shooting (D8) among all the disasters (both disasters belong to USA and man-made). In addition, Kaikoura Earthquake (D11) has the maximum similarity score with Cyclone Pam (D12) among all the disasters (both disasters belong to Oceania and Natural). Therefore, we term Dx and Dy as homogeneous disasters if they belong to the same continent and are of the same type. Subsequently, we term Dx and Dy as heterogeneous disasters if they do not satisfy either of these two conditions. We validate next the impact of training on a homogeneous disaster or a heterogeneous disaster on the summary quality. We show the results in Table VIII.

For each disaster Dx, we predict the number of tweets to be selected from each category as predicted by the linear regression model if it was trained on homogeneous disaster or heterogeneous disaster. We then select the representative tweets from each category based on the predicted number of tweets to create the summary and, finally, compare the generated summary with the ground-truth summary based on ROUGE-N scores. For the experiments, we randomly select eight disasters: D1, D2, D3, D7, D8, D10, D11, and D12. We show the precision, recall, and F1-score for three different variants of the ROUGE-N score, i.e., N = 1, 2, and L, respectively, in Table IX. Our observations indicate that there is around 34%–43%, 8%–30%, and 16%–59% increase in ROUGE-2 F1-score for any disaster in Asia, USA, and Oceania, respectively, for both man-made and natural disasters. Therefore, our observations indicate that an effective summary for Dx can be ensured if we identify the importance of categories from a Dy that is of the same type and from the same continent as Dy.

F. Discussion

To understand the effectiveness of OntoDSumm on unseen disaster categories, such as wildfire and landslides, we compare and validate the performance of OntoDSumm on a wildfire disaster event, i.e., Canada Wildfires, in which 1,456,810 acres were burned, including approximately 2,400 homes and buildings. This wildfire incident happened in May 2016. We have taken this dataset from [61] and performed all three sequential phases of OntoDSumm on this dataset. In Phase-I of OntoDSumm, we observe that around 87% of the tweets were correctly classified with respect to the dataset. For the rest of the 13% tweets, we manually checked these tweets, and we found that these tweets are irrelevant to the disaster. Hence, in the next phase, we do not consider these unclassified tweets. In Phase-II, we determine the importance of each category, which represents the number of tweets to be selected from each category in the final summary, using a linear regression model that trains on a very similar dataset to the Wildfire dataset. For this, we utilize the proposed disaster similarity index (discussed in Section V-B) and found that the Midwestern U.S. Floods (D10) dataset is very similar to this Wildfire dataset. Finally, we create the summary in Phase-III by selecting the tweets 20

20https://en.wikipedia.org/wiki/2016_Fort_McMurray_wildfire
from each category based on category importance using the proposed DMMR (discussed in Section V-C). Furthermore, we compare the generated summary by OntoDSumm, and the existing research works with the ground-truth summary based on ROUGE-N F1 scores. Our observations, as shown in Table X, indicate that OntoDSumm ensures better ROUGE-N F1-score in comparison with baselines. The improvement in summary scores of ROUGE-1 F1-score ranges from 4.08% to 16.33%, ROUGE-2 F1-score ranges from 16.67% to 44.44%, and ROUGE-L F1-score ranges from 10.71 to 28.57%. Therefore, based on our experiments, we can confirm that OntoDSumm ensures high efficiency on unseen disaster categories.

G. Limitations of OntoDSumm

We discuss the limitations of OntoDSumm that we have observed next.

1) Dependence on Existing Ontology: As OntoDSumm relies on an existing ontology to identify the categories of a tweet, it is not directly applicable to other summarization applications, such as news events and user opinions regarding products unless an existing ontology for that application is available. We are working toward developing ontology automatically from publicly available resources for any application as a future direction so that OntoDSumm is not dependent on an existing ontology.
TABLE X

| Approach | ROUGE-1 | ROUGE-2 | ROUGE-L |
|----------|---------|---------|---------|
| OntoDSumm | 0.49    | 0.18    | 0.28    |
| $B_1$     | 0.46    | 0.12    | 0.24    |
| $B_2$     | 0.47    | 0.13    | 0.24    |
| $B_3$     | 0.44    | 0.10    | 0.20    |
| $B_4$     | 0.41    | 0.15    | 0.25    |
| $B_5$     | 0.47    | 0.16    | 0.24    |
| $B_6$     | 0.42    | 0.13    | 0.22    |

2) **Identification of Category of a Tweet:** OntoDSumm cannot identify the categories of all tweets. For example, we found among all the disasters, and OntoDSumm performs the worst for $D_4$, where it could not identify the category for around 19.10%–31.59% of the tweets. Therefore, we believe that Phase-I of OntoDSumm could be further improved such that the category of more number of tweets could be effectively identified.

3) **Limitation of Disaster Dataset:** In Phase-II of OntoDSumm, we find the most similar disaster for finding out the category importance of the given disaster. However, it is quite possible that this phase will perform poorly when the existing dataset does not result in a high similarity score. To overcome this, a more diverse and more enriched disaster dataset is required.

VII. CONCLUSION AND FUTURE WORKS

In this article, we propose OntoDSumm, which can generate a tweet summary for a disaster with minimal human intervention. However, the summarization of tweets related to a disaster has several challenges, such as automatic identification of the category of a tweet, determination of representativeness of each category in summary respect to the disaster, and ensuring information coverage of each category while maintaining diversity in the summary. In order to handle these challenges, we propose a three-phase approach, which can handle each challenge systematically with high effectiveness and minimal human intervention. We believe the incorporation of domain knowledge through ontology for category identification, automatic understanding, and knowledge transfer across different disasters to gauge the importance of a category, and finally, a selection mechanism specifically designed for disasters ensures the high performance of OntoDSumm. Our experimental analysis shows that OntoDSumm can ensure the 2%–77% increase in ROUGE-N F1-score over existing research works. In addition, we validate experimentally each phase of OntoDSumm for generating a summary. As a future direction, we are working toward mitigation of all the limitations mentioned in Section VI-G.

REFERENCES

[1] H. Purohit, C. Castillo, and R. Pandey, “Ranking and grouping social media requests for emergency services using serviceability model,” *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–17, Dec. 2020.

[2] S. Priya, R. Sequeira, J. Chandra, and S. K. Dandapat, “Where should one get news updates: Twitter or Reddit,” *Online Social Netw. Media*, vol. 9, pp. 17–29, Jan. 2019.

[3] P. Goyal, P. Kaushik, P. Gupta, D. Vashisth, S. Agarwal, and N. Goyal, “Multilevel event detection, storyline generation, and summarization for tweet streams,” *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 1, pp. 8–23, Feb. 2020.

[4] X. Zheng, A. Sun, and K. Muthuswamy, “Tweet-aware news summarization with dual-attention mechanism,” in *Proc. Companion Web Conf.*, Apr. 2021, pp. 473–480.

[5] R. Chakraborty, M. Bhavars, S. K. Dandapat, and J. Chandra, “Tweet summarization of news articles: An objective ordering-based perspective,” *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 4, pp. 761–777, Aug. 2019.

[6] R. P. Narmadha and G. G. Sreeja, “A survey on online tweet segmentation for linguistic features,” in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2016, pp. 1–6.

[7] M. Imran and C. Castillo, “Towards a data-driven approach to identify crisis-related topics in social media streams,” in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 1205–1210.

[8] S. Roy, S. Mishra, and R. Matam, “Classification and summarization for informative tweets,” in *Proc. IEEE Int. students’ Conf. Elect., Electron. Comput. Sci. (SCCECS)*, Feb. 2020, pp. 1–4.

[9] M. Imran, P. Mitra, and C. Castillo, “Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages,” in *Proc. 10th Lang. Resour. Eval. Conf. (LREC)*, Paris, France: European Language Resources Association (ELRA), May 2016, pp. 1638–1643.

[10] S. Madichetty and M. Srivadi, “A novel method for identifying the damage assessment tweets during disaster,” *Future Gener. Comput. Syst.*, vol. 116, pp. 440–454, Mar. 2021.

[11] S. Priya, M. Bhanu, S. K. Dandapat, K. Ghosh, and J. Chandra, “TAQE: Tweet retrieval-based infrastructure damage assessment during disasters,” *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 2, pp. 389–403, Apr. 2020.

[12] M. Basu, A. Shandilya, K. Ghosh, and S. Ghosh, “Automatic matching of resource needs and availabilities in microblogs for post-disaster relief,” in *Proc. Companion Web Conf. Web Conf. (WWW)*, 2018, pp. 25–26.

[13] R. Dutt, M. Basu, K. Ghosh, and S. Ghosh, “Utilizing microblogs for assisting post-disaster relief operations via matching resource needs and availabilities,” *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1680–1697, Sep. 2019.

[14] M. Basu, A. Shandilya, P. Khosla, K. Ghosh, and S. Ghosh, “Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations,” *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 3, pp. 604–618, Jun. 2019.

[15] M. Imran, F. Oll, D. Caragea, and A. Torralba, “Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions,” *Inf. Process. Manage.*, vol. 57, no. 5, Sep. 2020, Art. no. 102261.

[16] L. Li, Z. Wang, Q. Zhang, and H. Wen, “Effect of anger, anxiety, and sadness on the propagation scale of social media posts after natural disasters,” *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102313.

[17] S. Ghosh et al., “Exploitation of social media for emergency relief and preparedness: Recent research and trends,” *Inf. Syst. Frontiers*, vol. 20, no. 5, pp. 901–907, Oct. 2018.

[18] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, “Extracting situational information from microblogs during disaster events: A classification-summarization approach,” in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 583–592.

[19] K. Rudra, A. Sharma, N. Ganguly, and M. Imran, “Classifying and summarizing information from microblogs during epidemics,” *Inf. Process. Manage.*, vol. 57, no. 5, Sep. 2020, pp. 25–26.

[20] M.-T. Nguyen, A. Kitamoto, and T.-T. Nguyen, “Sum4Act: A framework for retrieving and summarizing actionable tweets during a disaster for reaction,” in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, Cham, Switzerland: Springer, 2015, pp. 64–75.

[21] S. Dutta, S. Ghatak, M. Roy, S. Ghosh, and A. K. Das, “A graph based clustering technique for tweet summarization,” in *Proc. 4th Int. Conf. Rel., Infocom Technol. Optim. (ICRITO) (Trends Future Directions)*, Sep. 2015, pp. 1–6.

[22] S. Dutta et al., “Community detection based tweet summarization,” in *Emerging Technologies in Data Mining and Information Security*, Cham, Switzerland: Springer, 2019, pp. 797–808.
[68] Y. Sermet and I. Demir, “Towards an information centric flood ontology for information management and communication,” Earth Sci. Informat., vol. 12, no. 4, pp. 541–551, Dec. 2019.
[69] H. Yahya and R. Ramli, “Ontology for evacuation center in flood management domain,” in Proc. 8th Int. Conf. Inf. Technol. Multimedia (ICIMU), Aug. 2020, pp. 288–291.
[70] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, “CrisisLex: A lexicon for collecting and filtering microblogged communications in crises,” in Proc. Int. AAAI Conf. Web Social Media, vol. 8. Palo Alto, CA, USA: AAAI Press, 2014, pp. 1–10.
[71] H. V. Nguyen and L. Bai, “Cosine similarity metric learning for face verification,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2010, pp. 709–720.
[72] Y. Liu and M. Lapata, “Text summarization with pretrained encoders,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 3728–3738.
[73] J. MacQueen. “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp. Math. Statist. Probab., vol. 1, 1967, pp. 281–297.
[74] P. Bonacich, “Factoring and weighting approaches to status scores and clique identification,” J. Math. Sociol., vol. 2, no. 1, pp. 113–120, 1972.
[75] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” Comput. Netw. ISDN Syst., vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.
[76] V. A. Traag, “Faster unfolding of communities: Speeding up the Louvain algorithm,” Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 92, no. 3, Sep. 2015, Art. no. 032801.

Piyush Kumar Garg received the B.Tech. degree from the College of Technology and Engineering, Udaipur, India, in 2015, and the M.Tech. degree from IIT(ISM) Dhanbad, Dhanbad, India, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, IIT Patna, Patna, India.

His current research interests include social network analysis, crisis response, and information retrieval.

Roshni Chakraborty received the M.E. degree from the Indian Institute of Engineering Science and Technology Shibpur, Howrah, India, in 2014, and the Ph.D. degree from IIT Patna, Patna, India, in 2020.

She is an Assistant Professor with the Institute of Computer Science, University of Tartu, Tartu. She was a Post-Doctorate Research Fellow at Center for Data-Intensive Systems (Daisy), Aalborg University, Denmark, from November 2020 to November 2022. Her research interests include computational journalism, social computing, time-series analytics and signed networks.

Sourav Kumar Dandapat (Member, IEEE) received the B.E. degree from Jadavpur University, Kolkata, India, in 2002, and the M.Tech. and Ph.D. degrees from IIT Kharagpur, Kharagpur, India, in 2005 and 2015, respectively.

He has been an Assistant Professor at IIT Patna, Patna, India, since February 2016. His current research interests include computational journalism, social computing, information retrieval, and human–computer interaction.