6-14-2021

Information Overload in Crisis Management: Bilingual Evaluation of Embedding Models for Clustering Social Media Posts in Emergencies

Markus Bayer
Technical University of Darmstadt, bayer@peasec.tu-darmstadt.de

Marc-André Kaufhold
Technical University of Darmstadt, kaufhold@peasec.tu-darmstadt.de

Christian Reuter
Technical University of Darmstadt, reuter@peasec.tu-darmstadt.de

INFORMATION OVERLOAD IN CRISIS MANAGEMENT: BILINGUAL EVALUATION OF EMBEDDING MODELS FOR CLUSTERING SOCIAL MEDIA POSTS IN EMERGENCIES

Research Paper

Bayer, Markus, Technical University of Darmstadt, Darmstadt, Germany, bayer@peasec.tu-darmstadt.de

Kaufhold, Marc-André, Technical University of Darmstadt, Darmstadt, Germany, kaufhold@peasec.tu-darmstadt.de

Reuter, Christian, Technical University of Darmstadt, Darmstadt, Germany, reuter@peasec.tu-darmstadt.de

Abstract

Past studies in the domains of information systems have analysed the potentials and barriers of social media in emergencies. While information disseminated in social media can lead to valuable insights, emergency services and researchers face the challenge of information overload as data quickly exceeds the manageable amount. We propose an embedding-based clustering approach and a method for the automated labelling of clusters. Given that the clustering quality is highly dependent on embeddings, we evaluate 19 embedding models with respect to time, internal cluster quality, and language invariance. The results show that it may be sensible to use embedding models that were already trained on other crisis datasets. However, one must ensure that the training data generalizes enough, so that the clustering can adapt to new situations. Confirming this, we found out that some embeddings were not able to perform as well on a German dataset as on an English dataset.

Keywords: Social Media Clustering, Information Overload, Crisis Informatics, Unsupervised Machine Learning.

1 Introduction

In the past 20 years, social media was not only used in everyday life but also during almost every major natural and man-made crisis, including the 2001 September 11 attacks, 2012 Hurricane Sandy, 2013 European floods, or the ongoing COVID-19 pandemic, to gather and spread disaster-related information (Mirbabaie & Zapatka, 2017; Palen & Hughes, 2018; Reuter & Kaufhold, 2018). This user-generated content comprises multimedia files (e.g. audio, photo, video) and textual information (e.g. situational updates, public mood, specific information) that has the potential to increase situational awareness and improve crisis response for crisis volunteers, emergency personnel, and other involved persons (Hughes & Palen, 2009; Olteanu et al., 2015). However, emergency services face issues of information quality and overload under the time-critical constraints of large-scale emergencies, which requires an efficient and effective way to structure the incoming volumes of social big data (Imran et al., 2015; Olshannikova et al., 2017). To address this challenge, many algorithms, frameworks, methods, and tools arose from the research areas of machine learning (Alam et al., 2020; Imran et al., 2018), information systems (Eismann et al., 2018; Fischer et al., 2016), social media analytics (Fan & Gordon, 2014; Stieglitz, Mirbabaie, Ross, et al., 2018), and crisis informatics (Hagar, 2010; Palen & Anderson, 2016).

To tackle information overload, researchers came up with supervised machine learning classifiers to estimate the relevance of postings to the situation (Abel, Hauff, & Stronkman, 2012; Habdank et al., 2017) or to categorize the postings into groups of different information types (Caragea et al., 2011; Imran et al., 2017, 2013; Nguyen et al., 2016). Despite the value of such supervised approaches, the gathering and labelling of case-specific training data is costly and highly time-consuming, which is particularly problematic in disaster situations (Kaufhold, Bayer, et al., 2020). Little research, however, has focused on unsupervised techniques such as clustering, which utilize similarity measures to identify patterns in the data to form groups and do not need any training based on labelled data (Xu & Wunsch, 2005). To find similarities between social media messages, it is necessary to convert them into vectors, which ideally share a similar contextual meaning. The contextual conversion can be produced by embedding models, such as the Word2vec model (Mikolov et al., 2013).

In our literature review, we identified a variety of potentials for research. First, researchers have brought up both general and domain-dependent embedding models (Alam et al., 2020; Godin et al., 2015). However, there is a lack of knowledge on which ones perform better in emergencies. Furthermore, current clustering approaches are primarily learned on English data (Li et al., 2018).This calls for a cross-language evaluation of different embedding models. Since disasters are often characterized by time-critical constraints, performant clustering algorithms and embedding models are required to allow an almost real-time application. To further increase the interpretability and value of the built clusters in a disaster situation, a short description or a label for each would be desirable (Alam et al., 2020). In summary, the goal of this work is to establish an efficient and effective methodology for clustering social media posts in disaster situations, so that emergency personnel and other actors can gain a quick overview of the gathered data. The main aspect is to evaluate different state-of-the-art word and document embedding methods with respect to clustering and disaster situations. Thus, we seek to answer the following research questions: *To what degree are domain-dependent embeddings helpful for clustering the dynamic data in emergencies (RQ1)? Which embeddings are more invariant with respect to the language of the data (RQ2)? Which embedding methods are suitable for the time-critical analysis of Twitter data in emergency situations (RQ3)?*

In order to answer these questions, the paper is structured as follows: First, we present related work on the foundations and techniques regarding information overload and clustering before outlining the research gap (section 2). Based on these foundations, we present the method and implementation of embedding models and the clustering approach (section 3). Thereafter, we describe the selected datasets, evaluation criteria, and results of the evaluation (section 4). In summary, we evaluate 19 methods for creating the document embeddings on two different datasets and use k-means for clustering, from which the produced groups are evaluated with internal evaluation methods. Furthermore, we discuss the prospect for automatically labelling the clusters (section 5). The paper finishes with a discussion of the results and implications, the conclusion, and an outlook (section 6).

2 Related Work

In information systems, social media analytics is defined as “the process of social media data collection, analysis, and interpretation in terms of actors, entities, and relations” (Stieglitz et al., 2014). It aims to combine, extend, and adapt methods and tools for the analysis of social media data (Fan & Gordon, 2014; Stieglitz, Mirbabaie, Ross, et al., 2018). When applied to the domain of crisis informatics, it is often combined with interfaces for real-time analytics and machine learning algorithms (Imran et al., 2018; Onorati et al., 2018). This section presents techniques to mitigate information overload in crises and proposes clustering as a solution to reduce the amount of data presented to emergency personnel.

2.1 Foundations and Techniques to Mitigate Information Overload in Crises

When tens of thousands of social media messages are generated during large-scale emergencies (Reuter et al., 2019), authorities have to deal with the issue of information or social media overload (Lansmann & Klein, 2018). Amongst others, the concept is examined in management information systems and can be caused by personal factors, information characteristics, task and process parameters, organizational design, or information technology (Eppler & Mengis, 2004). With regard to characteristics and technology, information overload is often defined as “[too much] information presented at a rate too fast for a person to process” (Hiltz & Plotnick, 2013) and implies the danger of getting lost in data which may be irrelevant to the current task at hand and of data being processed and presented in an inappropriate way (Keim et al., 2008). In past crisis informatics research, several prototypes and techniques were explored to mitigate information overload in large-scale emergencies, whereof some are outlined in Table 1. The first intuitive step to find relevant (or to filter out irrelevant) information is the use of *search engines* that facilitate simple keyword-based or complex Boolean search queries. While these are often embedded in social media platforms such as Facebook or Twitter directly, developers can use platform search APIs to integrate their results into supportive third-party applications (Imran et al., 2015). In such applications, search engines are often combined with additional functionality allowing the *filtering of information* by metadata, such as language, location, social media platform, or time (Kaufhold, Rupp, et al., 2020). While this functionality is often provided by specific forms, *interactive visualizations*, such as charts, maps, timelines, or word clouds, can be used to reduce the displayed data by a specific gesture (Onorati et al., 2018). For instance, if a pie chart displays the numbers of positive, neutral, and negative sentiment messages, a click on the positive “wedge” could trigger attached list or map views that only show messages with positive sentiment.

Technique	**Description**
Search engine	Formulation of simple keyword search or complex Boolean search query engine, including operators such as “and”, “or”, and “not”, which are embedded into web interfaces or provided by search APIs (Imran et al., 2015).
Metadata filtering	Filtering of information by metadata, such as language, location, social media platform, or time, which is often combined with search functionality (Kaufhold, Rupp, et al., 2020).
Interactive visualizations	Use of interactive visualizations, such as charts, maps, timelines, or word clouds, to reduce the displayed data to a specific subset by a specific gesture (Onorati et al., 2018).
Message classification	Use of supervised machine learning models to classify information as relevant or irrelevant for a specific emergency (Habdank et al., 2017) or to categorize them into humanitarian information types (Alam et al., 2020).
Message clustering	Categorization of text documents into similar groups using similarity metrics and unsupervised machine learning techniques, which do not require labelled data for training (Fahad et al., 2014).
Information summarization	Automatic and real-time algorithms that use extraction or abstraction techniques to provide a general information summary of a disaster event (Rudra et al., 2018).

Table 1: Overview of different techniques deployed to mitigate information overload in crises, disasters, or emergencies.

While search engines and metadata filtering can be useful measures for reducing information overload, too restrictive search queries can lead to the problem that emergency services may miss out relevant information. This is especially true for location-based filtering since only a small amount of social data posts contain geocoordinates. Here, machine learning algorithms can help to find relevant information after data was collected. For instance, *message classification* techniques often apply supervised machine learning to binarily classify social media posts as relevant or irrelevant for a specific emergency (Habdank et al., 2017) or to categorize them into humanitarian information types, such as affected individuals, infrastructure and utilities, donations and volunteering, caution and advice, sympathy and support, other useful information, or not applicable (Alam et al., 2020). However, these techniques are often tailored to specific emergencies, thus not being generally applicable, and rely on the time-intensive labelling of data and model training. This stands in contrast to unsupervised *message clustering* techniques that categorize text documents into similar groups using similarity metrics and unsupervised machine learning techniques; hence, labelled data for training is not required (Fahad et al., 2014). Considering the human capacity of information processing, Miller (Miller, 1956) suggests "organizing or grouping the input into familiar units or chunks" to overcome such limitations. In accordance, further research suggests that 'chunking' social media messages by specific tools positively influences emergency managers' intention to use social media during emergencies (Rao et al., 2017). However, clusters might not be self-explanatory and require a useful summary of cluster content or at least describing labels (Gründer-Fahrer et al., 2018). On the one hand, information summarization approaches might be used for the automatic and real-time extraction or abstraction of a dataset or subset to provide an information summary (Rudra et al., 2018). On the other hand, the automatic labelling procedure outlined in section 5 can be relevant for obtaining a concise overview of the cluster contents.

2.2 Clustering, Embeddings, and their Application in Crisis Informatics

Clustering is performed by unsupervised machine learning methods, which can be applied when no class is to be predicted (Fahad et al., 2014). Thus, the data should rather be grouped into natural clusters (Witten et al., 2016). These groups are most frequently found with some kind of similarity measure for comparing the data. In contrast to supervised machine learning techniques, labelled data is not required, which makes it an interesting field of study. Authors like Imran et al. (Imran et al., 2018) already pointed out that the considerable amount of labelled data constitutes a challenge in disaster situations. There are many algorithms suited for clustering data, such as k-means (Hartigan & Wong, 1979) and mean-shift clustering (Cheng, 1995). Further existing work looked at clustering techniques in general (Xu & Wunsch, 2005), reviewed techniques in contrast to k-means (Jain, 2010), and with respect to Twitter data (Alnajran et al., 2017). However, k-means is a widely used algorithm which is also applied in crisis informatics (Alam et al., 2020), because it is simple and computationally efficient (Jain, 2010). Since clustering methods rely on numerical data as input, it is important to find a good numerical representation of the textual nature of a tweet. More precisely, it would be necessary to map the Twitter posts into a latent space that has an inherent structure based on contextual similarity. This means that similar posts get a similar number-vector and different posts are distant in this continuous space.

This challenge can be addressed with *embedding models*. The goal of representing text as numerical vectors with meaning can be abstracted as the goal to represent words as numerical vectors with meaning. This dates back to the 1960s and is based on the distributional hypothesis (Harris, 1954), which can be interpreted as "a word is characterized by the company it keeps" (Firth, 1957). With the Word2Vec model from Mikolov et al. (2013), word embeddings became one of the biggest trends in Natural Language Processing (NLP) research until the present. Mikolov et al. (2013) proposed a shallow neural network for building the embeddings, reaching state-of-the-art performance in various NLP tasks. As a next step, the words of a sentence or document can be further processed to get a single *sentence* or *document embedding*, respectively. More or less based on the Word2Vec model, various word and document embedding approaches have been proposed in recent years. Many embedding models are self-supervised, meaning that they create their own training labels without any human annotator (Liu et al., 2020). To model the distributional hypothesis, they need a large dataset for training. If a model has never seen a certain word (also called out-of-vocabulary word), it is hard to map it to a meaningful vector.

Training a new model for every new disaster situation and other fields of use is not feasible; therefore, people often rely on pretrained embeddings. A model like GloVe (Pennington et al., 2014) that has been trained on large common crawls, Wikipedia pages, and Twitter data, has vocabulary sizes varying from 400,000 to 2.2 million unique words. It seems worthwhile to examine if it would be more beneficial to train the embedding models using general or domain-related data, such as other past crisis datasets.

One of the main purposes of clustering is to gain insights from the underlying structure by discovering natural groupings (Jain, 2010). These groupings can reach from being very obvious to being latent, depending on the data and the observer. With regard to crisis informatics, a brief overview of the different categorization or grouping possibilities for social media posts is shown in Table 2. Most existing work categorizes posts as either relevant or not relevant for a specific emergency, while some try to map predefined humanitarian categories. The different groupings show that supervised classifiers cannot cover all different possibilities and an unsupervised clustering method might be helpful. Emergency situations are highly dynamic, so it is obvious that the goal is to not predefine the groups. Clusters from a clustering algorithm can and should be highly different, given new disaster situations.

Grouping possibilities
“damage”, “personal opinion”, “caution and advice”, “not relevant” (Alam et al., 2020)
“after effect”, “personal opinion”, “updates”, “other useful information”, “not relevant” (Alam et al., 2020)
“personal only”, “informative (direct)”, “informative (indirect)”, “informative (direct or indirect)”, “other” (Imran et al., 2013)
“caution and advice”, “casualties and damage”, “donations of money, goods or services”, “people missing, found or seen” “information source”, “other” (Imran et al., 2013)
“off-topic”, “on-topic and relevant to situational awareness”, “on-topic and not relevant to situational awareness” (Vieweg, 2012)
“off-topic”, “on-topic and not relevant to situational awareness”, “social environment”, “built environment”, “physical environment” (Vieweg, 2012)
“off-topic”, “on-topic and not relevant to situational awareness”, + 32 information types (Vieweg, 2012)
“relevant/informative”, “not relevant/informative” (Abel, Hauff, Houben, et al., 2012; Habdank et al., 2017; Kaufhold, Bayer, et al., 2020; Li et al., 2017; Nguyen et al., 2016; Spielhofer et al., 2016; Verma et al., 2011)

Table 2: Overview of different grouping possibilities proposed in various papers.

2.3 Research Objectives

The review identified a variety of measures to reduce information overload (Table 1), highlighting that chunking or clustering information positively influences emergency managers’ intention to use social media during emergencies (Eppler & Mengis, 2004; Rao et al., 2017). Apart from the emphasis on crisis situations, the evaluation performed in this paper can be seen as an intrinsic evaluation method for embedding-based clustering. Intrinsic evaluations are performed within the word vectors themselves, especially where no classifier is trained on them (Baroni et al., 2014; Schnabel et al., 2015). One of the most popular evaluation methods in this sector is the word similarity task, where the cosine similarity of word embeddings is compared to human judgements (Faruqui et al., 2016). Extrinsic evaluations on the other hand apply the embeddings on a downstream NLP task, for example sentiment analysis or POS tagging (Nayak et al., 2016). The contextually most related work to ours is conducted by Alam et al. (2019). They propose a system for clustering social media data in crisis situations using an embedding approach. The authors train a new Word2Vec model on a crisis dataset. These contextual embeddings are then averaged to get a document vector. Afterwards, the authors perform PCA to reduce the computational cost, which they unfortunately do not specify in numbers. On the lower dimensional data, they also perform k-means with an adaptive k-search approach. The resulting clusters are analysed and labelled manually by humans. In contrast to their work, we consider different approaches and take a deeper look into the assessment of them. Our goal is to evaluate and answer specific research questions concerning the embedding creation, since this is the most important step for clustering the posts.

Clustering in disaster situations is often found in relation to the temporal or spatial dimension (Lu & Zhou, 2016; Pohl et al., 2015; Sakai et al., 2015). In connection with textual data, state-of-the-art literature is, aside from Alam et al. (2019), lacking in this field. Yin et al. (2015) cluster event specific topics in emergencies but still use TF-IDF vectors instead of embeddings, which are in most cases proven to be superior. Abstracting the work from the crisis context, Dai, Bikdash and Meyer (2017) propose a clustering method in the area of public health surveillance. Their goal is not to find different groups, but to model a classifier that binarily decides if a post is related or unrelated to the health-topic. The authors are using the Word2Vec model to create embeddings, but do not consider other possible models. Our work can be compared to Li et al. (2018), where the authors conduct a large evaluation of different word and document embedding approaches for classification tasks in crisis situations. In a similar work, Khatua et al. (2019) compare word embeddings in the context of virus outbreaks. The difference of the work of both to ours lies in the task setting and evaluation method. Their goal is to classify posts in a supervised way that leads to an extrinsic evaluation method for assessing the embeddings, whereas we work in an unsupervised setting for clustering and propose an intrinsic evaluation method for embeddings. Similar to their work, we apply embeddings learned from previous disasters, to examine if they lead to a better separation of the data (RQ1). Beyond that, we put a special emphasis on language-invariance and time-criticality based on RQ2 and RQ3.

3 Architecture and Embedding Models

Our process of clustering social media posts is shown in Figure 1. The unstructured text is processed in an NLP-pipeline, which tokenizes the tweets according to the tokenizer used for the embedding model. This is necessary as different embedding models use different tokenization techniques. In the next step, an embedding creation approach is applied. These approaches comprise two differing groups of models. The sentence (or document) embedding models are directly building vectors for the whole tweet, while the word embedding models just create word vectors for every word in the tweet. If a sentence embedding model is chosen, the vectors are immediately fed into the clustering algorithm. If we take a word embedding model, the word embeddings of a tweet are getting averaged (Avg) or the minima and maxima (MinMax) vector are concatenated. We chose k-means (Hartigan and Wong, 1979) as clustering algorithm. K-means is dependent on the hyperparameter *k*, which specifies how many clusters should be formed.

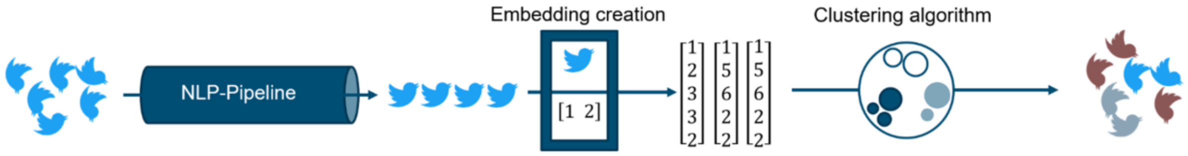

Figure 1: The clustering process, beginning with an NLP-pipeline for tokenization, over the embedding creation, to the clustering algorithm.

3.1 Software and Hardware Architectures

As a software basis for the clustering algorithm and evaluation metrics (see sections 4 and 5), we chose scikit-learn. The implementation of the embedding models depends on the individual instructions given by the authors or relevant libraries of the models. The same applies to the tokenization steps that are required for preprocessing the data according to the models. The system used for the implementation and evaluation has an Intel Xeon processor with a single core at 2.3 GHz. A Nvidia Tesla K80 with 12 gigabyte RAM is used for computing matrix-based calculations of several representation models. Furthermore, 25 gigabyte RAM are available for storing the representations efficiently.

3.2 Used Embedding Models for Comparison

There are many embedding approaches that have been developed since the proposal of Word2Vec. For a suitable selection of approaches, we considered the work of Li et al. (2018) and new emerging trends such as the state-of-the-art embedding model BERT by Devlin et al. (2018). The models used here are listed below with a brief description, while detailed descriptions can be found in the original papers.

- **Word2Vec** (W2V) – word embeddings (Mikolov et al., 2013). This model is one of the first that used shallow neural networks to create word embeddings. Due to its popularity, many different datasets were used by various authors to train the network. We are using three pretrained models:
  - Twitter Word2Vec model: Trained with 400 million tweets (Godin et al., 2015).
  - Crisis Twitter Word2Vec model: Trained with 52 million crisis-related tweets (Imran et al., 2016)
  - Crisis Twitter Word2Vec model: Trained with 364 million crisis-related tweets (Alam et al., 2018a, 2018b)
- **GloVe** – word embeddings (Pennington et al., 2014). This model combines the idea of Word2Vec with word occurrence statistics. The pretrained Twitter model with 2 billion tweets is used.
- **FastText** – word embeddings (Bojanowski et al., 2017; Joulin et al., 2017). The main difference of this Word2Vec extension is that it splits the words into n-grams, which has the advantage of getting good vectors for rare and out-of-vocabulary words. Another advantage is that FastText is already pretrained on 157 languages, which is especially important in disaster situations.
- **InferSent** – sentence embeddings (Conneau et al., 2017). This model is, in contrast to the previous models, trained on a supervised task. The authors are using a bidirectional LSTM, followed by a comparison layer and fully connected layers, to solve the SNLI task (Bowman et al., 2015). We are using the GloVe and FastText pretrained variants.
- **Universal sentence encoder** – sentence embeddings (Westerink & Vijverberg, 2018). This model extends the approach of InferSent to two more tasks (question answering and translation) with different architectures. We use the base and large pretrained model.
- **Sent2Vec** – sentence embeddings (Pagliardini et al., 2018). The goal of Sent2Vec is to assign each word a word embedding so that the average of all of them in the document constitutes a good vector. We are using the Twitter-learned embeddings.
- **Sentence-BERT** – sentence embeddings (Reimers & Gurevych, 2019; You et al., 2019). Bidirectional Encoder Representations from Transformers (Devlin et al., 2018) are considered to be the state of the art in the embedding creation task. The word embeddings are created in a contextualized manner, i.e., that not only the word itself is considered but also the context in which the word appears. Since finding the most similar sentence in 10,000 sentences requires about 65 hours (Reimers & Gurevych, 2019), we take advantage of the BERT extension from Reimers and Gurevych (2019), who state to solve this in 5 seconds, while maintaining the accuracy of BERT. We evaluate the base and large model.

4 Evaluation of Embedding Models

4.1 Datasets and Measurements

In order to answer our research questions, we decided to use a German and an English flooding dataset for evaluation. First, the German data is based on the 2013 European Floods, which had a severe impact on Germany (Kaufhold & Reuter, 2016) and contains about 4,000 posts related to the flooding. Second, the English set was crawled during the 2013 Colorado floods and contains about 1,000 posts (Olteanu et al., 2015). Both sets were chosen because they are referring to a similar scenario, are different in language, and also contain labels. Labelled data can be helpful for further work concerning an external evaluation. Furthermore, it is important that both sets are gathered during a flooding scenario in different languages, which is required to answer RQ2.

Evaluating clusters is a difficult task, because the formation of the clusters is ambiguous. We decided against an *external evaluation*. External evaluation assesses the clustering performance with an existing ground truth. By having such a ground truth, the benefits of clustering would be missed. Especially in the disaster setting it can be difficult, since several different groupings can be formed, as we have seen in section 2. Instead, we chose an *internal evaluation*, where the clusters themselves are analysed by the means of a measurement. This approach comes with the problem that the clustering algorithm can be optimized on this measurement. But since we are only using k-means, this concern is reduced. As measurement criteria, we use the Silhouette Coefficient (Rousseeuw, 1987), Calinski-Harabasz Index (Caliñski & Harabasz, 1974), and Davies-Bouldin Index (Davies & Bouldin, 1979). The Silhouette Coefficient measures how fitting an object is to its cluster, compared to the other clusters. The measure ranges from -1 to 1; a higher score is better. The Calinski-Harabasz Index measures the ratio of the sum of between-cluster distances and of inter-cluster distances. Again, a higher score relates to the better clustering performance. The Davies-Bouldin Index on the other hand indicates better clustering performance with a lower score, which is achieved when the groups are farther apart and less dispersed.

4.2 Results for the English and German Datasets

The results of the clustering task for the English and the German dataset are shown in Table 3 and Table 4. The three best scores for each evaluation metric are highlighted. We choose k based on section 2.2, where already several grouping possibilities are listed. To have some kind of adaptive evaluation, we test each model with a k ranging from 4 to 10. The majority of algorithms performed best when using a k of 5. The evaluation results shown in the tables are related to a k-means clustering with a k of 5, representing 5 clusters.

Embedding Model	**Variant**		**Embedding Creation**		**Cluster Creation**		**Silhouette**		**Calinski Harabasz**		**David Bouldin**	
W2V.Twitter	Avg	MinMax	99.06	38.75	0.47	0.96	0.07	0.05	102.93	55.50	1.71	2.85
W2V.Crisis1	Avg	MinMax	51.45	25.58	**0.28**	0.79	**0.70**	0.10	**172.99**	72.39	**0.22**	1.62
W2V.Crisis2	Avg	MinMax	127.86	131.08	**0.32**	0.68	**0.69**	0.09	**143.99**	76.61	**0.22**	**1.53**
Glove	Avg	MinMax	292.11	256.57	**0.40**	0.48	0.08	0.14	111.33	125.36	2.60	2.32
FastText	Avg	MinMax	166.44	163.73	0.54	0.76	0.06	**0.16**	64.51	**254.26**	2.67	1.61
USE	Base	Large	**1.36**	**0.08**	0.81	0.87	0.03	0.02	36.19	36.16	4.14	4.16
InferSent	Glove	FastText	112.99	91.49	4.15	5.33	0.05	0.05	48.08	45.87	3.37	3.56
Sent2Vec	Unigrams		260.64		0.99		0.03		43.77		3.47	
SBERT	Base	Large	25.68	70.76	0.87	1.24	0.06	0.07	65.07	66.48	2.83	2.99
SBERT.sts	Base	Large	**23.65**	70.26	1.02	1.20	0.04	0.05	49.42	47.32	3.74	3.89

Table 3: Evaluation of the English clustering task; the three best scores per metric (bold) and models with none (red), one (yellow), or multiple (green) high scores are marked.

For the English task, both Word2vec models based on a crisis dataset with the average sentence embeddings performed the best when inspecting the Silhouette Score (crisis dataset 1: 0.7) and the Davies Bouldin Score (crisis dataset 2: 0.22). The Calinski Harabasz Score of both is also particularly good, taking the second and third place behind the English FastText model (254.256). Regarding the overall clustering quality, it is also assumable that the FastText model performs good, even if the models that were learned on the crisis dataset seem to be better suited. Most of the other embedding techniques reach less promising results. Both variants of Universal Sentence Encoders (USE) are consistently the worst at creating embeddings that are good for forming clusters with the k-means method. In turn, they are the fastest algorithms with regard to the embedding creation time. Comparing the crisis Word2Vec models in relation to this evaluation metric, the first is more than two times faster than the second, resulting from the fact that it was trained on a smaller dataset. The FastText model is about 110 and 40 seconds slower than the two better performing models. In terms of cluster creation time, only InferSent is significantly deviating, but this can be considered negligible with about 5 seconds.

Embedding Model	Variant		Embedding Creation		Cluster Creation		Silhouette		Calinski Harabasz		David Bouldin	
W2V.Twitter	Avg	MinMax	101.67	71.82	1.31	2.73	0.23	0.23	549.62	639.45	**1.15**	2.06
W2V.Crisis1	Avg	MinMax	54.40	**50.17**	**1.15**	2.07	0.22	0.13	927.60	474.89	1.69	2.59
W2V.Crisis2	Avg	MinMax	128.56	129.21	**1.15**	1.88	**0.24**	0.15	912.67	518.65	**1.41**	2.43
Glove	Avg	MinMax	291.58	254.34	**0.91**	1.31	**0.28**	0.19	**1282.74**	542.23	1.87	2.24
FastText	Avg	MinMax	157.67	154.99	1.52	1.62	0.15	**0.27**	**1160.52**	**8873.64**	2.34	**1.18**
USE	Base	Large	**1.53**	**0.28**	2.03	2.04	0.11	0.11	357.96	358.17	2.95	2.96
InferSent	Glove	FastText	127.93	105.07	15.08	22.00	0.14	0.11	417.04	371.29	2.79	2.27
Sent2Vec	Unigrams		260.67		2.94		0.14		443.13		2.66	
SBERT	Base	Large	95.61	301.08	3.09	3.74	0.08	0.09	407.61	514.45	2.83	2.84
SBERT.sts	Base	Large	109.67	305.15	3.42	4.52	0.08	0.08	261.76	228.70	3.10	3.24

Table 4: Evaluation of the German clustering task; the three best scores per metric (bold) and models with none (red), one (yellow), or multiple (green) high scores are marked.

Inspecting the results of the German dataset, the FastText model again achieves the highest Calinski Harabasz Score. In contrast to the first task, both disaster-related embedding models are not as good with regard to the Silhouette Score and the Davies Bouldin Score. GloVe has the best Silhouette Score (0.276) and Word2Vec, which is based on a Twitter dataset, has the best Davies Bouldin Score (1.149). Inspecting the three best scores of each quality evaluation metric, it seems that the FastText model followed by the GloVe embeddings is best suited for the clustering task in German. As in the previous task, Universal Sentence Encoders are the fastest methods but come with a low clustering quality in comparison to most other embedding methods; only SBERT produces lower results. As observed in the previous task, the cluster creation time of all algorithms except InferSent is similar. Yet, the creation time of up to 22 seconds can no longer be considered as negligible.

It should be noted that the absolute results for the two different tasks cannot be compared with one another because the specific scores are not comparable for different datasets. For the English task, the two Word2Vec models, which were learned with crisis data, performed best. The clustering of the German dataset on the other hand showed that the domain dependent models were not able to maintain these results. This is most likely due to a loss in generalizability since the training was primarily performed with English crisis data. It can be observed that the model that was trained with more data (Word2Vec Crisis 2) also seems to be somewhat better when inspecting the language shift. This is especially important because messages in crisis situations are often communicated in different languages apart from English. To this criterion the FastText model might generalize best to, since it is available in 157 different languages. The model itself performed well on the first task, taking the third place behind the domain dependent models with regard to the clustering quality, and best on the second task. It is interesting to see that in the case of the Word2Vec model average embeddings seem to perform much better than the concatenation of the minimal and maximum word embedding, whereas in the case of the FastText model it is vice versa. This indicates that these specific embeddings learned by the FastText model are more meaningful and unique with regard to the whole tweet. Universal Sentence Encoders have by far the fastest embedding creation processes for both tasks. These speed advances are accompanied by poor cluster quality results. The well performing Word2Vec and FastText models require between 50 to 170 seconds, which might be fast enough for many applications. If a faster model is needed, it may be sensible to reduce the dataset on which, for example, the Word2Vec model is trained. However, such a reduction of the data leads to a reduction in the generalizability for new data.

5 Towards Automatic Cluster-Labelling

Mitigating information overload in emergencies is a complex problem. While it is required to improve the performance (i.e. to allow a near real-time application) and quality of clustering, emergency managers must be able to make sense of the generated clusters to facilitate the practical value of clustering in emergencies (Stieglitz, Mirbabaie, Fromm, et al., 2018). Descriptions or explanations of computed clusters can be achieved in several ways, such as information summaries, labels, or word clouds (see section 2.2). Therefore, we derived criteria that are necessary for effective techniques to describe and explain clusters in emergency situations in a human friendly way, which are enlisted Table 5. The criteria are based on the problem of information overload, the conditions in emergency situations and the human view of explanations, that are put into context with parts of our literature review.

Criteria	Description
Concise	Information explanations of the membership of tweets to a cluster need to be concise in the same way humans tend to prefer selected causation explanations over a complete list (Molnar, 2019).
Covering	A good representation of the cluster content should be comprehensive for the tweets in it. Otherwise, valuable information could be overlooked, which could cost lives in emergency situations (Habdank et al., 2017).
Non-redundant/unique	The explanation of one cluster should be unique, as it can be extracted from the cluster description process by Siroker and Miller (2008).
Non-laborious	Information explanations should at best be formed automatically. Manual actions require time, which is limited in disaster situations (Kaufhold, Bayer, et al., 2020).
Invariant	The explanations should be invariant to domain shifts, since emergency situations are highly dynamic (Li et al., 2017).
Truthful	The explanations of the clusters should be truthful. Truthfulness is important for human-friendly explanations but as stated by Molnar (2019) not as important as having concise explanations.

Table 5: Criteria for human friendly cluster explanations in emergency situations.

A simple approach to extracting an explanation is to show the user the tweet that is closest to the centroid of each cluster. However, this may not cover the other tweets in the cluster. Another approach would be to show the user several samples of posts in the cluster, leading to a higher coverage but in turn to a lower conciseness. An approach with very concise explanations of the clusters is present in the work of Alam et al. (2019), where emergency personnel have to label the different groups after the clustering process. These labels are short titles for the content present in the clusters, comparable to the grouping names in Table 2. However, this approach contradicts the non-laborious criterion. Since disaster situations have to be handled in a real-time manner by emergency personnel, we want to propose a suggestion to automate this process. The idea is to utilize the latent space of the embedding vectors, which is already created when the posts are clustered. For each group we try to construct posts that are as generic as possible but representative for the group as well (invariant and covering). These posts are mapped into the embedding space by using the same embedding method we used when performing the clustering. Then the artificial posts are vectors with the same dimensions as the vectors of the real posts. In this way, we can assign them to the most fitting clusters. In turn, this leaves us with the opportunity to provide the fitting cluster the label we already know from the corresponding self-created tweet (truthful criterion). For the demonstration of the approach, we created constructs of posts for the groups identified by Imran et al. (2013). To ensure that the posts fulfil the criteria to be relevant to the disaster and plausible to the group, we inspected the dataset used by Imran et al. (2013). We extracted the posts for each group, removed the stop words, and calculated the word frequencies. From the list of words and their frequencies, we deleted all posts that were too specific to the crisis. Based on the remaining ones we identified the words that were common in the tweets and sensible for the group to examine the posts containing them. This way we created the generic posts that are shown in Table 6.

Group names	**Generic posts**
Caution and advice	• [disaster_name] + “ warning - Stay safe - take precautions”
Casualties and damage	• “Buildings are damaged and destroyed. #” + [disaster_name] • “Several people were injured #” + [disaster_name] • “Several people were killed #” + [disaster_name]
Donations	• “You can make a donation to the “ + [disaster_name] + “ relief” • “Please provide goods, support or other donations for victims of #” + [disaster_name]
People missing, found, or seen	• “Several people are missing or unaccounted #” + [disaster_name] • “Please help find this person - contact us for any pointers #” + [disaster_name] • “Several people have been located and are alive #” + [disaster_name]
Information Source	• “Photos of “ + [disaster_name] + “ http:// t.co/random #report #documenting” • “News: A video of the “ + [disaster_name] “

Table 6: Generic posts for different humanitarian categories

For example, a generic tweet for “Casualties and damage” can contain important terms like “buildings”, “damaged”, and “destroyed”, which are very often stated in damage-related posts. This construct is ended by a hashtag concatenated with the disaster-related keyword that was chosen by the user or by the disaster name if an event detection system has been executed beforehand (invariant to other situations). This specific information is important because otherwise this tweet could have never been part of the original dataset. These constructed posts are particularly beneficial, since they contain all the necessary information and no human effort is needed by the time they are used.

Eventually, it is possible that one cluster has more than one label (multi-labelling), giving an overview over several topics that could be contained in it. In this way, we can have much more labels and self-constructed posts than groups for many different cases. In the end, we contradict the uniqueness criterion, by accepting different clusters to have the same label. If clusters have no label, the cluster and the posts in it are called “not identifiable” or “other”.

6 Discussion and Conclusion

In this paper, we examined techniques for clustering social media posts in emergencies based on their textual content as a means to reduce information overload. More specifically, we evaluated 19 different embedding methods that are suitable for building clusters of similar postings. The answers to the research questions that were proposed in the beginning are based on the evaluation in section 4.3.

To what degree are domain-dependent embeddings helpful for clustering the dynamic data in emergencies (RQ1)? To answer this question, we evaluated two Word2Vec models that were trained with domain-dependent crisis data and compared them to models that were trained without any specific context. Regarding the English task, the domain-dependent models are in fact in two of three clustering scores superior. However, inspecting the German task, the domain-independent, but language-specific model FastText is achieving higher scores. The two crisis Word2Vec models are still performing reasonably well on the German task, but it is noticeable that they are not able to generalize so satisfactory with regard to the language shift, since the training data primarily consisted of English data. This concerns the second research question, which asks for the best language-invariant models.

Which embeddings are more invariant with respect to the language of the data (RQ2)? It can be observed in both tables that FastText with MinMax sentence embeddings is a particularly good choice when dealing with the problems at hand, even reaching best scores for the German task. Pretrained FastText models are available in 157 different languages, which makes them, apart from its good score, a sensible option for this case. As stated before, the other models like the pretrained Word2Vec or SBERT variants produce inferior results on the German dataset, since they were trained on English data and have never seen most of the German words before. This means that they are not able to infer any meaningful vectors for non-English words, resulting in a worse clustering performance than a model that was trained with data from different languages.

Which embedding methods are suitable for the time-critical analysis of Twitter data in emergency situations (RQ3)? The third research question unites the other research questions and combines them with the time-criticality of disaster situations. To address this, we separately measured the embedding and cluster creation time in the evaluation. The embeddings created by InferSent set aside, cluster creation time seems to be almost invariant with respect to the other models. Inspecting the embedding creation times, USE was by far the fastest method for clustering the data, taking less than two seconds. But since this method does not perform well, we advise to take a disaster pretrained Word2Vec model when dealing with English data, and FastText when dealing with other languages. While the Word2Vec based models seem to be a bit faster, both can take up to three minutes in our evaluation.

6.1 Practical and Theoretical Implications

In this paper, we provide an overview and comparison of embedding models across two languages and propose an intrinsic embedding evaluation task (C1), give advice on the implementation of clustering and embedding approaches (C2), derive criteria for cluster explanations and propose a method for automatic post-labelling (C3), contribute with findings on the applicability of general and domain-dependent embedding models (C4), and propose a system for reducing information overload in social media streams (C5).

Comparison of embedding models across two languages and the proposal of an intrinsic embedding evaluation task (C1). The paper provides a comprehensive overview and comparison of the performance and quality of 19 embedding methods for clustering. By comparing datasets from two languages, it outlines language-specific conditions for the performance and quality of embedding models. While most existing approaches in crisis informatics examined clustering based on the temporal or spatial dimension (Lu & Zhou, 2016; Pohl et al., 2015; Sakai et al., 2015), this work contributes with the implementation and comparison of text-based clustering approaches. With this evaluation, we propose a generally new intrinsic evaluation task that gives insights on the embeddings and is easily adaptable for other domains.

Advice on the implementation of clustering and embedding approaches (C2). The evaluation of the proposed clustering method shows that the results are highly dependent on the input data, i.e., the embedding representations. This can be seen as an advantage of the proposed system as other works in this field do not evaluate multiple models or neglect a comparison of domain dependence and generalization capabilities of their systems (Comito et al., 2019; Curiskis et al., 2020; Dai et al., 2017). However, these methods consider other important factors that are further described in the next section. While existing work on embedding models focused an extrinsic evaluation method on classification tasks (Khatua et al., 2019; Li et al., 2018), our intrinsic evaluation on the clustering task revealed further advice for the implementation of embedding models. In our evaluation, disaster pretrained embeddings performed well but were not able to generalize to languages other than that of the pretrained data. Dealing with other languages it is sensible to either use a model being trained for this case (FastText) or to retrain the model with additional language data. In our tests, we evaluated the Sentence-BERT model by Reimers and Gurevych (2019), which failed to produce satisfactory results.

Derivation of criteria for cluster explanations and the proposal of a method for automatic cluster-labelling in emergency situations (C3). In section 5, we derived criteria for human friendly cluster explanations in emergency situations. Furthermore, in contrast to a manual labelling process (Gründer-Fahrer et al., 2018), we proposed a method for automatically labelling the clusters. This method relies on self-constructed tweets that could reflect humanitarian categories (Imran et al., 2013), are as general as possible with respect to the event, and are as fitting as possible to the group at the same time. These posts are then mapped into the embedding space, where they are assigned to the best-matching clusters.

Applicability of general and domain-dependent embedding models (C4). Domain-dependent embeddings are helpful for clustering if the domain is fitting. If the domain is not fitting (for example due to a language shift), a highly general model would be better suited. This stands in contrast to the findings of Khatua et al. (2019) who in their extrinsic evaluation come to the conclusion that domain dependent embeddings mostly outperform generic pre-trained embeddings. It is similar to the result of Li et al. (2018) who observed that their crisis-specific embeddings are more suitable for specific crisis tasks. We can deduce that depending on the unsupervised task, it is important to consider different embedding models. Concluding from a different path of reasoning by testing language invariance, we can also say that the currently existing crisis-specific embeddings are not able to generalize as well to other situations. This fact leads to further research opportunities in which an embedding training dataset consisting of various disaster situations and different languages could be created.

A system for reducing information overload in social media streams (C5). The proposed system (Figure 1) in combination with the insights from the evaluation and the automatic cluster-labelling procedure is suitable for reducing the overload of user-generated content in social media by finding meaningful and representative clusters. Crisis and emergency personnel can utilize this implementation to gain an overview over the online discourse and pick the right set of posts for their specific response activities (Kaufhold, Rupp, et al., 2020). In addition to this area of application, the system is also beneficial for all purposes that require compressed information of social media, such as brand management in the field of business analytics. Similarly, the clustering system can be beneficially employed in the cybersecurity domain to foster the process of detecting security events. Furthermore, our findings can enrich the clustering scheme by Alves et al. (2021) or similar works in this domain.

6.2 Limitations and Outlook

For future work it would be interesting to see if newer embedding models based on a large corpus of different disaster situations, such as CrisisLexT26 (Olteanu et al., 2015), would be beneficial to the proposed problem. However, as these are only available in a limited number of languages, it might be sensible to create a novel dataset containing different disaster situations combined with general Twitter data that may contain more languages. In this way, the new advances of the state-of-the-art embedding models can be combined with the domain dependence that works well according to our evaluation. Furthermore, the inclusion of the wider range of Twitter data leads to a higher generalizability for other languages. It would also be interesting to create more multilanguage evaluations, since we only evaluated German and English datasets. Especially evaluations for low-resource languages might lead to further insights that are important for dealing with the information overload in crisis situations. While this evaluation focuses on Twitter messages, which are inherently limited to 280 characters, the evaluation of larger texts from social media such as Facebook might lead to different results.

The general clustering framework can be enriched with insights from different works that emphasise other factors of analysis. For example, a combination of the incremental clustering approach proposed by Comito et al. (2019) should be beneficial when considering the data as incoming streams. Moreover, testing different clustering algorithms, as done by Curiskis et al. (2020), might even increase the performance. Concerning the prospect of automatically post-labelling the clusters, it would be helpful to build as many generic posts as possible. In our work, this method can only be seen as a prospect, requiring further evaluations and modifications to prove the concept. A possible way would be to measure the labelling quality externally with ground truth data. However, as emphasized earlier, this can lead to wrong conclusions. It could be combined with qualitative user research with emergency managers, for instance using interviews or scenario-based walkthroughs, to evaluate the efficiency – in terms of mitigating information overload – and usability of the clustering prototype. Finally, this paper focused on textual content and did not analyse information overload based on multimedia files. Existing studies highlight the application of image filtering techniques for deduplication and relevance assessment in crises (Alam et al., 2020) or unsupervised image segmentation for damage assessment (Küçük Matcı & Avdan, 2020), which could be used to complement text-based methods against information overload.

Acknowledgements. This work has been co-funded by the German Federal Ministry of Education and Research (BMBF) in the project CYWARN (13N15407) and by the LOEWE initiative (Hesse, Germany) within the emergenCITY center.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., Tao, K., & Stronkman, R. (2012). Semantics + Filtering + Search = Twitcident Exploring Information in Social Web Streams Categories and Subject Descriptors. *23rd ACM Conference on Hypertext and Social Media, HT'12*, 285–294. https://doi.org/10.1145/2309996.2310043

Abel, F., Hauff, C., & Stronkman, R. (2012). Twitcident: Fighting Fire with Information from Social Web Streams. *Proceedings of the 21st International Conference Companion on World Wide Web*, 5–8.

Alam, F., Joty, S., & Imran, M. (2018a). Domain adaptation with adversarial training and graph embeddings. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*. https://doi.org/10.18653/v1/p18-1099

Alam, F., Joty, S., & Imran, M. (2018b). Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*.

Alam, F., Ofli, F., & Imran, M. (2020). Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria. *Behaviour & Information Technology (BIT)*, *39*(3), 288–318. https://doi.org/10.1080/0144929X.2019.1610908

Alnajran, N., Crockett, K., McLean, D., & Latham, A. (2017). Cluster analysis of twitter data: A review of algorithms. *ICAART 2017 - Proceedings of the 9th International Conference on Agents and Artificial Intelligence*. https://doi.org/10.5220/0006202802390249

Alves, F., Bettini, A., Ferreira, P. M., & Bessani, A. (2021). Processing tweets for cybersecurity threat awareness. *Information Systems*. https://doi.org/10.1016/j.is.2020.101586

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. https://doi.org/10.3115/v1/p14-1023

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00051

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d15-1075

Caliñski, T., & Harabasz, J. (1974). A Dendrite Method Foe Cluster Analysis. *Communications in Statistics*. https://doi.org/10.1080/03610927408827101

Caragea, C., Mcneese, N., Jaiswal, A., Traylor, G., Kim, H., Mitra, P., Wu, D., Tapia, A. H., Giles, L., Jansen, B. J., & Yen, J. (2011). Classifying Text Messages for the Haiti Earthquake. *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM), May*, 1–10. https://doi.org/10.1.1.370.6804

Cheng, Y. (1995). Mean Shift, Mode Seeking, and Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/34.400568

Comito, C., Forestiero, A., & Pizzuti, C. (2019). Word embedding based clustering to detect topics in social media. *Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019*. https://doi.org/10.1145/3350546.3352518

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. https://doi.org/10.18653/v1/d17-1070

Curiskis, S. A., Drake, B., Osborn, T. R., & Kennedy, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing and Management*. https://doi.org/10.1016/j.ipm.2019.04.002

Dai, X., Bikdash, M., & Meyer, B. (2017). From social media to public health surveillance: Word embedding based clustering method for twitter classification. *Conference Proceedings - IEEE SOUTHEASTCON*. https://doi.org/10.1109/SECON.2017.7925400

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.1979.4766909

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Mlm*.

Eismann, K., Posegga, O., & Fischbach, K. (2018). Decision Making in Emergency Management: The Role of Social Media. *Proceedings of the 26th European Conference on Information Systems (ECIS)*.

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *Information Society*, *20*(5), 325–344. https://doi.org/10.1080/01972240490507974

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*(3), 267–279. https://doi.org/10.1109/TETC.2014.2330519

Fan, W., & Gordon, M. D. (2014). The Power of Social Media Analytics. *Communications of the ACM*, *57*(6), 74–81. https://doi.org/10.1145/2602574

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). *Problems With Evaluation of Word Embeddings Using Word Similarity Tasks*. https://doi.org/10.18653/v1/w16-2506

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*.

Fischer, D., Posegga, O., & Fischbach, K. (2016). Communication Barriers in Crisis Management: A Literature Review. *European Conference on Information Systems (ECIS)*, 1–18.

Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2015). *Multimedia Lab $@$ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations*. https://doi.org/10.18653/v1/w15-4322

Gründer-Fahrer, S., Schlaf, A., Wiedemann, G., & Heyer, G. (2018). Topics and topical phases in German social media communication during a disaster. In *Natural Language Engineering* (Vol. 24, Issue 2). https://doi.org/10.1017/S1351324918000025

Habdank, M., Rodehutskors, N., & Koch, R. (2017). Relevancy Assessment of Tweets using Supervised Learning Techniques Mining emergency related Tweets for automated relevancy classification. *2017 4th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*.

Hagar, C. (2010). Crisis Informatics: Introduction. *Bulletin of the American Society for Information Science and Technology*, *36*(5), 10–12. https://doi.org/10.1002/bult.2010.1720360504

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3). https://doi.org/10.1080/00437956.1954.11659520

Hartigan, A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*. https://doi.org/10.2307/2346830

Hiltz, S. R., & Plotnick, L. (2013). Dealing with Information Overload When Using Social Media for Emergency Management: Emerging Solutions. In T. Comes, F. Fiedrich, S. Fortier, J. Geldermann, & T. Müller (Eds.), *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM)* (pp. 823–827). ISCRAM Digital Library.

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. In J. Landgren & S. Jul (Eds.), *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM)* (Vol. 6, Issue 3/4). https://doi.org/10.1504/IJEM.2009.031564

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing Social Media Messages in Mass Emergency: A Survey. In *ACM Computing Surveys* (Vol. 47, Issue 4). ACM. https://doi.org/10.1145/2771588

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2018). Processing Social Media Messages in Mass Emergency: Survey Summary. *Companion Proceedings of the Web Conference 2018 (WWW '18)*, 507–511. https://doi.org/10.1145/3184558.3186242

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting Information Nuggets from Disaster-Related Messages in Social Media. In T. Comes, F. Fiedrich, S. Fortier, J. Geldermann, & L. Yang (Eds.), *Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM)* (pp. 791–800). https://doi.org/10.1145/2534732.2534741

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. *ArXiv Preprint ArXiv:1605.05894*. http://arxiv.org/abs/1605.05894

Imran, M., Mitra, P., & Srivastava, J. (2017). Enabling Rapid Classification of Social Media Communications During Crises. *International Journal of Information Systems for Crisis Response and Management*. https://doi.org/10.4018/ijiscram.2016070101

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. https://doi.org/10.1016/j.patrec.2009.09.011

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*. https://doi.org/10.18653/v1/e17-2068

Kaufhold, M.-A., Bayer, M., & Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing and Management*, *57*(1), 1–32. https://doi.org/10.1016/j.ipm.2019.102132

Kaufhold, M.-A., & Reuter, C. (2016). The Self-Organization of Digital Volunteers across Social Media: The Case of the 2013 European Floods in Germany. *Journal of Homeland Security and Emergency Management (JHSEM)*, *13*(1), 137–166. https://doi.org/10.1515/jhsem-2015-0063

Kaufhold, M.-A., Rupp, N., Reuter, C., & Habdank, M. (2020). Mitigating Information Overload in Social Media during Conflicts and Crises: Design and Evaluation of a Cross-Platform Alerting System. *Behaviour & Information Technology (BIT)*, *39*(3), 319–342. https://doi.org/10.1080/0144929X.2019.1620334

Keim, D., Andrienko, G., Fekete, J., Carsten, G., & Melan, G. (2008). Visual Analytics: Definition, Process and Challenges. *Information Visualization - Human-Centered Issues and Perspectives*, 154–175. https://doi.org/10.1007/978-3-540-70956-5_7

Khatua, A., Khatua, A., & Cambria, E. (2019). A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks. *Information Processing and Management*. https://doi.org/10.1016/j.ipm.2018.10.010

Küçük Matcı, D., & Avdan, U. (2020). Comparative analysis of unsupervised classification methods for mapping burned forest areas. *Arabian Journal of Geosciences*, *13*(15), 711. https://doi.org/10.1007/s12517-020-05670-7

Lansmann, S., & Klein, S. (2018). How Much Collaboration? Balancing the Needs for Collaborative and Uninterrupted Work. *European Conference on Information Systems (ECIS)*, 1–18.

Li, H., Caragea, D., Caragea, C., & Herndon, N. (2017). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 16–27. https://doi.org/10.1111/1468-5973.12194

Li, H., Li, X., Caragea, D., & Caragea, C. (2018). Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks. *Proceedings of the ISCRAM Asian Pacific 2018 Conference*.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., & Member, S. (2020). Self-supervised Learning: Generative or Contrastive. *ArXiv:2006.08218 [Cs.LG]*, 1–23.

Lu, X. S., & Zhou, M. (2016). Analyzing the evolution of rare events via social media data and k-means clustering algorithm. *ICNSC 2016 - 13th IEEE International Conference on Networking, Sensing and Control*. https://doi.org/10.1109/ICNSC.2016.7479041

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. In *Psychological Review* (Vol. 63, Issue 2, pp. 81–97). American Psychological Association. https://doi.org/10.1037/h0043158

Mirbabaie, M., & Zapatka, E. (2017). Sensemaking in Social Media Crisis Communication – A Case Study on the Brussels Bombings in 2016. *Twenty-Fifth European Conference on Information Systems (ECIS)*, 2169–2186.

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*.

Nayak, N., Angeli, G., & Manning, C. D. (2016). *Evaluating Word Embeddings Using a Representative Suite of Practical Tasks*. https://doi.org/10.18653/v1/w16-2504

Nguyen, D. T., Mannai, K. A. Al, Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2016). Rapid Classification of Crisis-Related Data on Social Networks using Convolutional Neural Networks. *International AAAI Conference on Web and Social Media*.

Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing Big Social Data. *Journal of Big Data*, *4*(1), 1–19. https://doi.org/10.1186/s40537-017-0063-x

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to Expect When the Unexpected Happens: Social Media Communications Across Crises. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 994–1009. https://doi.org/10.1145/2675133.2675242

Onorati, T., Díaz, P., & Carrion, B. (2018). From social networks to emergency operation centers: A semantic visualization approach. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2018.01.052

Pagliardini, M., Gupta, P., & Jaggi, M. (2018). *Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features*. https://doi.org/10.18653/v1/n18-1049

Palen, L., & Anderson, K. M. (2016). Crisis informatics: New data for extraordinary times. *Science*, *353*(6296), 224–225. https://doi.org/10.1126/science.aag2579

Palen, L., & Hughes, A. L. (2018). Social Media in Disaster Communication. In H. Rodríguez, W. Donner, & J. E. Trainor (Eds.), *Handbook of Disaster Research* (pp. 497–518). Springer International Publishing. https://doi.org/10.1007/978-3-319-63254-4_24

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/d14-1162

Pohl, D., Bouchachia, A., & Hellwagner, H. (2015). Social media for crisis management: clustering approaches for sub-event detection. *Multimedia Tools and Applications*, *74*(11), 3901–3932. https://doi.org/10.1007/s11042-013-1804-2

Rao, R., Plotnick, L., & Hiltz, S. R. (2017). Supporting the Use of Social Media by Emergency Managers: Software Tools to Overcome Information Overload. In *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. https://doi.org/10.18653/v1/d19-1410

Reuter, C., & Kaufhold, M.-A. (2018). Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics. *Journal of Contingencies and Crisis Management (JCCM)*, *26*(1), 41–57. https://doi.org/10.1111/1468-5973.12196

Reuter, C., Kaufhold, M. A., Schmid, S., Spielhofer, T., & Hahne, A. S. (2019). The impact of risk cultures: Citizens' perception of social media use in emergencies across Europe. *Technological Forecasting and Social Change (TFSC)*, *148*(119724). https://doi.org/10.1016/j.techfore.2019.119724

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. https://doi.org/10.1016/0377-0427(87)90125-7

Rudra, K., Goyal, P., Ganguly, N., Mitra, P., & Imran, M. (2018). Identifying Sub-events and Summarizing Disaster-Related Information from Microblogs. *SIGIR '18 The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 265–274. https://doi.org/10.1145/3209978.3210030

Sakai, T., Tamura, K., & Kitakami, H. (2015). Emergency situation awareness during natural disasters using density-based adaptive spatiotemporal clustering. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-22324-7_13

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d15-1036

Siroker, D., & Miller, S. (2008). Topical Clustering, Summarization, and Visualization. In *Screen*.

Spielhofer, T., Greenlaw, R., Markham, D., & Hahne, A. (2016). Data mining Twitter during the UK floods: Investigating the potential use of social media in emergency management. *2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, 1–6. https://doi.org/10.1109/ICT-DM.2016.7857213

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics: An Interdisciplinary Approach and Its Implications for Information Systems. *Business and Information Systems Engineering*, *6*(2), 89–96. https://doi.org/10.1007/s12599-014-0315-7

Stieglitz, S., Mirbabaie, M., Fromm, J., & Melzer, S. (2018). The Adoption of Social Media Analytics for Crisis Management - Challenges and Opportunities. *Proceedings of the 26th European Conference on Information Systems (ECIS)*.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., & Anderson, K. M. (2011). Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 385–392.

Vieweg, S. (2012). *Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications*. 1–300.

Westerink, R. H. S., & Vijverberg, H. P. M. (2018). Universal Sentence Encoder. *Toxicology and Applied Pharmacology*. https://doi.org/10.1006/taap.2002.9482

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*. https://doi.org/10.1016/c2009-0-19715-5

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. In *IEEE Transactions on Neural Networks*. https://doi.org/10.1109/TNN.2005.845141

Yin, J., Karimi, S., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2015). Using social media to enhance emergency situation awareness: *IJCAI International Joint Conference on Artificial Intelligence*, *2015-Janua*, 4234–4239. https://doi.org/10.1109/MIS.2012.6

You, Y., Li, J., Hseu, J., Song, X., Demmel, J., & Hsieh, C.-J. (2019). Reducing BERT Pre-Training Time from 3 Days to 76 Minutes. *ArXiv*.