Evaluation of Thematic Coherence in Microblogs

Iman Munire Bilal[1] **Bo Wang**[2,4] **Maria Liakata**[1,3,4]
Rob Procter[1,4] **Adam Tsakalidis**[3,4]
[1]Department of Computer Science, University of Warwick
[2]Department of Psychiatry, University of Oxford
[3]School of Electronic Engineering and Computer Science, Queen Mary University of London
[4]The Alan Turing Institute, London, UK
{iman.bilal|rob.procter}@warwick.ac.uk
{mliakata|bwang|atsakalidis}@turing.ac.uk

Abstract

Collecting together microblogs representing opinions about the same topics within the same timeframe is useful to a number of different tasks and practitioners. A major question is how to evaluate the quality of such thematic clusters. Here we create a corpus of microblog clusters from three different domains and time windows and define the task of evaluating thematic coherence. We provide annotation guidelines and human annotations of thematic coherence by journalist experts. We subsequently investigate the efficacy of different automated evaluation metrics for the task. We consider a range of metrics including surface level metrics, ones for topic model coherence and text generation metrics (TGMs). While surface level metrics perform well, outperforming topic coherence metrics, they are not as consistent as TGMs. TGMs are more reliable than all other metrics considered for capturing thematic coherence in microblog clusters due to being less sensitive to the effect of time windows.

1 Introduction

As social media gains popularity for news tracking, unfolding stories are accompanied by a vast spectrum of reactions from users of social media platforms. Topic modelling and clustering methods have emerged as potential solutions to challenges of filtering and making sense of large volumes of microblog posts (Rosa et al., 2011; Aiello et al., 2013; Resnik et al., 2015; Surian et al., 2016). Providing a way to access easily a wide range of reactions around a topic or event has the potential to help those, such as journalists (Tolmie et al., 2017), police (Procter et al., 2013), health (Furini and Menegoni, 2018) and public safety professionals (Procter et al., 2020), who increasingly rely on social media to detect and monitor progress of events, public opinion and spread of misinformation.

Recent work on grouping together views about tweets expressing opinions about the same entities has obtained clusters of tweets by leveraging two topic models in a hierarchical approach (Wang et al., 2017b). The theme of such clusters can either be represented by their top-N highest-probability words or measured by the semantic similarity among the tweets. One of the questions regarding thematic clusters is how well the posts grouped together relate to each other (*thematic coherence*) and how useful such clusters can be. For example, the clusters can be used to discover topics that have low coverage in traditional news media (Zhao et al., 2011). Wang et al. (2017a) employ the centroids of Twitter clusters as the basis for topic specific temporal summaries.

The aim of our work is to *identify reliable metrics for measuring thematic coherence in clusters of microblog posts*. We define thematic coherence in microblogs as follows: Given clusters of posts that represent a subject or event within a broad topic, with enough diversity in the posts to showcase different stances and user opinions related to the subject matter, thematic coherence is the extent to which posts belong together, allowing domain experts to easily extract and summarise stories underpinning the posts.

To measure thematic coherence of clusters we require robust domain-independent evaluation metrics that correlate highly with human judgement for coherence. A similar requirement is posed by the need to evaluate coherence in topic models. Röder et al. (2015) provide a framework for an extensive set of coherence measures all restricted to word-level analysis. Bianchi et al. (2020) show that adding contextual information to neural topic models improves topic coherence. However, the most commonly used word-level evaluation of topic coherence still ignores the local context of each word. Ultimately, the metrics need to achieve an opti-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6800–6814
August 1–6, 2021.

mal balance between coherence and diversity, such that resulting topics describe a logical exposition of views and beliefs with a low level of duplication. Here we evaluate thematic coherence in microblogs on the basis of topic coherence metrics, while also using research in text generation evaluation to assess semantic similarity and thematic relatedness. We consider a range of state-of-the-art text generation metrics (TGMs), such as *BERTScore* (Zhang et al., 2019), *MoverScore* (Zhao et al., 2019) and *BLEURT* (Sellam et al., 2020), which we repurpose for evaluating thematic coherence in microblogs and correlate them with assessments of coherence by journalist experts. The main contributions of this paper are:

- We define the task of assessing thematic coherence in microblogs and use it as the basis for creating microblog clusters (Sec. 3).
- We provide guidelines for the annotation of thematic coherence in microblog clusters and construct a dataset of clusters annotated for thematic coherence spanning two different domains (political tweets and COVID-19 related tweets). The dataset is annotated by journalist experts and is available [1] to the research community (Sec. 3.5).
- We compare and contrast state-of-the-art TGMs against standard topic coherence evaluation metrics for thematic coherence evaluation and show that the former are more reliable in distinguishing between thematically coherent and incoherent clusters (Secs 4, 5).

2 Related Work

Measures of topic model coherence: The most common approach to evaluating topic model coherence is to identify the latent connection between topic words representing the topic. Once a function between two words is established, **topic coherence** can be defined as the (average) sum of the function values over all word pairs in the set of most probable words. Newman et al. (2010) use Pointwise Mutual Information (PMI) as the function of choice, employing co-occurrence statistics derived from external corpora. Mimno et al. (2011) subsequently showed that a modified version of PMI correlates better with expert annotators. AlSumait et al. (2009) identified junk topics by measuring the distance between topic distribution and corpus-wide distribution of words. Fang et al. (2016a) model topic coherence by setting the distance between two topic words to be the cosine similarity of their respective embedded vectors. Due to its generalisability potential we follow this latter approach to topic coherence to measure thematic coherence in tweet clusters. We consider *GloVe* (Pennington et al., 2014) and *BERTweet* (Nguyen et al., 2020) embeddings, derived from language models pretrained on large external *Twitter* corpora. To improve performance and reduce sensitivity to noise, we followed the work of Lau and Baldwin (2016), who consider the mean topic coherence over several topic cardinalities $|W| \in \{5, 10, 15, 20\}$.

Another approach to topic coherence involves detecting intruder words given a set of topic words, an intruder and a document. If the intruder is identified correctly then the topic is considered coherent. Researchers have explored varying the number of 'intruders' (Morstatter and Liu, 2018) and automating the task of intruder detection (Lau et al., 2014). There is also work on topic diversity (Nan et al., 2019). However, there is a tradeoff between diversity and coherence (Wu et al., 2020), meaning high diversity for topic modelling is likely to be in conflict with thematic coherence, the main focus of the paper. Moreover, we are ensuring semantic diversity of microblog clusters through our sampling strategy (See Sec. 3.4).

Text Generation Metrics: **TGMs** have been of great use in applications such as machine translation (Zhao et al., 2019; Zhang et al., 2019; Guo and Hu, 2019; Sellam et al., 2020), text summarisation (Zhao et al., 2019) and image captioning (Vedantam et al., 2015; Zhang et al., 2019; Zhao et al., 2019), where a machine generated response is evaluated against ground truth data constructed by human experts. Recent advances in contextual language modeling outperform traditionally used *BLEU* (Papineni et al., 2002) and *ROUGE* (Lin, 2004) scores, which rely on surface-level n-gram overlap between the candidate and the reference.

In our work, we hypothesise that metrics based on contextual embeddings can be used as a proxy for microblog cluster thematic coherence. Specifically, we consider the following TGMs:
(a) BERTScore is an automatic evaluation metric based on BERT embeddings (Zhang et al., 2019). The metric is tested for robustness on adversarial paraphrase classification. However, it is based on a greedy approach, where every reference token is

[1] https://doi.org/10.6084/m9.figshare.14703471

linked to the most similar candidate token, leading to a time-performance trade-off. The harmonic mean F_{BERT} is chosen for our task due to its most consistent performance (Zhang et al., 2019).

(b) MoverScore (Zhao et al., 2019) expands from the *BERTScore* and generalises *Word Mover Distance* (Kusner et al., 2015) by allowing soft (many-to-one) alignments. The task of measuring semantic similarity is tackled as an optimisation problem with the constraints given by n-gram weights computed in the corpus. In this paper, we adopt this metric for unigrams and bigrams as the preferred embedding granularity.

(c) BLEURT (Sellam et al., 2020) is a state-of-the-art evaluation metric also stemming from the success of BERT embeddings, carefully curated to compensate for problematic training data. Its authors devised a novel pre-training scheme leveraging vast amounts of synthetic data generated through BERT mask-filling, back-translation and word dropping. This allows *BLEURT* to perform robustly in cases of scarce and imbalanced data.

3 Methodology

Notation We use $C = \{C_1, ..., C_n\}$ to denote a set of clusters C_i. Each cluster C_i is represented by the pair $C_i = (T_i, W_i)$, where T_i and W_i represent the set of tweets and top-20 topic words of the dominant latent topic in C_i, respectively.

The task of identifying thematic coherence in microblog clusters is formalised as follows: Given a set of clusters C, we seek to identify a metric function $f : C \rightarrow R$ s.t. high values of $f(C_i)$ correlate with human judgements for thematic coherence. Here we present *(a)* the creation of a corpus of topic clusters of tweets C and *(b)* the annotation process for thematic coherence. *(a)* involves a clustering (Sec. 3.2), a filtering (Sec. 3.3) and a sampling step (Sec. 3.4); *(b)* is described in (Sec. 3.5). Experiments to identify a suitable function f are in Sec. 4.

3.1 Data Sources

We used three datasets pertaining to distinct domains and collected over different time periods as the source of our tweet clusters.

The **COVID-19** dataset (Chen et al., 2020) was collected by tracking COVID-19 related keywords (e.g., *coronavirus*, *pandemic*, *stayathome*) and accounts (e.g., *@CDCemergency*, *@HHSGov*, *@DrTedros*) through the Twitter API from January to May 2020. This dataset covers specific recent events that have generated significant interest and its entries reflect on-going issues and strong public sentiment regarding the current pandemic.

The **Election** dataset was collected via the Twitter Firehose and originally consisted of all geo-located UK tweets posted between May 2014 and May 2016[2]. It was then filtered using a list of 438 election-related keywords relevant to 9 popular election issues[3] and a list of 71 political party aliases curated by a team of journalists (Wang et al., 2017c).

The **PHEME** dataset (Zubiaga et al., 2016) of rumours and non-rumours contains tweet conversation threads consisting of a source tweet and associated replies, covering breaking news pertaining to 9 events (i.e., Charlie Hebdo shooting, Germanwings airplane crash, Ferguson unrest, etc.).

These datasets were selected because they cover a wide range of topics garnering diverse sentiments and opinions in the Twitter sphere, capturing newsworthy stories and emerging phenomena of interest to journalists and social scientists. Of particular interest was the availability of stories, comprising groups of tweets, in the PHEME dataset, which is why we consider PHEME tweet clusters separately.

3.2 Tweet Cluster Generation

The task of thematic coherence evaluation introduced in this paper is related to topic modelling evaluation, where it is common practice (Mimno et al. (2011), Newman et al. (2010)) to gauge the coherence level of automatically created groups of topical words. In a similar vein, we evaluate thematic coherence in tweet clusters obtained automatically for the **Election** and **COVID-19** datasets. The clusters were created in the following way: Tweets mentioning the same keyword posted within the same time window (3 hours for *Election*, 1 hour for *Covid-19*) were clustered according to the two-stage clustering approach by Wang et al. (2017b), where two topic models (Yin and Wang, 2014; Nguyen et al., 2015) with a tweet pooling step are used. We chose this as it has shown competitive performance over several tweet clustering tasks, without requiring a pre-defined number of clusters.

[2]Unlike the Twitter API, the firehose provides 100% of the tweets that match user defined criteria, which in our case is a set of geo-location and time zone Twitter PowerTrack operators.

[3]EU and immigration, economy, NHS, education, crime, housing, defense, public spending, environment and energy.

The **PHEME** dataset is structured into conversation threads, where each source tweet is assigned a story label. We assume that each story and the corresponding source tweets form a coherent thematic cluster since they have been manually annotated by journalists. Thus the PHEME stories can be used as a gold standard for thematically coherent clusters. We also created artificial thematically incoherent clusters from PHEME. For this purpose we mixed several stories in different proportions. We designed artificial clusters to cover all types of thematic incoherence, namely: *Random*, *Intruded*, *Chained* (See Sec. 3.5 for definitions). For *Intruded*, we diluted stories by eliminating a small proportion of their original tweets and introducing a minority of foreign content from other events. For *Chained*, we randomly chose the number of subjects (varying from 2 to 5) to feature in a cluster, chose the number of tweets per subject and then constructed the 'chain of subjects' by sampling tweets from a set of randomly chosen stories. Finally, *Random* clusters were generated by sampling tweets from all stories, ensuring no single story represented more than 20% of a cluster. These artificial clusters from PHEME serve as ground-truth data for thematic incoherence.

3.3 Cluster Filtering

For automatically collected clusters (*COVID-19* and *Election*) we followed a series of filtering steps: duplicate tweets, non-English[4] tweets and ads were removed and only clusters containing 20-50 tweets were kept. As we sought to mine stories and associated user stances, opinionated clusters were prioritised. The sentiment analysis tool *VADER* (Gilbert and Hutto, 2014) was leveraged to gauge subjectivity in each cluster: a cluster is considered to be opinionated if the majority of its tweets express strong sentiment polarity.[5] *VADER* was chosen for its reliability on social media text and for its capacity to assign granulated sentiment valences; this allowed us to readily label millions of tweets and impose our own restrictions to classify neutral/non-neutral instances by varying the thresholds for the VADER compound score.

[4] https://pypi.org/project/langdetect/

[5] The absolute value of VADER compound score is required to be > 0.5, a much stricter condition than that used originally (Gilbert and Hutto, 2014).

3.4 Cluster Sampling

Work on assessing topic coherence operates on either the entire dataset (Fang et al., 2016b) or a random sample of it (Newman et al., 2010; Mimno et al., 2011). Fully annotating our entire dataset of thematic clusters would be too time-consuming, as the labelling of each data point involves reading dozens of posts rather than a small set of topical words. On the other hand, purely random sampling from the dataset cannot guarantee cluster diversity in terms of different levels of coherence. Thus, we opt for a more complex sampling strategy inspired by stratified sampling (Singh and Mangat, 2013), allowing more control over how the data is partitioned in terms of keywords and scores.
After filtering *Election* and *COVID-19* contained 46,715 and 5,310 clusters, respectively. We chose to sample 100 clusters from each dataset s.t. they:

- derive from a semantically diverse set of keywords (required for *Elections* only);
- represent varying levels of coherence (both);
- represent a range of time periods (both).

We randomly subsampled 10 clusters from each keyword with more than 100 clusters and keep all clusters with under-represented keywords (associated with fewer than 100 clusters). This resulted in 2k semantically diverse clusters for *Elections*.

TGM scores were leveraged to allow the selection of clusters with diverse levels of thematic coherence in the pre-annotation dataset. Potential score ranges for each coherence type were modelled on the PHEME dataset (See Sec. 3.2, 3.5), which is used as a gold standard for cluster coherence/incoherence. For each metric $\mathcal{M}$ and each coherence type CT, we defined the associated interval to be:

$$I(\mathcal{M})_{CT} = [\mu - 2\sigma, \mu + 2\sigma],$$

where μ, σ are the mean and standard deviation for the set of metric scores $\mathcal{M}$ characterising clusters of coherence type CT. We thus account for 95% of the data[6]. We did not consider metrics $\mathcal{M}$ for which the overlap between $I(\mathcal{M})_{\text{Good}}$, $I(\mathcal{M})_{\text{Intruded-Chained}}$[7] and $I(\mathcal{M})_{\text{Random}}$ was significant as this implied the metric was unreliable.

[6] Both the *Shapiro-Wilk* and *Anderson-Darling* statistical tests had showed the PHEME data is normally distributed.

[7] *Intruded* and *Chained* clusters mostly define the intermediate level of coherence, so their score ranges are similar, hence the two groups are unified.

As we did not wish to introduce metric bias when sampling the final dataset, we subsampled clusters across the intersection of all suitable metrics for each coherence type CT. In essence, our final clusters were sampled from each of the sets $\mathcal{C}_{CT} = \{C_i | \mathcal{M}(C_i) \in I(\mathcal{M})_{CT} \forall \text{ metric } \mathcal{M}\}$. For each of *COVID-19* and *Elections* we sampled 50 clusters $\in \mathcal{C}_{\text{Good}}$, 25 clusters $\in \mathcal{C}_{\text{Intruded-Chained}}$ and 25 clusters $\in \mathcal{C}_{\text{Random}}$.

3.5 Coherence Annotation Process

Coherence annotation was carried out in four stages by three annotators. We chose experienced journalists as they are trained to quickly and reliably identify salient content. An initial pilot study including the journalists and the research team was conducted; this involved two rounds of annotation and subsequent discussion to align the team's understanding of the guidelines (for the guidelines see Appendix B).

The first stage tackled tweet-level annotation within clusters and drew inspiration from the classic task of word intrusion (Chang et al., 2009): annotators were asked to group together tweets discussing a common subject; tweets considered to be 'intruders' were assigned to groups of their own. Several such groups can be identified in a cluster depending on the level of coherence. This grouping served as a building block for subsequent stages. This sub-clustering step offers a good trade-off between high annotation costs and manual evaluation since manually creating clusters from thousands of tweets is impractical. We note that agreement between journalists is not evaluated at this first stage as obtaining exact sub-clusters is not our objective. However, vast differences in sub-clustering are captured in the next stages in quality judgment and issue identification (See below).The second stage concerned cluster quality assessment, which is our primary task. Similar to Newman et al. (2010) for topic words, annotators evaluated tweet cluster coherence on a 3-point scale (*Good*, *Intermediate*, *Bad*). *Good* coherence is assigned to a cluster where the majority of tweets belong to the same theme (sub-cluster), while clusters containing many unrelated themes (sub-clusters) are assigned *bad* coherence.

The third stage pertains to issue identification of low coherence, similar to Mimno et al. (2011). When either *Intermediate* or *Bad* are chosen in stage 2 annotators can select from a list of issues to justify their choice:

- **Chained**: several themes are identified in the cluster (with some additional potential random tweets), without clear connection between any two themes.
- **Intruded**: only one common theme can be identified among some tweets in the cluster and the rest have no clear connection to the theme or to each other.
- **Random**: no themes can be identified and there is no clear connection among tweets in the cluster.

Inter-annotator agreement (IAA) was computed separately for the second and third stages as they serve a different purpose. For the second stage (cluster quality), we obtain average Spearman correlation $r_s = 0.73$ which is comparable to previous coherence evaluation scores in topic modelling literature ((Newman et al., 2010) with rs = 0.73 / 0.78 and (Aletras and Stevenson, 2013) with rs = 0.70 / 0.64 / 0.54) and average Cohen's Kappa $\kappa = 0.48$ (moderate IAA). For the third stage (issue identification), we compute average $\kappa = 0.36$ (fair IAA).

Analysis of pairwise disagreement in stage 2 shows only 2% is due to division in opinion over Good-Bad clusters. Good-Intermediate and Intermediate-Bad cases account for 37% and 61% of disagreements respectively. This is encouraging as annotators almost never have polarising views on cluster quality and primarily agree on the coherence of a good cluster, the main goal of this task. For issue identification the majority of disagreements ($\%49$) consists in distinguishing Intermediate-Chained cases. This can be explained by the expected differences in identifying subclusters in the first stage. For the adjudication process, we found that a majority always exists and thus the final score was assigned to be the majority label (2/3 annotators). Table 1 presents a summary of the corpus size, coherence quality and issues identified for *COVID-19* and *Election* (See Appendix C for a discussion).

4 Experiments

Our premise is that a pair of sentences scoring high in terms of TGMs means that the sentences are semantically similar. When this happens across many sentences in a cluster then this denotes good cluster coherence. Following Douven and Meijs (2007), we consider three approaches to implementing and adapting TGMs to the task of measuring thematic

	General			Cluster Quality			Cluster Issue		
Dataset	Clusters	Tweets	Tokens	Good	Intermediate	Bad	Intruded	Chained	Random
COVID-19	100	2,955	100K	18	31	51	32	25	25
Election	100	2,650	52K	25	50	25	28	33	14

Table 1: Statistics of the annotated clusters where the final label is assigned to be the majority label.

coherence. The differences between these methods consist of: (a) the choice of the set of tweet pairs $S \subset T \times T$ on which we apply the metrics and (b) the score aggregating function $f(C)$ assigning coherence scores to clusters. The TGMs employed in our study are *BERTScore* (Zhang et al., 2019), *MoverScore* (Zhao et al., 2019) for both unigrams and bigrams and *BLEURT* (Sellam et al., 2020). We also employed a surface level metric based on cosine similarity distances between TF-IDF representations[8] of tweets to judge the influence of word co-occurrences in coherence analysis. Each approach has its own advantages and disadvantages, which are outlined below.

4.1 Exhaustive Approach

In this case $S = T \times T$, i.e., all possible tweet pairs within the cluster are considered. The cluster is assigned the mean sum over all scores. This approach is not biased towards any tweet pairs, so is able to penalise any tweet that is off-topic. However, it is computationally expensive as it requires $O(|T|^2)$ operations. Formally, given a TGM $\mathcal{M}$, we define this approach as:

$$f(C) = \frac{1}{\binom{|T|}{2}} \cdot \sum_{\substack{\text{tweet}_i, \text{tweet}_j \in T \\ i<j}} \mathcal{M}(\text{tweet}_i, \text{tweet}_j).$$

4.2 Representative Tweet Approach

We assume there exists a representative tweet able to summarise the content in the cluster, denoted as the *representative tweet* (i.e. tweet_{rep}). This is formally defined as:

$$\text{tweet}_{rep}(C) = arg \min_{\text{tweet}_i \in C} D_{KL}(\theta, \text{tweet}_i),$$

where we compute the *Kullback–Leibler* divergence (D_{KL}) between the word distributions of the topic θ representing the cluster C and each tweet in C (Wan and Wang, 2016); we describe the computation of D_{KL} in Appendix A. We also considered other text summarisation methods (Basave et al., 2014; Wan and Wang, 2016) such as *MEAD* (Radev et al., 2000) and *Lexrank* (Erkan and Radev, 2004) to extract the best representative tweet, but our initial empirical study indicated D_{KL} consistently finds the most appropriate representative tweet. In this case cluster coherence is defined as below and has linear time complexity $O(|T|)$:

$$f(C) = \frac{1}{|T|} \sum_{\text{tweet}_i \in T} \mathcal{M}(\text{tweet}_i, \text{tweet}_{rep}).$$

As $S = \{(\text{tweet}, \text{tweet}_{rep}) | \text{ tweet} \in T\} T \times T$, the coherence of a cluster is heavily influenced by the correct identification of the representative tweet.

4.3 Graph Approach

Similar to the work of Erkan and Radev (2004), each cluster of tweets C can be viewed as a complete weighted graph with nodes represented by the tweets in the cluster and each edge between tweet_i, tweet_j assigned as weight: $w_{i,j} = \mathcal{M}(\text{tweet}_i, \text{tweet}_j)^{-1}$. In the process of constructing a complete graph, all possible pairs of tweets within the cluster are considered. Hence $S = T \times T$ with time complexity of $O(|T|^2)$ as in Section 4.1. In this case, the coherence of the cluster is computed as the average *closeness centrality* of the associated cluster graph. This is a measure derived from graph theory, indicating how 'close' a node is on average to all other nodes; as this definition intuitively corresponds to coherence within graphs, we included it in our study. The closeness centrality for the node representing tweet_i is given by:

$$CC(\text{tweet}_i) = \frac{|T| - 1}{\sum_{\text{tweet}_j \in T} d(\text{tweet}_j, \text{tweet}_i)},$$

where $d(\text{tweet}_j, \text{tweet}_i)$ is the shortest distance between nodes tweet_i and tweet_j computed via Dijkstra's algorithm. Note that as Dijkstra's algorithm only allows for non-negative graph weights and *BLEURT*'s values are mostly negative, we did not include this TGM in the graph approach implementation. Here cluster coherence is defined as the average over all closeness centrality scores of the nodes in the graph:

$$f(C) = \frac{1}{|T|} \sum_{\text{tweet} \in T} CC(\text{tweet}_i).$$

[8] Tweets are embedded into a vector space of TF-IDF representations within their corresponding cluster.

	Election	COVID-19	PHEME
	r_s / ρ / τ	r_s / ρ / τ	r_s / ρ / τ
Exhaustive TF-IDF	**0.62** / 0.62 / **0.49**	**0.68** / **0.72** / **0.53**	0.81 / 0.73 / 0.67
Graph TF-IDF	**0.62** / **0.63** / 0.48	0.66 / **0.72** / 0.52	0.74 / 0.71 / 0.60
Exhaustive BLEURT	0.49 / 0.48 / 0.37	0.66 / 0.65 / 0.52	**0.84** / **0.86** / **0.69**
Exhaustive BERTScore	0.58 / 0.57 / 0.44	0.62 / 0.64 / 0.49	0.83 / 0.80 / 0.68
Topic Coherence Glove	-0.25 / -0.27 / -0.19	0.04 / 0.02 / 0.03	N/A
Avg Topic Coherence Glove	-0.22 / -0.23 / -0.17	-0.03 / -0.03 / -0.02	N/A
Topic Coherence BERTweet	-0.23 / -0.22 / -0.18	0.10 / 0.11 / 0.08	N/A
Avg Topic Coherence BERTweet	-0.17 / -0.16 / -0.14	0.04 / 0.04 / 0.03	N/A

Table 2: Agreement with annotator ratings across the *Election*, *COVID-19* and *PHEME* datasets. The metrics are Spearman's rank correlation coefficient (r_s), Pearson Correlation coefficient (ρ) and Kendall Tau (τ).

5 Results

5.1 Quantitative Analysis

Table 2 presents the four best and four worst performing metrics (for the full list of metric results refer to Appendix A). *MoverScore* variants are not included in the results discussion as they only achieve average performance.

Election and COVID-19 *Exhaustive TF-IDF* and *Graph TF-IDF* consistently outperformed TGMs, implying that clusters with a large overlap of words are likely to have received higher coherence scores. While TF-IDF metrics favour surface level co-occurrence and disregard deeper semantic connections, we conclude that, by design all posts in the thematic clusters (posted within a 1h or 3 h window) are likely to use similar vocabulary. Nevertheless, TGMs correlate well with human judgement, implying that semantic similarity is a good indicator for thematic coherence: *Exhaustive BERTScore* performs the best of all TGMs in *Election* while *Exhaustive BLEURT* is the strongest competitor to TF-IDF based metrics for *COVID-19*.

On the low end of the performance scale, we have found topic coherence to be overwhelmingly worse compared to all the TGMs employed in our study. *BERTweet* improves over *Glove* embeddings but only slightly as when applied at the word level (for topic coherence) it is not able to benefit from the context of individual words. We followed Lau and Baldwin (2016), and computed average topic coherence across the top $5, 10, 15, 20$ topical words in order to obtain a more robust performance (see *Avg Topic Coherence Glove*, *Avg Topic Coherence BERTweet*). Results indicate that this smoothing technique correlates better with human judgement for *Election*, but lowers performance further in *COVID-19* clusters.

In terms of the three approaches, we have found that the *Exhaustive* and *Graph* approaches perform similarly to each other and both outperform the *Representative Tweet* approach. Sacrificing time as trade off to quality, the results indicate that metrics considering all possible pairs of tweets account for higher correlation with annotator rankings.

PHEME The best performance on this dataset is seen with TGM *BLEURT*, followed closely by *BERTScore*. While TF-IDF based metrics are still in the top four, surface level evaluation proves to be less reliable: PHEME stories are no longer constrained by strict time windows[9], which allows the tweets within each story to be more lexically diverse, while still maintaining coherence. In such instances, strategies depending exclusively on word frequencies perform inconsistently, which is why metrics employing semantic features (*BLEURT*, *BERTScore*) outperform TF-IDF ones. Note that PHEME data lack the topic coherence evaluation, as these clusters were not generated through topic modelling (See *Subsection* 3.2).

5.2 Qualitative Analysis

We analysed several thematic clusters to get a better insight into the results. Tables 3 and 4 show representative fragments from 2 clusters labelled as 'good' in the *COVID-19* dataset. The first cluster contains posts discussing the false rumour that bleach is an effective cure to COVID-19, with the majority of users expressing skepticism. As most tweets in this cluster directly quote the rumour and thus share a significant overlap of words, not surprisingly, TF-IDF based scores are high *Exhaustive*

[9]Stories were generated across several days, rather then hours, by tracking on-going breaking news events on Twitter.

Cluster Annotation: Good	Common Keyword: 'coronavirus'
Trump-loving conspiracy nuts tout drinking bleach as a 'miracle' cure for coronavirus - "They may have found a cure for Trump lovers and MAGA but not anything else" #MAGAIDIOTS #TestOnDonJr #OneVoice	
Pro-**Trump conspiracy** theorists **tout drinking bleach as a 'miracle' cure for coronavirus**	
Trump-loving conspiracy nuts tout drinking bleach as a 'miracle' cure for coronavirus – DRINK UP, MAGAts!	
Isn't this just a problem solving itself? #Darwinism **Trump-loving conspiracy nuts tout drinking bleach as a 'miracle' cure for coronavirus**	
Trump-loving conspiracy nuts tout drinking bleach as a 'miracle' cure for coronavirus... Is a quart each sufficient? Will go multiple gallons-gratis.	

Table 3: Cluster fragment from COVID-19 dataset, Exhaustive TF-IDF = 0.109 and Exhaustive BLEURT = -0.808.

Cluster Annotation: Good	Common Keyword: 'pandemic'
@CNN @realDonaldTrump administration recently requested $2.5 billion in emergency funds to prepare the U.S. for a possible widespread outbreak of coronavirus. Money isnt necessary if the Trump past 2 years didnt denudate government units that were designed to protect against pandemic	
@realDonaldTrump @VP @SecAzar @CDCgov @CDCDirector I bet that was the case of 3 people who had gone no where. You cutting CDC, Scientists & taking money that was set aside for pandemic viruses that Obama set aside has not helped. You put Pence in charge who did nothing for IN aids epidemic because he said he was a Christian.	
Trump fired almost all the pandemic preparedness team that @BarackObama put in place and his budget promised cutting $ 1.3 billion from @CDC. With 'leadership' like that, what could anyone expect except dire preparation in America? # MAGA2020 morons: be careful at his rallies	
@USER Democrats DO NOT want mils of Americans to die from coronavirus. They aren't the ones who fired the whole pandemic team Obama put in place. It was Trump. He left us unprepared. All he's interested in is the stock market, wealthy donors & getting re-elected.	
@USER , Obama set up a pandemic reaction force, placed higher budgets for the CDC AND health and Human Services. Trump on the other hand, have significantly cut the budgets to HHS and the CDC. They disbanded the White House pandemic efforts. With a politician, not a Scientist	

Table 4: Cluster fragment from COVID-19 dataset, Exhaustive TF-IDF = 0.040 and Exhaustive BLEURT = -0.811.

TF-IDF = 0.109. In the second cluster, however, users challenge the choices of the American President regarding the government's pandemic reaction: though the general feeling is unanimous in all posts of the second cluster, these tweets employ a more varied vocabulary. Consequently, surface level metrics fail to detect the semantic similarity *Exhaustive TF-IDF = 0.040*. When co-occurrence statistics are unreliable, TGMs are more successful for detecting the 'common story' diversely expressed in the tweets: in fact, Exhaustive BLEURT assigns similar scores to both clusters (-0.808 for Cluster 1 and -0.811 for Cluster 2) in spite of the vast difference in their content intersection, which shows a more robust evaluation capability.

We analyse the correlation between topic coherence and annotator judgement in *Tables* 5 and 6. Both are illustrative fragments of clusters extracted from the *Election* dataset. Though all tweets in *Table* 5 share the keyword 'oil', they form a bad random cluster type, equivalent to the lowest level of coherence. On the other hand, *Table* 6 clearly presents a good cluster regarding an immigration tragedy at sea. Although this example pair contains clusters on opposite sides of the coherence spectrum, topic coherence metrics fail to distinguish the clear difference in quality between the two. Moreover, *Table* 6 receives lower scores (TC Glove = 0.307) than its incoherent counterpart (TC Glove = 0.330) for Glove Topic Coherence. However, TGM metric BERTScore and surface-level metric TF-IDF correctly evaluate the two clusters by penalising incoherence (Exhaustive BERTScore = 0.814 and Exhaustive TF-IDF = 0.024) and awarding good clusters (Exhaustive BERTScore = 0.854 and Exhaustive TF-IDF = 0.100).

6 Conclusions and Future Work

We have defined the task of creating topic-sensitive clusters of microblogs and evaluating their thematic coherence. To this effect we have investigated the efficacy of different metrics both from the topic modelling literature and text generation metrics TGMs. We have found that TGMs correlate much better with human judgement of thematic coherence compared to metrics employed in topic model evaluation. TGMs maintain a robust performance across different time windows and are generalisable across several datasets. In future work we plan to use TGMs in this way to identify thematically

Cluster Annotation: Bad Random	Common Keyword: 'oil'
M'gonna have a nap, I feel like I've drank a gallon of like grease or oil or whatever bc I had fish&chips like 20 minutes ago	
Check out our beautiful, nostalgic oil canvasses. These stunning images will take you back to a time when life...	
Five years later, bottlenose dolphins are STILL suffering from BP oil disaster in the Gulf. Take action!	
Once the gas and oil run out countries like Suadia Arabia and Russia won't be able to get away with half the sh*t they can now	
Ohhh this tea tree oil is burning my face off	

Table 5: Cluster fragment from Election dataset, TC Glove = 0.330, Exhaustive BERTScore = 0.814 and Exhaustive TF-IDF = 0.024.

Cluster Annotation: Good	Common Keyword: 'migrants'
Up to 300 migrants missing in Mediterranean Sea are feared dead #migrants.	
NEWS: More than 300 migrants feared drowned after their overcrowded dinghies sank in the Mediterranean	
Imagine if a ferry sunk with 100s dead - holiday makers, kids etc. Top story everywhere. 300 migrants die at sea and it doesn't lead.	
@bbc5live Hi FiveLive: you just reported 300 migrants feared dead. I wondered if you could confirm if the MIGRANTS were also PEOPLE? Cheers.	
If the dinghies were painted pink would there be as much uproar about migrants drowning as the colour of a f**king bus?	

Table 6: Cluster fragment from Election dataset, TC Glove = 0.307, Exhaustive BERTScore = 0.854 and Exhaustive TF-IDF = 0.100.

coherent clusters on a large scale, to be used in downstream tasks such as multi-document opinion summarisation.

Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant no. EP/V030302/1) and The Alan Turing Institute (grant no. EP/N510129/1). We would like to thank our 3 annotators for their invaluable expertise in constructing the datasets. We also thank the reviewers for their insightful feedback. Finally, we would like to thank Yanchi Zhang for his help in the redundancy correction step of the pre-processing.

Ethics

Ethics approval to collect and to publish extracts from social media datasets was sought and received from Warwick University Humanities & Social Sciences Research Ethics Committee. During the annotation process, tweet handles, with the except of public figures, organisations and institutions, were anonymised to preserve author privacy rights. In the same manner, when the datasets will be released to the research community, only tweets IDs will be made available along with associated cluster membership and labels.

Compensation rates were agreed with the annotators before the annotation process was launched. Remuneration was fairly paid on an hourly rate at the end of task.

References

Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. 2013. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *IWCS*, pages 13–22.

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of lda generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer.

Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. 2014. Automatic labelling of topic models learned from twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking social media discourse about the

covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill*, 6(2):e19273.

Igor Douven and Wouter Meijs. 2007. Measuring coherence. *Synthese*, 156(3):405–425.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016a. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1057–1060.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016b. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1057–1060.

Marco Furini and Gabriele Menegoni. 2018. Public health and social media: Language analysis of vaccine conversations. In *2018 international workshop on social sensing (SocialSens)*, pages 50–55. IEEE.

CHE Gilbert and Erric Hutto. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, volume 81, page 82.

Yinuo Guo and Junfeng Hu. 2019. Meteor++ 2.0: Adopt syntactic level paraphrase knowledge into machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 501–506, Florence, Italy. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

Jey Han Lau and Timothy Baldwin. 2016. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–487, San Diego, California. 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 *EACL*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

Fred Morstatter and Huan Liu. 2018. In search of coherence and consensus: Measuring the interpretability of statistical topics. *Journal of Machine Learning Research*, 18(169):1–32.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy. 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, page 100–108, USA. Association for Computational Linguistics.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Dat Quoc Nguyen, Thanh Vu, and A. Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *ArXiv*, abs/2005.10200.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rob Procter, Shuaib Choudhry, Jim Smith, Martine Barons, and Adam Edwards. 2020. Roadmapping uses of advanced analytics in the uk food and drink sector.

Rob Procter, Jeremy Crump, Susanne Karstedt, Alex Voss, and Marta Cantijoch. 2013. Reading the riots: What were the police doing on twitter? *Policing and society*, 23(4):413–436.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

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*, pages 399–408.

Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2011. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, 63.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ravindra Singh and Naurang Singh Mangat. 2013. *Elements of survey sampling*, volume 15. Springer Science & Business Media.

Didi Surian, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, and Adam G Dunn. 2016. Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. *Journal of medical Internet research*, 18(8):e232.

Peter Tolmie, Rob Procter, David William Randall, Mark Rouncefield, Christian Burger, Geraldine Wong Sak Hoi, Arkaitz Zubiaga, and Maria Liakata. 2017. Supporting the use of user generated content in journalistic practice. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3632–3644. ACM.

R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Xiaojun Wan and Tianming Wang. 2016. Automatic labeling of topic models using text summaries. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2305.

Bo Wang, Maria Liakata, Adam Tsakalidis, Spiros Georgakopoulos Kolaitis, Symeon Papadopoulos, Lazaros Apostolidis, Arkaitz Zubiaga, Rob Procter, and Yiannis Kompatsiaris. 2017a. TOTEMSS: Topic-based, temporal sentiment summarisation for Twitter. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 21–24, Tapei, Taiwan. Association for Computational Linguistics.

Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017b. A hierarchical topic modelling approach for tweet clustering. In *International Conference on Social Informatics*, pages 378–390. Springer International Publishing.

Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017c. TDParse: Multi-target-specific sentiment recognition on Twitter. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 483–493, Valencia, Spain. Association for Computational Linguistics.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782, Online. Association for Computational Linguistics.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 233–242, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*, pages 338–349. Springer.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

Appendix A

Representative-Tweet Selection

As described in Section 4.2, we select the tweet that has the lowest divergence score to the top topic words of the cluster. Following (Wan and Wang, 2016), we compute the *Kullback–Leibler* divergence (D_{KL}) between the word distributions of the topic θ the cluster C represents and each tweet in C as follows:

$$D_{KL}(\theta, \text{tweet}_i) = \sum_{w \in TW \bigcup SW} p_\theta(w) * log \frac{p_\theta(w)}{tf(w, \text{tweet}_i)/len(\text{tweet}_i)}$$

where $p_\theta(w)$ is the probability of word w in topic θ. TW denotes top 20 words in cluster C according to the probability distribution while SW denotes the set of words in tweet$_i$ after removing stop words. $tf(w, \text{tweet}_i)$ denotes the frequency of word w in tweet$_i$, and $len(\text{tweet}_i)$ is the length of tweet$_i$ after removing stop words. For words that do not appear in SW, we set $tf(w, \text{tweet}_i)/len(\text{tweet}_i)$ to 0.00001.

Timings of Evaluation Metrics

Metric	Time (in seconds)
Exhaustive BERTScore	10.84
Exhaustive BLEURT	234.10
Exhaustive MoverScore1	11.73
Exhaustive MoverScore2	11.42
Exhaustive TF-IDF	0.05
Graph TF-IDF	0.12

Table A1: Average timings of metric performances per 1 cluster

Complete Results

The complete results of our experiments are in *Table* A2. The notation is as follows:

- **Exhaustive** indicates that the *Exhaustive Approach* was employed for the metric.
- **Linear** indicates that the *Representative Tweet Approach* was employed for the metric.
- **Graph** indicates the the *Graph Approach* was employed for the metric.

  Shortcuts for the metrics are: **MoverScore1** = MoverScore applied for unigrams; **MoverScore2** = MoverScore applied for bigrams

PHEME data coherence evaluation

As original PHEME clusters were manually created by journalists to illustrate specific stories, they are by default coherent. Hence, according to the guidelines, these clusters would be classified as "Good". For the artificially created clusters, PHEME data is mixed such that different stories are combined in different proportions (See 3.2). Artificially intruded and chained clusters would be classed as 'Intermediate' as they have been generated on the basis that a clear theme (or themes) can be identified. Finally, an artificially random cluster was created such that there is no theme found in the tweets as they are too diverse; this type of cluster is evaluated as 'Bad'.

	Election	COVID-19	PHEME
	r_s / ρ / τ	r_s / ρ / τ	r_s / ρ / τ
Exhaustive TF-IDF	**0.62** / 0.62 / **0.49**	**0.68** / **0.72** / **0.53**	0.81 / 0.73 / 0.67
Linear TF-IDF	**0.51** / 0.48 / **0.39**	**0.36** / **0.45** / **0.27**	N/A
Graph TF-IDF	**0.62** / **0.63** / 0.48	0.66 / **0.72** / 0.52	0.74 / 0.71 / 0.60
Exhaustive BLEURT	0.49 / 0.48 / 0.37	0.66 / 0.65 / 0.52	**0.84** / **0.86** / **0.69**
Linear BLEURT	0.41 / 0.40 / 0.32	0.34 / 0.34 / 0.26	N/A
Exhaustive BERTScore	0.58 / 0.57 / 0.44	0.62 / 0.64 / 0.49	0.83 / 0.80 / 0.68
Linear BERTScore	0.49 / 0.50 / 0.38	0.50 / 0.53 / 0.38	N/A
Graph BERTScore	0.57 / 0.57 / 0.44	0.62 / 0.64 / 0.49	0.83 / 0.73 / 0.68
Exhaustive MoverScore1	0.56 / 0.55 / 0.43	0.46 / 0.56 / 0.35	0.56 / 0.56 / 0.44
Linear MoverScore1	0.54 / 0.52 / 0.41	0.36 / 0.39 / 0.28	N/A
Graph MoverScore1	0.53 / 0.53 / 0.42	0.37 / 0.44 / 0.29	0.52 / 0.56 / 0.40
Exhaustive MoverScore2	0.46 / 0.46 / 0.35	0.35 / 0.46 / 0.27	0.40 / 0.35 / 0.30
Linear MoverScore2	0.47 / 0.46 / 0.35	0.26 / 0.31 / 0.20	N/A
Graph MoverScore2	0.47 / 0.49 / 0.36	0.42 / 0.50 / 0.32	0.36 / 0.39 / 0.27
Topic Coherence Glove	-0.25 / -0.27 / -0.19	0.04 / 0.02 / 0.03	N/A
Avg Topic Coherence Glove	-0.22 / -0.23 / -0.17	-0.03 / -0.03 / -0.02	N/A
Topic Coherence BERTweet	-0.23 / -0.22 / -0.18	0.10 / 0.11 / 0.08	N/A
Avg Topic Coherence BERTweet	-0.17 / -0.16 / -0.14	0.04 / 0.04 / 0.03	N/A

Table A2: Agreement with human annotators across the *Election*, *COVID-19* and *PHEME* datasets. The metrics are Spearman's rank correlation coefficient (r_s), Pearson Correlation coefficient (ρ) and Kendall Tau (τ).

Appendix B: Annotation Guidelines

Overview

You will be shown a succession of clusters of posts from Twitter (tweets), where the posts originate from the same one hour time window. Each cluster has been generated by software that has decided its tweets are variants on the same 'subject'. You will be asked for your opinion on the quality ('coherence') of each cluster as explained below. As an indication of coherence quality consider how easy it would be to summarise a cluster.

Steps

In the guidelines below, a *subject* is a group of at least three tweets referring to the same topic. **Marking common subjects**: In order to keep track of each subject found in the cluster, label it by entering a number into column **Subject Label** and then assign the same number for each tweet that you decide is about the same subject. **Note**, the **order** of the tweets will automatically change as you enter each number so that those assigned with the same subject number will be listed **together**.

1. **Reading a Cluster of Tweets**
   (a) Carefully read each tweet in the cluster with a view to uncovering overlapping concepts, events and opinions (if any).

   (b) Identify the common keyword(s) present in all tweets within the cluster. Note that common keywords across tweets in a cluster are present in all cases by design, so by itself it is not a sufficient criterion to decide on the quality of a cluster.

   (c) Mark tweets belonging to the same subject as described in the paragraph above.

2. **Cluster Annotation : What was your opinion about the cluster?**
   (a) Choose '**Good**' if you can identify one subject within the cluster to which most tweets refer (you can count these based on the numbers you have assigned in the column *Subject Label*). This should be a cluster that you would find it easy to summarise. Proceed to **Step 4**.

   (b) Choose '**Intermediate**' if you are uncertain that the cluster is good, you would find it difficult to summarise its information or you find that there are a small number (e.g., one, two or three) of unrelated subjects being discussed that are of similar size (*chained*, See issues in Step 3) or one clear subject with a mix of other unrelated tweets (intruded, See issues in Step 3). Additionally, if there is one significantly big subject and one or more other 'small' subjects (small 2,3 tweets), this cluster should be *Intermediate Intruded*. Proceed to **Step 3**.

   (c) Choose '**Bad**' if you are certain that the cluster is not good and the issue of fragmented subjects within the cluster is such that many unrelated subjects are being discussed (heavily *chained*) or there is one subject with a mix of unrelated tweets but the tweets referring to one subject are a minority. Proceed to **Step 3**.

3. **Issue Identification: What was wrong with the cluster?**
   (a) Choose '**Chained**' if several subjects can be identified in the cluster (with some potential random tweets that belong to no subject), but there are no clear connections between any two subjects. This issue can describe both an Intermediate and a Bad cluster.
   (b) Choose '**Intruded**' if only one common subject can be identified in some tweets in the cluster and the rest of tweets have no clear connections to the subject or between each other. This issue can describe both an Intermediate and a Bad cluster.
   (c) Choose '**Random**' if no subjects can be identified at all as there are no clear connections between the tweets in the cluster. Usually 'Random' will be a property of a Bad cluster.

4. **Cluster Summarisation** You are asked to provide a brief summary (20-40 words) for each **Good** cluster you had identified in **Step 2**.

Appendix C: Corpus Statistics

Size

In terms of size, we observe that the average tweet in *Election* data is significantly shorter (20 tokens) than its correspondent in the *COVID-19* corpus which is 34 tokens long. We observe that the former's collection period finished before Twitter platform doubled its tweet character limit which would be confirmed by the figures in the table. Further work will tackle whether tweet length in a cluster has any impact on the coherence of its message.

Score differences

We believe differences in the application of the clustering algorithm influenced the score differences between *Election* and *COVID-19* datasets. The clustering algorithm we employed uses a predefined list of keywords that partitions the data into sets of tweets mentioning a common keyword as a first step. The keyword set used for the *Election* dataset contains 438 keywords, while the *COVID-19* dataset contains 80 keywords used for Twitter API tracking (Chen et al., 2020). We also note that the different time window span can impact the quality of clusters.