information

MDPI

Article

Automated Seeded Latent Dirichlet Allocation for Social Media Based Event Detection and Mapping

Cornelia Ferner [1,*], Clemens Havas [2], Elisabeth Birnbacher [1], Stefan Wegenkittl [1] and Bernd Resch [2,3,*]

1 Information Technology and Systems Management (ITS), Salzburg University of Applied Sciences, Urstein Sued 1, 5412 Puch/Hallein, Austria; ebirnbacher@fh-salzburg.ac.at (E.B.); stefan.wegenkittl@fh-salzburg.ac.at (S.W.)
2 Department of Geoinformatics—Z_GIS, University of Salzburg, Schillerstrasse 30, 5020 Salzburg, Austria; clemensrudolf.havas@sbg.ac.at (C.H.)
3 Center for Geographic Analysis, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA
* Correspondence: cornelia.ferner@fh-salzburg.ac.at (C.F.); bernd.resch@sbg.ac.at (B.R.)

Received: 18 June 2020; Accepted: 22 July 2020; Published: 25 July 2020

Abstract: In the event of a natural disaster, geo-tagged Tweets are an immediate source of information for locating casualties and damages, and for supporting disaster management. Topic modeling can help in detecting disaster-related Tweets in the noisy Twitter stream in an unsupervised manner. However, the results of topic models are difficult to interpret and require manual identification of one or more "disaster topics". Immediate disaster response would benefit from a fully automated process for interpreting the modeled topics and extracting disaster relevant information. Initializing the topic model with a set of seed words already allows to directly identify the corresponding disaster topic. In order to enable an automated end-to-end process, we automatically generate seed words using older Tweets from the same geographic area. The results of two past events (Napa Valley earthquake 2014 and hurricane Harvey 2017) show that the geospatial distribution of Tweets identified as disaster related conforms with the officially released disaster footprints. The suggested approach is applicable when there is a single topic of interest and comparative data available.

Keywords: topic modeling; social media; geospatial analysis; disaster management

1. Introduction

Social media channels like Twitter have become an established communication channel for various actors, private and public. The restriction of Tweets to a length of 140 characters until 2017 and 280 characters since leads to concise messages that are available in near real-time and at no cost. Twitter started service in 2006, and in 2009, when an airplane crashed on the Hudson river in New York City, the news was spread on Twitter 15 minutes before mainstream media caught up [1]. With currently around 6000 Tweets sent every second and about 500 million per day [2], events all around the world are reported and made public. Monitoring public sentiment to predict election results [3] or even riots [4] or to analyze urban processes [5] is one aspect of Twitter based event detection.

Another line of research is concerned with detecting natural disasters by analyzing Tweets that share information in near real-time [6]. Especially geo-tagged Tweets can be a valuable source for disaster response. Twitter users as "social sensors" [7,8] immediately deliver in-situ information at currently no additional cost. Traditional remote sensing approaches such as satellite or drone images suffer from lower coverage (both spatial and temporal) and a temporal lag before the data is available [9]. Although social media data have been shown to be a valuable addition, their analysis is difficult due to the data's noisiness, its unstructured and diverse nature and the lack of labeling.

Annotating and classifying the data is not feasible in a timely manner. An unsupervised alternative are topic models, allowing for organizing document collections into general themes, so called topics. As an example, Latent Dirichlet Allocation (LDA) [10] models topics as distributions over a fixed vocabulary and documents as a mixture of topics.

The output of an LDA is difficult to process. Topics can be noisy, not well separated and identifying a topic with desired content can be infeasible even for humans [11]. The fictitious example in Table 1 (left) with five topics illustrates that there might be not a single disaster-related topic, but three (②, ④, and ⑤). Automating the detection of relevant topics with a keyword-based approach could fail, as topic ⑤ lacks the term "earthquake". For real-world and more diverse datasets, the difficulties even increase. Guided [12] and seeded [13] LDA variants are a first step towards automated extraction of the disaster-related topic. Both suggest methods to incorporate prior knowledge about the topics' term distribution. Instead of initializing the topics with random terms, the so called seed words are assigned a high probability for the specific topic. This seeding guides the LDA towards a desirable output during inference, meaning that terms that occur in the same context will eventually also have a high probability in the same topic. Topic ① in Table 1 (right) could be an example for a topic seeded with disaster-related terms, e.g., "earthquake" or "shaking". Tweets corresponding to the seeded topic can then automatically be detected.

Table 1. Fictitious example to illustrate the output of a normal, unseeded (left) and a guided LDA (right). The five topics could be detected by a topic model in a dataset containing Tweets related to an earthquake. Note that for guided LDA, all earthquake-related terms are in a single topic ①, whereas the event is covered in the topics ②, ④, ⑤ by the LDA.

LDA	Guided LDA
① love thank happy	① earthquake sleep damage
② earthquake woke felt	② show watch life
③ game stadium sunday	③ game stadium today
④ earthquake last night	④ school tomorrow start
⑤ damage street hope	⑤ love thank happy

One remaining question is how to automatically determine meaningful seed words. In [14], domain experts manually define the seeds. Another possibility is to apply external sources to incorporate word correlation knowledge, such as WordNet [15] or pretrained word embeddings [16]. Jagarlamudi et al. [13] rely on labeled data to extract seed words by applying feature selection. We aim to close this gap between currently existing approaches, where manual interference is needed especially to adapt to a new event, and the applicability in real world scenarios that often requires immediate action. We propose a method to automatically determine seed words for the disaster-related topic by comparing the vocabulary of the day when the disaster took place with that of a preceding, typical day in the same area. The resulting seed words are used to initialize a single topic of the LDA. After modeling the dataset, Tweets having assigned their maximum value at the specific topic are labeled as related to the event.

This paper investigates the potential of the fully automated seeding for the topic modeling of Tweets. We compare the performance against a basic LDA model and against a single pre-determined seed word for two different Twitter datasets: one covering the Napa valley earthquake in 2014 and the other covering hurricane Harvey in Texas in 2017. Besides an intrinsic evaluation of the coherence of the modeled topics, we determine the classification performance on a small, labeled subset of the data to assess the semantic congruence of the extracted relevant Tweets. Furthermore, we apply a geospatial analysis to determine the exact locations and "hot spots" affected by the events. Validating the spatial distributions of the Tweets against the official disaster footprints allows to generate additional value for disaster management.

In the following, Section 2 summarizes existing work on using topic models for event and disaster detection in Tweets. Our proposed method to automatically seed a LDA model is introduced in Section 3. Section 4 presents the datasets used to run the corresponding experiments and the generated seed words. Intrinsic and extrinsic evaluations are conducted in Sections 5 and 6, respectively, followed by a discussion of the results in Section 7 and a conclusion in Section 8.

2. Background and Related Work

The use of Tweets to gain relevant information for disaster management is an increasingly popular field of study. Some approaches for exploiting Twitter data for disaster management extract messages based on predefined hashtags and keywords only [17], without applying topic models. An issue with working with hashtags and keywords only is that the approach loses its immediacy. Hashtags are defined by the Twitter community and evolve over time. Topic models capture more of the semantics of a Tweet and are less dependent on single terms. Gründer-Fahrer et al. [18] assess the possibilities of the LDA model for the detection of a flood in Germany and Austria, based on German language Tweets. Their claim is that LDA combined with optimization methods shows high applicability for disaster management. Before discussing the various modifications of the model meant to improve the performance, we formally introduce the basic LDA and its key concepts.

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative probabilistic model where it is assumed that words observed in a number of documents of a corpus are generated by latent topics. The assumption is that each document, and a single Tweet is considered a document, is a mixture of topics. A topic is a distribution over terms of a fixed vocabulary and is usually represented by its terms with the highest probabilities. Note that we use "term" to denote a single class or entity, while word describes a specific occurrence of a term in a document.

The plate notation in Figure 1 illustrates the generative process of the LDA for M documents in the corpus, N words in a document with V different terms in the vocabulary and K topics. Each observed word w_i has a unique topic assignment z_i. The assignment is drawn from the multinomial distributions θ (distributions of topics in a document) and β (distribution of terms in a topic) with parameters α and η, respectively. θ and β are Dirichlet distributions. With $\theta \in \mathbb{R}^{M \times K}$ and $\beta \in \mathbb{R}^{K \times V}$ known, each document can be represented in K-dimensional space by the topics it contains.

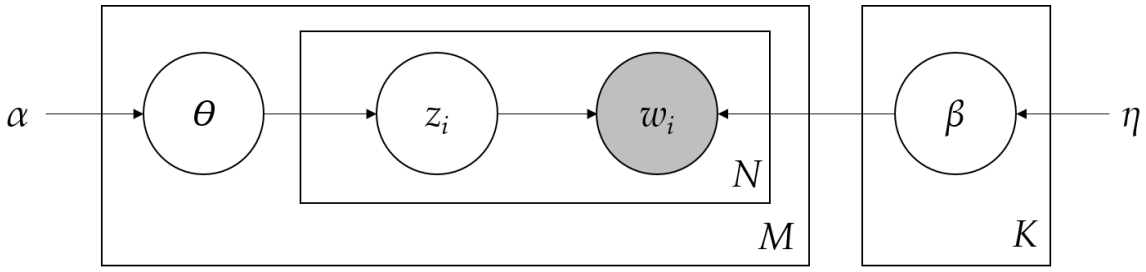

Figure 1. Plate notation of the LDA (based on [19]). The shaded node represents the observed variable, a word w_i in a document. The latent variables z_i for the topic assignment, θ for the document-topic distribution and β for the topic-term distribution are shown as clear nodes.

In a topic modeling application, the variables θ, β and z are unobserved, or latent. The goal is to learn the posterior probability

$$p(\theta, \beta, \boldsymbol{z} \mid \alpha, \eta) = \frac{p(\theta, \beta, \boldsymbol{z}, \boldsymbol{w} \mid \alpha, \eta)}{p(\boldsymbol{w} \mid \alpha, \eta)} \tag{1}$$

and thus computing the latent variables. As there is no closed-form solution to this problem, a number of approximation methods are used instead. Variational inference is applied in the original LDA

paper [19]. Hoffman et al. [20] apply variational Bayes inference. Another approach is to use Gibbs sampling, where the topic assignments per word are iteratively refined [21].

The evaluation of topic models is not straightforward. Intrinsic measurement scores such as log-likelihood and perplexity are shown to be negatively correlated with human judgement [11]. Thus, Mimno et al. [22] suggest an alternative intrinsic measure for assessing the coherence of a topic model: UMass coherence. In contrast to other coherence metrics, UMass coherence does not rely on external data sources [23]. It is instead based on the co-occurrence of the top n terms of each topic in the modeled documents:

$$coherence\left(T, W^{(T)}\right) = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \log \frac{D\left(w_i^{(T)}, w_j^{(T)}\right) + 1}{D\left(w_j^{(T)}\right)} \tag{2}$$

$D(w_i, w_j)$ is the number of documents both terms w_i and w_j occur in, $D(w_j)$ is the document frequency of term w_j. The UMass coherence indicates how well the topics are separable based on their vocabularies. Higher values indicate higher topic coherence. Besides the intrinsic measures, topic models can be evaluated with task-specific measures or based on classification performance for a held-out, annotated part of the data.

Incorporating prior knowledge in the model can improve the model output, especially when used for follow-up tasks. Wang et al. [24] introduce a targeted topic model (TTM) that uses seed words to guide the decision about the relevance of a document for a topic. The relevance is represented by a Bernoulli distribution included in the generative process.

Guided [12] and seeded [13] LDA variants do not change the generative process of the original LDA, but adapt the initialization of the topic-term distribution to reflect the importance of the seed words. Another option to recognize prior knowledge would be to change the Dirichlet prior of the topic-term distribution from symmetric to (heavily) skewed towards the seeds.

2.2. LDA for Disaster Detection in Tweets

Various efforts have been made to adapt and improve the basic LDA model to meet the requirements for specifically detecting disaster-related content in Twitter data. One line of research builds on iteratively refining the model by manually selecting the topic attributed to the disaster.

Kireyev et al. [25] use a general-purpose dataset to model the LDA and apply it afterwards to their Twitter data (approx. 20,000 Tweets each) of two disasters: a tsunami and an earthquake. By analyzing the topic distribution of disaster-related Tweets, similar documents in the larger original corpus are selected to further refine the topic model. Resch et al. [26] use a cascaded approach to select disaster related Tweets from a dataset on the Napa valley earthquake in a first step and separate into more refined topics afterwards. The selection of relevant topics is done manually in both stages and the performance is assessed on a small, labeled part of the dataset. They introduce a method to assess the spatial distribution of relevant Tweets.

Using labeled Tweets is another method to improve the basic LDA model, yet a labor-intensive one. Imran and Castillo [27] have experts label a set of Tweets on different crisis in 2012 and 2013 with predefined, relevant topics. Tweets that fit neither of these topics are further refined using LDA. Both works use qualitative analysis, i.e., visualization and human judgment, to assess their results.

Supervised LDA (sLDA) was introduced by Blei and McAuliffe [28] to predict document classes and is trained on a labeled dataset. Ashktorab et al. [29] use sLDA to provide relevant information extracted from social media to first responders. They experimented with datasets on twelve different crisis events in North America with approximately 1000 annotated Tweets each. Other classification methods proved to be more accurate than the sLDA and achieve higher evaluation scores (including F_1, precision and recall) for the binary classification task.

In more recent work, topic models capable of handling seed words to produce a more desired and better interpretable output have been used. To the best of our knowledge, those seed words have so far been only defined manually.

Kirsch et al. [30] use a targeted topic model (TTM) with predefined keywords in an emergency context. The results are evaluated by qualitative analyses such as the top words and hashtags of the topics or related sentiment.

Yang et al. [31] suggest a location-based dynamic sentiment-topic model (LDST) to make use of the geographic information of Tweets directly in the topic model. They apply the LDST to a dataset of approx. 160,000 Tweets with geo-reference related to hurricane Sandy and manually define seed words for five different topics. The evaluation of the geographic accuracy and relevance is done subjectively by assessing the shift in topics in different areas and states.

Table 2 gives an overview of the methods discussed above. While all approaches succeed in extracting relevant information, all rely on manually provided prior information (either hand-crafted seed words, manual data annotation or human assignment of the relevant topic). This delays the reaction to a disaster and minimizes generalization over different disaster types which both can be avoided by automatically extracting the initial seed words.

Table 2. Comparison of our approach (bottom) to related work concerning the methods used to incorporate prior information to the various model, dataset sizes and evaluation methods.

Type	Authors	Model	Number of Tweets	Evaluation method
manual selection	Resch et al. [26]	LDA	95,000	classification scores; spatial mapping
	Kireyev et al. [25]	LDA	20,000	qualitative analysis
manual labeling	Imran and Castillo [27]	LDA	1000	clustering scores (intra- and inter-similarity; volume)
	Ashktorab et al. [29]	sLDA	1000	classification scores (F_1 score, precision, recall)
manually defined seed words	Kirsch et al. [30]	TTM	10,000	qualitative analysis
	Yang et al. [31]	LDST	160,000	spatial mapping
automatically defined seed words	Ferner et al.	GLDA	95,000 (earthquake); 8000 (hurricane)	intrinsic analysis; classification scores; spatial mapping

3. Automated Generation of Seed Words

The automated generation of seed words is based on the assumption that in the case of a natural disaster, as with any other event, Tweets from within the affected area will differ significantly from those on "normal" days both concerning the discussed topics (i.e., terms used) and the frequency of certain terms. However, extracting the most frequent terms of the Tweets posted on the day of the event might not be significant enough. Thus, we propose to compare the vocabularies of two different days and suggest two variants: the first one computes term frequencies based on the union set of the two vocabularies, the second based on their difference set.

3.1. Union Vocabulary

Let C and D be the set of Tweets from the normal ("comparison") and event ("disaster") day, respectively, and $V_1 = V_C \cup V_D$ the shared vocabulary from both corpora. $tf_{t,D}$ denotes the absolute term frequency of term $t \in V_1$ in corpus D.

$$s_t = \frac{tf_{t,D}}{\sum_{w \in V_1} tf_{w,D}} - \frac{tf_{t,C}}{\sum_{w \in V_1} tf_{w,C}} \tag{3}$$

The score $s_t \in [-1, 1]$ for each term is the difference between the normalized term frequencies in D and C. Terms only occurring in V_D have a high positive score, while terms only occurring in V_C have a negative score. For terms being equally present in both vocabularies, the score is 0. The set of n seed words $S_1 = \{t_1, \ldots, t_n\}$ consists of the n highest positive scores. This method allows to also determine the n most negative values in v_1 for seeding a second topic with non-disaster-related terms. This possibility is currently not further explored.

3.2. Difference Vocabulary

The second method computes the frequency of terms that only occur in the disaster vocabulary $V_2 = V_D \setminus V_C, t \in V_2$.

$$s_t = \frac{tf_{t,D}}{\sum_{w \in V_2} tf_{w,D}} \tag{4}$$

The set of n seed words $S_2 = \{t_1, \ldots, t_n\}$ consists of the n terms with highest scores $s_t \in (0, 1]$.

Because of the second method being very restrictive and potentially yielding an empty set V_2 when only considering single terms, biterms are added to the vocabularies for both methods.

3.3. Topic Initialization

Once the seed words are available, they can be incorporated into the LDA model. The LDA algorithm is given a document-term matrix (the observed words $\boldsymbol{w}$ in M documents) as input and produces a topic-term matrix β and a document-topic matrix θ in the run of several iterative inference steps as output. Each Tweet is treated as a single document. The topic-term matrix β is usually initialized by a constant [19] or randomly [20]. For guiding the LDA, we explicitly initialize the selected seed terms with 1 for the disaster topic (usually the first topic) and 0 else, while randomly initializing all other terms, which is similar to the method in [12].

3.4. Tweet Extraction

During several training iterations (inference), the distributions for the topic-term matrix β and the document-topic matrix θ are learned. The result is a modeled LDA, meaning that both distributions β and θ are as close as possible to the mathematically exact solution of $C = \beta \cdot \theta$, where C is the normalized co-occurrence matrix of terms in documents.

The modeled matrices allow for an intuitive interpretation: The topic-term matrix reveals the probability of each term occurring in a specific topic, such that we can deduce the top terms for each topic. The document-topic matrix models the probability distribution of topics in a Tweet so that we can assign the most probable topic to a Tweet if we need a single label only.

4. Experiments

The aim of the experiments is to show that (i) the suggested methods extract suitable seed words that (ii) improve the performance of a baseline LDA implementation. The proposed method for an automated topic modeling process is evaluated on two Twitter datasets covering different natural disasters. The first one contains 94,458 geo-referenced Tweets captured on the 24th of August, 2014 in the area of the Napa valley in California, where an earthquake occurred in the early morning. The second dataset covers geo-referenced Tweets in the area around Houston, Texas from 27 August 2017, when a hurricane hit. It contains 8078 Tweets. Both datasets contain Tweets in English only and were crawled using Twitter's Streaming and REST API [32] restricted to extracting only Tweets with geo-reference.

The difference of the two datasets lies not only in their size, but also in the different nature of the two disasters: While early-warning systems can predict hurricanes in advance, earthquakes hit unexpectedly. Smaller earthquakes can follow the initial one, whereas hurricanes are often accompanied by heavy rainfall and flooding [33]. Both the preparedness to a hurricane and its more diverse nature will affect the coverage and terminology of the disaster in social media and thus the topic models. This also lead to different approaches in selecting a day for comparison: While for the earthquake, a day from a week before the event is considered, the day of comparison for the hurricane lies one month ahead. By being this restrictive, we aim at ruling out possible vocabulary overlaps, as the imminent weather changes might be announced in the days following up to the hurricane.

The Tweets are preprocessed in the following order before passing them to the LDA:

1. Removal of URLs, user names, mentions and email addresses;

2. Lower casing;
3. Removal of numbers and special characters;
4. Stemming;
5. Removal of (stemmed) stopwords;
6. Removal of words shorter than four characters.

For assessing the term frequencies, scikit-learn's [34] `CountVectorizer` is used. The parameter `ngram_range` is set to `(1,2)` to also consider biterms and `min_df` is set to 2 to disregard terms that only occur once in the dataset. Applying this range of preprocessing steps is in line with findings by Denny and Spirling [35] and Maier et al. [36] who investigated the effects of different preprocessing steps and their ordering.

4.1. Earthquake

The 17th of August, 2014 is considered as normal day for comparison with the earthquake on the 24th of August, 2014. After preprocessing, 85,311 Tweets from the day of the disaster and 72,320 Tweets from the normal day are available. Note that the Tweets from the normal day are not used for topic modeling, but only for the seed word generation.

Figure 2 lists the ten most frequent terms in the Tweets on the event day and compares them to the frequencies in the Tweets on the normal day. "earthquake" is the most frequent term and could be considered a good seed word for the disaster topic. The remaining terms, except for the swear words and "night" to some extent, are unrelated to the earthquake. Moreover, their frequencies do not differ from the normal day. The most frequent terms thus cannot be expected to help in guiding the LDA.

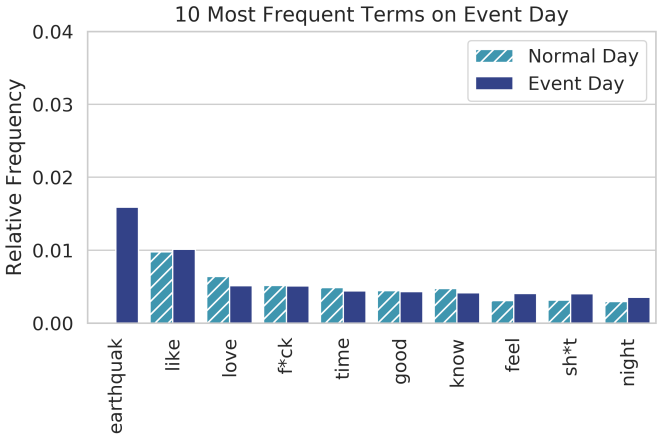

Figure 2. Relative frequency of the ten most frequent terms in the earthquake dataset on the day of the event compared to their relative frequency in the comparison dataset of a normal day.

In order to extract a wider range of seed words, it is inevitable to set the frequencies of the two days in relation: Applying the two presented methods V1 (union vocabulary) and V2 (difference vocabulary) with $n = 10$ yields the terms listed in Table 3. A number of terms is directly related to the earthquake ("earthquak", "felt", "quak", "shake"), others indirectly, referring to the night-time of the event ("woke", "sleep"). Some terms refer to a music event happening at the same day: "beyonc" for the singer Beyoncé and "voteso" as the preprocessed version of "#vote5sos".

Table 3. Results of the two proposed variants V1 (union vocabulary) and V2 (difference vocabulary) for automatically determining seed words on the earthquake dataset.

V1	V2
earthquak	magnitud
napa	aftershock
felt	napaquak
quak	magnitud earthquak
woke	first earthquak
beyonc	felt earthquak
shake	feel earthquak
sleep	napa earthquak
california	earthquak last
voteso	earthquak woke

As expected, V2 contains more biterms. All terms are directly related to the earthquake. The term "napaquak" is in fact a hashtag, with the character "#" removed during preprocessing. For simplicity, the biterms are split up (with duplicates removed) to only provide single terms to the LDA. Table 4 lists the final set of 10 seed terms for both variants.

Table 4. Final set of seed words for the variants V1 (union vocabulary) and V2 (difference vocabulary) for the earthquake dataset. The number of seed words can differ as the biterms from Table 3 are split up.

Earthquake	**V1**	earthquak, napa, felt, quak, woke, beyonc, shake, sleep, california, voteso
	V2	magnitud, aftershock, napaquak, earthquak, first, felt, feel, napa, last, woke

4.2. Hurricane

Hurricane Harvey swept over Houston, Texas particularly from 26th to 28th of August, 2017 [37]. We consider the 27th of August as day of the disaster, and the 28th of July, 2017 as normal day for comparison. After preprocessing, 7043 Tweets from the disaster day and 7020 from the normal day are available.

As with the earthquake dataset, extracting the most frequent terms on event day as seed words would be misleading. Most of the terms listed in Figure 3 are related to job postings ("hire", "careerarc", "open") in the Houston area. Except for the terms "houston" and "texa", the term frequencies are also comparable to the normal day.

The methods V1 and V2 comparing the two datasets yield different terms. Table 5 displays the results of V1 and V2 with $n = 10$. Although biterms are less frequent for this dataset, again most occur in V2, while only one is included in V1. Both methods cover a number of related terms to the hurricane (e.g., "flood", "rain") and also the most prominent hashtag "hurricaneharvey".

Table 6 lists the final set of seed terms after splitting the biterms, resulting in nine seeds for V1 and 12 for V2.

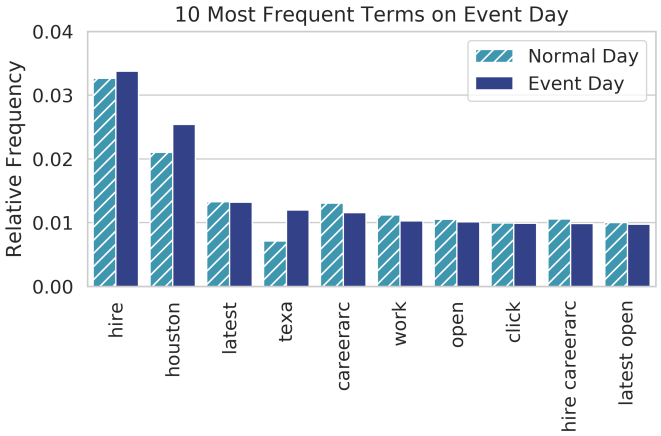

Figure 3. Relative frequency of the ten most frequent terms in the hurricane dataset on the day of the event compared to their relative frequency in the comparison dataset of a normal day.

Table 5. Results of the two proposed variants V1 (union vocabulary) and V2 (difference vocabulary) for automatically determining seed words on the hurricane dataset.

V1	**V2**
flood	flood
hurricaneharvey	hurricaneharvey
harvey	harvey
texa	tornado
houston	hurrican
water	high water
rain	storm
houston texa	warn includ
tornado	close flood
hurrican	flash flood

Table 6. Final set of seed words for the variants V1 (union vocabulary) and V2 (difference vocabulary) for the hurricane dataset (bottom). The number of seed words can differ as the biterms from Table 5 are split up.

Hurricane	**V1**	flood, hurricaneharvey, harvey, texa, houston, water, rain, tornado, hurrican
	V2	flood, hurricaneharvey, harvey, tornado, hurrican, high, water, storm, warn, includ, close, flash

5. Intrinsic Evaluation

As there are no other variants that allow for a fully automated process, we assess the competitiveness of our suggested methods (referred to as GLDA V1 and GLDA V2 in short) against a basic, unseeded LDA (LDA in short) and a baseline guided LDA (GLDA Baseline) that only uses the terms "earthquake" and "hurricane" as seeds, respectively. The selection of the disaster types as seed words could be considered as an almost automatic method, as the manual effort is negligible. That said, the dataset is not investigated any further to come up with more manual seed words for the GLDA Baseline. For all experiments, the number of topics K are varied from 2 to 100 to find an optimum.

Scikit-learn's `LatentDirichletAllocation` was used as LDA implemenation, where we overwrite the initialization method as described in Section 3.3 and allow for 25 iterations. The implemented inference method is the variational Bayes algorithm applied in batch mode, i.e., using all training data at once in each update step. The parameters α and η, prior parameters for the document-topic distribution and the topic-term distribution (see Figure 1), are set to $1/10,000$ and $1/K$. $1/K$ is the default value in scikit-learn, whereas α is deliberately set to a low value to account for the short text length in Tweets instead of applying a specific Twitter LDA model as in [38,39].

For our experiment, we compute the topic coherence for the disaster topic T based on the list of $n = 20$ top terms $W^{(T)} = (w_1, \ldots, w_S)$ (see Equation (2)). Figure 4 (left) illustrates the topic coherence of the disaster topic for varying K on the earthquake and the hurricane dataset, respectively. We run the LDA 5 times for each K to assess the average performance. For all methods, higher values of K are better. Obviously, the coherence of the LDA and the baseline GLDA improve significantly while the values of the GLDA with automatically extracted seed words remain low. However, for the overall topic coherence, i.e., the mean over all topics, the GLDA V1 and V2 are competitive (see Figure 4 (middle)). The variance over five runs is much higher for the GLDA Baseline and the LDA, indicating that the guidance by seed words adds stability to the model.

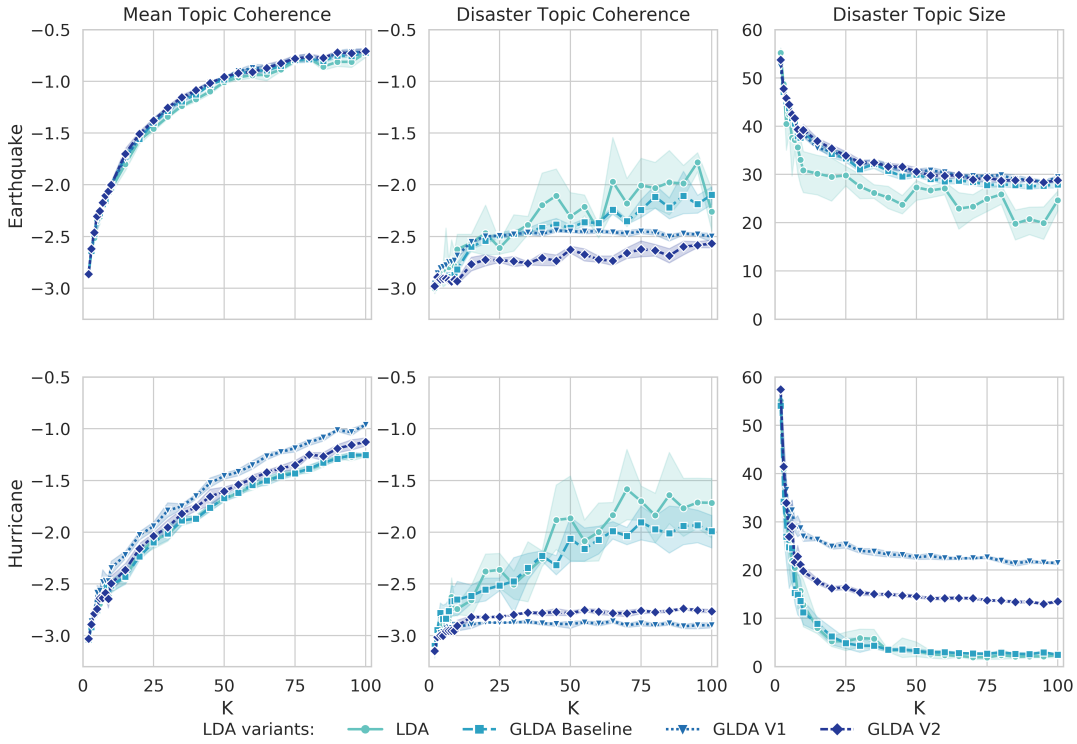

Figure 4. Average coherence over all topics (**left**), topic coherence for the disaster topic (**middle**) and the percentage of Tweets from the test set remaining in the disaster topic (**right**) for all four model variants and varying numbers of topics K. Markers indicate the mean over five runs with same K, while the upper and lower limits refer to the 95% confidence interval.

The disaster topic's size in Figure 4 (right) shows an opposing trend to its coherence and is lower for the LDA and GLDA Baseline variants than for the GLDA versions 1 and 2. As the topic coherence increases for the LDA, the topic becomes very restrictive and only covers a small fraction of Tweets. The remaining fraction of Tweets in the disaster topic is close to zero for the hurricane dataset. For the topic classification, this could suggest that the recall is low for increasing K.

6. Extrinsic Evaluation

Besides topic coherence, we use a small set of annotated data for an extrinsic evaluation and define a binary classification task: The test sets comprise 1331 manually labeled Tweets (1.6%) from the earthquake dataset and 993 (14.1%) from the hurricane dataset. Almost half of the labeled Tweets are disaster related for the earthquake dataset and about a quarter for the hurricane dataset.

Tweets having their maximum at the disaster topic in the document-topic matrix are considered as disaster related, all other as not related. For the LDA, we define the disaster topic as topic for which the term "earthquake" or "hurricane" has the highest value. For the guided variants, the seeded topic

is the disaster topic. All variants in the experiments model the unlabeled Tweets, and the results are used to "classify" the held-out set of labeled Tweets.

6.1. Tweet Classification

As performance measure for the binary classification, we use the F_1 score, the harmonic mean of precision and recall [40]. Precision is the measure for the number of Tweets correctly classified as disaster related. Recall defines the ratio of Tweets classified as disaster related from all Tweets labeled as disaster related. In other words, precision gives an idea of how many relevant Tweets are extracted, while recall assesses the completeness of retrieved relevant Tweets.

The resulting mean F_1 scores and their 95% confidence interval over five runs for both datasets are illustrated in Figure 5 (left). The guided variants outperform the LDA. For the hurricane dataset, the GLDA Baseline is also clearly inferior to the variants V1 and V2. Moreover, the performance of the variants V1 and V2 is more stable and stays above 60% F_1 score for all K. Again, the variance over the five runs is lower for the variants V1 and V2. As expected, the recall (see Figure 5 (middle)) correlates with the disaster topic size and is lower for the LDA. While the recall is better for smaller values of K, the precision (Figure 5 (right)) increases with higher K.

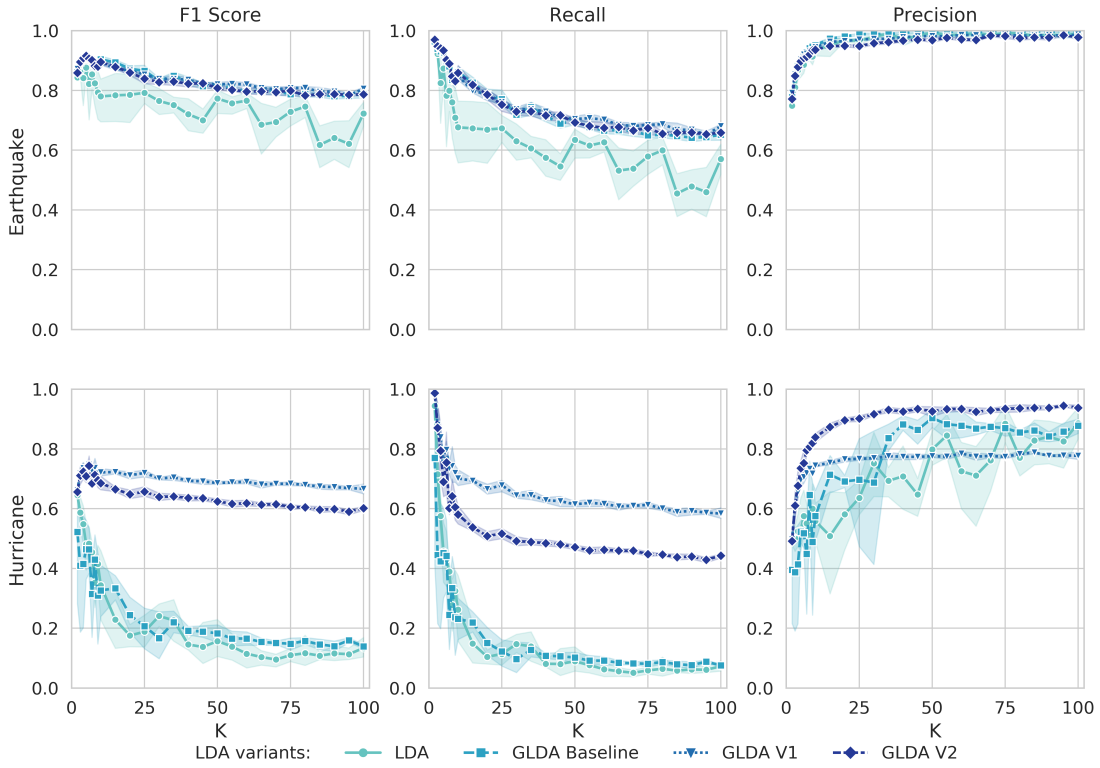

Figure 5. Results of the Tweet classification on the test set: F_1 score (**left**), recall (**middle**) and precision (**right**) for all four model variants and varying numbers of topics K. Markers indicate the mean over five runs with same K, while the upper and lower limits refer to the 95% confidence interval.

Table 7 lists the detailed classification performance measures for both datasets and all variants corresponding to the best F_1 score each. With all models, the best performance is achieved with $K < 8$ on both datasets. With this optimal setting, the performance of the three guided LDA variants is almost on a par for the earthquake dataset, but better than that of the LDA. For the hurricane dataset, the performance in terms of F_1 score of GLDA V1 and V2 outperforms the LDA by approximately 10% and the GLDA Baseline by more than 20%.

Table 8 highlights the resulting top 10 terms of the disaster topic for both GLDA variants, and $K = 5$ or $K = 6$ for the earthquake and hurricane dataset, respectively. Most of the terms occur for both variants (90%, unique terms in italic), although the ordering differs. It is interesting to note that the terms differ from the initial seed words shown in Tables 4 and 6. This underlines the idea that the initial setting only guides the LDA but is not restrictive. Note, on the one hand, that the term "hurrican" is missing, as well as "shake" for the earthquake dataset. On the other hand, the unrelated terms "beyonc" and "voteso" have vanished.

Table 7. Detailed classification results including precision and recall on the earthquake dataset (top) and the hurricane dataset (bottom) based on the best average F_1 score of the five runs for each variant.

	Variant	K	Precision	Recall	F_1 score
Earthquake	LDA	5	88.43%	87.32%	87.63%
	GLDA Baseline	5	90.59%	92.68%	91.62%
	GLDA V1	5	90.31%	92.71%	91.49%
	GLDA V2	5	89.75%	93.36%	91.52%
Hurricane	LDA	2	49.06%	94.41%	64.56%
	GLDA Baseline	2	39.53%	76.99%	52.23%
	GLDA V1	6	70.58%	79.72%	74.77%
	GLDA V2	6	75.22%	75.45%	74.41%

Table 8. Top 10 terms in the disaster-related topic (i.e., highest probability in the topic-term matrix β) for both datasets based on the best value for K for the two automated variants GLDA V1 and V2. The ordering corresponds to the term probability. Terms that are unique for one variant are highlighted in italic.

Earthquake	**V1**	$K = 5$	earthquak, sleep, napa, felt, night, california, like, quak, feel, *woke*
	V2	$K = 5$	earthquak, feel, last, night, napa, *first*, felt, like, california, sleep
Hurriance	**V1**	$K = 6$	houston, texa, flood, hurricaneharvey, harvey, *rain*, traffic, water, tornado, close
	V2	$K = 6$	houston, flood, hurricaneharvey, harvey, traffic, water, texa, tornado, close, *high*

6.2. Geospatial Analysis

In addition to the extraction of semantically relevant Tweets, their geospatial distribution is analyzed in Figures 6–9. For both datasets, we compare the results based on the output of the LDA and GLDA V2. Cells in red and blue signify the hot and cold spots of the disaster related Tweets in relation to the total amount of Tweets in the given cell area as proposed by Resch et al. [20]. All Tweets are aggregated in a fishnet that overlaps with the area of interest and the ratio between the number of total Tweets and the disaster related Tweets in each grid cell is computed. The fishnet consists of squared cells with side length $l = \sqrt{2\frac{A}{n}}$, where A is the size of the study area, and n is the number of points in the study area [41]. The informativeness is high if the ratio of disaster related Tweets is high which is especially true in rural areas where the Tweet activity is usually lower than in urban areas.

To identify geospatial clusters, we applied a hot spot analysis based on Getis-Ord GI* [42] that determines statistically significant hot and cold spots by including the neighborhood of the analyzed cell. The hypothesis is that disaster related Tweets will be more frequent in affected areas. Thus, the identified hot spots are a strong indicator for areas impacted by a natural disaster. In order to evaluate that the automatically generated seed words yield an accurate disaster topic, we compare the hot and cold spots with the official disaster footprints. For the Napa valley earthquake in Figures 6 and 7, the results are matched against the earthquake's footprint as assessed by the US Geological Survey (USGS) [43]. We excluded values that would fall in the categories "Not felt" and "Weak" in the USGS Peak Ground Acceleration (PGA) dataset [43]. Those categories would not add significant value to the interpretation of the map as those values signify that people do not notice the appearance of an earthquake and do not recognize it as such [44].

The hot and cold spots as identified based on the disaster related Tweets from the LDA in Figure 6 reveal mismatches: North of Santa Rosa, where the earthquake still was felt, the analysis reveals cold spots. For the area south of Oakland that lies outside of the affected area, no significant results were obtained. Most importantly, the epicenter around Napa is not correctly detected, as no hot spots are identified in this area. However, the hot spots that are detected mostly are located within the official footprint.

The results based on the disaster related Tweets from GLDA V2 in Figure 7 reveal some major improvements. First of all, the epicenter is correctly identified by a cluster of hot spots with highest confidence. The area south of Oakland down to San Jose that lies outside of the official footprint is clearly highlighted as cold spot. The area north of Santa Rosa is now correctly identified as hot spot.

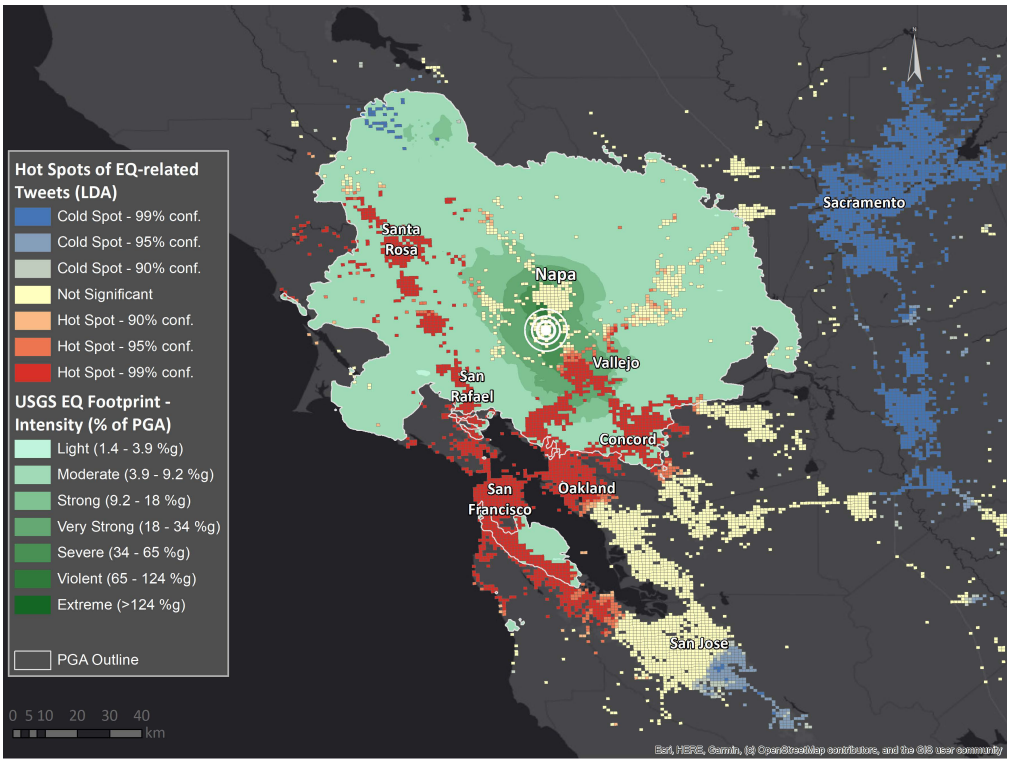

Figure 6. Earthquake hot and cold spots obtained using plain LDA compared to the US Geological Survey (USGS) [43] footprint measuring the intensity in per cent of the peak ground acceleration (PGA).

The official data available for hurricane Harvey assesses the extend of the flood that followed in the area [45]. Figure 8 shows the results based on the disaster related Tweets as detected by the LDA. Almost no significant spots could be detected, although the most affected area should be in and around Houston, while San Antonio, Austin and Lafayette are hardly affected.

Figure 9 illustrates the results based on GLDA V2 that now are in line with the flooded areas. Not only the hot spot in Houston is correctly detected, but also the area further east. The surrounding cities (San Antonio, Austin and Lafayette) are identified as cold spots this time.

The accordance between the computed hot spots and the official footprints for both disasters suggests that the automatically extracted information from Tweets can be an immediate evidence to support disaster management.

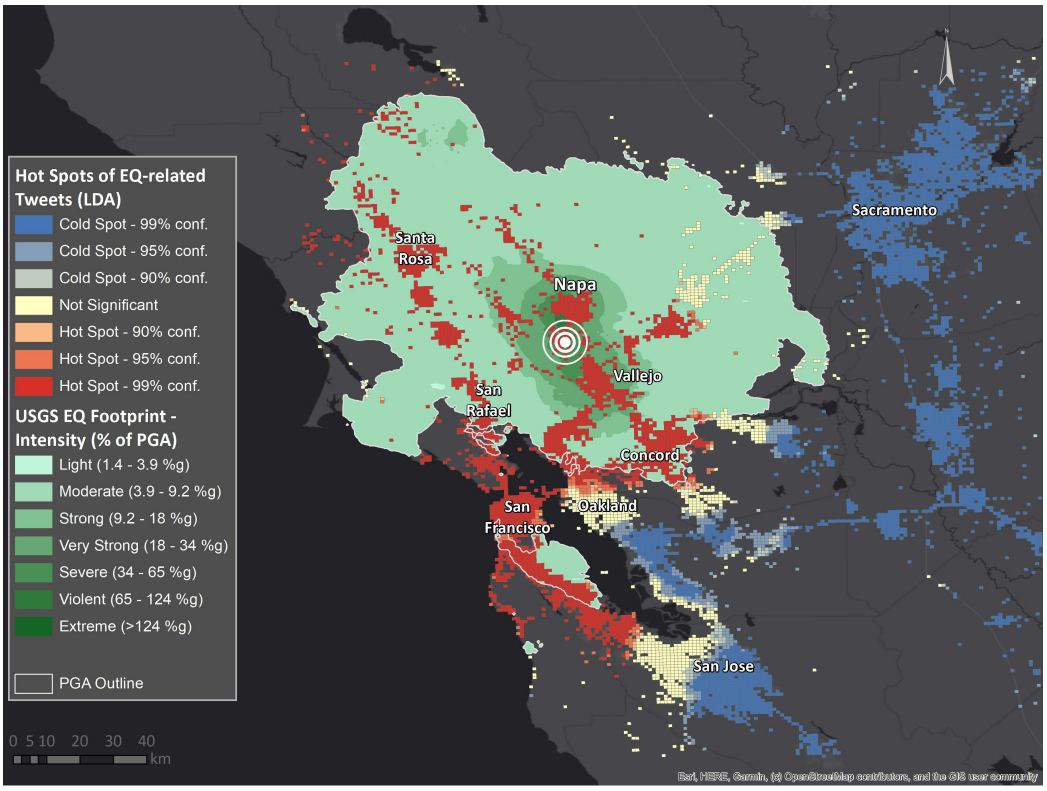

Figure 7. Earthquake hot and cold spots obtained using GLDA V2 compared to the US Geological Survey (USGS) [43] footprint measuring the intensity in per cent of the peak ground acceleration (PGA).

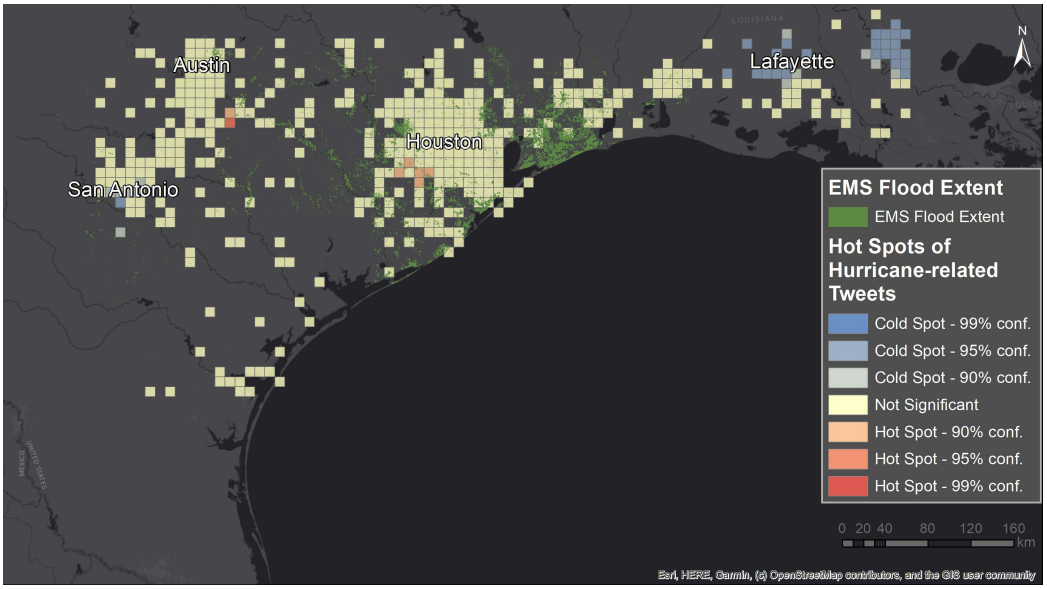

Figure 8. Hurricane hot and cold spots obtained using plain LDA compared to COPERNICUS Emergency Management Service (EMS) [45] mapping data depicting flooded areas six days after the landfall.

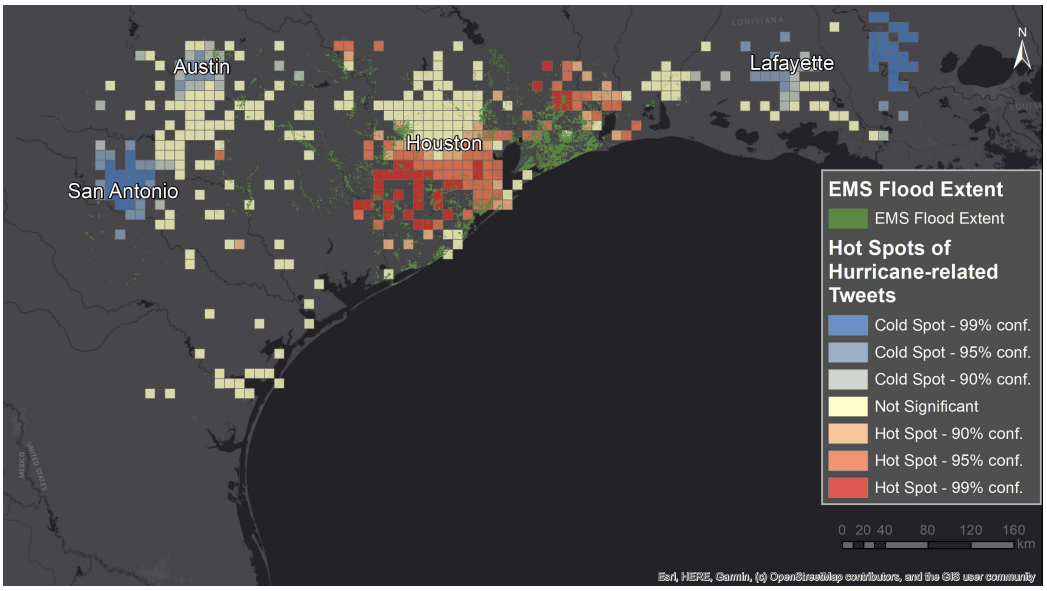

Figure 9. Hurricane hot and cold spots obtained using GLDA V2 compared to COPERNICUS Emergency Management Service (EMS) [45] mapping data depicting flooded areas six days after the landfall.

7. Discussion

Comparing Tweets from two different days to extract meaningful seed words is a feasible way to automate the topic modeling process. Moreover, the experiments show that guiding the LDA towards those seed words for topics of interest improves the performance—only slightly in terms of overall topic coherence, but significantly for extracting disaster related Tweets, which is the central goal in supporting disaster management [26].

When comparing the two datasets, it is clear that by fully automating the topic modeling process, the classification performance does not suffer. On the contrary, the classification of the hurricane dataset improves considerably. Keeping in mind that the nature of a hurricane is more diverse than that of an earthquake, it seems that the standard methods are limited. While using a "dummy" keyword (GLDA Baseline) as seed is competitive on the earthquake dataset, this variant also performs poorly on the hurricane dataset. Automatically generating seed words thus will also be helpful in case of unknown, diverse or rare events that are hard to describe by a small set of manually defined keywords.

As the results of five different runs each imply, the performance of the guided LDA with automatically generated seed words is more stable. As the variation over the runs is low, there is no need for further hyper-parameter tuning or experimenting with the right number of topics, which is beneficial in real-world scenarios. For the LDA, even with the best performing K, the standard deviation is 3.9% on the earthquake dataset, while only 0.5% for both GLDA variants. On the hurricane dataset, the standard deviation is 1.7% for LDA, 1.5% for GLDA V1 and 0.8% for GLDA V2. At least for the LDA, this result could also mean that the model did not fully converge yet and more iteration steps (and more time) are needed.

According to our experiments, meaningful and disaster related Tweets are extracted with a small number of topics, i.e., $K <= 6$. However, the F_1 score might not always be a good proxy for assessing the retrieval. Tasks focusing on displaying the retrieved Tweets in text-based form might require smaller, but highly precise result sets and thus favor precision over recall. For quantity-focused tasks, a higher recall might be preferable. The experiments suggest that the optimal value for K then differs: smaller K for better recall, higher K for improved precision.

The geospatial visualization of hot and cold spots reveals an impressive alignment with the actual, official disaster footprints when computed based on the automatically extracted Tweets with GLDA

V2. Correctly distinguishing between affected and non-affected areas allows to direct aid and rescue efforts to places where help is needed, and not to densely populated, urban areas by default.

Concerning the generalizability of these observations, experiments with further datasets and different events would be needed. Future experiments could also serve to investigate the effect of multilingualism on our approach. As both catastrophes took place in the United States, the predominant number of Tweets is in English. Although LDA can model topics over documents in different languages, more sophisticated topic models have been introduced to explicitly handle multilingualism. Existing methods for multilingual topic modeling [46] would need to be adapted to also handle seed words.

In this study, the Twitter dataset was collected with a geo-crawler software that can crawl the Streaming and REST API of Twitter besides multiple other social media network APIs. Using the API, Tweets can currently be crawled up to a week back. In case of a natural disaster or other event, this allows for a comparison of two different days a week apart. For dates further back, online repositories such as the Internet Archive's Twitter stream [47] that collect data continuously can be consulted. Authorities or disaster response organizations might even have an interest in monitoring a geographic region by crawling the data regularly by themselves.

The geo-crawler focuses on geo-referenced social media posts that can include the precise coordinates of a GPS-enabled device. Although in June 2019, Twitter announced a fundamental change in adding precise location information to a Tweet [48], Tweets with precise geolocation information still can be collected. Furthermore, the extraction of location information from text has made substantial progress in the last years and has consequently opened new opportunities for geospatial analysis on Twitter data [49–51]. Therefore, the methodology developed in this paper can also be applied to other use cases which has been tested with success outside the scope of this study.

8. Conclusions

Information on social media is available in a timely manner, thus being an important source for detecting and responding to events. Especially for disaster management, a system monitoring the Twitter stream in a given region and automatically detecting natural disasters or other crises would be useful. Such a system can be implemented based on our fully automated approach to incorporate data-intrinsic a-priori knowledge into a topic model (LDA) without any need for manual interference. By comparing Twitter vocabularies from different days, seed words indicating the events can be generated. These extracted seed words are used to guide the LDA to model a single disaster-related topic. The Tweets corresponding to this topic can then be mapped to get an estimate for affected areas.

The method is tested on Tweets from two datasets covering an earthquake and a hurricane, respectively. Two approaches for comparing the Tweets' vocabularies from a preceding day with the day of the disaster event are presented. The first approach is to extract the most frequent terms over the union vocabulary of both days and subtracting the normal day's term counts from that of the disaster day. The second possibility is to extract the most frequent terms over the difference set.

Both methods yield event-related seed words that improve the F_1 score of a basic LDA or a baseline guided LDA on a small set of labeled Tweets. The resulting topics are similarly coherent as with a standard topic model. Moreover, the detected hot and cold spots based on the geo-references of the Tweets are aligned with the official footprints of the examined disasters. This underlines the value of Tweets as decision support for disaster management. The model parameter K, i.e., the number of topics to model, can be used to trade off recall against precision.

Author Contributions: Conceptualization, C.F., C.H., B.R. and S.W.; methodology, C.F., E.B., S.W.; software, C.F., C.H., E.B.; validation, C.F., C.H. and E.B.; formal analysis, C.F., C.H. and E.B.; investigation, C.F., C.H. and E.B.; data curation, C.H. and E.B.; writing–original draft preparation, C.F.; writing–review and editing, C.F., C.H., E.B., B.R. and S.W.; visualization, C.F., C.H.; supervision, B.R. and S.W.; project administration, B.R. and S.W.; funding acquisition, B.R. and S.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This study has been carried out in the HUMAN+ project, which has been funded by the Austrian security research programme KIRAS of the Federal Ministry of Agriculture, Regions and Tourism (BMLRT), project number 865697. Additional funding was granted by the European Regional Development Fund (ERDF) for project number AB215 in the INTERREG program Austria-Bavaria 2014-2020.

Acknowledgments: We would like to thank Harvard University's Center for Geographic Analysis for their support by providing us with the Twitter data for our study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hodge, K. Techradar 10 News Stories That Broke on Twitter First. Available online: https://www.techradar.com/news/world-of-tech/internet/10-news-stories-that-broke-on-twitter-first-719532 (accessed on 24 July 2020).
2. Internet Live Stats. Twitter Usage Statistics. Available online: https://www.internetlivestats.com/twitter-statistics/ (accessed on 24 July 2020).
3. Wang, H.; Can, D.; Kazemzadeh, A.; Bar, F.; Narayanan, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations, Stroudsburg, PA, USA, 8–14 July 2012; pp. 115–120.
4. Alsaedi, N.; Burnap, P.; Rana, O. Can We Predict a Riot? Disruptive Event Detection Using Twitter. *ACM Trans. Internet Technol.* **2017**, *17*, 1–26. [CrossRef]
5. Resch, B.; Summa, A.; Zeile, P.; Strube, M. Citizen-centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-time-linguistics Algorithm. *Urban Plan.* **2016**, *1*, 114–127. [CrossRef]
6. Niles, M.T.; Emery, B.F.; Reagan, A.J.; Dodds, P.S.; Danforth, C.M. Social Media Usage Patterns During Natural Hazards. *PLoS ONE* **2019**, *14*, e0210484. [CrossRef] [PubMed]
7. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010, pp. 851–860.
8. Resch, B. People as Sensors and Collective Sensing-contextual Observations Complementing Geo-sensor Network Measurements. In *Progress in Location-Based Services*; Springer: New York, NY, USA, 2013; pp. 391–406.
9. Havas, C.; Resch, B.; Francalanci, C.; Pernici, B.; Scalia, G.; Fernandez-Marquez, J.; Van Achte, T.; Zeug, G.; Mondardini, M.; Grandoni, D.; et al. E2mc: Improving Emergency Management Service Practice through Social Media and Crowdsourcing Analysis in Near Real time. *Sensors* **2017**, *17*, 2766. [CrossRef] [PubMed]
10. Blei, D.M. Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
11. Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J.L.; Blei, D.M. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*; Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2009; pp. 288–296.
12. Li, J.; Ge, Y.; Hong, Y.; Cheema, A.; Gu, B. Textual Review Dimensionality and Helpfulness: A Multi-Method Study. 2017. Available online: https://abstract=2931934 (accessed on 18 June 2020).
13. Jagarlamudi, J.; Daumé, H., III; Udupa, R. Incorporating Lexical Priors into Topic Models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 204–213.
14. Andrzejewski, D.; Zhu, X. Latent Dirichlet Allocation with Topic-in-Set Knowledge. In Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, Boulder, Colorado, 4 June 2009; pp. 43–48.
15. Hu, Y.; Boyd-Graber, J. Efficient Tree-based Topic Modeling. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Jeju Island, Korea, 8–14 July 2012; Association for Computational Linguistics: Boulder, Colorado, 2012; pp. 275–279.
16. Xie, P.; Yang, D.; Xing, E. Incorporating Word Correlation Knowledge into Topic Modeling. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 725–734.

17. Guan, X.; Chen, C. Using Social Media Data to Understand and Assess Disasters. *Nat. Hazards* **2014**, *74*, 837–850. [CrossRef]
18. Gründer-Fahrer, S.; Schlaf, A.; Wiedemann, G.; Heyer, G. Topics and topical phases in German social media communication during a disaster. *Nat. Lang. Eng.* **2018**, *24*, 221–264. [CrossRef]
19. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
20. Hoffman, M.; Bach, F.R.; Blei, D.M. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*; Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2010; pp. 856–864.
21. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
22. Mimno, D.; Wallach, H.M.; Talley, E.; Leenders, M.; McCallum, A. Optimizing Semantic Coherence in Topic Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 262–272.
23. Stevens, K.; Kegelmeyer, P.; Andrzejewski, D.; Buttler, D. Exploring Topic Coherence Over Many Models and Many Topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 2012 pp. 952–961.
24. Wang, S.; Chen, Z.; Fei, G.; Liu, B.; Emery, S. Targeted Topic Modeling for Focused Analysis. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1235–1244.
25. Kireyev, K.; Palen, L.; Anderson, K. Applications of Topics Models to Analysis of Disaster-related Twitter Data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*; JMLR.org: Whistler, BC, Canada, 2009; Volume 1.
26. Resch, B.; Usländer, F.; Havas, C. Combining Machine-learning Topic Models and Spatiotemporal Analysis of Social Media Data for Disaster Footprint and Damage Assessment. *Cartogr. Geogr. Inf. Sci.* **2018**, *45*, 362–376. [CrossRef]
27. Imran, M.; Castillo, C. Towards a Data-driven Approach to Identify Crisis-related Topics in Social Media Streams. In Proceedings of the 24th ACM International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 1205–1210.
28. Blei, D.M.; McAuliffe, J.D. Supervised topic models. In Proceedings of the 20th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 3–7 December 2007; pp. 121–128.
29. Ashktorab, Z.; Brown, C.; Nandi, M.; Culotta, A. Tweedr: Mining Twitter to Inform Disaster Response. In Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management, University Park, PA, USA, 18–21 May 2014; ISCRAM Association: Centre County, PA, USA, 2014.
30. Kirsch, B.; Giesselbach, S.; Knodt, D.; Rüping, S. Robust End-User-Driven Social Media Monitoring for Law Enforcement and Emergency Monitoring. In *Community-Oriented Policing and Technological Innovations*; Leventakis, G., Haberfeld, M.R., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 29–36.
31. Yang, M.; Mei, J.; Ji, H.; Zhao, W.; Zhao, Z.; Chen, X. Identifying and Tracking Sentiments and Topics from Social Media Texts during Natural Disasters. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 527–533.
32. Twitter. Docs—Twitter Developer. Available online: https://developer.twitter.com/en/docs (accessed on 24 July 2020).
33. Carter, W.N. *Disaster Management: A Disaster Manager's Handbook*; Asian Development Bank: Metro Manila, Philippines, 2008.
34. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
35. Denny, M.J.; Spirling, A. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do About It. *Political Anal.* **2018**, *26*, 168–189. [CrossRef]
36. Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; et al. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Commun. Methods Meas.* **2018**, *12*, 93–118. [CrossRef]

37. van Oldenborgh, G.J.; van der Wiel, K.; Sebastian, A.; Singh, R.; Arrighi, J.; Otto, F.; Haustein, K.; Li, S.; Vecchi, G.; Cullen, H. Attribution of Extreme Rainfall from Hurricane Harvey, August 2017. *Environ. Res. Lett.* **2017**, *12*, 124009. [CrossRef]
38. Lin, T.; Tian, W.; Mei, Q.; Cheng, H. The Dual-sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text. In Proceedings of the 23rd ACM International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; pp. 539–550.
39. Zhao, W.X.; Jiang, J.; Weng, J.; He, J.; Lim, E.P.; Yan, H.; Li, X. Comparing Twitter and Traditional Media using Topic Models. In *European Conference on Information Retrieval*; Springer: New York, NY, USA, 2011; pp. 338–349.
40. Manning, C.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. *Nat. Lang. Eng.* **2010**, *16*, 100–103.
41. Wong, D.W.S.; Lee, J. *Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS*; John Wiley & Sons: Hoboken, NJ, USA, 2005.
42. Ord, J.K.; Getis, A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geogr. Anal.* **1995**, *27*, 286–306. [CrossRef]
43. US Geological Survey. Earthquake Hazards Program. Available online: https://earthquake.usgs.gov/earthquakes/eventpage/nc72282711/executive#general_executive (accessed on 18 June 2020).
44. US Geological Survey. The Modified Mercalli Intensity Scale. Available online: https://www.usgs.gov/natural-hazards/earthquake-hazards/science/modified-mercalli-intensity-scale?qt-science_center_objects=0#qt-science_center_objects (accessed on 18 June 2020).
45. COPERNICUS Emergency Management Service. COPERNICUS EMS-Mapping. EMSR229: Hurricane Harvey in Texas. Available online: https://emergency.copernicus.eu/mapping/list-of-components/EMSR229 (accessed on 18 June 2020).
46. Boyd-Graber, J.; Blei, D.M. Multilingual Topic Models for Unaligned Text. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June 2009; AUAI Press: Arlington, VA, USA, 2009; pp. 75–82.
47. Internet Archive. Archive Team: The Twitter Stream Grab. Available online: https://archive.org/details/twitterstream (accessed on 24 July 2020).
48. Twitter Support. Announcement. Available online: https://twitter.com/TwitterSupport/status/1141039841993355264?s=20 (accessed on 24 July 2020).
49. Kumar, A.; Singh, J.P. Location Reference Identification from Tweets during Emergencies: A Deep Learning Approach. *Int. J. Disaster Risk Reduct.* **2019**, *33*, 365–375. [CrossRef]
50. Limsopatham, N.; Collier, N. Bidirectional LSTM for Named Entity Recognition in Twitter Messages. In Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT), Osaka, Japan, 11 December 2016; pp. 145–152.
51. Inkpen, D.; Liu, J.; Farzindar, A.; Kazemi, F.; Ghazi, D. Location Detection and Disambiguation from Twitter Messages. *J. Intell. Inf. Syst.* **2017**, *49*, 237–253. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).