Template-based Abstractive Microblog Opinion Summarization

Iman Munire Bilal[1,4], **Bo Wang**[2,4], **Adam Tsakalidis**[3,4], **Dong Nguyen**[5], **Rob Procter**[1,4], **Maria Liakata**[1,3,4]

[1]Department of Computer Science, University of Warwick, UK
[2]Center for Precision Psychiatry, Massachusetts General Hospital, USA
[3]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[4]The Alan Turing Institute, London, UK
[5]Department of Information and Computing Sciences, Utrecht University, The Netherlands
{iman.bilal|rob.procter}@warwick.ac.uk bwang29@mgh.harvard.edu
{atsakalidis|mliakata}@qmul.ac.uk d.p.nguyen@uu.nl

Abstract

We introduce the task of microblog opinion summarization (MOS) and share a dataset of 3100 gold-standard opinion summaries to facilitate research in this domain. The dataset contains summaries of tweets spanning a 2-year period and covers more topics than any other public Twitter summarization dataset. Summaries are abstractive in nature and have been created by journalists skilled in summarizing news articles following a template separating factual information (main story) from author opinions. Our method differs from previous work on generating gold-standard summaries from social media, which usually involves selecting representative posts and thus favors extractive summarization models. To showcase the dataset's utility and challenges, we benchmark a range of abstractive and extractive state-of-the-art summarization models and achieve good performance, with the former outperforming the latter. We also show that fine-tuning is necessary to improve performance and investigate the benefits of using different sample sizes.

1 Introduction

Social media has gained prominence as a means for the public to exchange opinions on a broad range of topics. Furthermore, its social and temporal properties make it a rich resource for policy makers and organizations to track public opinion on a diverse range of issues (Procter et al., 2013; Chou et al., 2018; Kalimeri et al., 2019). However, understanding opinions about different issues and entities discussed in large volumes of posts in platforms such as Twitter is a difficult task. Existing work on Twitter employs extractive summarization (Inouye and Kalita, 2011; Zubiaga et al., 2012; Wang et al., 2017a; Jang and Allan, 2018) to filter through information by ranking and selecting tweets according to various criteria. However, this approach unavoidably ends up including incomplete or redundant information (Wang and Ling, 2016).

To tackle this challenge we introduce *Microblog opinion summarization* (MOS), which we define as a multi-document summarization task aimed at capturing diverse reactions and stances (opinions) of social media users on a topic. While here we apply our methods to Twitter data readily available to us, we note that this summarization strategy is also useful for other microblogging platforms. An example of a tweet cluster and its opinion summary is shown in Table 1. As shown, our proposed summary structure for MOS separates the factual information (story) from reactions to the story (opinions); the latter is further divided according to the prevalence of different opinions. We believe that making combined use of stance identification, sentiment analysis and abstractive summarization is a challenging but valuable direction in aggregating opinions expressed in microblogs.

The availability of high quality news article datasets has meant that recent advances in text summarization have focused mostly on this type of data (Nallapati et al., 2016; Grusky et al., 2018; Fabbri et al., 2019; Gholipour Ghalandari et al., 2020). Contrary to news article summarization, our task focuses on summarizing an event as well as ensuing public opinions on social media. Review opinion summarization (Ganesan et al., 2010; Angelidis and Lapata, 2018) is related to MOS and faces the same challenge of filtering through large volumes of user-generated content. While

Transactions of the Association for Computational Linguistics, vol. 10, pp. 1229–1248, 2022. https://doi.org/10.1162/tacl_a_00516
Action Editor: Ivan Titov. Submission batch: 2/2022; Revision batch: 6/2022; Published 11/2022.

Human Summary
Main Story: The UK government faces intense backlash after its decision to fund the war in Syria. **Majority Opinion:** The majority of users criticise UK politicians for not directing their efforts to more important domestic issues like the NHS, education and homelessness instead of the war in Syria. **Minority Opinion:** Some users accuse the government of its intention to kill innocents by funding the war.
Tweet Cluster
It is shocking to me how the NHS is on its knees and the amount of homeless people that need help in this country...but we have funds for war!..SAD The government cannot even afford to help the homeless people of Britain yet they can afford to fund a war? It makes no proper sense at all They spend so much on sending missiles to murder innocent people and they complain daily about homeless on the streets? Messed up. Also, no money to resolve the issues of the homeless or education or the NHS. Yet loads of money to drop bombs? #SyriaVote

Table 1: Abridged cluster of tweets and its corresponding summary. Cluster content is color-coded to represent information overlap with each summary component: blue for Main Story, red for Majority Opinion, and green for Minority Opinion.

recent work (Chu and Liu, 2019; Bražinskas et al., 2020) aims to produce review-like summaries that capture the consensus, MOS summaries inevitably include a spectrum of stances and reactions. In this paper we make the following contributions:

1. We introduce the task of microblog opinion summarization (MOS) and provide detailed guidelines.

2. We construct a corpus[1] of tweet clusters and corresponding multi-document summaries produced by expert summarizers following our detailed guidelines.

3. We evaluate the performance of existing state-of-the-art models and baselines from three summarization domains (news articles, Twitter posts, product reviews) and four model types (abstractive vs. extractive, single document vs. multiple documents) on our corpus, showing the superiority of neural abstractive models. We also investigate the benefits of fine-tuning with various sample sizes.

[1]This is available at https://doi.org/10.6084/m9.figshare.20391144.

2 Related Work

Opinion Summarization has focused predominantly on customer reviews with datasets spanning reviews on Tripadvisor (Ganesan et al., 2010), Rotten Tomatoes (Wang and Ling, 2016), Amazon (He and McAuley, 2016; Angelidis and Lapata, 2018) and Yelp (Yelp Dataset Challenge; Yelp).

Early work by Ganesan et al. (2010) prioritized redundancy control and concise summaries. More recent approaches (Angelidis and Lapata, 2018; Amplayo and Lapata, 2020; Angelidis et al., 2021; Isonuma et al., 2021) employ aspect driven models to create relevant topical summaries. While product reviews have a relatively fixed structure, MOS operates on microblog clusters where posts are more loosely related, which poses an additional challenge. Moreover, while the former generally only encodes the consensus opinion (Bražinskas et al., 2020; Chu and Liu, 2019), our approach includes both majority and minority opinions.

Multi-document summarization has gained traction in non-opinion settings and for news events in particular. DUC (Dang, 2005) and TAC conferences pioneered this task by introducing datasets of 139 clusters of articles paired with multiple human-authored summaries. Recent work has seen the emergence of larger scale datasets such as WikiSum (Liu et al., 2018), Multi-News (Fabbri et al., 2019), and WCEP (Gholipour Ghalandari et al., 2020) to combat data sparsity. Extractive (Wang et al., 2020b,c; Liang et al., 2021) and abstractive (Jin et al., 2020) methods have followed from these multi-document news datasets.

Twitter Summarization is recognised by Cao et al. (2016) to be a promising direction for tracking reaction to major events. As tweets are inherently succinct and often opinionated (Mohammad et al., 2016), this task is at the intersection of multi-document and opinion summarization. The construction of datasets (Nguyen et al., 2018; Wang and Zhang, 2017) usually requires a clustering step to group tweets together under specific temporal and topical constraints, which we include within our own pipeline. Work by Jang and Allan (2018) and Corney et al. (2014) makes use of the subjective nature of tweets by identifying two stances for each topic to be summarized; we generalize this idea and do not impose a restriction on the number of possible opinions on a topic. The

lack of an abstractive gold standard means that the majority of existing Twitter models are extractive (Alsaedi et al., 2021; Inouye and Kalita, 2011; Jang and Allan, 2018; Corney et al., 2014). Here we provide such an abstractive gold standard and show the potential of neural abstractive models for microblog opinion summarization.

3 Creating the MOS Dataset

3.1 Data Sources

Our MOS corpus consists of summaries of microblog posts originating from two data sources, both involving topics that have generated strong public opinion: **COVID-19** (Chen et al., 2020) and **UK Elections** (Bilal et al., 2021).

- **COVID-19**: Chen et al. (2020) collected tweets by tracking COVID-19 related keywords (e.g., *coronavirus*, *pandemic*, *stayathome*) and accounts (e.g., *@CDCemergency*, *@HHSGov*, *@DrTedros*). We use data collected between January 2020 and January 2021, which at the time was the most complete version of this dataset.
- **UK Elections**: The **Election** dataset consists of all geo-located UK tweets posted between May 2014 and May 2016. The tweets were filtered using a list of 438 election-related keywords and 71 political party aliases curated by a team of journalists.

We follow the methodology in Bilal et al. (2021) to obtain opinionated, coherent clusters of between 20 and 50 tweets: The clustering step employs the GSDMM-LDA algorithm (Wang et al., 2017b), followed by thematic coherence evaluation (Bilal et al., 2021). The latter is done by aggregating exhaustive metrics BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2020), and TF-IDF to construct a random forest classifier to identify coherent clusters. Our final corpus is created by randomly sampling 3100 clusters,[2] 1550 each from the COVID-19 and Election datasets.

3.2 Summary Creation

The summary creation process was carried out in 3 stages on the Figure Eight platform by 3 journalists experienced in sub-editing. Following Iskender et al. (2021), a short pilot study was followed by a meeting with the summarizers to ensure the task and guidelines were well understood. Prior to this, the design of the summarization interface was iterated to ensure functionality and usability (See *Appendix A* for interface snapshots).

In the first stage, the summarizers were asked to read a cluster of tweets and state whether the opinions within it could be easily summarized by assigning one of three cluster types:

1. **Coherent Opinionated**: there are clear opinions about a common main story expressed in the cluster that can be easily summarized.
2. **Coherent Non-opinionated**: there are very few or no clear opinions in the cluster, but a main story is clearly evident and can be summarized.
3. **Incoherent**: no main story can be detected. This happens when the cluster contains diverse stories to which no majority of tweets refers, hence it cannot be summarized.

Following Bilal et al. (2021) on thematic coherence, we assume a cluster is coherent if and only if its contents can be summarized. Thus, both Coherent Opinionated and Coherent Non-opinionated can be summarized, but are distinct with respect to the level of subjectivity in the tweets, while Incoherent clusters cannot be summarized.

In the second stage, *information nuggets* are defined in a cluster as important pieces of information to aid in its summarization. The summarizers were asked to highlight information nuggets when available and categorise their aspect in terms of: WHAT, WHO, WHERE, REACTION, and OTHER. Thus, each information nugget is a pair consisting of the text and its aspect category (see *Appendix A* for an example). Inspired by the pyramid evaluation framework (Nenkova and Passonneau, 2004) and extractive-abstractive two-stage models in the summarization literature (Lebanoff et al., 2018; Rudra et al., 2019; Liu et al., 2018), information nuggets have a dual purpose: (1) helping summarizers create the final summary and (2) constituting an extractive reference for summary informativeness evaluation (See 5.2.1).

In the third and final stage of the process, the summarizers were asked to write a short

[2]Limited resources available for annotation determined the size of the MOS corpus.

	Total	COVID-19	Election
Size (#clusters)	3100	1550	1550
Coherent Opinionated	42%	41%	43%
Coherent Non-opinionated	30%	24%	37%
Incoherent	28%	35%	20%

Table 2: Annotation statistics of our MOS corpus.

	R-1$_{f_1}$	R-2$_{f_1}$	R-L$_{f_1}$	BLEURT
Summary	37.46	17.91	30.16	−.215
Main Story	35.15	12.98	34.59	−.324
Majority Opinion	27.53	6.15	25.95	−.497
Minority Opinion(s)	22.90	5.10	24.39	−.703

Table 3: Agreements between summarizers wrt to final summary, main story, majority opinion and minority opinions using ROUGE-1,2,L and BLEURT.

template-based summary for coherent clusters. Our chosen summary structure diverges from current summarization approaches that reconstruct the ''most popular opinion'' (Bražinskas et al., 2020; Angelidis et al., 2021). Instead, we aim to showcase a spectrum of diverse opinions regarding the same event. Thus, the summary template comprises three components: *Main Story*, *Majority Opinion*, *Minority Opinion(s)*. The component *Main Story* serves to succinctly present the focus of the cluster (often an event), while the other components describe opinions about the main story. Here, we seek to distinguish the most popular opinion (*Majority opinion*) from ones expressed by a minority (*Minority opinions*). This structure is consistent with the work of Gerani et al. (2014) in template-based summarization for product reviews, which quantifies the popularity of user opinions in the final summary.

For ''Coherent Opinionated clusters'', summarizers were asked to identify the majority opinion within the cluster and, if it exists, to summarize it, along with any minority opinions. If a majority opinion could not be detected, then the minority opinions were summarized. The final summary of ''Coherent Opinionated clusters'' is the concatenation of the three components: *Main story* + *Majority Opinion* (if any) + *Minority Opinion(s)* (if any). In 43% of opinionated clusters in our MOS corpus a majority opinion and at least one minority opinion were identified. Additionally, in 12% of opinionated clusters, 2 or more main opinions were identified (See Appendix C, Table 13), but without a majority opinion as there is a clear divide between user reactions. For clusters with few or no clear opinions (Coherent Non-opinionated), the final summary is represented by the *Main Story* component. Statistics regarding the annotation results are shown in Table 2.

Agreement Analysis

Our tweet summarization corpus consists of 3100 clusters. Of these, a random sample of 100 clusters was shared among all three summarizers to compute agreement scores. Each then worked on 1000 clusters.

We obtain a Cohen's Kappa score of $\kappa = 0.46$ for the first stage of the summary creation process, which involves categorising clusters as either Coherent Opinionated, Coherent Non-opinionated or Incoherent. Previous work (Feinstein and Cicchetti, 1990) highlights a paradox regarding Cohen's kappa in that high levels of agreement do not translate to high kappa scores in cases of highly imbalanced datasets. In our data, at least 2 of the 3 summarizers agreed on the type of cluster in 97% of instances.

In addition, we evaluate whether the concept of 'coherence/summarizability' is uniformly assessed, that is, we check whether summarizers agree on what clusters can be summarized (Coherent clusters) and which clusters are too incoherent. We find that 83 out of 100 clusters were evaluated as coherent by the majority, of which 65 were evaluated as uniformly coherent by all.

ROUGE-1,2,L and BLEURT (Sellam et al., 2020) are used as proxy metrics to check the agreement in terms of summary similarity produced between the summarizers. We compare the consensus between the complete summaries as well as individual components such as the main story of the cluster, its majority opinion and any minority opinions in Table 3. The highest agreement is achieved for the Main Story, followed by Majority Opinion and Minority Opinions. These scores can be interpreted as upper thresholds for the lexical and semantic overlap later in Section 6.

3.3 Comparison with Other Twitter Datasets

We next compare our corpus against the most recent and popular Twitter datasets for summarization in Table 4. To the best of our knowledge there are currently no abstractive summarization Twitter datasets for either event or opinion summarization. While we primarily focussed on the

Dataset	Time span	#keywords	#clusters	Avg. Cluster Size (#posts)	Summary	Avg. Summary Length (#tokens)
COVID-19	1 year	41	1003	31	Abstractive	42
Election	2 years	112	1236	30	Abstractive	36
Inouye and Kalita (2011)	5 days	50	200	25	Extractive	17
SMERP (Ghosh et al., 2017)	3 days	N/A	8	359	Extractive	303
TSix (Nguyen et al., 2018)	26 days	30	925	36	Extractive	109

Table 4: Overview of other Twitter datasets.

collection of opinionated clusters, some of the clusters we had automatically identified as opinionated were not deemed to be so by our annotators. Including the non-opinionated clusters helps expand the depth and range of Twitter datasets for summarization.

Compared to the summarization of product reviews and news articles, which has gained recognition in recent years because of the availability of large-scale datasets and supervised neural architectures, Twitter summarization remains a mostly uncharted domain with very few datasets curated. Inouye and Kalita (2011)[3] collected the tweets for the top ten trending topics on Twitter for 5 days and manually clustered these. The SMERP dataset (Ghosh et al., 2017) focuses on topics on post-disaster relief operations for the 2016 earthquakes in central Italy. Finally, TSix (Nguyen et al., 2018) is the dataset most similar to our work as it covers, but on a smaller scale, several popular topics that are deemed relevant to news providers.

Other Twitter summarization datasets include: (Zubiaga et al., 2012; Corney et al., 2014) on summarization of football matches, (Olariu, 2014) on real-time summarization for Twitter streams. These datasets are either publicly unavailable or unsuitable for our summarization task.[4]

Summary Type. These datasets exclusively contain extractive summaries, where several tweets are chosen as representative per cluster. This results in summaries which are often verbose, redundant and information-deficient. As shown in other domains (Grusky et al., 2018; Narayan et al., 2018), this may lead to bias towards extractive summarization techniques and hinder progress for abstractive models. Our corpus on COVID-19 and Election data aims to bridge this gap and introduces an abstractive gold standard generated by journalists experienced in sub-editing.

Size. The average number of posts in our clusters is 30, which is similar to the TSix dataset and in line with the empirical findings by Inouye and Kalita (2011), who recommend 25 tweets/cluster. Having clusters with a much larger number of tweets makes it harder to apply our guidelines for human summarization. To the best of our knowledge, our combined corpus (COVID-19 and Election) is currently the biggest human-generated corpus for microblog summarization.

Time-span. Both COVID-19 and Election partitions were collected across year-long time spans. This is in contrast to other datasets, which have been constructed in brief time windows, ranging from 3 days to a month. This emphasizes the longitudinal aspect of the dataset, which also allows topic diversity as 153 keywords and accounts were tracked through time.

4 Defining Model Baselines

As we introduce a novel summarization task (MOS), the baselines featured in our experiments are selected from domains tangential to microblog opinion summarization, such as news articles, Twitter posts, and product reviews (See Section 2). In addition, the selected models represent diverse summarization strategies: abstractive or extractive, supervised or unsupervised, multi-document (MDS) or single-document summarization (SDS). Note that most SDS models enforce a length limit (1024 characters) over the input, which makes it impossible to summarize the whole cluster of tweets. We address this issue by only considering the most relevant tweets ordered by topic relevance. The latter is computed using the Kullback-Leibler divergence with respect to the topical word

[3]It is unclear whether the full corpus is available: Our statistics were calculated based on a sample of 100 posts for each topic, but the original paper mentions that 1500 posts for each topic were initially collected.

[4]Comparing to live stream summarization where millions of posts are used as input, we focus on summarization of clusters of maximum 50 posts.

distribution of the cluster in the GSDMM-LDA clustering algorithm (Wang et al., 2017b).

The summaries were generated such that their length matches the average length of the gold standard. Some model parameters (such as Lexrank) only allow sentence-level truncation, in which case the length matches the average number of sentences in the gold standard. For models that allow a word limit to the text to be generated (BART, Pegasus, T5), a minimum and maximum number of tokens was imposed such that the generated summary would be within [90%, 110%] of the gold standard length.

4.1 Heuristic Baselines

Extractive Oracle: This baseline uses the gold summaries to extract the highest scoring sentences from a cluster of tweets. We follow Zhong et al. (2020) and rank each sentence by its average ROUGE-{1,2,L} recall scores. We then consider the highest ranking 5 sentences to form combinations of k^5 sentences, which are re-evaluated against the gold summaries. k is chosen to equal the average number of sentences in the gold standard. The highest scoring summary with respect to the average ROUGE-{1,2,L} recall scores is assigned as the oracle.

Random: k sentences are extracted at random from a tweet cluster. We report the mean result over 5 iterations with different random seeds.

4.2 Extractive Baselines

LexRank (Erkan and Radev, 2004) constructs a weighed connectivity graph based on cosine similarities between sentence TF-IDF representations.

Hybrid TF-IDF (Inouye and Kalita, 2011) is an unsupervised model designed for Twitter, where a post is summarized as the weighted mean of its TF-IDF word vectors.

BERTSumExt (Liu and Lapata, 2019) is an SDS model comprising a BERT (Devlin et al., 2019)-based encoder stacked with Transformer layers to capture document-level features for sentence extraction. We use the model trained on CNN/Daily Mail (Hermann et al., 2015).

HeterDocSumGraph (Wang et al., 2020b) introduces the heterogenous graph neural network, which is constructed and iteratively updated using both sentence nodes and nodes representing other semantic units, such as words. We use the MDS model trained on Multi-News (Fabbri et al., 2019).

Quantized Transformer (Angelidis et al., 2021) combines Transformers (Vaswani et al., 2017) and Vector-Quantized Variational Autoencoders for the summarization of popular opinions in reviews. We trained QT on the MOS corpus.

4.3 Abstractive Baselines

Opinosis (Ganesan et al., 2010) is an unsupervised MDS model. Its graph-based algorithm identifies valid paths in a word graph and returns the highest scoring path with respect to redundancy.

PG-MMR (Lebanoff et al., 2018) adapts the single document setting for multi-documents by introducing 'mega-documents' resulting from concatenating clusters of texts. The model combines an abstractive SDS pointer-generator network with an MMR-based extractive component.

PEGASUS (Zhang et al., 2020) introduces gap-sentences as a pre-training objective for summarization. It is then fine-tuned for 12 downstream summarization domains. We chose the model pre-trained on Reddit TIFU (Kim et al., 2019).

T5 (Raffel et al., 2020) adopts a unified approach for transfer learning on language-understanding tasks. For summarization, the model is pre-trained on the Colossal Clean Crawled Corpus (Raffel et al., 2020) and then fine-tuned on CNN/Daily Mail.

BART (Lewis et al., 2020) is pre-trained on several evaluation tasks, including summarization. With a bidirectional encoder and GPT2, BART is considered a generalization of BERT. We use the BART model pre-trained on CNN/Daily Mail.

SummPip (Zhao et al., 2020) is an MDS unsupervised model that constructs a sentence graph following Approximate Discourse Graph and deep embedding methods. After spectral clustering of the sentence graph, summary sentences are generated through a compression step of each cluster of sentences.

Copycat (Bražinskas et al., 2020) is a Variational Autoencoder model trained in an unsupervised setting to capture the consensus opinion in

[5]For opinionated clusters, we set k=3 and for non-opinionated k=1.

product reviews for Yelp and Amazon. We train it on the MOS corpus.

5 Evaluation Methodology

Similar to other summarization work (Fabbri et al., 2019; Grusky et al., 2018), we perform both automatic and human evaluation of models. Automatic evaluation is conducted on a set of 200 clusters: Each partition of the test (COVID-19 Opinionated, COVID-19 Non-opinionated, Election Opinionated, Election Non-opinionated) contains 50 clusters uniformly sampled from the total corpus. For the human evaluation, only the 100 opinionated clusters are evaluated.

5.1 Automatic Evaluation

Word overlap is evaluated according to the harmonic mean F_1 of ROUGE-1, 2, L[6] (Lin, 2004) as reported elsewhere (Narayan et al., 2018; Gholipour Ghalandari et al., 2020; Zhang et al., 2020). Work by Tay et al. (2019) acknowledges the intractability of ROUGE in opinion text summarization as sentiment-rich language uses a vast vocabulary that does not rely on word matching. This issue is mitigated by Kryscinski et al. (2021) and Bhandari et al. (2020), who use semantic similarity as an additional assessment of candidate summaries. Similarly, we use text generation metrics BLEURT (Sellam et al., 2020) and BERTScore[7] (Zhang et al., 2020) to assess semantic similarity.

5.2 Human Evaluation

Human evaluation is conducted to assess the quality of summaries with respect to three objectives: 1) linguistic quality, 2) informativeness, and 3) ability to identify opinions. We conducted two human evaluation experiments: the first (5.2.1) assesses the gold standard and non-fine-tuned model summaries on a rating scale, and the second (5.2.2) addresses the advantages and disadvantages of fine-tuned model summaries via Best-Worst Scaling. Four and three experts were employed for the two experiments, respectively.

[6]We use ROUGE-1.5.5 via the *pyrouge* package: https://github.com/bheinzerling/pyrouge.

[7]BERTScore has a narrow score range, which makes its interpretation more difficult than for BLEURT. Because both metrics produce similar rankings, BERTScore can be found in Appendix C.

5.2.1 Evaluation of Gold Standard and Models

The first experiment focused on assessing the gold standard and best models from each summarization type: Gold, LexRank (best extractive), SummPip (best unsupervised abstractive), and BART (best supervised).

Linguistic quality measures 4 syntactic dimensions, which were inspired by previous work on summary evaluation. Similar to DUC (Dang, 2005), each summary was evaluated with respect to each criterion below on a 5-point scale.

- *Fluency* (Grusky et al., 2018): Sentences in the summary ''should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.''
- *Sentential Coherence* (Grusky et al., 2018): A sententially coherent summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.
- *Non-redundancy* (Dang, 2005): A non-redundant summary should contain no duplication, that is, there should be no overlap of information between its sentences.
- *Referential Clarity* (Dang, 2005): It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role is in the story.

Informativeness is defined as the amount of factual information displayed by a summary. To measure this, we use a Question-Answer algorithm (Patil, 2020) as a proxy. Pairs of questions and corresponding answers are generated from the information nuggets of each cluster. Because we want to assess whether the summary contains factual information, only information nuggets belonging to the 'WHAT', 'WHO', 'WHERE' are selected as input. We chose not to include the entire cluster as input for the QA algorithm, as this might lead the informativeness evaluation to

prioritize irrelevant details in the summary. Each cluster in the test set is assigned a question-answer pair and each system is then scored based on the percentage of times its generated summaries contain the answer to the question. Similar to factual consistency (Wang et al., 2020a), informativeness penalizes incorrect answers (hallucinations), as well as the lack of a correct answer in a summary.

As **Opinion** is a central component for our task, we want to assess the extent to which summaries contain opinions. Assessors report whether summaries identify any majority or minority opinions.[8] A summary contains a majority opinion if most of its sentences express this opinion or if it contains specific terminology ('The majority is/ Most users think...', etc.), which is usually learned during the fine-tuning process. Similarly, a summary contains a minority opinion if at least one of its sentences expresses this opinion or it contains specific terminology ('A minority/ A few users', etc.). The final scores for each system are the percentage of times the summaries contain majority or minority opinions, respectively.

5.2.2 Best-Worst Evaluation of Fine-tuned Models

The second human evaluation assesses the effects of fine-tuning on the best supervised model, BART. The experiments use non-fine-tuned BART (BART), BART fine-tuned on 10% of the corpus (BART FT10%) and BART fine-tuned on 70% of the corpus (BART FT70%).

As all the above are versions of the same neural model, Best-Worst scaling is chosen to detect subtle improvements, which cannot otherwise be quantified as reliably by traditional ranking scales (Kiritchenko and Mohammad, 2017). An evaluator is shown a tuple of 3 summaries (BART, BART FT70%, BART FT30%) and asked to choose the best/worst with respect to each criteria. To avoid any bias, the summary order is randomized for each document following van der Lee et al. (2019). The final score is calculated as the percentage of times a model is scored as the best, minus the percentage of times it was selected as the worst (Orme, 2009). In this setting, a score of 1 represents the unanimously best, while -1 is unanimously the worst.

[8]Note that whether the identified minority or majority opinions are correct is not evaluated here. This is done in Section 5.2.2.

The same criteria as before are used for **linguistic quality** and one new criterion is added to assess **Opinion**. We define *Meaning Preservation* as the extent to which opinions identified in the candidate summaries match the ones identified in the gold standard. We draw a parallel between the *Faithfulness* measure (Maynez et al., 2020), which assesses the level of hallucinated information present in summaries, and *Meaning Preservation*, which assesses the extent of hallucinated opinions.

6 Results

6.1 Automatic Evaluation

Results for the automatic evaluation are shown in Table 5.

Fine-tuned Models Unsurprisingly, the best performing models are ones that have been fine-tuned on our corpus: *BART (FT70%)* and *BART (FT10%)*. Fine-tuning has been shown to yield competitive results for many domains (Kryscinski et al., 2021; Fabbri et al., 2021), including ours. In addition, one can see that only the fine-tuned abstractive models are capable of outperforming the *Extractive Oracle*, which is set as the upper threshold for extractive methods. Note that on average, the Oracle outperforms the Random summarizer by a 59% margin, which only fine-tuned models are able to improve on, with 112% for *BART (FT10%)* and 114% for *BART (FT70%)*. We hypothesize that our gold summaries' template format poses difficulties for off-the-shelf models and fine-tuning even on a limited portion of the corpus produces summaries that follow the correct structure (See Table 9 and Appendix C for examples). We include comparisons between the performance of *BART (FT10%)* and *BART (FT70%)* on the individual components of the summary in Table 6.[9]

Non-Fine-tuned Models Of these, *SummPip* performs the best across most metrics and datasets with an increase of 37% in performance over the random model, followed by *LexRank* with an increase of 29%. Both models are designed for the multi-document setting and benefit from the common strategy of mapping each sentence

[9]We do not include other models in the summary component-wise evaluation because it is impossible to identify the Main Story, Majority Opinion, and Minority Opinions in non-fine-tuned models.

Models	COVID-19 Opinionated (CO)				COVID-19 Non-opinionated (CNO)				Election Opinionated (EO)				Election Non-opinionated (ENO)			
	R-1 f_1	R-2 f_1	R-L f_1	BLEURT	R-1 f_1	R-2 f_1	R-L f_1	BLEURT	R-1 f_1	R-2 f_1	R-L f_1	BLEURT	R-1 f_1	R-2 f_1	R-L f_1	BLEURT
Heuristics																
Gold (195 char)																
Random Sentences (204 char)	13.55	1.09	9.22	−.660	7.30	0.70	5.97	−.968	11.82	0.80	8.27	−.576	6.75	0.89	5.69	−.592
Extractive Oracle (289 char)	15.45	1.67	10.29	−.382	11.80	1.38	9.27	−.510	15.33	1.60	10.12	**−.146**	10.06	**2.15**	8.46	**−.056**
Extractive Unsupervised Models																
LexRank (265 char)	16.41	1.48	10.89	−.560	10.87	1.01	8.76	−.849	14.27	1.15	9.62	−.418	9.11	1.08	7.41	−.456
Hybrid TF-IDF (277 char)	12.87	1.26	8.85	−.608	9.33	0.83	7.51	−.745	12.06	1.12	8.42	−.430	7.93	1.13	6.56	−.298
Quantized Transformer (273 char)	14.23	1.03	9.55	−.621	9.85	0.96	7.83	−.857	14.78	1.08	9.45	−.468	8.69	0.81	6.79	−.668
Extractive Supervised Models																
BERTSumExt (225 char)	14.22	1.31	9.68	−.571	9.78	1.11	7.70	−.699	11.93	1.10	8.47	−.384	8.06	1.00	6.63	−.407
HeterDocSumGraph (295 char)	15.13	1.19	9.79	−.748	10.05	0.88	7.79	−.867	14.28	0.96	9.15	−.564	8.40	0.72	6.86	−.626
Abstractive Unsupervised Models																
Opinosis (215 char)	12.45	1.14	8.86	−.534	8.35	0.73	6.99	−.673	11.34	1.00	8.15	−.537	6.69	0.95	5.66	−.518
SummPip (236 char)	12.96	1.37	9.32	−.488	11.30	1.46	9.09	−.559	13.05	1.15	8.90	−.409	9.93	1.36	7.74	−.228
Copycat (153 char)	12.47	1.31	9.41	−.552	10.99	1.32	9.25	−.621	14.05	1.56	10.25	−.503	7.48	1.10	6.36	−.316
Abstractive Supervised Models																
PG-MMR (238 char)	11.93	1.08	8.93	−.450	9.68	1.37	8.01	−.578	12.36	1.07	8.73	−.400	8.14	1.04	6.86	−.302
Pegasus (216 char)	13.78	1.40	9.78	−.535	10.37	1.41	8.61	−.616	12.68	1.23	9.28	−.481	9.12	1.11	7.34	−.283
T5 (206 char)	14.25	1.31	9.97	−.530	9.11	1.21	7.72	−.669	12.99	1.06	8.82	−.470	8.59	1.15	7.06	−.347
BART (237 char)	15.95	1.46	10.74	−.521	10.41	1.55	8.48	−.576	13.71	1.18	9.09	−.409	9.11	1.15	7.37	−.372
Fine-tuned Models																
BART (FT 10%) (245 char)	21.53	**3.86**	**14.76**	**−.257**	**15.49**	**2.61**	12.04	−.449	19.77	2.99	13.11	−.209	12.31	**1.87**	**9.62**	−.081
BART (FT 70%) (246 char)	**21.54**	3.74	14.54	−.259	15.31	2.54	**12.09**	**−.439**	**20.59**	**3.42**	**13.63**	**−.183**	**12.37**	1.72	9.58	**−.071**

Table 5: Performance on the **test set** of baseline models evaluated with automatic metrics: ROUGE-n (R-n) and BLEURT. The best model from each category (Extractive, Abstractive, Fine-tuned) and **overall** are highlighted.

Models	COVID-19 Opinionated (CO)				Election Opinionated (EO)			
	R-1 f_1	R-2 f_1	R-L f_1	BLEURT	R-1 f_1	R-2 f_1	R-L f_1	BLEURT
Main Story								
BART (FT 10%)	**11.43**	**2.49**	**9.95**	**−.082**	**9.82**	**1.72**	**8.31**	−.185
BART (FT 70%)	11.18	2.29	9.57	−.137	9.55	**1.70**	8.19	**−.104**
Majority Opinion								
BART (FT 10%)	**20.25**	**4.28**	**16.86**	**−.487**	17.88	3.11	14.57	−.442
BART (FT 70%)	16.18	−.505	**19.13**	**3.74**	**15.60**	**−.392**		
19.74 4.06								
Minority Opinion(s)								
BART (FT 10%)	**19.05**	4.66	**15.87**	**−.544**	15.26	3.97	13.34	−.791
BART (FT 70%)	18.70	**4.81**	**15.83**	−.643	**15.98**	**4.63**	**14.01**	**−.604**

Table 6: Performance of fine-tuned models per each summary component (Main Story, Majority Opinion, Minority Opinion(s)) on the **test set** evaluated with automatic metrics: ROUGE-n (R-n) and BLEURT.

in a tweet from the cluster into a node of a sentence graph. However, not all graph mappings prove to be useful: Summaries produced by *Opinosis* and *HeterDocSumGraph*, which employ a word-to-node mapping, do not correlate well with the gold standard. The difference between word and sentence-level approaches can be partially attributed to the high amount of spelling variation in tweets, making the former less reliable than the latter.

ROUGE vs BLEURT The performance on ROUGE and BLEURT is tightly linked to the data differences between COVID-19 and Election partitions of the corpus. Most models achieve higher ROUGE scores and lower BLEURT scores on the COVID-19 than on the Election dataset. An inspection of the data differences reveals that COVID-19 tweets are much longer than Election ones (169 *vs* 107 characters), as the latter had been collected before the increase in length limit from 140 to 280 characters in Twitter posts. This is in line with findings by Sun et al. (2019), who revealed that high ROUGE scores are mostly the result of longer summaries rather than better quality summaries.

6.2 Human Evaluation

Evaluation of Gold Standard and Models Table 7 shows the comparison between the gold standard and the best performing models against a set of criteria (See 5.2.1). As expected, the human-authored summaries (Gold) achieve the highest scores with respect to all linguistic quality and structure-based criteria. However, the gold standard fails to capture informativeness as well as its automatic counterparts, which are, on average, longer and thus may include more information. Since *BART* is previously pre-trained

Model	Fluency	Sentential Coherence	Non-redundancy	Referential Clarity	Informativeness	Majority	Minority
Gold	4.52	4.63	4.85	4.31	57%	86%	64%
Lexrank	3.03	2.43	3.10	2.55	58%	15%	**62%**
BART	**3.24**	**2.76**	**3.46**	3.01	67%	8%	60%
SummPip	2.73	2.70	2.53	**3.37**	**69%**	**32%**	36%

Table 7: Evaluation of Gold Standard and Models: Results.

Model	Fluency	Sentential Coherence	Non-redundancy	Referential Clarity	Meaning Preservation
BART	−0.76	−0.65	**0.15**	−0.42	−0.54
BART FT 10%	0.30	0.22	−0.11	**0.25**	0.14
BART FT 70%	**0.44**	**0.43**	−0.04	0.17	**0.40**

Table 8: Best-Worst Evaluation of Fine-tuned models: Results.

on CNN/DM dataset of news articles, its output summaries are more fluent, sententially coherent and contain less duplication than the unsupervised models *Lexrank* and *SummPip*. We hypothesize that *SummPip* achieves high referential clarity and majority scores as a trade-off for its very low non-redundancy (high redundancy).

Best-Worst Evaluation of Fine-tuned Models The results for our second human evaluation are shown in Table 8 using the guidelines presented in 5.2.2. The model fine-tuned on more data *BART (FT70%)* achieves the highest fluency and sentential coherence scores. As seen in Table 9, the summary produced by *BART (FT70%)* contains complete and fluent sentences, unlike its counterparts. Most importantly, fine-tuning yields better alignment with the gold standard with respect to meaning preservation, as the fine-tuned models *BART (FT70%)* and *BART (FT10%)* learn how to correctly identify and summarize the main story and the relevant opinions in a cluster of tweets. In the specific example, non-fine-tuned *BART* introduces a lot of irrelevant information ('industrial air pollution','google, apple rolling out covid') to the main story and offers no insight into the opinions found in the cluster of tweets, whereas both fine-tuned models correctly introduce the Main Story and both partially identify the Majority Opinion ('great idea' for anti-maskers 'to dig graves'). However, we note that the fine-tuning process does not lead to increased performance with respect to all criteria; non-redundancy is compromised and referential clarity stops improving after a certain amount of training data. As observed in the example, *BART (FT70%)* contains duplicated content 'think this is a great idea. What a great idea!'. Wilcoxon signed rank tests

Human Summary: Anti-maskers are forced to dig graves for Covid19 victims in Indonesia. The majority of Twitter users think it is a good idea that will cause a reality check and make people rethink their ideas.

BART FT 70%: Anti-maskers in Indonesia are forced to dig graves for Covid-19 victims as punishment for not wearing mask. The majority think this is a great idea. What a great idea! It's a good way to get people to reconsider their misinformed opinions. We should do this in the US.

BART FT 10%: Anti-maskers forced to dig graves for Covid-19 victims in Indonesia as punishment for refusing to wear a mask. The majority believe that this is a good idea, and that it will hopefully bring about an end to the need for masks. A minority do not believe it is necessary to wear a

BART: Covid-19 can be more deadly for people living in areas with industrial air pollution. Anyone refusing to wear a mask is made to dig graves for covid-19 victims as punishment in east java. as domestic violence spikes, many victims and their children have nowhere to live. google, apple rolling out covid-

Table 9: BART Summary Examples for the same cluster of tweets.

with $p < 0.05$ and $p < 0.10$ are used for significance testing between all pairs of models. We note that most pairwise differences are significant at $p < 0.05$, while *BART (FT70%)* and BART NFT differences are significant at $p < 0.10$ for non-redundancy. The only two exceptions are referential clarity and non-redundancy between *BART (FT70%)* and *BART (FT10%)* where both fine-tuned models perform similarly.

7 Error Analysis

Error analysis is carried out on 30 fine-tuned BART summaries from a set of 15 randomly sampled clusters. The results are found in Table 10.

Hallucination Fine-tuning on the MOS corpus introduces hallucinated content in 8 out of 30 manually evaluated summaries. Generated summaries contain opinions that prove to be either false or unfounded after careful inspection of the cluster of tweets. We follow the work of Maynez et al. (2020) in classifying hallucinations as either intrinsic (incorrect synthesis of information in the source) or extrinsic (external information

Error type	Freq.	Example
Intrinsic Hallucination	4/30	**Example 1** **Generated Summary**: United States surpasses six million coronavirus cases and deaths and remains at the top of the global list of countries with the most cases and deaths The majority are pleased to see the US still leads the world in terms of cases and deaths, with 180,000 people succumbing to Covid-19.
Extrinsic Hallucination	4/30	**Example 2** **Generated Summary**: Sex offender Rolf Harris is involved in a prison brawl after absconding from open jail. The majority think Rolf Harris deserves to be spat at and called a ''nonce'' and a ''terrorist'' for absconding from open prison. A minority are putting pressure on
Information Loss	12/30	**Example 3** **Human Summary**: Miley Cyrus invited a homeless man on stage to accept her award. Most people thought it was a lovely thing to do and it was emotional. A minority think that it was a publicity stunt. **Generated Summary**: Miley Cyrus had homeless man accept Video of the Year award at the MTV Video Music Awards. The majority think it was fair play for Miley Cyrus to allow the homeless man to accept the award on her behalf. She was emotional and selfless. The boy band singer cried and thanked him for accepting the

Table 10: Error Analysis: Frequency of errors and representative summary examples for each error type.

not found in the source). Example 1 in Table 10 is an instance of an intrinsic hallucination: The majority opinion is wrongly described as 'pleased', despite containing the correct facts regarding US coronavirus cases. Next, Example 2 shows that Rolf Harris 'is called a terrorist', which is confirmed to be an extrinsic hallucination as none of the tweets in the source cluster contain this information.

Information Loss Information loss is the most frequent error type. As outlined in Kryscinski et al. (2021), the majority of current summarization models face length limitations (usually 1024 characters) which are detrimental for long-input documents and tasks. Since our task involves the detection of all opinions within the cluster, this weakness may lead to incomplete and less informative summaries, as illustrated in Example 3 from Table 10. The candidate summary does not contain the minority opinion identified by the experts in the gold standard. An inspection of the cluster of tweets reveals that most posts expressing this opinion are indeed not found in the first 1024-character allowed limit of the cluster input.

8 Conclusions and Future Work

We have introduced the task of Twitter opinion summarization and constructed the first abstractive corpus for this domain, based on template-based human summaries. Our experiments show that existing extractive models fall short on linguistic quality and informativeness while abstractive models perform better but fail to identify all relevant opinions required by the task. Fine-tuning on our corpus boosts performance as the models learn the summary structure.

In the future, we plan to take advantage of the template-based structure of our summaries to refine fine-tuning strategies. One possibility is to exploit style-specific vocabulary during the generation step of model fine-tuning to improve on capturing opinions and other aspects of interest.

Acknowledgments

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) through project funding and its Enrichment PhD Scheme. We are grateful to our reviewers and action editor for reading our paper carefully and critically and thank them for their insightful comments and suggestions. We would also like to thank our annotators for their invaluable expertise in constructing the corpus and completing the evaluation tasks.

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. When the corpus will be released to the research community, only tweet IDs will be made available along with associated cluster membership and summaries. 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.

Appendix A

Summary Annotation Interface

Stage 1: Reading and choosing cluster type

The majority of the tweets in the cluster revolve around the subject of Trident nuclear submarines. The cluster contains many opinions which can be summarized easily, hence this cluster is *Coherent Opinionated*. Choose 'Yes' and proceed to the next step.

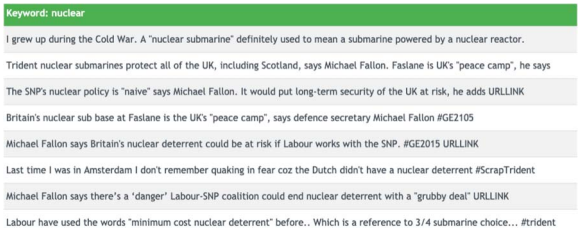

Figure 1: Fragment of a cluster of tweets for keyword 'nuclear'.

Figure 2: Choose type of cluster 'Coherent Opinionated'.

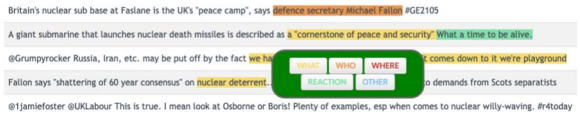

Figure 3: Example of information nuggets: 'a cornerstone of peace and security' describes the nuclear submarine (WHAT information nugget), while 'defence secretary Michael Fallon' describes a person (WHO information nugget).

Stage 2: Highlighting information nuggets

Highlight important information and select the relevant aspect each information nugget belongs to.

Stage 3: Template-based Summary Writing

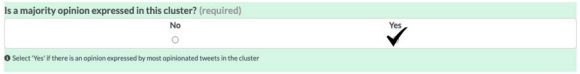

Figure 4: Choose whether there exists a majority opinion in the cluster.

Most user reactions dismiss the Trident plan and view it as an exaggerated security measure. This forms the *Majority Opinion*. A few users express fear for UK's potential future in a nuclear war. This forms a *Minority Opinion*.

Write cluster summary following the structure: Main Story + Majority Opinion (+ Minority Opinions).

Appendix B

Complete Results: BERTScore Evaluation

Model Implementation Details

T5, Pegasus, and BART were implemented using the HuggingFace Transformer package (Wolf et al., 2020) with max sequence length of 1024 characters.

Fine-tuning parameters for BART are: 8 batch size, 5 training epochs, 4 beams, enabled early stopping, 2 length penalty, and no trigram repetition for the summary generation. The rest of the parameters are set as default following the configuration of *BartForConditionalGeneration*: activation function gelu, vocabulary size 50265, 0.1 dropout, early stopping, 16 attention heads, 12 layers with feed forward layer dimension set as 4096 in both decoder and encoder. Quantized Transformer and Copycat models are trained for 5 epochs.

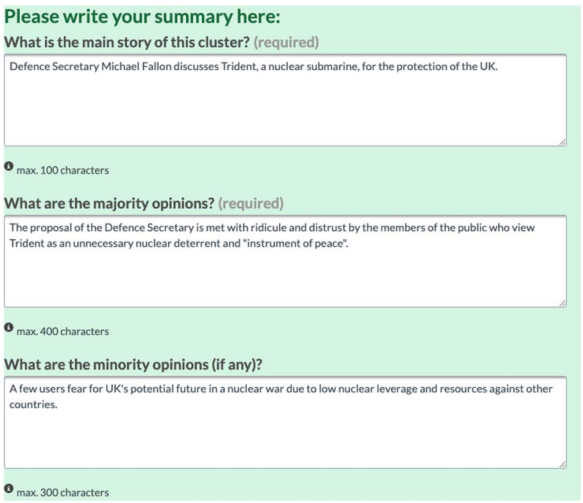

Figure 5: Summary template of a Coherent Opinionated cluster with a majority opinion.

Models	COVID-19 Opinionated	COVID-19 Non-opinionated	Election Opinionated	Election Non-opinionated
Heuristics				
Random Sentences	0.842	0.838	0.846	0.861
Extractive Oracle	0.858	0.867	0.871	**0.904**
Extractive Models				
LexRank	0.851	0.849	0.856	0.868
Hybrid TF-IDF	0.851	0.853	0.856	0.879
BERTSumExt	0.848	0.851	0.859	0.874
HeterDocSumGraph	0.839	0.840	0.847	0.853
Quantized Transformer	0.840	0.827	0.850	0.856
Abstractive Models				
Opinosis	0.845	0.853	0.846	0.860
PG-MMR	0.853	0.857	0.851	0.863
Pegasus	0.850	0.856	0.852	0.869
T5	0.850	0.851	0.853	0.872
BART	0.852	0.854	0.856	0.868
SummPip	0.852	0.858	0.854	0.878
Copycat	0.848	0.852	0.848	0.872
Fine-tuned Models				
BART (FT 10%)	**0.873**	**0.870**	0.875	**0.893**
BART (FT 70%)	**0.873**	**0.870**	**0.878**	0.892

Table 11: Performance on **test set** of baseline models evaluated with BERTScore.

Appendix C

Cluster examples and summaries from the MOS Corpus

Tweet cluster fragment for keyword ‘‘CDC’’
Gosh i hope these cases are used for the negligent homicide class action suit that’s being constructed against trump. cdc warns against drinking hand sanitizer amid reports of deaths
the cdc has also declared, b̈eing stupid is hazardous to your health.ÜRLLINK
cdc warning! do not drink hand sanitizer! what the hell! people be idiots!
cdc warns against drinking hand sanitizer amid reports of deaths seriously omg?!
if the cdc has to put out a health bulletin to inform people not to try drinking hand sanitizers, how stupid are those people?
from the ‘‘if you had any doubt’’ department: the cdc is alerting your fellow americans not to drink hand sanitizer. obviously more than a couple of people have had to be treated for it. I wonder were they poisoned in the womb, too many concussions, mt. dew in their milk bottle when they were babies?
oh my...the cdc actually had to warn people not to drink hand sanitizer. only under a trump presidency have people acted so stupidly.
@realdonaldtrump you should try drinking the hand sanitizer. After your ridiculous suggestion to inject disinfectants, people have decided to drink it and are dying. CDC now issued a warning not to drink it. since u don’t believe anything the scientists say go ahead and drink it. First get kids out of cages
@USER i think this actually speaks more to the stupidity of the cdc.
@USER trump is in control of the cdc. don’t believe a single word that they are saying
this is sadly what happens when you put an idiot like @realdonaldtrump in the white house...people had seizures, lost vision and dead after drinking hand sanitizer, cdc warns URLLINK
@cdcgov @usfda @USER is it really necessary to tell people not to ingest hand sanitizer????? jesus
cdc gives out warning that people are going blind and are dying from drinking hand sanitizer. #darwinism URLLINK
Human Written Summary **Main Story**: CDC warns people not to drink hand sanitiser after people fall ill and die after doing so. **Majority Opinion**: The majority are shocked by the stupidity of people thinking it’s OK to drink hand sanitiser. **Minority Opinions**: A minority blame the Trump administration for putting forward the idea of people disinfecting their insides. A minority think the warning is an act of stupidity by the CDC.
Lexrank Summary: cdc just issued a warning not to drink it. the cdc just issued an official warning to never drink hand sanitizer and i just. . . . people are having seizures, losing vision and dying after drinking hand sanitizer, cdc warns via @yahoo. if you need a warning from the cdc not to drink hand sanitizer. . . please go ahead and drink hand sanitizer.
SummPip Summary: people have decided to drink it and are dying after swallowing hand sanitizer . . . children masks . people are having seizures, losing vision and dying after drinking hand sanitizer, cdc warns via @yahoo .
BART Summary: People are getting sick and even dying after swallowing hand sanitizer, the us centers for disease control and prevention said wednesday. four died and others have suffered impaired vision or seizures, the cdc says. The cdc has also declared, ‘‘being stupid is hazardous to your health’’ @realdonaldtrump you
BART FT 10% **Summary**: People are dying and even dying after drinking hand sanitizer, the Centers for Disease Control warns. The majority are shocked and appalled by the CDC’s warning against people drinking hand Sanitizer. A minority believe the CDC is exaggerating the danger of drinking hand Sanitizinger to distract people from the
BART FT 70% **Summary**: CDC warns that people are dying and losing vision and blindness after drinking hand sanitiser. The majority believe that people are extremely stupid for even thinking about doing this, and it should not be allowed to happen. A minority do not believe the CDC’s warning and think it is not necessary to take any action

Table 12: Example of excerpt from tweet cluster ‘‘CDC’’, human summary and best generated summary candidates.

Tweet cluster for keyword ‘‘mental health’’
A ’landmark moment’? Nick Clegg (Lib Dems) promise to put mental health on par with physical #health URLLINK #inclusion #care
All of a sudden, Nick Clegg is concerned about people with mental health issues. Nothing at all to do with trying to win voters and save his job.
Delighted that nick is finally doing something about mental health in our nhs
Nick Clegg promises ’dignity and respect’ in NHS mental health treatment video URLLINK	Guardian
I have been hearing very positive noises on the radio today from Lib Dems re: mental health treatment. Certainly long overdue but great to hear!
But if you are patting Nick Clegg on the back for new mental health reforms, consider this:
Mate, Clegg could have stood up to Cameron before his harmful reductive mental health policies got implemented.
Awesome that Clegg highlighted mental health to rapturous applause, but sure he did that with tuition fees once.
.nickclegg speech #libdemconf focusing on mental health was cool. Araith Nick Clegg yn canolpwyntio ar iechyd meddyliol yn wych.
Nick Clegg’s pandering towards the treatment of mental health illness is kinda sad and pathetic#hecantbuyavote
One immediate victory of Clegg’s speech; putting mental health issues on the agenda and in the media. #ldconf #bbcnews
LibDems are back to promising the unachievable because they know they’re safe away from power. Shame because mental health is in dire state.
His position in government could have been used to stop the reductive mental health reforms Cameron put in years back. Did he? no.
Human Written Summary
Main Story: Nick Clegg promises to focus on mental health provision in the NHS. **Minority Opinions**: Some Twitter users are pleased something is ’finally’ being done about it and that it is great, it is highlighting mental health. Others are asking why he didn’t do it when he was in power and say that Clegg is doing it for personal gain.
Lexrank Summary: Nick Clegg promises ’dignity and respect’ in NHS mental health treatment video Speaking before his speech to the. . . Been hearing very positive noises on the radio today from Lib Dems re: mental health treatment. One immediate success of Clegg’s speech; getting mental health issues on the agenda and in the media. nickclegg a six week wait for mental health related treatment, but didn’t hear how you’ll resource the #NHS to achieve the needed care!
SummPip Summary: happy about nick clegg could have been used to stop the reductive mental health treatment . but if you are patting nick clegg is all of a sudden concerned about people with mental health issues . nick clegg promises ’ dignity and respect ’ in nhs mental health treatment video speaking before his speech to the . . . been hearing very positive noises on the radio today from lib dems re: mental health treatment .
BART Summary: Lib Dems promise to put mental health on par with physical health. Nick Clegg promises ’dignity and respect’ in NHS mental health treatment video. But if you are patting Nick Clegg on the back for new mental health reforms, consider this: Feeling blessed, trying to eradicate mental health stigma and getting lifetime opportunities
BART FT 10% **Summary**:Lib Dem Nick Clegg makes a speech about mental health in the NHS. The majority are pleased that the Lib Dem leader is trying to tackle the stigma attached to mental health. A minority are disappointed that he is pandering to the far right and anti-gay groups. A minority believe he is setting us up for a
BART FT 70% **Summary**: Lib Dem leader Nick Clegg makes a speech about putting mental health on a par with physical health in the manifesto. The majority are pleased that Nick Clegg is taking a lead on mental health and saying that mental health needs to be treated with dignity and respect. A minority are dismayed by Nick Clegg

Table 13: Example of excerpt from tweet cluster ‘‘mental health’’, human summary and best generated summary candidates.

References

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2021. Automatic summarization of real world events using Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):511–514.

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.175

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1403

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.751

Iman Munire Bilal, Bo Wang, Maria Liakata, Rob Procter, and Adam Tsakalidis. 2021. Evaluation of thematic coherence in microblogs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6800–6814, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.461

Ziqiang Cao, Chengyao Chen, Wenjie Li, Sujian Li, Furu Wei, and Ming Zhou. 2016. Tgsum: Build tweet guided multi-document summarization dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30. https://doi.org/10.1609/aaai.v30i1.10376

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. https://doi.org/10.2196/19273, PubMed: 32427106

Wen-Ying Sylvia Chou, April Oh, and William MP Klein. 2018. Addressing health-related misinformation on social media. *JAMA*, 320(23):2417–2418. https://doi.org/10.1001/jama.2018.16865, PubMed: 30428002

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *ICML*.

D. Corney, Carlos Martin, and Ayse Göker. 2014. Two sides to every story: Subjective event summarization of sports events using twitter. In *SoMuS@ICMR*.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Günes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479. https://doi.org/10.1613/jair.1523

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad.

2021. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.57

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1102

Alvan R. Feinstein and Dominic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–9. https://doi.org/10.1016/0895-4356(90)90158-L

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1168

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.120

Saptarshi Ghosh, Kripabandhu Ghosh, Tanmoy Chakraborty, Debasis Ganguly, Gareth Jones, and Marie-Francine Moens. 2017. First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP). In *Proceedings of the 39th European Conference on IR Research – J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193*, ECIR 2017, pages 779–783. Springer International Publishing AG. https://doi.org/10.1145/3130332.3130338

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1065

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883037

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

David Inouye and Jugal K. Kalita. 2011. Comparing Twitter summarization algorithms for multiple post summaries. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 298–306.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and

challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised abstractive opinion summarization by generating sentences with tree-structured topic guidance.

Myungha Jang and James Allan. 2018. Explaining controversy on social media via stance summarization. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1221–1224, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3209978.3210143

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.556

Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2019. Human values and attitudes towards vaccination in social media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 248–254.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *NAACL-HLT*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2074

Wojciech Kryscinski, Nazneen Fatema Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R. Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *CoRR*, abs/2105.08209.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1446

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.147

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.

Peter J. Liu, Mohammad Saleh, E. Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam M. Shazeer. 2018. Generating Wikipedia by summarizing long sequences. *International Conference on Learning Representations*, abs/1801.10198.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1387

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1003

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290. https://doi.org/10.18653/v1/K16-1028

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1206

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152.

Minh-Tien Nguyen, Dac Viet Lai, Huy-Tien Nguyen, and Le-Minh Nguyen. 2018. TSix: A human-involved-creation dataset for tweet summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Andrei Olariu. 2014. Efficient online summarization of microblogging streams. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 236–240, Gothenburg, Sweden. Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-4046

Bryan K. Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and hb.

Suraj Patil. 2020. Question generation. https://github.com/patil-suraj/question_generation.

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. https://doi.org/10.1080/10439463.2013.780223

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Koustav Rudra, Pawan Goyal, Niloy Ganguly, Muhammad Imran, and Prasenjit Mitra. 2019. Summarizing situational tweets in crisis scenarios: An extractive-abstractive approach. *IEEE Transactions on Computational Social Systems*, 6(5):981–993. https://doi.org/10.1109/TCSS.2019.2937899

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. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.704

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? Pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenyi Tay, Aditya Joshi, Xiuzhen Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-faced ROUGE: Examining the suitability of ROUGE for opinion summary evaluation. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 52–60, Sydney, Australia. Australasian Language Technology Association.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-8643

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.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.450

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 summarization for Twitter. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 21–24.

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. https://doi.org/10.1007/978-3-319-67256-4_30

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020b. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.553

Kexiang Wang, Baobao Chang, and Zhifang Sui. 2020c. A spectral method for unsupervised multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 435–445, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.32

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1007

Zhongqing Wang and Yue Zhang. 2017. A neural model for joint event detection and summarization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4158–4164. https://doi.org/10.24963/ijcai.2017/581

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yelp. Yelp dataset challenge. https://www.yelp.com/dataset.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *ICML*, abs/1912.08777.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 1949–1952, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3397271.3401327

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.552

Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. 2012. Towards real-time summarization of scheduled events from twitter streams. https://doi.org/10.1145/2309996.2310053