Bridging the gap between supervised classification and unsupervised topic modelling for social-media assisted crisis management

Mikael Brunila* **Rosie Zhao*** **Andrei Mircea*** **Sam Lumley** **Renee Sieber**
McGill University / Montreal, QC, Canada
{rosie.zhao,andrei.romascanu,sam.lumley}@mail.mcgill.ca
mikael.brunila@gmail.com
renee.sieber@mcgill.ca

Abstract

Social media such as Twitter provide valuable information to crisis managers and affected people during natural disasters. Machine learning can help structure and extract information from the large volume of messages shared during a crisis; however, the constantly evolving nature of crises makes effective domain adaptation essential. Supervised classification is limited by unchangeable class labels that may not be relevant to new events, and unsupervised topic modelling by insufficient prior knowledge. In this paper, we bridge the gap between the two and show that BERT embeddings finetuned on crisis-related tweet classification can effectively be used to adapt to a new crisis, discovering novel topics while preserving relevant classes from supervised training, and leveraging bidirectional self-attention to extract topic keywords. We create a dataset of tweets from a snowstorm to evaluate our method's transferability to new crises, and find that it outperforms traditional topic models in both automatic, and human evaluations grounded in the needs of crisis managers. More broadly, our method can be used for textual domain adaptation where the latent classes are unknown but overlap with known classes from other domains.

1 Introduction

1.1 Social Media for Crisis Management

As climate change increases the frequency of extreme weather events and the vulnerability of affected people, effective crisis management is becoming increasingly important for mitigating the negative effects of these crises (Keim, 2008). In the general crisis management literature, social media has been identified as a useful source of information for crisis managers to gauge reactions from and communicate with the public, increase situational awareness, and enable data-driven decision-making (Tobias, 2011; Alexander, 2014; Jin et al., 2014).

1.2 The Need For NLP

The large volume and noise-to-signal ratio of social media platforms such as Twitter makes it difficult to extract actionable information, especially at a rate suited for the urgency of a crisis. This has motivated the application of natural language processing (NLP) techniques to help automatically filter information in real-time as a crisis unfolds (Imran et al., 2013; Emmanouil and Nikolaos, 2015).

Other work has investigated the use of NLP models to automatically classify tweets into finer-grained categories that can be more salient to crisis managers and affected people in rapidly evolving situations (Ragini and Anand, 2016; Schulz et al., 2014). Training such classification models typically requires large-scale annotated corpora of crisis-related tweets such as that made available by Imran et al. (2016), which covers a variety of countries and natural disasters including flooding, tropical storms, earthquakes, and forest fires.

1.3 Limitations of Current Methods

Whereas supervised approaches work well for classifying tweets from the same event as their training data, they often fail to generalize to novel events (Nguyen et al., 2017). A novel event may differ from past events in terms of location, type of event, or event characteristics; all of which can change the relevance of a tweet classification scheme.

Various methods of domain adaptation have been suggested for addressing this issue (Li et al., 2018; Sopova, 2017; Alrashdi and O'Keefe, 2020). However, this type of supervised classification assumes that relevant classes remain the same from event to event. Probabilistic topic modelling approaches such as Latent Dirichlet Allocation (LDA) can overcome this limitation and identify novel categorizations (Blei et al., 2003). Unfortunately, these unsupervised methods are typically difficult to apply to tweets due to issues of document length and non-standard language (Hong and Davison, 2010).

*Equal contribution.

Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 33–49
April 20, 2021.

Furthermore, the categorizations produced by these models can be difficult to interpret by humans, limiting their usefulness (Blekanov et al., 2020).

1.4 Our Contributions

To address these issues, we propose a method for the unsupervised clustering of tweets from novel crises, using the representations learned from supervised classification. Specifically, we use the contextual embeddings of a pretrained language model finetuned on crisis tweet classification.

Our method bridges the gap between supervised approaches and unsupervised topic modelling, improving domain adaptation to new crises by allowing classification of tweets in novel topics while preserving relevant classes from supervised training. Our model is robust to idiosyncrasies of tweet texts such as short document length and non-standard language, and leverages bi-directional self-attention to provide interpretable topic keywords.

We assess our approach's transferability to novel crises by creating a dataset of tweets from Winter Storm Jacob, a severe winter storm that hit Newfoundland, Canada in January 2020. This event differs significantly from past crisis tweet classification datasets on which we finetune our model, and allows us to evaluate domain adaptation for novel events. We find that our approach indeed identifies novel topics that are distinct from the labels seen during supervised training.

In line with human-centered machine learning principles (Ramos et al., 2019), we also create a novel human evaluation task aligned with the needs of crisis managers. We find that, with high inter-rater reliability, our model provides consistently more interpretable and useful topic keywords than traditional approaches, while improving cluster coherence as measured by intruder detection. Automated coherence measures further support these findings. Our code and dataset are available at https://github.com/smacawi/bert-topics.

2 Related Work

2.1 Topic Modelling

Topic modelling in a variety of domains is widely studied. Although there are many existing approaches in the literature, most innovations are compared to the seminal Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). LDA is a generative probabilistic model that considers the joint distribution of observed variables (words) and hidden variables (topics). While LDA has well-known issues with short text, approaches such as Biterm Topic Modelling (BTM) have been developed to address these (Yan et al., 2013). BTM specifically addresses the sparsity of word co-occurrences: whereas LDA models word-document occurrences, BTM models word co-occurrences ('biterms') across the entire corpus.

2.2 Clustering

Topic modelling can be distinguished from clustering, where documents are usually represented in a multi-dimensional vector space and then grouped using vector similarity measures. These representations are typically formed through matrix factorization techniques (Levy and Goldberg, 2014) that compress words (Mikolov et al., 2013b; Pennington et al., 2014) or sentences (Mikolov et al., 2013a; Lau and Baldwin, 2016) into "embeddings". Recently, language representation models building on Transformer type neural networks (Vaswani et al., 2017) have upended much of NLP and provided new, "contextualized" embedding approaches (Devlin et al., 2019; Peters et al., 2018). Among these is the Bidirectional Encoder Representations from Transformers (BERT) model, which is available pretrained on large amounts of text and can be fine-tuned on many different types of NLP tasks (Devlin et al., 2019). Embeddings from BERT and other Transformer type language models can potentially serve as a basis for both topic modelling (Bianchi et al., 2020) and clustering (Reimers et al., 2019), but many questions about their usefulness for these tasks remain open.

The use of embeddings for clustering in the field of crisis management has been explored by Demszky et al. (2019) using trained GloVe embeddings. Zahera et al. (2019) used contextualized word embeddings from BERT to train a classifier for crisis-related tweets on a fixed set of labels. Using BERT to classify tweets in the field of disaster management was also studied by Ma (2019) by aggregating labelled tweets from the CrisisNLP and CrisisLexT26 datasets. While the aforementioned work requires data with a gold standard set of labels, our proposed clustering approach using finetuned BERT embeddings is applied in an unsupervised environment on unseen data — it invites domain expertise to determine an appropriate set of labels specific to the crisis at hand.

2.3 Topic Keyword Extraction

A significant issue when clustering text documents is how the keywords of each cluster or topic are determined. Unlike standard topic models such as LDA, clustering approaches do not jointly model the distributions of keywords over topics and of topics over documents. In other words, clusters do not contain any obvious information about which keywords should represent the clusters as topics. While the previous generation of embedding models has been leveraged for interpretable linguistic analysis in a wide variety of settings (Garg et al., 2018; Hamilton et al., 2016; Kozlowski et al., 2019), the interpretability of language models in general and BERT in particular remains a contested issue (Rogers et al., 2020).

One promising line of research has been on the attention mechanism of Transformer models (Bahdanau et al., 2015; Jain and Wallace, 2019). Clark et al. (2019) found that the attention heads of BERT contained a significant amount of syntactic and grammatical information, and Lin et al. (2019) concluded that this information is hierarchical, similar to syntactic tree structures. Kovaleva et al. (2019) noted that different attention heads often carry overlapping and redundant information. However, if attention is to be useful for selecting topic keywords, the crucial question is whether it captures semantic information. Jain and Wallace (2019) found that attention heads generally correlated poorly with traditional measures for feature importance in neural networks, such as gradients, while Serrano and Smith (2019) showed that attention can "noisily" predict the importance of features for overall model performance and Wiegreffe and Pinter (2019) argued that attention can serve plausible, although not faithful explanations of models. To the best of our knowledge, there is no previous work on leveraging attention to improve topic modelling interpretability.

2.4 Model Coherence

Whether based on clustering or probabilistic approaches, topic models are typically evaluated by their coherence. While human evaluation is preferable, several automated methods have been proposed to emulate that of human performance (Lau et al., 2014). In both cases, coherence can be thought of formally as a measure of the extent to which keywords in a topic relate to each other as a set of semantically coherent *facts* (Röder et al., 2015; Aletras and Stevenson, 2013; Mimno et al., 2011). If a word states a fact, then the coherence of a set of topic keywords can be measured by computing how strongly a word W' is confirmed by a conditioning set of words W^*. This can be done either directly where W' is a word in a set of topic words and W^* are the other words in the same set (e.g. Mimno et al., 2011), or indirectly by computing context vectors for both W' and W^* and then comparing these (e.g. Aletras and Stevenson, 2013). In a comprehensive comparison of different coherence measures, Röder et al. (2015) found that, when comparing the coherence scores assigned by humans to a set of topics against a large number of automated metrics, indirect confirmation measures tend to result in a higher correlation between human and automated coherence scores.

2.5 Human Evaluation of Topic Models

While automated coherence measures can be used to rapidly evaluate topic models such as LDA, studies have shown that these metrics can be uncorrelated — or negatively correlated — to human interpretability judgements. Chang et al. (2009) demonstrated that results given by perplexity measures differed from that of their proposed 'intrusion' tasks, where humans identify spurious words inserted in a topic, and mismatched topics assigned to a document. Tasks have been formulated in previous works to meaningfully enable human judgement when analyzing the topics (Chuang et al., 2013; Lee et al., 2017).

The latent topics given by these methods should provide a semantically meaningful decomposition of a given corpus. Formalizing the quality of the resulting latent topics via qualitative tasks or quantitative metrics is even less straightforward in an applied setting, where it is particularly important that the semantic meaning underlying the topic model is relevant to its users. Due to our focus on the comparison between the labels in the CrisisNLP dataset and novel topics discovered by our model, we restrict this part of the analysis to nine topics.

Our work is motivated by structuring crude textual data for practical use by crisis managers, guiding *corpus exploration* and efficient information retrieval. It remains difficult to verify whether the latent space discovered by topic models is both interpretable and useful without a gold standard set of labels.

3 Methodology

In this section we describe the dataset of snowstorm-related tweets we created to evaluate our model's ability to discover novel topics and transfer to unseen crisis events. We then outline the process by which our model learns to extract crisis-relevant embeddings from tweets, and clusters them into novel topics from which it then extracts interpretable keywords.

3.1 Snowstorm Dataset

On January 18 2020, Winter Storm Jacob hit Newfoundland, Canada. As a result of the high winds and severe snowfall, 21,000 homes were left without power. A state of emergency was declared in the province as snowdrifts as high as 15 feet (4.6 m) trapped people indoors (Erdman, 2020).

Following the Newfoundland Snowstorm, we collected 21,797 unique tweets from 8,471 users between January 17 and January 22 using the Twitter standard search API with the following search terms: `#nlwhiteout`, `#nlweather`, `#Newfoundland`, `#nlblizzard2020`, `#NLStorm2020`, `#snowmaggedon2020`, `#stormageddon2020`, `#Snowpocalypse2020`, `#Snowmageddon`, `#nlstorm`, `#nltraffic`, `#NLwx`, `#NLblizzard`. Based on past experience with the Twitter API, we opted to use hashtags to limit irrelevant tweets (e.g. searching for `blizzard` resulted in half the collected tweets being about the video game company with the same name). We filter retweets to only capture unique tweets and better work within API rate limits. We make the dataset publicly available with our code.

3.2 Finetuned Tweet Embeddings Model

Our proposed approach, Finetuned Tweet Embeddings (FTE) involves training a model with bidirectional self-attention such as BERT (Devlin et al., 2019) to generate embeddings for tweets so that these can be clustered using common off-the-shelf algorithms such as K-Means (Lloyd, 1982; Elkan, 2003). We then combine activations from the model's attention layers with Term frequency-Inverse document frequency (Tf-Idf) to identify keywords for each cluster and improve model interpretability.

3.2.1 Model Finetuning

To build a model that extracts tweet embeddings containing information relevant to crisis management, we finetune a pretrained BERT language representation model on classifying tweets from various crisis events. Similar to Romascanu et al. (2020), we finetune on CrisisNLP, a dataset of Crowdflower-labeled tweets from various types of crises, aimed at crisis managers (Imran et al., 2016). These show a significant class imbalance with large amounts of tweets shunted into uninformative categories such as `Other useful information`, further motivating the need for unsupervised topic discovery. CrisisNLP label descriptions and counts are included in Appendix A for context. To address the issue of class imbalance, we create a random stratified train-validation split of 0.8 across the datasets, preserving the same proportions of labels.

To finetune BERT, we add a dropout layer and a linear classification layer on top of the `bert-base-uncased` model, using the 768-dimensional `[CLS]` last hidden state as input to our classifier. We train the model using the Adam optimizer with the default fixed weight decay, and a batch size of four over a single epoch. Our model obtains an accuracy of 0.78 on the withheld validation dataset. One advantage of BERT is its subword tokenization which can dynamically build representations of out-of-vocabulary words from subwords, allowing a robust handling of the non-standard language found in tweets.

3.2.2 Tweet Embedding

Once the model is trained, we use a mean-pooling layer across the last hidden states to generate a tweet embedding, similar to Reimers and Gurevych (2019) who found mean-pooling to work best for semantic similarity tasks. Whereas the hidden state for the `[CLS]` token contains sufficient information to separate tweets between the different CrisisNLP labels, the hidden states for the other tokens in the tweet allow our model to capture token-level information that can help in identifying novel topics beyond the supervised labels. For example, tweets that use similar words — even those not occurring in the CrisisNLP dataset — will have more similar embeddings and thus be more likely to cluster in the same topic.

3.2.3 Clustering

Given the embeddings for each tweet, we apply an optimized version of the K-Means clustering algorithm to find our candidate topics (Elkan, 2003). We use the K-Means implementation in `Sklearn` with the default 'k-means++' initialization and n_int equal to 10.

3.2.4 Keyword Extraction

To extract keywords from topics generated by our model and ensure their interpretability, we experiment with two approaches and their combination.

We first identify relevant keywords for each cluster using Tf-Idf (Spärck Jones, 2004), combining each cluster into one document to address the issue of low term frequencies in short-text tweets. During automatic evaluation, we perform a comprehensive grid search over Tf-Idf and other hyperparameters:

1. maximum document frequency (mdf) between 0.6 and 1.0 with intervals of 0.1 (to ignore snowstorm related terms common to many clusters);
2. sublinear Tf-Idf (shown to be advantageous by Paltoglou and Thelwall (2010));
3. phrasing (grouping of frequently co-occurring words proposed by Mikolov et al. (2013b))

We find an mdf of 0.6 and sublinear Tf-Idf perform best for FTE with our number of topics, and we use these hyperparameters in our experiments. Phrasing makes no significant difference and we only include unigrams in our keywords.

However, Tf-Idf only uses frequency and does not leverage the crisis-related knowledge learned during finetuning. Based on the observation by Clark et al. (2019), we use BERT's last layer of attention for the `[CLS]` token to identify keywords that are important for classifying tweets along crisis management related labels. For each cluster, we score keywords by summing their attention values (averaged across subwords) across tweets where they occur, better capturing the relevance of a keyword to crisis management.

We also experiment with the combination of Tf-Idf and attention by multiplying the two scores for each token, allowing us to down-weight frequent but irrelevant words and up-weight rarer but relevant words. For all three approaches, we drop stopwords, hashtags, special characters, and URLs based on preliminary experiments that found these contributing substantially to noise in the topic keywords.

4 Evaluation

In this section we describe the baselines, as well as the automatic and human evaluations used.

4.1 Baselines

We report on two standard topic modelling techniques: BTM and LDA, for which we train models ranging from five to fifteen topics under 10 passes and 100 iterations, following the work of Blei et al. (2003). For LDA, we generate clusters of tweets by giving a weighted topic assignment to each word present in a given document according to the topic distribution over all words present in the corpus. Clusters can be generated similarly with BTM, but according to the topic distribution over biterms (with a window size of 15).

We also report on BERT, which is simply our method without the finetuning step, i.e. using vanilla pretrained BERT embeddings with K-Means clustering and keyword extraction based on Tf-Idf and attention. For further comparison with the FTE model, we focus our analysis on trained baselines with nine topics, the number of labels in the CrisisNLP dataset (Imran et al., 2016).

4.2 Automatic Evaluation

To evaluate topics we calculate C_{NPMI} and C_V, two topic coherence metrics based on direct and indirect confirmation respectively (Röder et al., 2015). These are described in Appendix B. For subsequent human evaluation, we select the configuration (§3.2.4) of each model with the highest C_V for nine topics. This allows us to directly compare with the nine labels from the CrisisNLP dataset and assess our model's ability to learn novel topics. We focus on C_V as Röder et al. (2015) found it to correlate best with human judgements (in contrast to $UMASS$, another commonly used coherence metric that was not included in our analysis due to poor correlation with human judgements).

4.3 Human Evaluation

We performed anonymous evaluation through four annotators[1]. Since these models are primarily for use by crisis managers, we aim to concentrate our evaluation from the perspective of annotators working in the field. Specifically, we propose two evaluation methods focused on (1) topic keywords and (2) document clustering within topics.

[1] Student researchers familiar with the crisis management literature and the needs of crisis managers as described in §1.1

4.3.1 Keyword Evaluation

To assess the quality of topic keywords, annotators were presented with the top 10 keywords for each topic (Table 1) and asked to assign an interpretability score and a usefulness score on a three-point scale. Following the criteria of Rosner et al. (2013), we define interpretability as **good** (eight to ten words are related to each other), **neutral** (four to seven words are related), or **bad** (at most three words are related). Usefulness in turn considers the ease of assigning a short label to describe a topic based on its keywords, similar to Newman et al. (2010) except we further require the label should be useful for crisis managers. We score usefulness on a three-point scale: **useful**, **average**, or **useless**.

4.3.2 Cluster Evaluation

The second task assesses — from the perspective of a crisis manager — the interpretability and usefulness of the actual documents clustered within a topic, instead of only analyzing topic keywords as done in previous work.

Given an anonymized model, for each topic we sample 10 sets of four documents within its cluster along with one document — the 'intruder' — outside of that topic. For each set of documents, all four annotators were tasked with identifying the intruder from the sample of five documents, as well as assigning an interpretability score and a usefulness score to each sample.

The task of intrusion detection is a variation of Chang et al. (2009). However, instead of intruder topics or topic words, we found that assessing intruder tweets would give us a better sense of the differences in the clusters produced by our models. Participants were also given the option of labeling the intruder as 'unsure' to discourage guessing.

The interpretability score was graded on a three-point scale: **good** (3-4 tweets seem to be part of a coherent topic beyond "snowstorm"), **neutral**, and **bad** (no tweets seem to be part of a coherent topic beyond "snowstorm"). The cluster usefulness score was similar to the keyword usefulness score, but formulated as a less ambiguous binary assignment of **useful** or **useless** for crisis managers wanting to filter information during a crisis.

4.4 Model agreement

While our human cluster evaluation provides a good estimate of how topic clusters appear to humans, it does not necessarily establish the difference between two models' document clusters due to the random sampling involved. In other words, two models may have different cluster evaluation results, but similar topic clusters. We define 'agreement' as a measure of the overlap between two unsupervised classification models.

Given a model A, its agreement Agr_A with a model B is

$$\mathrm{Agr}_A(B) = \frac{\sum_{i=0}^{N} \max_j \mathrm{p}(A_i, B_j)}{N} \quad (1)$$

where A_i is the set of documents in the i^{th} cluster of A and $\mathrm{p}(A_i, B_j)$ is the proportion of documents in A_i that are also in B_j. We further define model agreement between A and B as the average of $\mathrm{Agr}_A(B)$ and $\mathrm{Agr}_B(A)$.

5 Results

5.1 Automatic Evaluation

Figure 1 shows that combining attention and Tf-Idf produces the highest automated C_V coherence scores for our method, across a range of topic numbers. However, the improvement of adding attention is marginal and attention alone performed much worse, suggesting it is suboptimal for identifying keywords. Recent work by Kobayashi et al. (2020) proposes a norm-based analysis which may improve upon this.

Figure 2 shows that FTE significantly outperforms the LDA and BTM baselines, with similar scores to the BERT baseline. We observed similar trends for C_{NPMI}. However, despite BERT's high C_V scores, we found that the topics it generated were of very low quality, as described below.

5.2 Qualitative Analysis of Keywords

Topic keywords are shown in Table 1. By restricting the number of topics to the number of labels in the CrisisNLP dataset, we were able to ask our annotators to identify overlap with these original labels. Conversely, this allows us to show that our approach indeed bridges the gap between supervised classification and unsupervised topic modelling by identifying novel topics in addition to salient topics from supervised training.

5.2.1 FTE Keywords

Annotators identified overlap between generated topics and relevant classes from the CrisisNLP dataset: Topic 4 capturing donation needs and volunteering services, Topic 5 expressing sympathy

Model	Topic 1	2	3	4	5	6	7	8	9
FTE	reporting	ivyparkxadidas	outage	assistance	prayer	blowingsnow	trapped	monster	bread
	monster	mood	campus	assist	praying	alert	stranded	meteorologist	song
	snowiest	song	widening	troop	pray	advisory	hydrant	drifting	coffee
	recorded	blackswan	advisory	volunteer	wish	caution	ambulance	perspective	milk
	peak	le	reported	providing	wishing	advised	dead	stormofthecentury	feelin
	temperature	snowdoor	impassable	relief	humanity	stormsurge	garbage	mood	pin
	cloudy	perspective	remaining	aid	brave	wreckhouse	rescue	snowdrift	enjoying
	reported	ode	thousand	request	surviving	surge	permitted	climate	laugh
	equivalent	music	reporting	offering	loved	drifting	body	windy	favorite
	meteorologist	adidasxivypark	suspended	rescue	kindness	avoid	helped	snowdoor	girl
BTM	emergency	eminem	safe	cbcnl	people	like	nltraffic	closed	cm
	state	photo	stay	today	today	storm	road	st	st
	st	click	blizzard	thank	storm	snow	power	tomorrow	winds
	city	learn	newfoundland	people	need	time	st	john	pm
	cityofstjohns	saveng	storm	day	like	day	street	remain	today
	says	michelleobama	canada	home	day	newfoundland	drive	january	km
	declared	sexeducation	weather	work	grocery	going	pearl	update	airport
	john	ken	nlstorm	help	food	blizzard	roads	emergency	yyt
	mayor	starr	warm	storm	open	house	line	state	snowfall
	roads	pin	snowstorm	nltraffic	know	today	mount	today	blizzard
BERT	shareyourweather	pin	thankful	lay	metro	mood	glad	thankful	monster
	ivyparkxadidas	feelin	bank	justintrudeau	campus	hutton	looked	glad	stormsurge
	saturdaythoughts	taxi	yo	save	provincial	cute	apartment	sharing	le
	saturdaymotivation	mayor	mayor	suck	mayor	compound	cloudy	favourite	wreckhouse
	snowdoor	metro	neighbor	radiogregsmith	remain	lovely	law	shareyourweather	newfoundlandlabrador
	bcstorm	en	walked	kettle	operation	spread	snowblowing	grateful	ode
	titanscollections	blowingsnow	wa	anthonygermain	pharmacy	stream	honestly	monster	snowiest
	badboysforlife	ivyparkxadidas	bus	kilbride	region	design	as	feelin	ottawa
	bingo	bus	glad	kaylahounsell	taxi	ivyparkxadidas	eat	neighbor	historic
	blackswan	dannybreennl	pharmacy	campus	advisory	crisis	weird	mom	explorenl

Table 1: Topic keywords for our FTE model and the BTM and BERT baselines used in human evaluation.

and emotional support, Topic 7 covering missing or trapped people, and Topic 2 seemingly covering unrelated information. Distinct novel topics were also identified in the meteorological information in Topic 1, and information about power outages and closures in Topic 3. Topics 8 and 9 were less clear to annotators, but the former seemed to carry information about how extreme the storm was thought to be and the latter about citizens bundling up indoors with different foods and activities.

5.2.2 BTM Keywords

The topics in BTM were less semantically meaningful to annotators, although they found interesting topics there as well, with Topic 5 showing information about the need to stockpile provisions, Topic 7 relating to traffic conditions, and Topic 8 potentially providing information about a state of emergency and closed businesses.

5.2.3 BERT Keywords

The topics in BERT were largely incoherent for annotators, with the exceptions being Topic 9 (positive sentiment) and Topic 5 (services and advisories). This is in stark contrast to the large automated coherence scores obtained by this method, indicating the importance of pairing automatic evaluation with human evaluation.

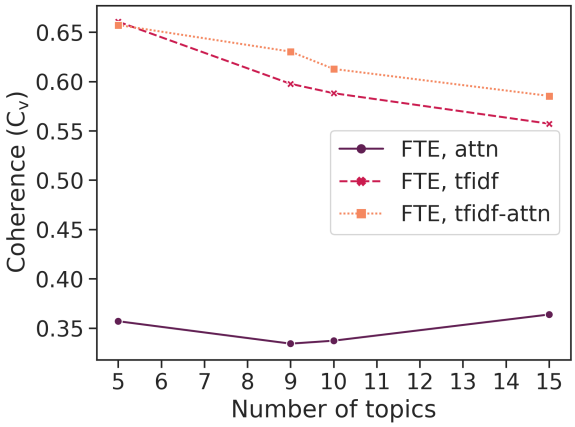


Figure 1: Effect of keyword extraction strategies.

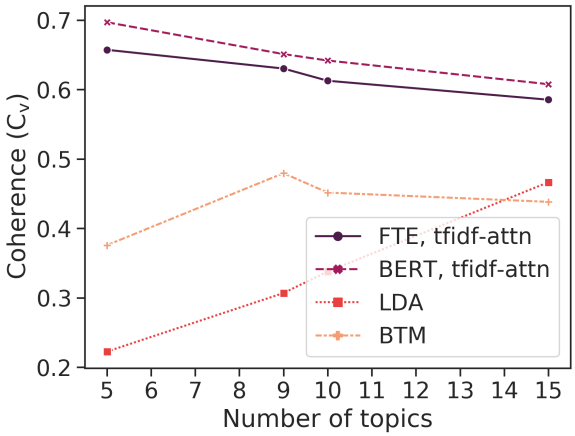
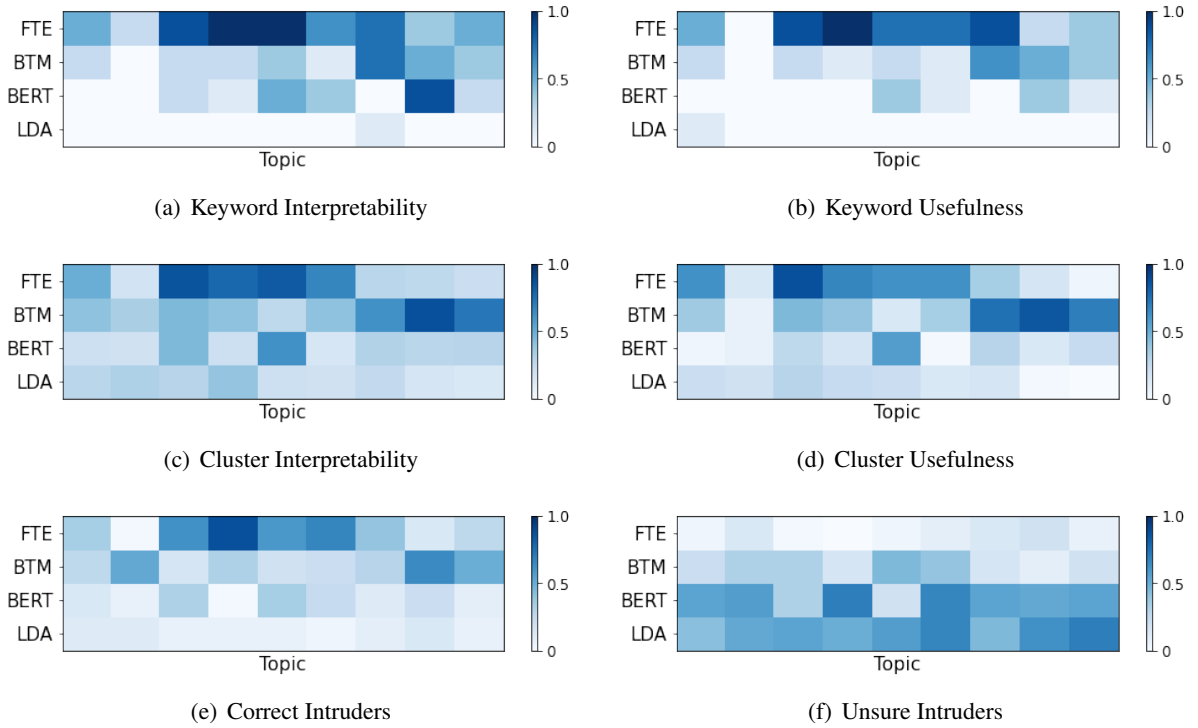


Figure 2: Comparison of FTE with baselines.



(a) Keyword Interpretability

(b) Keyword Usefulness

(c) Cluster Interpretability

(d) Cluster Usefulness

(e) Correct Intruders

(f) Unsure Intruders

Figure 3: Topic-level results for keyword and cluster evaluations, aligned with topics from Table 1. All scores are rescaled to values between 0 and 1, then averaged across annotators and samples.

	Average Score		Topic Count		Fleiss' κ	
Score	BTM	FTE	BTM	FTE	BTM	FTE
Interpretability	31.94	**65.28**	1	**5**	15.01	**17.97**
Usefulness	27.78	**59.72**	1	**5**	12.36	**21.55**

Table 2: Keyword Evaluation scores averaged across topics, number of topics with average scores greater than 0.5, and inter-rater agreements (Fleiss' κ).

	Average Score		Topic Count		Fleiss' κ	
Score	BTM	FTE	BTM	FTE	BTM	FTE
Interpretability	50.28	**51.53**	3	**4**	11.05	**23.45**
Usefulness	45.46	**46.11**	3	**5**	**21.82**	21.60
Correct Intruders	35.28	**44.17**	2	**4**	25.78	**31.50**
Unknown Intruders	26.39	**8.89**	0	0	-	-

Table 3: Cluster Evaluation scores averaged across topics, number of topics with average scores greater than 0.5, and inter-rater agreements (Fleiss' κ).

5.3 Human evaluation

In Figure 3, we compare topic-level results for keyword evaluations (averaged across annotators), and for cluster evaluations (averaged across samples and annotators). We summarize these results for BTM and FTE in Table 2 and Table 3, leaving out LDA and BERT due to significantly lower scores.

We find that our method outperforms the various baselines on both the keyword and cluster evaluations. In particular, the improvement over BERT further confirms the importance of the finetuning step in our method. In contrast, the improvements in average scores over BTM reported in Table 3 are marginal. Nevertheless, we find that the number of interpretable and useful topic clusters was greater for our approach. Indeed, while the BTM baseline had more semi-interpretable (i.e. only a subset of the sampled tweets seemed related) but non-useful topics, our method had a much clearer distinction between interpretable/useful and non-interpretable/non-useful topics, suggesting that tweets marked as hard to interpret and not useful are consistently irrelevant. This may be preferable for downstream applications, as it allows users to better filter our irrelevant content.

The annotators also identified intruder tweets in topic samples from FTE more reliably and with less uncertainty, as measured by the number of correct intruders predicted and the number of times an intruder could not be predicted. Interestingly, BTM topics rated for high interpretability had lower rates of correct intruder detection, suggesting that these topics may seem misleadingly coherent to annotators. Inter-rater agreements as measured by Fleiss' κ further confirm that annotators more often dis-

agreed on intruder prediction and interpretability scoring for BTM topics. This is undesirable for downstream applications, where poor interpretability of topics can lead to a misinterpretation of data with real negative consequences.

5.4 Model agreement

Agreement for the models was 32.7% between FTE and BERT, 26.5% between FTE and BTM, 19.9% between FTE and LDA and 20.7% between BTM and LDA. This confirms that the different models also generate different clusters.

6 Conclusion

This paper introduces a novel approach for extracting useful information from social media and assisting crisis managers. We propose a simple method that bridges the gap between supervised classification and unsupervised topic modelling to address the issue of domain adaptation to novel crises.

Our model (FTE, Finetuned Tweet Embeddings) incorporates crisis-related knowledge from supervised finetuning while also being able to generate salient topics for novel crises. To evaluate domain adaptation, we create a dataset of tweets from a crisis that significantly differs from existing crisis Twitter datasets: the 2020 Winter Storm Jacob.

Our paper also introduces human evaluation methods better aligned with downstream use cases of topic modelling in crisis management, emphasizing human-centered machine learning. In both these human evaluations and traditional automatic evaluations, our method outperforms existing topic modelling methods, consistently producing more coherent, interpretable and useful topics for crisis managers. Interestingly, our annotators reported that several coherent topics seemed to be composed of related subtopics. In future work, the number of topics as a hyper-parameter could be explored to see if our approach captures these salient subtopics.

Our method, while simple, is not specific to crisis management and can be more generally used for textual domain adaptation problems where the latent classes are unknown but likely to overlap with known classes from other domains.

Acknowledgements

We are grateful for the funding from Environment and Climate Change Canada (ECCC GCXE19M010). Mikael Brunila also thanks the Kone Foundation for their support.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

David E. Alexander. 2014. Social Media in Disaster Risk Reduction and Crisis Management. *Science and Engineering Ethics*, 20(3):717–733.

Reem Alrashdi and Simon O'Keefe. 2020. Automatic Labeling of Tweets for Crisis Response Using Distant Supervision. In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 418–425, New York, NY, USA. Association for Computing Machinery.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. *arXiv:2004.03974 [cs]*. ArXiv: 2004.03974.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Ivan S. Blekanov, Svetlana S. Bodrunova, Nina Zhuravleva, Anna Smoliarova, and Nikita Tarasov. 2020. The Ideal Topic: Interdependence of Topic Interpretability and Other Quality Features in Topic Modelling for Short Texts. In *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis*, Lecture Notes in Computer Science, pages 19–26, Cham. Springer International Publishing.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.

Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. In *International Conference on Machine Learning*, pages 612–620. PMLR. ISSN: 1938-7228.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Charles Elkan. 2003. Using the triangle inequality to accelerate k-means. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 147–153, Washington, DC, USA. AAAI Press.

Dontas Emmanouil and Doukas Nikolaos. 2015. Big data analytics in prevention, preparedness, response and recovery in crisis and disaster management. In *The 18th International Conference on Circuits, Systems, Communications and Computers (CSCC 2015), Recent Advances in Computer Engineering Series*, volume 32, pages 476–482.

Jonathan Erdman. 2020. Crippling Newfoundland, Canada, Blizzard From Bomb Cyclone Smashes All-Time Daily Snow Record. *The Weather Channel*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644. ISBN: 9781720347118 Publisher: National Academy of Sciences Section: PNAS Plus.

Sedigheh Khademi Habibabadi and Pari Delir Haghighi. 2019. Topic Modelling for Identification of Vaccine Reactions in Twitter. In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW 2019, pages 1–10, New York, NY, USA. Association for Computing Machinery.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA. Association for Computing Machinery.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Díaz, and Patrick Meier. 2013. Extracting information nuggets from disaster- Related messages in social media. In *ISCRAM 2013 Conference Proceedings – 10th International Conference on Information Systems for Crisis Response and Management*, pages 791–801, KIT; Baden-Baden. Karlsruher Institut fur Technologie. ISSN: 2411-3387 Journal Abbreviation: ISCRAM 2013.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1638–1643, Portorož, Slovenia. European Language Resources Association (ELRA).

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Yan Jin, Brooke Fisher Liu, and Lucinda L. Austin. 2014. Examining the Role of Social Media in Effective Crisis Management: The Effects of Crisis Origin, Information Form, and Source on Publics' Crisis Responses. *Communication Research*, 41(1):74–94. ZSCC: 0000391 Publisher: SAGE Publications Inc.

Mark E. Keim. 2008. Building Human Resilience: The Role of Public Health Preparedness and Response As an Adaptation to Climate Change. *American Journal of Preventive Medicine*, 35(5):508–516.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*, 84(5):905–949.

Jey Han Lau and Timothy Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2177–2185, Cambridge, MA, USA. MIT Press.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-5973.12194.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

S. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137. Conference Name: IEEE Transactions on Information Theory.

Guoqin Ma. 2019. Tweets Classification with BERT in the Field of Disaster Management. page 15.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, USA. Association for Computational Linguistics.

Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017, pages 632–635. AAAI press.

Georgios Paltoglou and Mike Thelwall. 2010. A Study of Information Retrieval Weighting Schemes for Sentiment Analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395, Uppsala, Sweden. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

J. Rexiline Ragini and P. M. Rubesh Anand. 2016. An empirical analysis and classification of crisis related tweets. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–4. ISSN: 2473-943X.

Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. 2019. Emerging Perspectives in Human-Centered Machine Learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

A. Romascanu, H. Ker, R. Sieber, R. Zhao, M. Brunila, S. Greenidge, S. Lumley, D. Bush, and S. Morgan. 2020. Using deep learning and social network analysis to understand and manage extreme flooding. *Journal of Contingencies and Crisis Management*.

Frank Rosner, Alexander Hinneburg, Michael Röder, Martin Nettling, and Andreas Both. 2013. Evaluating topic coherence measures. In *Topic Models: Computation, Application, and Evaluation*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, Shanghai, China. Association for Computing Machinery.

Axel Schulz, Eneldo Loza Mencía, Thanh-Tung Dang, and Benedikt Schmidt. 2014. Evaluating multi-label classification of incident-related tweet. In *# MSM*, pages 26–33. Citeseer.

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Oleksandra Sopova. 2017. Domain adaptation for classifying disaster-related Twitter data. Report, Kansas State University.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60(5):493–502. Publisher: Emerald Group Publishing Limited.

Ed Tobias. 2011. Using Twitter and other social media platforms to provide situational awareness during an incident. *Journal of Business Continuity & Emergency Planning*, 5(3):208–223. Publisher: Henry Stewart Publications LLP.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Günter Wallner, Simone Kriglstein, and Anders Drachen. 2019. Tweeting your Destiny: Profiling Users in the Twitter Landscape around an Online Game. In *2019 IEEE Conference on Games (CoG)*, pages 1–8. ISSN: 2325-4289.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 1445–1456, Rio de Janeiro, Brazil. Association for Computing Machinery.

Hamada M Zahera, Rricha Jalota, Ibrahim Elgendy, and Mohamed Ahmed Sherif. 2019. Fine-tuned BERT Model for Multi-Label Tweets Classification. page 7.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018. Topic Memory Networks for Short Text Classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3120–3131, Brussels, Belgium. Association for Computational Linguistics.

A CrisisNLP Dataset

Here we include class distributions (Table 4) and descriptions (Table 5) from the CrisisNLP Dataset.

	Label Id								
Crisis Dataset	1	2	3	4	5	6	7	8	9
2013_pak_eq	351	5	16	29	325	75	112	764	336
2014_cali_eq	217	6	4	351	83	84	83	1028	157
2014_chile_eq	119	6	63	26	10	250	541	634	364
2014_odile	50	39	153	848	248	77	166	380	52
2014_india_floods	959	14	27	67	48	44	30	312	502
2014_pak_floods	259	117	106	94	529	56	127	698	27
2014_hagupit	66	8	130	92	113	349	290	732	233
2015_pam	143	18	49	212	364	93	95	542	497
2015_nepal_eq	346	189	85	132	890	35	525	639	177
Total	2510	402	633	1851	2610	1063	1969	5729	2345

Table 4: Label counts for the different datasets labeled by Crowdflower workers in (Imran et al., 2016)

B Automated Coherence Metrics

The direct confirmation C_{NPMI}, uses a token-by-token ten word sliding window, where each step determines a new virtual document. Co-occurrence in these documents is used to compute the normalized pointwise mutual information (NPMI) between a given topic keyword W' and each member in the conditioning set of other topic keywords W^*, such that:

$$NPMI = (\frac{PMI(W', W^*)}{-log(P(W', W^*) + \epsilon)})^{\gamma}$$

$$PMI = log \frac{P(W', W^*) + \epsilon}{P(W') * P(W^*)}$$

The coherence of a topic is then calculated by taking the arithmetic mean of these confirmation values, with ϵ as a small value for preventing the log of zero.

The indirect confirmation C_V is instead based on comparing the contexts in which W' and W^* appear. W' and W^* are represented as vectors of the size of the total word set W. Each value in these vectors consist of a direct confirmation between the word that the vector represents and the words in W. However, now the context is just the tweet that each word appears in. The indirect confirmation between each word in the topic is the cosine similarity of each pair of context vectors such that

$$\cos(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot u_i}{||\vec{u}||_2 ||\vec{w}||_2}$$

where $\vec{u} = \vec{v}(W')$ and $\vec{w} = \vec{v}(W^*)$. Once again, the arithmetic mean of these similarity values gives the coherence of the topic.

C_V was found by Röder et al. (2015) to be the most interpretable topic coherence metric when compared to human judgement and has later been used extensively on assessing the coherence of short texts like tweets as well (Zeng et al., 2018; Habibabadi and Haghighi, 2019; Wallner et al., 2019). We also use C_{NPMI}, which has been one of the most successful topic coherence measures based on direct confirmation (Röder et al., 2015; Lau et al., 2014). See Röder et al. (2015) for further details on both C_V and C_{NPMI}.

Label	Id	Description
Injured or dead people	1	Reports of casualties and/or injured people due to the crisis
Missing, trapped, or found people	2	Reports and/or questions about missing or found people
Displaced people and evacuations	3	People who have relocated due to the crisis, even for a short time (includes evacuations
Infrastructure and utilities damage	4	Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored
Donation needs or offers or volunteering services	5	Reports of urgent needs or donations of shelter and/or supplies such as food, water, clothing, money, medical supplies or blood; and volunteering services
Caution and advice	6	Reports of warnings issued or lifted, guidance and tips
Sympathy and emotional support	7	Prayers, thoughts, and emotional support
Other useful information	8	Other useful information that helps one understand the situation
Not related or irrelevant	9	Unrelated to the situation or irrelevant

Table 5: Label descriptions and id's in (Imran et al., 2016)

C Human evaluation tweet samples

Here we present one of the set of tweets presented to human annotators for each model and topic. We also show the ground truth intruder tweet which the annotators were asked to predict. Non-ASCII characters were removed here, but included in tweets shown to human annotators.

C.1 FTE

C.1.1 Topic 0

- How bad was the blizzard in St. John's, Newfoundland? Here's what a seniors home looks like the day after. photo: https://t.co/6n8txqnuWl
- East End of St. Johns 4 days post blizzard. #NLwx #NLtraffic #snowmegeddon2020 https://t.co/5s2p9Ivejf
- Well here's the big mother storm en route. Currently the size of Nova Scotia, nbd. #nlwx https://t.co/CFi9szzunK
- INTRUDER: I hope I don't have to go to work on Monday because I don't remember my password anymore. #nlwx #stormageddon2020 https://t.co/BswQppEFVh
- #StJohns declares #StateOfEmergency and #Newfoundland and #Labrador get pounded with #SnowFall with more to come https://t.co/JNxnIF5mNx

C.1.2 Topic 1

- Gentle heart of Jesus #nlwx https://t.co/cEbvv3it5f
- This was the moment I fell in love with #Newfoundland, #Canada when I first entered the Gros Morne National Park. https://t.co/2mSD79u6qC
- #nlwx https://t.co/md4pRSafW5
- @UGEABC families-send a pic of what youre reading!!! @NLESD @PowersGr6 @AndreaCoffin76 @MrBlackmoreGr1 https://t.co/FwEbRe9YJs
- INTRUDER: So much for the snow I was really hoping to make a snowman Did anyone get their snowman built? #wheresthesnow https://t.co/FT3IJqIVsS

C.1.3 Topic 2

- Garbage collection in the City of St. John's is cancelled for the rest of the week. It will resume on Monday, Jan 2 https://t.co/nvWHGHcrwJ
- City of St. Johns Snow Clearing on Standby :Due to deteriorating conditions and reduced visibility snow clearing o https://t.co/UC7GV2wFrw
- Memorial Universitys St. Johns, Marine Institute and Signal Hill campuses will remain closed all day. #nlwx
- Current outages. St. Johns area. https://t.co/wEwJKWKxpF #nlwx https://t.co/FsIieBLzGw
- INTRUDER: $SIC Sokoman Minerals Provides Winter 2020 Exploration Update at Moosehead Gold Project, Central Newfoundland https://t.co/ma9Ai4PLeE

C.1.4 Topic 3

- Not everyone has the funds to go to the grocery store (not just during a SOE) Hats off to the food banks that are o https://t.co/sljOgZZUmf
- Up to 300 troops from across Canada will be asked to work on the response to the unprecedented #nlblizzard2020. Gag https://t.co/Yw7jUAWUJE
- INTRUDER: Beautiful @DowntownStJohns the morning after #Snowmageddon https://t.co/HS7jhLIt4l
- So pleased to see the joint efforts of @GovNL with the cities/towns during #snowmaggedon2020. From calling it a SOE https://t.co/0oKer9CLmo
- Updated story: Five days is a long time without food for people who can't afford to stock up. Staff at one St. John https://t.co/gNJSXMsv3B

C.1.5 Topic 4

- Were quite buried in Paradise right now! Luckily still have power for now. Stay safe everyone! #nlwx https://t.co/sOC7TL4Al7

- Hoping everyone's pets are safe inside your homes. #nlblizzard2020
- @IDontBlog Yikes! Hoping everyone there stays safe.... #nlblizzard2020 #NLStorm2020
- @DanKudla The weather channel is calling for snow in Toronto, but nothing like that. Hope youre all safe in Newfou https://t.co/0KN2AD3JrS
- INTRUDER: BREAKING — Province is calling in the military @NTVNewsNL #Nlwx https://t.co/ggfOlrbYCz

C.1.6 Topic 5

- We've made it through the worst of the storm, but #yyt's State of Emergency remains in effect. Pls stay inside & sa https://t.co/YGVxNOEzf1
- INTRUDER: How COLD is it? Coby took less than 30 seconds to use the facilities this morning! #snowmaggedon2020 #snowstorm https://t.co/f4zZ38MJol
- Stay safe St. Johns! #Newfoundland #snowmaggedon2020
- Take your time cleaning this up, Newfoundland friends! It is a brutal amount of snow! Be safe! #nlwx https://t.co/POwfBYWHFy
- Meanwhile in ON, 15-25cm w/ 50km wind forecast and asked to stay off streets. #nlblizzard2020 https://t.co/wwKVnqmM3F

C.1.7 Topic 6

- @VOCMNEWS This is amazing. In the next couple of days we could really get so many of our neighbourhood hydrants dug https://t.co/7Z2TOfwpwB
- @weathernetwork batteries charging and camera gear drying out after 16 hours shooting in blizzard. @MurphTWN #nlwx https://t.co/yGUMV93Yrw
- So we visited friends last night and left at the height of the storm ... somewhere under there are a couple of cars https://t.co/NYHHxXT4EI
- @KrissyHolmes #nltraffic just like any other weekday morning coming out of Cbs. Two solid lines of traffic
- INTRUDER: @BrianWalshWX so its 7:00 pm , how much more snow potentially will fall before noon tomorrow? #nlblizzard2020 #nlwx

C.1.8 Topic 7

- @StormchaserUKEU A glimpse into the future @yyt #nlwx
- Its official - weve named this storm #BettyWhiteOut2020 in honour of #BettyWhites-Birthday #nlwx
- Even for just a moment with the front door open, snow is hitting you in the face and the wind is taking your breath https://t.co/pJkpwwqDfF
- INTRUDER: #nlwx https://t.co/2mnaF7TkSl
- Open those curtains, let all the sunshine heat in that you can! Solar gain will help us through. #nlwx https://t.co/AFAubBvWwg

C.1.9 Topic 8

- INTRUDER: Plow came by and then the neighbours started a snow clearing party. So thankful #nlblizzard2020 #nlwx https://t.co/QpIapG1N4Y
- Y seguimos con la supernevada (fuera de lo comn, tambin hay que decirlo) de #Newfoundland , en #Canad ! Laia ll https://t.co/O1adKSXbFa
- If you have to wear a full snowsuit to go shopping, STAY HOME. You do NOT need to be out. You do NOT need a fondue https://t.co/qig24xkrK5
- @NewfieScumbag Pretty much... love of God! In case you didnt get the memo, keep your packin vehicle off the roads https://t.co/AJe6RpgWBS
- So our front door looks like a neat little burrow now, at least. #nlwx #Snowmageddon2020 https://t.co/qUCdntRJgE

C.2 BTM

C.2.1 Topic 0

- THIS!! #nlwx #nltraffic https://t.co/fGPl0DDJWF

- 2/2 During a State of Emergency the public is advised to contact their nearest emergency hospital department for em https://t.co/MTbO1MHF2K

- there's a house in there somewhere #snowmaggedon2020 https://t.co/r36lWxJPzd

- INTRUDER: @PaulDoroshenko Yup, sorry as a proud Newfoundlander living on Vancouver Island for the last ten years they totally https://t.co/eNdNvtnJUI

- Move over, CBC announcers! #NLStorm2020 https://t.co/TTvE0LZgIG

C.2.2 Topic 1

- INTRUDER: Snowmageddon 2. The return of the snowstorm! Now playing...#nlwx #nlsnowstorm2020 #snowdoor #blizzard https://t.co/EvBlOdrg23

- If you're NOT feelin this? Then you have NO "PULSE" @therealbigthump #SaturdayMorning #Eminem https://t.co/r1orKgKGar

- Click On Photo To Learn More #michelleobama Eminem #SAvENG Ken Starr #SexEducation #nlwx How Cheap https://t.co/jCYbHGDKM1

- If you're NOT feelin this? Then you have NO "PULSE" @therealbigthump #SaturdayMorning #Eminem #WomensMarch https://t.co/su8qA0qjP0

- If you're NOT feelin this? Then you have NO "PULSE" @therealbigthump #SaturdayMorning #Eminem #WomensMarch https://t.co/pORhRKunhs

C.2.3 Topic 2

- Stay safe, Newfoundland. Stay warm. Stay home. Stay off the roads. #Newfoundland

- Newfoundland peeps - the CBC in Toronto is warning that Toronto is about to be walloped with snow, and that Envir https://t.co/NmWRu8t6rL

- INTRUDER: #NLStorm2020 #nltraffic Thanks for coming and doing some cleanup on Gower St right now. Many fans in windows watch https://t.co/FZBsIaN7v1

- My thoughts go to those who need to serve the communities during #Snowmageddon in NFLD today. Stay safe while you do what you need to do.

- @DrAJHalifax They're getting absolutely walloped! #nlstorm

C.2.4 Topic 3

- Have woken up ... and grateful that the power is still on in the house. I'm also grateful that the noise of the w https://t.co/xQjXoz3qED

- A huge THANK YOU to all of the deputies, officers, troopers, medics, firefighters, dispatchers, plow operators and https://t.co/mCnFbDUHd8

- Watch what happened at 1:56:51 in @DaHonestyPolicy's broadcast: Come Talk With the Ladies #DaAngels #DaNation https://t.co/L4zsDuzTHe

- What do you expect when graduate with good cgpa don't have a job after graduating, Efcc free guys to hustle jor. Dr https://t.co/DYGlpBBmHe

- INTRUDER: Weve made the Washington Post #nlwx https://t.co/sjqdJkR7XL

C.2.5 Topic 4

- INTRUDER: Its a brand new day #nlwx https://t.co/l30WBqLHIE

- Whos up for a twitter game: give us your best movie title related to #snowmaggedon2020 pls use #snowmoviesnl so we can retweet you #nlwx

- somebody should make a video of #snowmaggedon2020 with Greata saying how dare you

- Doing my civic duty as a Minnesotan today....storming @Target and battling the masses for the last cart to stock up https://t.co/D72nJ0I3em

- My 13-yr old got tickets to tomorrow night's @Raptors game for Xmas but will miss b/c all flights are grounded at https://t.co/NsvWNwcIix

C.2.6 Topic 5

- Mom look! One of my videos made the @BBCWorld news! Bys, tis a rough way to get famous! #nlwx #blizzard2020 https://t.co/Y0jx2pycO7
- Very cool! #Newfoundland $gold $SIC.B #mooseheadmadness https://t.co/YKfIBIMCaA
- The morning after. Pictures from my Daughter and my Sister. Both live in St. John's NL #NLWX #Blizzard #NL https://t.co/UhtFlddTXn
- It's #SaturdayMorning and the storm has died down here in Newfoundland. Still blowing hard out there but not as bad https://t.co/Qkoj6XjBc3
- INTRUDER: We are still in a SOE - STAY OFF THE ROADS so all of the emergency & essential workers can do their jobs! Even thou https://t.co/2TSmfKYlnZ

C.2.7 Topic 6

- @590VOCM several reports of very slippery conditions at the intersection of Carrick Drive and Stavanger Drive. Ple https://t.co/pDlCBjvLfi
- Dave says there's a car off the road and now on top of a snowbank on the ORR WB just before Paradise turnoff #nltraffic
- Is @MaryBrowns open in Mount Pearl? #nlwx #companylunch
- Power outage here in Bonavista. #nlwx #nlblizzard2020 #snowmaggedon2020
- INTRUDER: The #nlwx hashtag is making me really appreciate our mildly inconvenient snowstorm here in Wisconsin. 27 inches of https://t.co/dzV0elcy6y

C.2.8 Topic 7

- INTRUDER: @CBCNews #nlwx #snowmaggedon2020 might be a while before things return to anything close to normal. Main road throu https://t.co/1vHX2B9oPo
- MUN's St. John's, Signal Hill campuses and Marine Institute will be closed until Jan. 27 #nlschools #YYTSOE #nlstorm2020 #nlwx
- #SOEYYT remains in place for tomorrow, Jan 23. #nlwx
- Bell island ferry update #nltraffic https://t.co/pTFo4FuxfM
- MUN closed METROBUS OFF THE ROADS SCHOOLS CLOSED https://t.co/G3rvNoXVt5

C.2.9 Topic 8

- With a 9:30pm snow total of 20cm, today is #Gander's snowiest day so far this winter. #NLWx https://t.co/ZYhyV2Xpsm
- #nlblizzard2020 #nlstorm look at the winds... the gusts are Cat 4 hurricane equivalent https://t.co/JjFbiMJAdg
- 10min avg wind speeds of 132.0km/h with max gust of 167.4km/h through 12:10pm at Green Island, Fortune Bay. #nlwx https://t.co/GtisE875av
- INTRUDER: This is Lovely. I have always had a soft spot in my heart for #Newfoundland and the wonderful people there. https://t.co/CqPyzW3vLH
- There has been the equivalent of 32.7 mm of precipitation since Fri 04:30 at "ST JOHNS WEST CLIMATE" #NLStorm