Combining Supervised and Unsupervised Learning to Detect and Semantically Aggregate Crisis-Related Twitter Content

Jens Kersten*
German Aerospace Center – Jena, Germany†
jens.kersten@dlr.de

Jan Bongard
German Aerospace Center – Jena, Germany
jan.bongard@dlr.de

Friederike Klan
German Aerospace Center – Jena, Germany
friederike.klan@dlr.de

ABSTRACT

Twitter is an immediate and almost ubiquitous platform and therefore can be a valuable source of information during disasters. Current methods for identifying and classifying crisis-related content are often based on single tweets, i.e., already known information from the past is neglected. In this paper, the combination of tweet-wise pre-trained neural networks and unsupervised semantic clustering is proposed and investigated. The intention is to (1) enhance the generalization capability of pre-trained models, (2) to be able to handle massive amounts of stream data, (3) to reduce information overload by identifying potentially crisis-related content, and (4) to obtain a semantically aggregated data representation that allows for further automated, manual and visual analyses. Latent representations of each tweet based on pre-trained sentence embedding models are used for both, clustering and tweet classification. For a fast, robust and time-continuous processing, subsequent time periods are clustered individually according to a Chinese restaurant process. Clusters without any tweet classified as crisis-related are pruned. Data aggregation over time is ensured by merging semantically similar clusters. A comparison of our hybrid method to a similar clustering approach, as well as first quantitative and qualitative results from experiments with two different labeled data sets demonstrate the great potential for crisis-related Twitter stream analyses.

Keywords

Information overload reduction, semantic clustering, crisis informatics, Twitter stream.

INTRODUCTION

Twitter is an immediate social media platform. This real-time character, paired with a huge and highly active user community distributed over the world, enables public instant discourses about the latest topics and events. Twitter can therefore be a valuable source of information in case of natural and man-made events and disasters (Wiegmann, Kersten, Senaratne, et al. 2020).

Current state-of-the-art methods for filtering or classifying crisis- and event-related messages are often based on supervised and semi-supervised machine learning techniques (Kruspe et al. 2020), for example pre-trained convolutional neural networks (Burel and Alani 2018), models trained from scratch utilizing ad-hoc labeled data (Kaufhold et al. 2020; Snyder et al. 2019) and domain adaptation (Mazloom et al. 2019). In-depth cross-event and cross-event type experiments using pre-trained deep learning models in (Kersten et al. 2019; Wiegmann, Kersten, Klan, et al. 2020a) revealed, that generic models trained with data from various events are suitable to

*corresponding author
†www.dlr.de/dw/en

filter Twitter stream data with an average misclassification rate of 3.8 % (false positive rate). However, for the task of identifying crisis-related messages, average F1 drops of ≈ 0.4 were observed in exhaustive cross-event type experiments.

One reason for the limited generalization capability might be the fact that the analysis is approached on microblog-level. This is known to be challenging, for example due to abbreviations, potential misspellings, informal language, and a limited number of characters. Recently proposed clustering methods for Twitter novelty detection (Kruspe 2020) and event detection based on the full Twitter firehose (Fedoryszak et al. 2019) impressively demonstrate the value of incorporating contextual information by semantically and temporally aggregating microblog messages. Motivated by this, we propose to combine semantic clustering with pre-trained models for classifying crisis-related tweets in order to enhance the performance of these models and to automatically detect crises-focused, semantically aggregated data. With the ability of spawning and merging clusters, we furthermore seek to capture the unfolding and development of crisis (sub-) events, even if they occur in parallel.

A combination of unsupervised and supervised methods is commonly used for tasks like event detection (Angaramo and Rossi 2018) and topic detection and tracking (TDT) (Li et al. 2021). With special emphasis on the Twitter overload reduction problem by identifying clusters of potentially crisis-related messages in unfiltered Twitter streams, our main contributions are: (1) We investigate, how static pre-trained and general models to classify crisis-related tweets can be enhanced by unsupervised clustering; (2) in contrast to often utilized word embeddings, we use sentence embeddings to represent tweets; (3) we propose an incremental CRP clustering method to enable the processing of continuous stream data. A secondary intention of our approach is to obtain a semantically aggregated data representation that allows for further automated analysis steps, like summarization or localization, as well as easier manual inspections, visualizations and eventually understanding of the contents. These aspects will be addressed in future works.

The article is organized as follows. After a discussion of Related Work in the following section, our Proposed Method is outlined. The data sets used for evaluation are described in section Data Sets and Experimental Results. Furthermore, first qualitative and quantitative results obtained with our method are summarized and discussed. Finally, conclusions are drawn and steps for future works are pointed out.

RELATED WORK

Text clustering is utilized for various applications, like topic modeling (Viegas et al. 2019), topic detection (Miranda et al. 2020; S. T. Nguyen et al. 2019) and topic tracking (Fedoryszak et al. 2019), (sub-) event detection (Angaramo and Rossi 2018; Jiang et al. 2019), and data aggregation in general (Qiang et al. 2019; S. Yang et al. 2019). An overview of current approaches in this field grouped by the features used for clustering is provided in table 1.

Spatio-temporal approaches usually involve density-based methods, like STDBSCAN (Ester et al. 1996), but have to be further complemented with methods, like Latent Dirichlet Attocation (LDA), in order to analyze message contents (M. D. Nguyen and Shin 2017). However, processing large amounts of stream data, for example Twitter microblogs, requires computationally inexpensive approaches, like burst detection (Fedoryszak et al. 2019), K-means (Miranda et al. 2020; Singh and Shashi 2019), or infinite mixture modeling (Dai et al. 2017; Kruspe 2020).

In the majority of identified research works, word embeddings are utilized to represent text (Dai et al. 2017; Ertugrul et al. 2017; Jiang et al. 2019; Mendonça et al. 2019; Miranda et al. 2020; Singh and Shashi 2019; Zhou et al. 2019). Especially for short texts, word embeddings have shown to be useful to augment traditional features (Comito et al. 2019) or to expand a small set of representative terms with semantically similar words (Qiang et al. 2019; Viegas et al. 2019). Recently proposed sentence embeddings, like Google's Universal Sentence Encoder (USE) (Cer et al. 2018), provide a latent representation of whole sentences and are able to capture semantic as well as contextual information. In (Kruspe 2020), USE is utilized to represent and directly cluster microblog messages for the detection of novelty in social media messages during emerging crisis events.

The addressed research problem in this work is strongly related to the task of topic detection and tracking, where clustering techniques are often involved. Challenges in TDT can be the sparsity and complexity of data, varying topic granularities, unexpectedness of emerging topics, and the unpredictability of topic evolution (W. Liu et al. 2020). With emphasis on distributed and parallel processing, Li et al. 2021 utilize DBSCAN to detect and a parallel K-Nearest-Neighbor algorithm to track topics. In order to robustly detect new emerging events, W. Liu et al. 2020 propose to involve domain-specific hierarchical ontolgies. A new event is identified according to similarities between the linguistic expression of ontology nodes and vectorized news events. Focusing on multimodal Twitter data, where noisy sentences and images, misspellings, new invented words, or informal language are common challenges, a transformer-based approach for topic detection is proposed in (Asgari-Chenaghlu et al. 2020). Named

Table 1. Related work grouped by features used for clustering.

Reference	*Methods*	*Application*	*Data*
		Traditional Features	
(Angaramo and Rossi 2018)	Bag-of-Words	Event detection	Twitter
(S. T. Nguyen et al. 2019)	Central centroids	Hot topic detection	Twitter, Events2012 (McMinn et al. 2013)
(Qiang et al. 2019)	Dirichlet multinomial mixture, text expansion	Text clustering	TREC (2011/2012) (TextREtrievalConference 2015), (Yin and J. Wang 2014)
(Viegas et al. 2019)	Non-negative matrix factorization, text expansion	Topic modelling, document clustering	Various
(S. Yang et al. 2019)	Word graphs, similarity	Text clustering	(Yin and J. Wang 2014), (Yin, Chao, et al. 2018), (J. Yang and Leskovec 2011), (TextREtrievalConference 2015)
(Fedoryszak et al. 2019)	Named entity recognition, burst detection	Real-time event detection	Twitter firehose
		Density: Time, Location, Content	
(M. D. Nguyen and Shin 2017)	Extension of DBSCAN (Ester et al. 1996)	Text clustering	Twitter
(S. Xu et al. 2020)	LDA, ST-DBSCAN (Birant and Kut 2007)	Event detection	Twitter, own labels
(Zhang and Eick 2019)	LDA, contour-based	Event tracking	Twitter
		Word Embeddings	
(Dai et al. 2017)	Chinese restaurant process, smilarity	Public health	Twitter, own labels
(Ertugrul et al. 2017)	Hierarchical clustering	Event detection	Twitter
(Comito et al. 2019)	Chinese restaurant	Online topic detection	Tweets, crowdsourcing
(Jiang et al. 2019)	Keyword-pairs, hypothesis test	Crisis sub-event detection	Twitter
(Mendonça et al. 2019)	K-Means, EM, Mean Shift, DBSCAN	Text document clustering	News articles, medical abstracts
(Singh and Shashi 2019)	K-Means	Topic clustering, summarization	Newswire (Nenkova 2005)
(Zhou et al. 2019)	Parallel K-Means	Text similarity measurement	Website customer comments
(Miranda et al. 2020)	Self-organizing maps, K-Means	Topic detection	20 Newsgroup (Dua and Graff 2017)
		Sentence Embeddings	
(Hadifar et al. 2019)	Auto-encoder, K-Means	Short text clustering	Stackoverflow, SearchSnippets, PubMed (J. Xu et al. 2017)
(Kruspe 2020)	Chinese restaurant	Novel topic detection	Twitter: TREC-IS 2019A (McCreadie et al. 2019)

entity recognition in images and texts in combination with fine-tuned word embeddings to analyze semantic relations between words are used for graph-based clustering and TDT.

The evaluation of clustering methods is challenging and often manual inspections and qualitative interpretations are carried out (see for example (Zhang and Eick 2019; Miranda et al. 2020)). Currently, there is no consistent benchmarking data set available by which event detection systems or clusterings of Twitter data can be measured (Fedoryszak et al. 2019). Resources, like CrisisLex (Olteanu et al. 2015) or CrisisNLP (Imran et al. 2016) are quite useful for classification tasks, but only contain a focused subset of tweets. Simulating realistic stream conditions can be done by augmenting labeled data with a large amount of off-topic messages, as done in (Wiegmann, Kersten, Klan, et al. 2020b). A better but more demanding setting would be to provide a full stream and corresponding labels for all the tweets relevant for a specific application. One example for this is the Events2012 data set (McMinn et al. 2013).

PROPOSED METHOD

Since a Chinese restaurant-based clustering (CRP) is able to robustly handle massive data volumes (Dai et al. 2017), we favor this method over other methods, like K-means, where the number of clusters has to be defined in advance. The data stream-character is taken into account by clustering subsequent time intervals and by merging semantically similar clusters across interval borders. The complete workflow is depicted in figure 1.

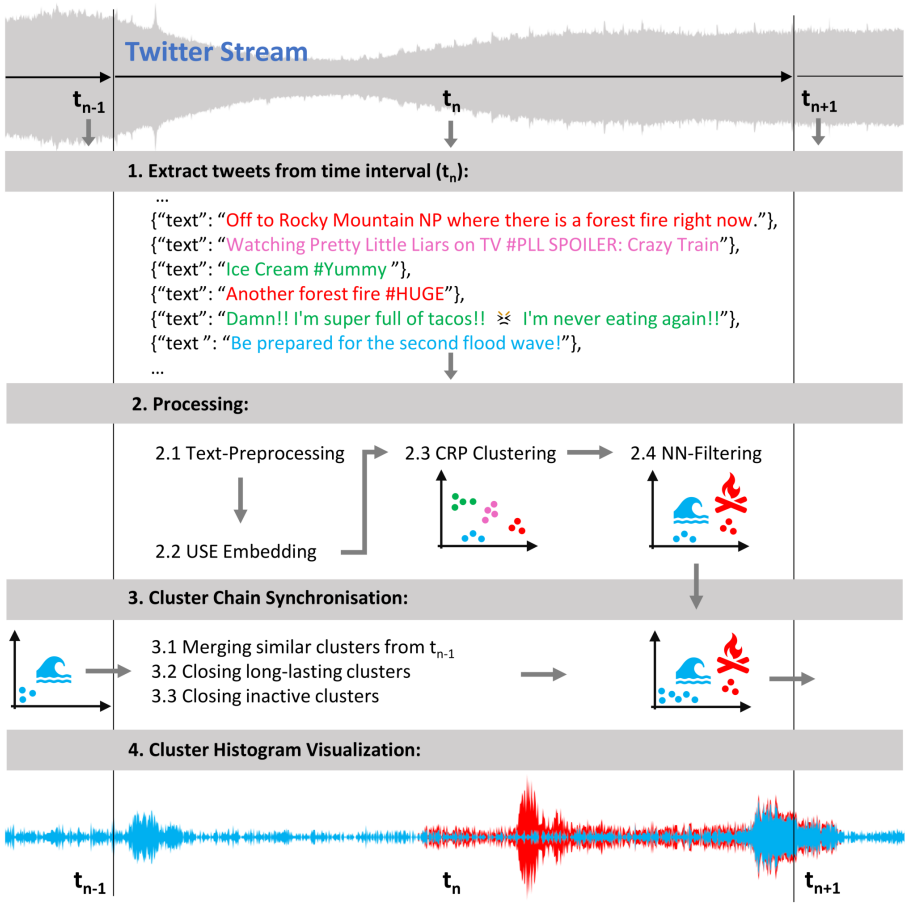


Figure 1. Proposed workflow for Twitter stream clustering.

After a sequence of standard pre-processing steps (e.g. lowercasing, language filtering, tokenization, removing URLs, numbers, and hashtags), a latent space representation is obtained for each tweet with a pre-trained transformer model[1]. A new tweet is assigned to the closest cluster as long as the cosine similarity is within a certain threshold. Otherwise, a new cluster is spawned. In order to speed up the process, clusters are represented by central centroids (S. T. Nguyen et al. 2019), which are obtained by averaging the fixed number of cluster representatives that are most similar to each other. Furthermore, this ensures robustness towards outliers.

Adding a tweet to a cluster is followed by updating its representatives and the central centroid. Clusters from the same as well as from different time periods can be merged according to the cosine distance between the central centroids. Inactive and long-lasting clusters can be pruned based on time-thresholds. Our method differs to (Kruspe 2020) in the following ways: (1) We analyze time-intervals instead of complete data sets to handle stream data, (2) a constant cosine similarity threshold is used instead of estimating and adapting an Euclidean distance-based threshold, (3) a K-means-based initialization of the process is therefore not required, (4) the concept of central centroids is utilized, and (5) the option to merge similar clusters occurring over time is given.

Unsupervised clustering is complemented by applying a tweet-wise pre-trained model (Wiegmann, Kersten, Klan, et al. 2020a) for assigning crisis-relatedness. A cluster that does not contain at least one tweet classified as crisis-related is pruned. This is intended to reduce information overload by only retaining potentially crisis-related clusters. In the following section, this proposed automated method is evaluated.

[1] https://tfhub.dev/google/universal-sentence-encoder-large/4

Even though not addressed at this early stage of research, some application-related aspects and motivations of our method shall be pointed out. An aggregated and reduced crisis-related data representation is intended to provide a sound foundation for fast manual content inspections as well as for further automated analyzes, like the detection, tracking, and localization of parallel events or discussions. An overview of the topics contained in the clusters can be summarized, for example by determining the most significant keywords. This allows for fast customization by pruning or emphasizing clusters related to specific topics of interest. Hence, the combination with an interactive interface might enable a near real-time and flexible stream analysis for practical applications. This interactive approach is expected to mitigate lower accuracies of pre-trained models in case of new events in the following ways. First, a lower recall could be mitigated by retaining clusters in which only a few related tweets were found according to the model but more actually related tweets might be contained. Second, a lower precision could be mitigated by manually inspecting (or further analyzing) and discarding clusters of low model-confidence, irrelevant topics or identified false positives. An in-depth investigation of these hypotheses will be part of our future work.

DATA SETS AND EXPERIMENTAL RESULTS

Data Sets

We use two different data sets for a first evaluation of our approach. For a comparison to (Kruspe 2020), the same resource of labeled tweets (24 crisis-related information classes, multiple labels allowed) from the TExt Retrieval Conference (TREC) (McCreadie et al. 2019) 2019A edition is investigated. It contains around 20,000 labeled tweets covering 15 different crisis events (natural and man-made). Since this set does not contain a realistic amount of off-topic messages, the Events2012 data set is utilized in a second experiment. It represents a four week Twitter stream containing over 120 million tweets, where relevance judgments for a subset of 150,000 tweets are available covering more than 500 events from a broad range of topics, including natural disasters. Due to deletions and privacy changes on Twitter, only around 50 % of all tweets and around 1/3 of the labeled tweets from Events2012 are currently available.

Experiment 1: Class-wise Cluster Purity

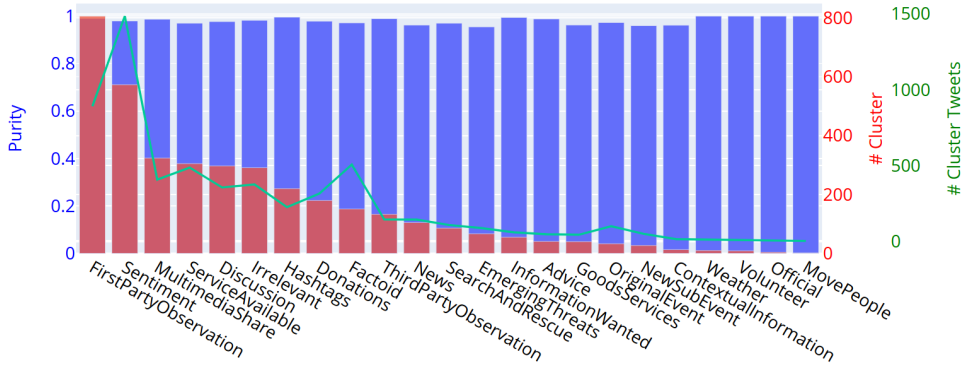
In the first experiment, we evaluate the CRP clustering based on cluster purities, i.e., without involving pre-trained models and without analyzing multiple time intervals. Hence, for each TREC-event, an independent clustering of the corresponding, chronologically organized data is conducted. The cluster purity reflects, how well classes can be isolated and distinguished. Ideally, a cluster or a set of clusters contains all tweets of the same thematic class, such as "Donations", "Sentiment", or "ServiceAvailable". Similar to (Kruspe 2020), we therefore determine the cluster purity individually for each event and class. Example results for the Nepal earthquake in 2015 with cosine similarity thresholds of $t = 0.7$, 0.5 and 0.2 are depicted in figure 2.

Our results with $t = 0.7$ coincides with those reported in (Kruspe 2020), i.e., the dominant classes found most often relate to the generally most dominant classes in the data set and clusters dominated by rare, but very important classes (for example “EmergingThreats” and “SearchAndRescue”) are also detected and display high purity. However, the number of tweets per class (green line in figure 2 (a)) compared to the number of clusters (red) reflects, that the clusters tend to be very small. For instance, $\approx 1,000$ Tweets of class FirstPartyObservation are grouped into ≈ 800 clusters. Hence, a large fraction of the result clusters is represented by a single tweet. In contrast, lower thresholds t produce larger clusters and introduce a trade-off between cluster size and purity. Furthermore, the detection of rare classes appears to be more difficult in case of a low threshold. As a consequence, less result classes are present in figure 2 (c).

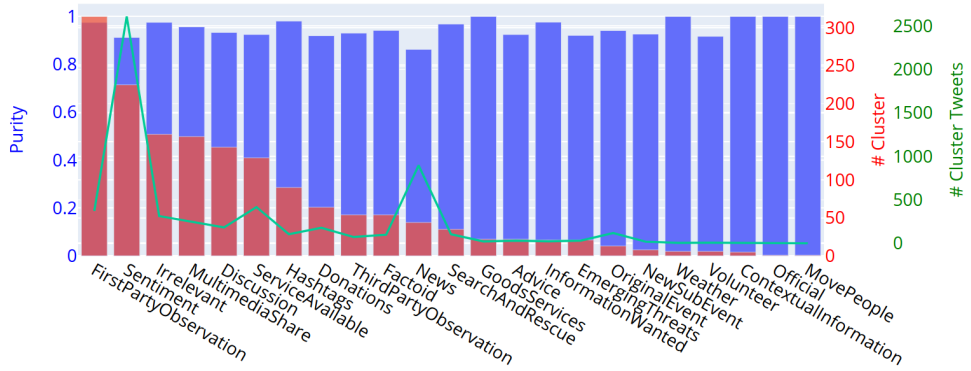
A further important observation is that, compared to the method in (Kruspe 2020), our approach leads to a significant speed-up of up to 500. Reasons for this are the fixed cosine similarity threshold and the utilized central centroid concept for representing clusters. Even though this might not be representative due to possibly inefficient experimental implementations, this shows a clear trend.

Experiment 2: Stream Analysis

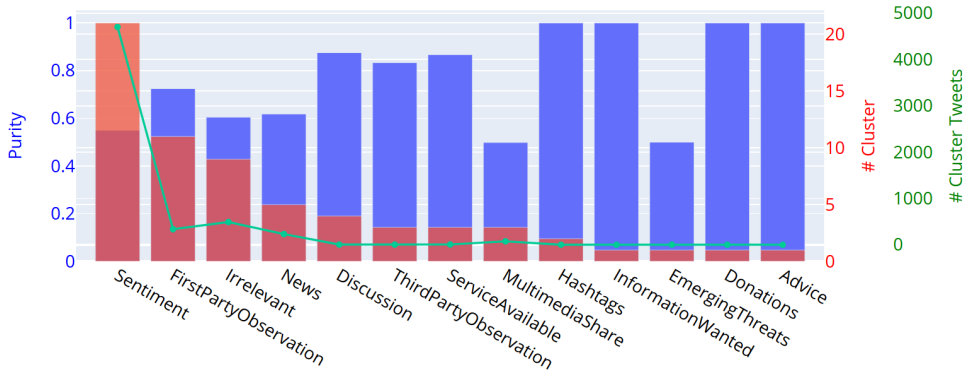
In our second experiment, we use selected subsets on a daily basis from Events2012 in order to evaluate the clustering performance in case of stream data. Furthermore, we want to evaluate, if our method is able to automatically identify and isolate crisis-related tweets representing specific events, and if parallel events can be distinguished. For all experiments, a cosine similarity threshold of $t = 0.7$ and a time interval of one hour turned out to be a good



(a) $t = 0.7$



(b) $t = 0.5$



(c) $t = 0.2$

Figure 2. Results for the Nepal earthquake (2015) data: Label purities (blue), number of found clusters (red) and accumulated number of tweets (green) by most frequent class for different cosine similarity thresholds t.

choice. In contrast to the first experiment, the obtained cluster sizes are much larger with the Events20212 data. Results for three different days are summarized in table 2.

For the hospital fire on October 24, ground truth labels for 25 tweets are available. The pre-trained DNN was able to correctly classify only two of them. In contrast, 13 relevant tweets were identified by our hybrid method. Since the DNN is trained to detect any crisis-related messages for a broad range of man-made and natural disasters, the overall numbers of detected tweets from the whole day are quite high (DNN: 51k, Hybrid: 85k). Since the average total amount of tweets per day is around 2.1 million, this means a significant reduction to potentially relevant messages, where the hybrid approach leads to a nearly doubled amount of potentially relevant tweets. Compared to the DNN, it is likely, that a larger fraction of false positives is contained here. On the other hand, the fact that more GT tweets could be identified by the hybrid method is also a hint for a better recall. This leaves ample room for further exhaustive quantitative investigations.

Through manual inspections we realized, that, in addition to the labeled tweets provided in Events2012, there are many more messages "hidden" in the data that are actually somehow related to the investigated events. We therefore

Table 2. Selected results for the Events2012 data set. For each day and event, the available number of ground truth tweets (GT), the number of GT tweets found correctly (GT-F), and the overall number of tweets identified as crisis-related for the whole day are provided for the pre-trained model (DNN) and our hybrid method. Additionally, the number of all clusters that contain GT labels (GT-C), the corresponding number of tweets (T) and the mean purity (MP) for these clusters are provided.

			DNN	*Hybrid*			
Date	*Event*	*GT*	*GT-F/Overall*		*GT-C*	*T*	*MP*
Oct 14	Tornado outside of Walsh	156	153/49,346	153/97,337	60	1,825	0.92
Oct 24	Hospital on fire, Taiwan	25	2/51,837	**13**/85,284	2	19	1.0
	Explosion and fire, Khartoum	63	3	4	3	10	1.0
	Surfer killed by shark, CA	96	15	**66**	2	98	1.0
Nov 02	Marathon canceled, Sandy	695	540/74,516	**677**/112,783	17	2,293	0.93
	Hurricane Sandy death toll	25	23	24	7	2,774	0.85
	Landfall hurricane Sandy	56	55	55	16	2,722	0.91

carefully inspected all clusters that contain at least one ground truth label and, for instance, found them distributed over 60 clusters that contain around 1,800 tweets for the tornado event on October 14. A manual labeling of these reveals a quite high cluster purity of around 90 %. This is also true for the other events and therefore indicates a good capability of our method to isolate event-related, semantically similar microblogs.

On October 24 and November 02, three disasters or sub-events took place each. A significant increase of the detection rate by our hybrid method can be observed in case of a fire, a killed surfer and a marathon canceled due to hurricane Sandy. Since almost all relevant tweets were already found by the DNN in some cases, no gain can be measured here. In case of the explosion and a fire on October 24, both methods perform very poor. This is due to the fact that this type of event (explosion) is only covered by few corresponding training samples (see (Wiegmann, Kersten, Klan, et al. 2020a) for more details) and indicates a strong dependence on the DNN quality and thematic coverage. However, if a specific event type is covered by the DNN, our hybrid method is able to significantly enhance the results in the investigated cases. It is worth noting that even though some of the related tweets could not be found automatically, they still might be grouped in one ore more clusters that could be identified by further analyzes or manual inspections.

In order to get a rough impression about the precision of our method, we investigated all clusters with more than three tweets that were identified as relevant but do not contain any ground truth label. The corresponding results in table 3 demonstrate, that the average purity of the large amount of clusters found on each day is significantly lower compared to the clusters in table 2. This indicates, that our method tends to identify and group a broad range of content, that might potentially be crisis-related, but also has a relatively high false positive rate. However, we believe that this behavior is well suited as an initial setting for the task of overload reduction in general. At this stage, no further specifications regarding the topics or keywords of interest were made. According to (Stieglitz et al. 2018), a reliable overload reduction is the first essential step for practical applications. The usually low veracity of information was identified as one of the key challenge for the adoption of social media analytics. Furthermore, “a customization of filtering algorithms for the needs of emergency management agencies might be needed for them to respond to their specific crisis situations” (Stieglitz et al. 2018). We argue that our proposed reduction and semantic aggregation of stream data provides a sound foundation in terms of overload reduction and data preparation in this context. Our semantic aggregation may therefore be a good starting point for the development of further analysis methods that could be used in real scenarios, for example in context of further overload reduction, trustworthiness analysis, localization, summarization, and adaptive situation monitoring.

Qualitative Analysis

As an example for the aforementioned required customization to specific crisis situations, we generated a UMAP (McInnes et al. 2020) visualization for all clusters of October 24 that contain the keyword "Long Island". The well distinguished clusters at the top of figure 3 demonstrate, that the embedding-based clustering has the potential to well isolate and aggregate (sub-) events and discussed topics. Furthermore, this interactive visualization, where for example the most dominant keywords are used to characterize a cluster and tweet texts can be shown on

Table 3. Amount of clusters (>3 tweets) without any ground truth label that contain at least one tweet classified as related, the corresponding total amount of identified related tweets per day, and the average cluster purity.

Date	*# Cluster*	*# Tweets*	*Average Purity*
Oct 14	1,528	61,915	0.19
Oct 24	1,539	44,725	0.16
Nov 02	2,068	59,414	0.34

mouse-over, enables to immediately capture a good overview of the current situation and past developments. For instance, the information about crowds at petrol stations might be valuable, since this reflects security-related issues and required services of citizens affected by power supply interruptions. A further important observation is that of course not only physical events, but also discussions related to these events can be found. We can see that there was a vital discussion about the NY marathon, that finally leads to an Twitter activity peak after it was officially canceled, even though the mayor tried to avoid the canceling.

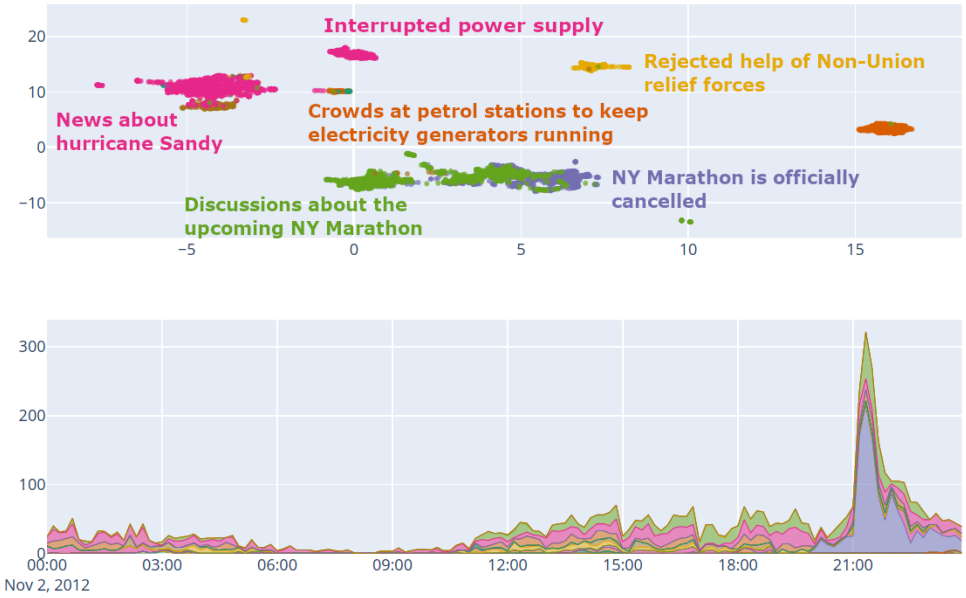


Figure 3. Top: 2D UMAP-based visualization of selected clusters containing the keyword "Long Island" identified on October 14, 2012. Bottom: Tweet counts over time (GMT) per cluster/topic.

CONCLUSION AND FUTURE WORK

In this paper, the task of Twitter overload reduction through automatically detecting and semantically aggregating crisis-related Twitter content is addressed. Current tweet-wise pre-trained models for this task are known to have a limited generalization capability in case of new events and event types. The combination of such a model with an unsupervised semantic tweet clustering is intended to mitigate this effect. As demonstrated qualitatively, taking into account contextual information helps to provide an immediate overview of discussed topics and events as well as their development over time.

We use latent space representations of tweets based on pre-trained sentence encoders for both, tweet classification and clustering. A stream of incoming tweets is clustered based on a Chinese restaurant process, where a central centroid cluster representation allows for fast computations and ensures robustness against outliers. Clusters without crisis-related tweets are discarded. A semantic merging of the remaining clusters over time takes the stream character of the data into consideration. In future works, we will also investigate the effect of linking clusters to cluster chains, like done by Fedoryszak et al. 2019. This approach is likely to be helpful to detect topic shifts.

Our experimental results provide some first valuable insights but also demonstrates, that further exhaustive quantitative investigations are required. The analysis of cluster purities indicates, that our method is able to isolate and identify crisis- and event-related Twitter content. Clusters that actually represent event-related content show a

high purity. However, since our method in its initial state is designed to identify any content that might be somehow related to a crisis event, also a large amount of clusters with lower purities can be expected. In the second experiment (stream analysis), our method was able to significantly enhance the detection rate of crisis-related tweets at the cost of a significantly higher false positive rate.

From a practical perspective, the results indicate a need for further cluster content analysis steps that allow to tailor the overload reduction process. We argue that the obtained aggregated data representation is a well suited starting point to summarize the course of discussions related to events – even if they occur in parallel – as well as for further customized filtering. Our current work focuses on the preparation of a comprehensive labeled and realistic Twitter stream data set that covers various different disaster events and types. Further experiments and developments of our method will be based on this resource.

REFERENCES

Angaramo, F. and Rossi, C. (2018). "Online clustering and classification for real-time event detection in Twitter". In: *15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Ed. by K. Boersma and B. Tomaszeski. Rochester, NY (USA): Rochester Institute of Technology.

Asgari-Chenaghlu, M., Feizi-Derakhshi, M.-R., Farzinvash, L., Balafar, M., and Motamed, C. (2020). "TopicBERT: A Transformer transfer learning based memory-graph approach for multimodal streaming social media topic detection". In: *ArXiv* abs/2008.06877.

Birant, D. and Kut, A. (2007). "ST-DBSCAN: An algorithm for clustering spatial-temporal data". In: *Data and Knowledge Engineering* 60.1, pp. 208–221.

Burel, G. and Alani, H. (2018). "Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media". In: *Proceedings of the 15th ISCRAM*.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). "Universal Sentence Encoder". In: *CoRR* abs/1803.11175. arXiv: 1803.11175.

Comito, C., Forestiero, A., and Pizzuti, C. (2019). "Word Embedding Based Clustering to Detect Topics in Social Media". In: *IEEE/WIC/ACM International Conference on Web Intelligence*. WI'19. Thessaloniki, Greece: Association for Computing Machinery, pp. 192–199.

Dai, X., Bikdash, M., and Meyer, B. (2017). "From social media to public health surveillance: Word embedding based clustering method for twitter classification". In: *SoutheastCon 2017*, pp. 1–7.

Dua, D. and Graff, C. (2017). *UCI Machine Learning Repository*.

Ertugrul, A. M., Velioglu, B., and Karagoz, P. (2017). "Word Embedding Based Event Detection on Social Media". In: *Hybrid Artificial Intelligent Systems*. Ed. by F. J. Martínez de Pisón, R. Urraca, H. Quintián, and E. Corchado. Cham: Springer International Publishing, pp. 3–14.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). "A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, pp. 226–231.

Fedoryszak, M., Frederick, B., Rajaram, V., and Zhong, C. (2019). "Real-Time Event Detection on Social Data Streams". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, pp. 2774–2782.

Hadifar, A., Sterckx, L., Demeester, T., and Develder, C. (Aug. 2019). "A Self-Training Approach for Short Text Clustering". In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Florence, Italy: Association for Computational Linguistics, pp. 194–199.

Imran, M., Mitra, P., and Castillo, C. (2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portoroz, Slovenia: European Language Resources Association (ELRA).

Jiang, S., Groves, W., Anzaroot, S., and Jaimes, A. (2019). "Crisis Sub-Events on Social Media: A Case Study of Wildfires". In: *Proceedings of the AI for Social Good Workshop at the 36th International Conference on Machine Learning (AISG@ICML)*. Long Beach, CA, USA.

Kaufhold, M.-A., Bayer, M., and Reuter, C. (2020). "Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning". In: *Information Processing & Management* 57.1.

Kersten, J., Kruspe, A., Wiegmann, M., and Klan, F. (2019). "Robust Filtering of Crisis-related Tweets". In: *Proceedings of the 16th ISCRAM, May 19-22*. ISCRAM. Valencia, Spain.

Kruspe, A., Kersten, J., and Klan, F. (2020). "Review article: Detection of informative tweets in crisis events". In: *Natural Hazards and Earth System Sciences Discussions* 2020, pp. 1–18.

Kruspe, A. (2020). "Detecting novelty in social media messages during emerging crisis events". In: *Proceedings of the 17th ISCRAM, May 24-27*. ISCRAM. Blacksburg, Virginia (USA).

Li, C., Liu, M., Cai, J., Yu, Y., and Wang, H. (2021). "Topic Detection and Tracking Based on Windowed DBSCAN and Parallel KNN". In: *IEEE Access* 9, pp. 3858–3870.

Liu, W., Jiang, L., Wu, Y., Tang, T., and Li, W. (2020). "Topic Detection and Tracking Based on Event Ontology". In: *IEEE Access* 8, pp. 98044–98056.

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. (July 2019). "A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets". In: *International Journal of Information Systems for Crisis Response and Management* 11.

McCreadie, R., Buntain, C., and Soboroff, I. (2019). "TREC Incident Streams: Finding Actionable Information on Social Media". In: *International Conference on Information Systems for Crisis Response and Management (ISCRAM). Valencia, Spain*.

McInnes, L., Healy, J., and Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv: 1802.03426 [stat.ML].

McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). "Building a Large-scale Corpus for Evaluating Event Detection on Twitter". In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM)*. San Francisco, California, USA: ACM, pp. 409–418.

Mendonça, I., Trouvé, A., Fukuda, A., Murakami, K., Tsai, C.-F., Hu, Y.-H., Wang, M.-C., Liu, K. E., Yu, X., Yuan, Y., et al. (2019). "On Clustering Algorithms: Applications in Word-Embedding Documents." In: *JCP* 14.2, pp. 88–92.

Miranda, G. R. de, Pasti, R., and Castro, L. N. de (2020). "Detecting Topics in Documents by Clustering Word Vectors". In: *Distributed Computing and Artificial Intelligence, 16th International Conference*. Ed. by F. Herrera, K. Matsui, and S. Rodríguez-González. Cham: Springer International Publishing, pp. 235–243.

Nenkova, A. (2005). "Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference". In: *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*. AAAI'05. Pittsburgh, Pennsylvania: AAAI Press, pp. 1436–1441.

Nguyen, M. D. and Shin, W.-Y. (2017). "DBSTexC: Density-Based Spatio-Textual Clustering on Twitter". In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ASONAM '17. Sydney, Australia: Association for Computing Machinery, pp. 23–26.

Nguyen, S. T., Ngô, B. C., Vo, C., and Cao, T. H. (2019). "Hot Topic Detection on Twitter Data Streams with Incremental Clustering Using Named Entities and Central Centroids". In: *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 1–6.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. Vancouver, BC, Canada: ACM, pp. 994–1009.

Qiang, J., Li, Y., Yuan, Y., Liu, W., and Wu, X. (2019). "A practical algorithm for solving the sparseness problem of short text clustering". In: *Intell. Data Anal.* 23, pp. 701–716.

Singh, A. K. and Shashi, M. (2019). "Vectorization of Text Documents for Identifying Unifiable News Articles". In: *International Journal of Advanced Computer Science and Applications* 10.7.

Snyder, L. S., Lin, Y., Karimzadeh, M., Goldwasser, D., and Ebert, D. S. (2019). "Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness". In: *IEEE Transactions on Visualization and Computer Graphics*.

Stieglitz, S., Mirbabaie, M., Fromm, J., and Melzer, S. (2018). "The Adoption of Social Media Analytics for Crisis Management – Challenges and Opportunities". In: *Twenty-Sixth Eur. Conf. Inf. Syst. (ECIS2018)*.

TextREtrievalConference (2015). *Microblog Track*. https://trec.nist.gov/data/microblog.html.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). "CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling". In: *Proceedings of the*

Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19. Melbourne VIC, Australia: Association for Computing Machinery, pp. 753–761.

Wiegmann, M., Kersten, J., Senaratne, H., Potthast, M., Klan, F., and Stein, B. (2020). "Opportunities and Risks of Disaster Data from Social Media: A Systematic Review of Incident Information". In: *Natural Hazards and Earth System Sciences Discussions* 2020, pp. 1–16.

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., and Stein, B. (2020a). "Analysis of Filtering Models for Disaster-Related Tweets". In: *Proceedings of the 17th ISCRAM, May 24-27*. ISCRAM. Blacksburg, Virginia (USA).

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., and Stein, B. (Mar. 2020b). *Disaster Tweet Corpus 2020*. https://doi.org/10.5281/zenodo.3713920. Version 1.0.0.

Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., and Zhao, J. (2017). "Self-Taught Convolutional Neural Networks for Short Text Clustering". In: *CoRR* abs/1701.00185. arXiv: 1701.00185.

Xu, S., Li, S., and Huang, W. (2020). "A spatial-temporal-semantic approach for detecting local events using geo-social media data". In: *Transactions in GIS* 24.1, pp. 142–173. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12589.

Yang, J. and Leskovec, J. (2011). "Patterns of Temporal Variation in Online Media". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: Association for Computing Machinery, pp. 177–186.

Yang, S., Huang, G., and Cai, B. (2019). "Discovering Topic Representative Terms for Short Text Clustering". In: *IEEE Access* 7, pp. 92037–92047.

Yin, J., Chao, D., Liu, Z., Zhang, W., Yu, X., and Wang, J. (2018). "Model-Based Clustering of Short Text Streams". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, pp. 2634–2642.

Yin, J. and Wang, J. (2014). "A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: Association for Computing Machinery, pp. 233–242.

Zhang, Y. and Eick, C. F. (2019). "Tracking Events in Twitter by Combining an LDA-Based Approach and a Density-Contour Clustering Approach". In: *International Journal of Semantic Computing* 13.01, pp. 87–110.

Zhou, S., Xu, X., Liu, Y., Chang, R., and Xiao, Y. (2019). "Text Similarity Measurement of Semantic Cognition Based on Word Vector Distance Decentralization With Clustering Analysis". In: *IEEE Access* 7, pp. 107247–107258.