Analyzing the Dialect Diversity in Multi-document Summaries

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, Ameeta Agrawal
Portland State University
{oolabisi,ahuds2,ajetter,ameeta}@pdx.edu

Abstract

Social media posts provide a compelling, yet challenging source of data of diverse perspectives from many socially salient groups. Automatic text summarization algorithms make this data accessible at scale by compressing large collections of documents into short summaries that preserve salient information from the source text. In this work, we take a complementary approach to analyzing and improving the quality of summaries generated from social media data in terms of their ability to represent salient as well as diverse perspectives. We introduce a novel dataset, *DivSumm*, of dialect diverse tweets and human-written extractive and abstractive summaries[1]. Then, we study the extent of dialect diversity reflected in human-written reference summaries as well as system-generated summaries. The results of our extensive experiments suggest that humans annotate fairly well-balanced dialect diverse summaries, and that cluster-based pre-processing approaches seem beneficial in improving the overall quality of the system-generated summaries without loss in diversity.

1 Introduction

Since the launch of Twitter, its short, informal, creative, albeit noisy, social media posts called tweets, have been collected, labeled, and studied in numerous natural language processing (NLP) tasks, ranging from identifying topics and places of rising interest to sentiment analysis, and more. These easily accessible user-generated tweets, produced contemporaneously as world and private events unfold, provide insights into the perspective of diverse social groups but are too manifold for humans to interpret at scale. In response, automatic text summarization algorithms aim to compress long pieces of text into short, fluent, and consistent summaries while preserving the most salient information from the source text (Meng et al., 2012; Lin et al., 2021a; Amplayo et al., 2021).

Like other NLP models, summarization algorithms run the risk of perpetuating unintentional social biases against diverse groups (e.g., race or gender) and promoting structures and practices that systematically limit some groups' access to resources and decision-making power (Blodgett et al., 2020). This is because the collections of online texts such as news or Wikipedia articles, typically used for developing NLP algorithms, reflect the interests, language patterns, and structured writing style of their author demographics, which differ from those of other communities. Algorithms trained on such datasets may produce synopses in which diverse perspectives are systematically excluded. This means that groups who manage to overcome existing barriers to participation, for instance, via social media posts, and who speak up and offer their perspectives may still not be heard.

Fairness definitions for NLP models are generally based on the notion of equal treatment – an algorithm is considered fair when it performs the same for mainstream and underrepresented groups (i.e., group fairness) or delivers the same conclusions about an individual, regardless of the group they belong to (i.e., individual fairness) (Czarnowska et al., 2021). Recent works have demonstrated the disparate performance of tools on sensitive subpopulations in domains (Tatman, 2017; Buolamwini and Gebru, 2018). In this work, we consider the notion of group fairness and interpret it in terms of representation distribution of some socially salient attribute (dialect) in the summary.

Our goal is to investigate the ability of existing models of summarization to reflect the diversity of input text in the generated summaries, and propose a simple yet effective approach for improving the group-level diversity of summaries generated from noisy social media tweets, for which very little prior work exists to date (Dash et al., 2019;

[1] *DivSumm* dataset is available at https://github.com/PortNLP/DivSumm

Proceedings of the 29th International Conference on Computational Linguistics, pages 6208–6221
October 12–17, 2022.

Dataset	domain	attribute (#groups)	#topics	#summ/topic	#docs/input	#sent/summ	Ext?	Abs?
MeToo (Dash et al., 2019)	tweets	gender (2)	1	3	488	-	Y	N
Claritin (Dash et al., 2019)	tweets	gender (2)	1	3	4037	100	Y	N
US-Election (Dash et al., 2019)	tweets	political leaning (3)	1	3	2120	-	Y	N
DivSumm (ours)	tweets	dialect (3)	25	2	90	5	Y	Y

Table 1: Statistics of some social multi-document extractive summarization datasets with socially salient user group attributes and human written summaries.

Keswani and Celis, 2021). To this end, we seek to answer two specific questions with respect to summarization of dialect diverse tweets: (Q1) how diverse are summaries generated by humans? and, (Q2) how diverse are summaries generated by automatic summarization models with and without the proposed fairness interventions?

Our work makes several contributions:

- we introduce and comprehensively analyze a novel dataset, *DivSumm*, of diverse dialect tweets across several topics and corresponding extractive and abstractive human-written reference summaries;
- we study the group diversity of reference summaries and investigate two simple yet effective approaches for applying diversity interventions at the pre-processing stage of the summarization process;
- we conduct an extensive set of experiments using six existing extractive as well as abstractive summarization models as black-boxes, and report the results in terms of three types of metrics (reference-less, reference-based, and representation).

2 Related Work

In this section we discuss two relevant areas of prior work: multi-document summarization for social media data, and an overview of existing summarization datasets.

Multi-document social media summarization

Summarizing of social media data remains an interesting area of research with numerous approaches focused on optimizing the textual quality, factuality, fluency, and many other properties of the summaries (Li and Zhang, 2020; He et al., 2020; Dusart et al., 2021). However, unlike other text input (e.g., news articles), data from social media are also incredibly diverse consisting of opinions and perspectives from people from many walks of life, and while reflecting this richness of diversity in the summaries generated from them is an important goal, there have been few notable efforts in this direction.

One of the early works to study the notion of fairness in summaries generated from social media data used extractive summarization models and noted that the generated summary is not always a fair representation of the input data, even though the tweets written by different social groups (gender and political leaning in this case) are of comparable quality (Shandilya et al., 2018; Dash et al., 2019). Following these assertions, they proposed three fairness-preserving algorithms that can be applied during the pre-processing, in-processing, and post-processing stages. Keswani and Celis (2021) investigated the role of extractive summarization models in the context of dialect diversity of summaries, but did so without access to manually annotated summaries and proposed a bias mitigation model at the post-processing stage.

Summarization datasets

Developing summarization datasets is a challenging task and as such, researchers often utilize creative methods for automatically obtaining summaries of documents (Nallapati et al., 2016; Xu et al., 2021; Perez-Beltrachini and Lapata, 2021; Varab and Schluter, 2021). In cases where automatic document/summary pairings cannot be obtained, human annotation, usually through crowdsourcing, is often used (Khalman et al., 2021; Lin et al., 2021b). Specifically in the domain of social media data, many multi-document datasets have been recently introduced, some with extractive summaries, others with abstractive summaries, including TSix (Nguyen et al., 2018), Amazon and Yelp (He and McAuley, 2016; Bražinskas et al., 2020), SPACE (Angelidis et al., 2021), and ISSumSet (Dusart et al., 2021), to name just a few, but as none of these contain explicit markers of diverse social groups, we are motivated to develop a novel dataset of dialect diverse summaries.

3 Problem Description

For the task of multi-document summarization of social media data, the input typically consists of dozens to hundreds of documents (e.g., tweets) about the same topic that are all considered in generating a single summary. Formally, the input is a set of documents further split into n sentences, $\mathcal{D} = \{d_1, ..., d_n\}$. Given the sentences in $\mathcal{D}$, the goal of a multi-document summarization model is to generate a summary $\mathcal{S} = \{s_1, \ldots, s_k\}$ consisting of k sentences, where $k << n$ is usually a hyper-parameter. There are two types of summarization models, extractive models where $\mathcal{S} \subset \mathcal{D}$, and abstractive models which extract and rewrite salient pieces of text.

In multi-user settings where users belong to multiple social groups, each sentence d_i may be additionally accompanied by social attribute g_j, $g_j \in \mathcal{G}$, where $\mathcal{G}$ is a set of attributes of a demographic group such as gender, race, or dialect. To adequately accommodate the variety of perspectives expressed by multiple diverse groups, the goal of a *diversity-preserving* multi-document summarization algorithm then is to generate a summary $\mathcal{S}$ with the goal of not only optimizing textual quality but also satisfying some fairness constraint such as group fairness, which compare quantities at group level (Czarnowska et al., 2021). In the context of summarization, this problem formulation naturally lends itself well to extractive models where sentences from the final summaries can be traced back to source documents and their user group labels more so than to abstractive models where text usually gets rewritten making group label attribution challenging to ascertain.

One simple approach of computing group representation distribution in extractive summaries, which we denote as $\mathcal{R}(\mathcal{S})$, is by calculating the number of tweets belonging to each user group in the summary. Assuming m distinct and disjoint groups, $\mathcal{R}(\mathcal{S}) = \{\frac{|\mathcal{S}_1|}{k}, \ldots, \frac{|\mathcal{S}_m|}{k}\}$ where $|\mathcal{S}_j|$ is the number of sentences within the summary $\mathcal{S}$ from group j, and k is the total number of sentences in the summary.

Under the notion of equal representation, where the representation of all groups should be equal, regardless of their distribution in the input data, one can compute an aggregate diversity score. For instance, we can compute the **Representation Gap**, where a lower gap score would imply a more balanced distribution. Several metrics can be used to estimate the level of dispersion[2], and we choose the range, $\mathcal{R}_{rg}(\mathcal{S}) = \max\{\mathcal{R}(\mathcal{S})\} - \min\{\mathcal{R}(\mathcal{S})\}$. In other words, for well-balanced group distributions, a smaller $\mathcal{R}_{rg}(\mathcal{S})$ score indicates a more diverse summary. As an example, if a summary of 5 lines contains 3 sentences from group A, and a sentence each from group B and group C, then $\mathcal{R}(\mathcal{S}) = \{0.6, 0.2, 0.2\}$, and $\mathcal{R}_{rg}(\mathcal{S}) = 0.4$.

4 Dialect Diverse Summarization Dataset (*DivSumm*)

In order to study the diversity-preserving capabilities of summarization algorithms, we need a suitable dataset, and although numerous summarization datasets have become available in recent years, only a handful of them contain explicit diverse social group information (see Table 1), which motivates us to develop and contribute a novel dataset – *DivSumm*. Our dataset consists of input-summary pairs on a set of 25 topics of tweets written in three different dialects. We obtain their corresponding human-generated extractive and abstractive summaries. Table 2 presents an instance from *DivSumm*, while the following subsections outline our process of creating and exploring this dataset.

4.1 Obtaining human-written reference summaries

To obtain a large number of tweets of diverse dialectal language for annotating and creating *DivSumm*, we turn to a corpus of English tweets automatically labeled with dialect information by inferring three demographic dialect proportions, namely, *African-American (AA)* English, *Hispanic* English, and *White* English, using a model trained with census data (Blodgett et al., 2016). All the preprocessing details are included in Appendix A.

Our annotation study was designed to obtain both extractive as well as abstractive summaries, with the input consisting of multiple documents (randomly selected and shuffled 90 tweets on a given topic, with 30 tweets per dialect group) to generate topic-wise summaries (e.g., summary for `NBA`, `Netflix`, `Beyonce`, and so on). In other words, we feature all dialects with equal proportion, an approach typical of datasets that are meant to be evaluation benchmarks (Fabris et al., 2022). For generating the summaries, we invited a diverse group of ten volunteers familiar with tweets and

[2]We also computed the standard deviation of $\mathcal{R}(\mathcal{S})$ which showed a high correlation with range.

Input Documents (Tweets)
G3: If Lakers play like that every game no chance for anyone else in the NBA #NoExceptions G2: The Fan Fictions and Imagines other Beliebers come up with , OMFG I LOVE THIS FANBASE ! G2: Remember when they the NBA play-offs were boring and disappointing-not so much anymore #nbafantastic G1: NBATV shittin on u niggas.lol. Melo and Monroe doing work doe. G3: Going to take in some NBA tonight. Pacers/Knicks. NBA games are much better in person. G2: Just fast forward to the trophy presentation and the sucking off of LeBron by your company. NBA is garbage G1: No point guard in the NBA can hold Russell Westbrook G1: The Black Mamba 81 point showcase in showing now on NBATV !! G2: NBA reffs.. Responsible for half of Miami heats points since 2010. -.-t #refsforMVP G3: And te NBA was in a lockout last year too.....NOBODY complained about that last year. Bc frankly, nobody cares. ... ... G3: Is it possible that boozer is the worst player in the NBA? #GETOUT G3: NBA rule change I'm shocked has never happened: An assist if the player you passed to makes both free throws. G3: If this was an NBA game, Michael Carrera would be walking away from it with a $100,000 fine. #flopcity G1: Lakers??? Is that even a NBA team? I thought they were D-League.....

Extractive Summaries	
Annotator 1	Annotator 2
Going to take in some NBA tonight. Pacers/Knicks. NBA games are much better in person. There's a difference btw NBA ready talent & NBA prospects. UK has a ton of prospects this yr. Seeing the difference this yr. And the #Lakers get back in the game by slowing it down, grinding on 'D', and going at the basket. Who knew? 0_0 #NBA. ***It's pathetic that everyone thinks that the NBA is better than the NHL...*** *Free NBA League Pass Preview till tomorrow, Dam I'm sure I'm staying in this Saturday night!*	*Remember when they the NBA play-offs were boring and disappointing-not so much anymore #nbafantastic. Lebron is still the best player in the NBA. Put your mouth on it.". Bucks vs Sixers. This is why I love NBA TV. #FanNight. Looks like I'm watching the sugar bowl game cz its halftime for the NBA.* ***It's pathetic that everyone thinks that the NBA is better than the NHL...***
Abstractive Summaries	
Annotator 1	Annotator 2
It's an NBA game night. Many people are tuned in because it's the first game of the season. There are different reactions to the game because some think it's awesome and some think it's whack. It would greatly have to do with fans(the team they support). Viewers also gave opinions of different players they consider to be the best and Lebron is thought to be overhyped.	*All Tweets seem to contain 'NBA' regardless of placement or context, often even including tweets where NBA is part of a word such as "FANBASE". Overall the majority of tweets seem to be in-regards to the National Basketball Association, NBA, and include more often critiques or 'insults' of competing teams. It is also evident that the 'NBA', as a brand, include different ancillary businesses such as 'NBA League Pass', NBATV, and NBA2K13 (video game) in conjunction with the obvious basketball games themselves. A portion of the sample includes a notable use of racial epitaphs or slurs that may or may not be used in derision. The overwhelming bulk of the tweets is of a negative (critiquing, admonishing) nature as opposed to a positive (hopeful, cheering) message.*

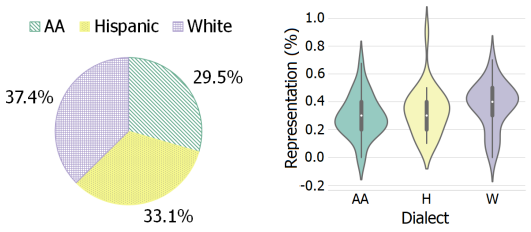
Table 2: Example instance from our *DivSumm* dataset showing input documents with corresponding reference extractive and abstractive summaries generated by two annotators. For the extractive summaries, overlapping text between the annotators is denoted in red text and bold font.

their idiosyncrasies, and more importantly, spanning a range of diversity across dialects, gender, and ethnicity. *Note that we did not mention anything about the dialects of the tweets before sharing the files for annotation, so the annotators had no background information about the goals of this study other than the fact that we seek to summarize the tweets* – this was done to mitigate any form of potential bias. We provided concise and clear guidelines for generating the summaries, along with an example set of tweets and corresponding extractive and abstractive summaries. For the extractive summaries, the annotators were requested to select 5 tweets that they believe to capture the salient points from the set of documents, while for the abstractive summaries, the annotators were asked to write 5 sentences in their own words to summarize the important points across all the documents. Every set of input tweets was summarized by two annotators, thus helping us develop a dataset of 25 topics, with a total of 100 pairs of input (sets of documents) and output (human-written summaries).

It is worth mentioning that the diversity of the annotator pool will undoubtedly have an impact on the ultimate annotations (summaries) obtained. The inherent subjective nature of summarization process suggests that different annotators will approach it from different perspectives, which is a strength but also a weakness of this process and

	R1	RL
S1, random	0.234	0.210
S2, random	0.220	0.193
S1, S2	**0.315**	**0.301**

Table 3: ROUGE scores comparing human-human summaries (S1 and S2) and human-random summaries.



(a) (b)

Figure 1: *Diversity Analysis of Human-written Reference Extractive Summaries in* DivSumm. The plot (a) displays the average representation $\mathcal{R}(\mathcal{S})$ of each dialect with each dialect's $\mathcal{R}(\mathcal{S})$ in the range of 29.5 - 37.4%, indicating a fairly balanced representation. In plot (b), we present the violin plots of the distribution scores noting few outliers.

has been extensively studied in recent literature (Clark et al., 2021; Gehrmann et al., 2022). To help account for these limitations, during evaluation we report the results using not only metrics that rely on these reference summaries, but also those that do not require reference summaries.

As summarizing informal user-generated data is a particularly challenging task, we also compute the inter-annotator agreement to measure the lexical overlap of extractive reference summaries written by both annotators. As shown in Table 3, the average ROUGE-1 and ROUGE-L F1 scores for pairs of human annotated summaries was much higher (∼0.31) than that of a randomly generated summary compared against the human-generated summaries (∼0.21), suggesting that the human-written summaries are more similar, and arguably reliable.

4.2 Analyzing diversity in human-written summaries

Before studying how well automatic summarization models reflect diversity in system summaries, a natural question to ask is how well do humans summarize such diverse data. To answer this question, we conduct a thorough analysis of the reference summaries of *DivSumm* dataset in an attempt to uncover any interesting insights into the way that humans (specifically, our annotators) approach dialect-diverse summarization process. Importantly, our answers to these analyses will serve as a principled baseline when we later evaluate the model-generated summaries.

The extractive summaries provide a uniquely interesting opportunity to explore the question of how diverse human-written summaries are by allowing us to compare the proportion of representation of each dialect group within the summaries. Recall that our dataset features an equal number of tweets for each of the three dialect groups in the input, and, therefore, for equal or proportional representation in the output summary, a well-balanced summary would contain equal number of tweets from each group. As the plots in Figure 1 show, we found a fairly balanced representation for each dialect group in the summary ($\mathcal{R}(\mathcal{S})$ ranges between 29.5% to 37.4%).

5 Modeling Diversity in System Summaries

In this section, we explore the diversity-preserving qualities of recently proposed summarization algorithms using three approaches described below and visualized in Figure 2.

VANILLA: This standard baseline approach of summarizing uses a single aggregated set of randomized documents from all the dialect groups, e.g., a total of 90 tweets, as the input to the summarization model without any pre-processing.

CLUSTER-HEURISTIC (CLUSTER-H) : Similar to the pre-processing approach of Dash et al. (2019), this method first heuristically partitions the set of input documents into group-based subsets ($\mathcal{D} = \{\mathcal{D}_1 \cup ... \cup \mathcal{D}_m\}$, each $\mathcal{D}_j$ containing a set of documents from group j) before passing them to the summarization model to generate separate group summaries ($\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_m\}$) – one for each of the three group-specific tweets. However, instead of concatenating these summaries to generate the final summary, we shuffle and combine these m group-level summaries into a single document and pass that again to the summarization model to generate a new, combined summary. In doing so, we seek to first preserve group-level salient information before aggregating the most informative units from such individual summaries into a unified summary.

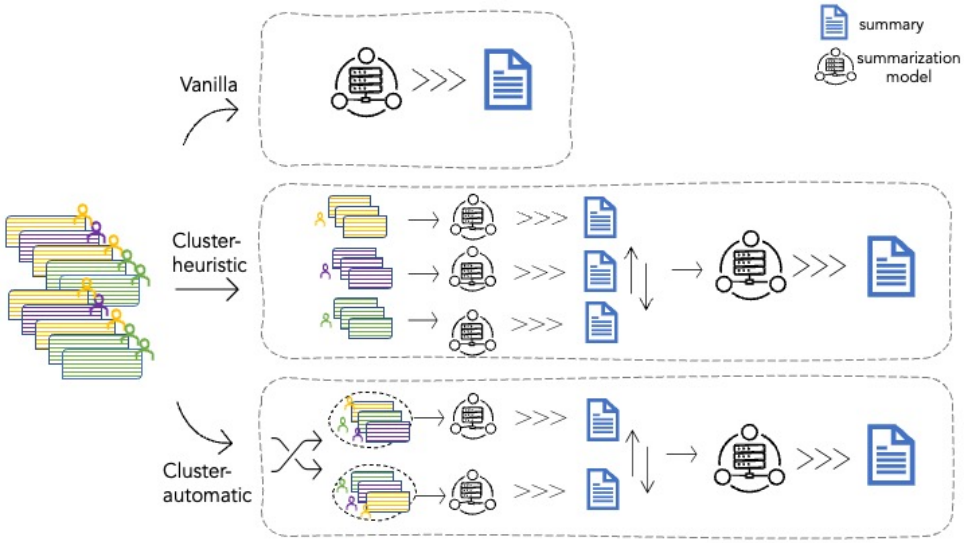


Figure 2: Illustration showing the overview of VANILLA, CLUSTER-HEURISTIC and CLUSTER-AUTOMATIC.

CLUSTER-AUTOMATIC (CLUSTER-A): Conversely, in a more pragmatic scenario, the sensitive group attribute may not be reliably observable, or inferring it is not possible due to some reason, including ethical reason. In such a case, we also investigate an attribute-agnostic approach based on automatic clustering, as follows: (i) Generate p clusters of $\mathcal{D}$ via some clustering algorithm (e.g., k-means) with the optimal value of p determined through silhouette score, yielding $\mathcal{D} = \{\mathcal{D}_1 \cup ... \cup \mathcal{D}_p\}$. (ii) Generate a set of corresponding summaries $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_p\}$. (iii) Concatenate all the p summaries into a single document and pass it again to the summarization algorithm to generate a final summary.

6 Experiments

The three approaches are studied in the context of recently proposed summarization models, both extractive and abstractive, and the quality of the summaries is evaluated using reference-based metrics, reference-less metrics, and representation metric.

6.1 Summarization models

We explore six recent extractive and abstractive summarization models described below (we refer to Appendix B for full implementation details).

Extractive methods: **TEXTRANK** (Mihalcea and Tarau, 2004), an unsupervised graph-based ranking method, determines the most important sentences in a document based on information extracted from the document itself, and therefore performs well even without any form of domain knowledge or pre-training. **BERT-EXT** (Miller, 2019), an extractive summarization model, uses pretrained embeddings from BERT (Devlin et al., 2018) and k-means clustering to select sentences closest to the centroid as the summaries, and similarly, **LONGFORMER-EXT** (Miller, 2019) which uses pretrained embeddings from LongFormer (Beltagy et al., 2020).

Abstractive methods: **BART** (Lewis et al., 2019) is a sequence-to-sequence model combining a bidirectional encoder with an auto-regressive decoder and trained by corrupting the document with an arbitrary noisy function. **T5** (Raffel et al., 2019), an encoder-decoder model trained using teacher forcing, modifies the original transformer architecture to convert language problems into a text-to-text format. **LED** (Longformer Encoder-Decoder) (Beltagy et al., 2020), a variant of the Longformer model with both encoder and decoder transformer stacks, has shown to improve modeling long sequences for sequence-to-sequence learning.

6.2 Evaluation metrics

To evaluate the quality of the system-generated summaries, we consider three types of metrics.

Reference-based: **ROUGE** (Lin, 2004) calculates the lexical overlap between the machine-generated output and the human-written reference summaries. For our experiments, we report the F1 scores of ROUGE-1 (overlapping unigrams) and ROUGE-L (longest common subsequences). To compute the final scores, a system-generated summary is compared with each of the two reference summaries, and their average score is reproted.

Extractive Models

Model	TextRank					BERT-EXT					LONGFORMER-EXT				
	R1	RL	SQA	B	$\mathcal{R}_{rg}\downarrow$	R1	RL	SQA	B	$\mathcal{R}_{rg}\downarrow$	R1	RL	SQA	B	$\mathcal{R}_{rg}\downarrow$
VANILLA	0.25	0.23	0.06	0.11	0.15	0.22	0.21	0.04	0.13	0.04	0.22	0.20	0.04	0.12	0.17
CLUSTER-H	0.26	0.24	0.06	0.10	0.18	0.22	0.21	0.05	0.12	0.07	0.22	0.20	0.05	0.11	0.20
CLUSTER-A	0.25	0.23	0.07	0.10	0.16	0.22	0.20	0.05	0.12	0.06	0.23	0.21	0.05	0.11	0.16

Abstractive Models

Model	BART					T5					LED				
	R1	RL	SQA	B	$\mathcal{R}_{rg}\downarrow$	R1	RL	SQA	B	$\mathcal{R}_{rg}\downarrow$	R1	RL	SQA	B	$\mathcal{R}_{rg}\downarrow$
VANILLA	0.16	0.14	0.06	0.09	-	0.15	0.13	0.07	0.08	-	0.13	0.12	0.06	0.07	-
CLUSTER-H	0.16	0.15	0.06	0.09	-	0.14	0.13	0.05	0.07	-	0.14	0.13	0.05	0.08	-
CLUSTER-A	0.17	0.15	0.06	0.09	-	0.12	0.11	0.06	0.06	-	0.14	0.14	0.05	0.07	-

Table 4: Results of the three approaches (VANILLA, CLUSTER-H, and CLUSTER-A) across three extractive summarization models (TextRank, BERT-Ext, LongFormer-Ext) and three abstractive summarization models (BART, T5, LED) using the *DivSumm* summarization dataset. The metrics reported include ROUGE-1 (R1), ROUGE-L (RL), SummaQA (SQA), BLANC (B), and for the extractive summaries $\mathcal{R}_{rg}(\mathcal{S})$ denoting the Representation Gap. All scores are averaged over two runs. The best scores per model and per metric have been underlined. For reference, the Representation Gap in human summaries $\mathcal{R}_{rg}(\mathcal{S})$ = 0.08.

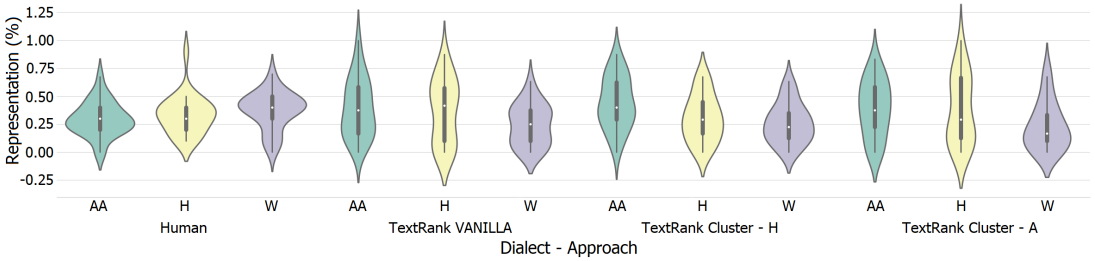


Figure 3: Violin plots of $\mathcal{R}(\mathcal{S})$ per dialect and per approach for TEXTRANK over *DivSumm* dataset. It is noticed that the violins for both *AA* and *Hispanic* tweets are considerably thinner as compared to the human summaries indicating many outliers on both ends of the spectrum.

Reference-less: **SummaQA** (Scialom et al., 2019) evaluates the quality of a text summary without relying on reference summaries, making it a practical choice for assessing summaries generated from large collections of tweets. Instead, it uses a question-answering model based on BERT to answer cloze-style questions using the system-generated summaries. **BLANC** (Vasilyev et al., 2020), another reference-less metric, measures how the performance of a pretrained language model improves during language understanding tasks when the model is given access to a summary, and correlates with informativeness (Iskender et al., 2021).

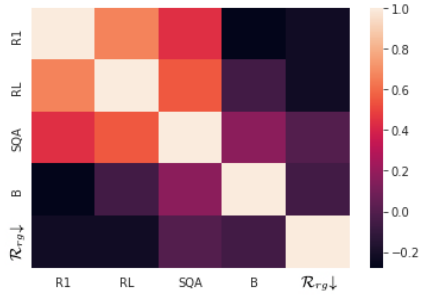
Representation Gap: Finally, we also report the $\mathcal{R}_{rg}(\mathcal{S})$ of the extractive summaries by calculating the range of the representation distribution in the summary.

7 Results and Discussion

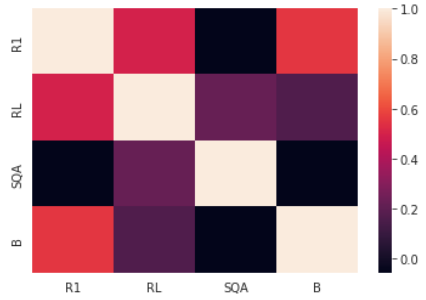
Table 4 presents the detailed results of our experiments across extractive and abstractive models, and some samples of system-generated summaries are included in Appendix C. In looking at the representation gap scores of extractive models, we note that the VANILLA approach without any intervention does well in terms of $\mathcal{R}_{rg}$ on 2 out of 3 datasets, while in the case of CLUSTER-H performance is strictly worse despite that model being designed to consider group-level information. In terms of summary quality, all three approaches perform comparably. Of note, however, is the performance of BERT-Ext model which yields impressive representation gap scores, suggesting that centroid-based approaches in particular could be effective as unsupervised diversity-preserving models.

Model	*Extractive* AVERAGE					*Abstractive* AVERAGE			
	R1	RL	SQA	B	$\mathcal{R}_{rg}\downarrow$	R1	RL	SQA	B
VANILLA	0.232	0.217	0.053	**0.123**	**0.12**	0.147	0.136	**0.068**	**0.083**
CLUSTER-H	0.236	0.218	0.057	0.117	0.15	**0.151**	**0.140**	0.059	**0.083**
CLUSTER-A	**0.237**	**0.219**	**0.062**	0.115	0.13	0.150	0.139	0.061	0.080

Table 5: Averaged results of the three approaches (VANILLA, CLUSTER-H, and CLUSTER-A) across three extractive summarization models and three abstractive summarization models evaluated using the *DivSumm* summarization dataset. The metrics reported include ROUGE-1 (R1), ROUGE-L (RL), SummaQA (SQA), BLANC (B), and for the extractive summaries, $\mathcal{R}_{rg}(\mathcal{S})$.



(a)



(b)

Figure 4: Pairwise Kendall's Tau correlations for summary evaluation metrics, (a) extractive models, and (b) abstractive models. A higher score (shown in lighter color) indicates higher correlation between the rankings provided by a pair of metrics, whereas a smaller score (shown in darker color) indicates weaker correlation.

Figure 3 presents the violin plots for one of the summarization models, TEXTRANK, to allow us to further investigate the differences between the three approaches (please see Appendix D for analysis of the other two extractive models). For all the approaches, the violins for both *AA* and *Hispanic* summaries are noticeably thinner than the corresponding human summaries, suggesting that the $\mathcal{R}(\mathcal{S})$ representation distribution of system summaries contains many outliers on either end of the spectrum.

Table 5 presents the averaged results of all the extractive methods and abstractive methods, which confirm that VANILLA generates more diverse summaries, while CLUSTER-A does generally better on the other metrics related to summary quality, hinting at a plausible trade off between the two dimensions (Celis et al., 2018). Considering the results of extractive and abstractive models together, it appears that some sort of clustering of documents before passing them to the summarization model remains beneficial in improving the overall quality of summaries.

Next, given the multiple evaluation metrics along dimensions of quality and representation, we further study the correlations between these metrics and present some heatmap visualizations generated using the pairwise Kendall's Tau correlation values. Figure 4 summarizes the correlations computed using the average results. Each approach was first sorted from best to worst based on the scores provided by each metric, and these rankings were then used to calculate Kendall's Tau to represent how well the rankings correlate between metrics. A score closer to 1.0 indicates high correlation (that is, the metrics ranked approaches in similar order), while a score closer to -1.0 indicates poor correlation (the metrics ranked approaches in a different order).

We observe that in the case of extractive summaries, metrics R1, RL, and SummaQA have high to moderate inter-correlation, while BLANC and $\mathcal{R}_{rg}(\mathcal{S})$ do not correlate with any other metrics. For the abstractive models, R1 shows moderate correlation with RL and BLANC, but SummaQA shows poor correlation with all the other three metrics. Overall, we conclude that (i) both the reference-based metrics R1 and RL correlate well as expected, (ii) R1 correlates well with reference-less metric SummaQA for extractive summaries, and with BLANC for abstractive summaries, (iii)

Model	BART				T5				LED			
	R1	RL	SQA	B	R1	RL	SQA	B	R1	RL	SQA	B
VANILLA	0.202	0.184	0.212	0.201	0.194	0.181	0.269	0.281	0.160	0.148	0.181	0.172
CLUSTER-A	0.172	0.157	0.229	0.270	0.170	0.158	0.185	0.224	0.176	0.161	0.217	0.271

Table 6: Results of VANILLA and CLUSTER-A as applied to another dataset, DialogSumm (Chen et al., 2021)

the two reference-less metrics (SummaQA and BLANC) do not correlate with each other, and (iv) finally, no metrics seem to be correlating well with the representation metric ($\mathcal{R}_{rg}$), suggesting the need for new metrics that can measure representation of diversity in addition to other dimensions of quality.

Finally, since clustering-based approaches improved over the baseline approach in terms of summary quality, we conduct one more investigation to evaluate whether CLUSTER-A generalizes to another dataset involving multiple users such as *DialogSumm* dataset (Chen et al., 2021). The results presented in Table 6 indicate that VANILLA performs better when using T5 model, CLUSTER-A brings additional gains to LED model, while remaining comparable on the third model.

8 Conclusions

In this work, we investigate whether, and to what extent, do system-generated summaries reflect the diversity of socially salient groups present in the input data. To answer this question, we first develop a novel summarization dataset, *DivSumm*, by obtaining human-written reference summaries, of both extractive and abstractive sort, for dialect diverse tweets. In analyzing the human-written reference summaries, we were encouraged to note that on average humans generated reasonably well-balanced dialect diverse summaries. This was followed by an extensive evaluation exploring the diversity reflected in system summaries by experimenting with three approaches as applied to six summarization algorithms, and evaluated using multiple metrics of summary quality and representation. Future avenues of work include expanding our dataset to consider other diverse social attributes and improving the summarization models along dimensions of both quality as well as representation.

9 Ethical Considerations

Tweets provide a rich and diverse source of natural language data but in working with unfiltered social media data, we also run the risk of encountering unconventional or in some cases what may be considered as offensive language. Being sensitive to these limitations, before undertaking the annotation process, we carefully informed the annotators of some of the inherent risks of annotating tweets and provided them with the option of withdrawing from the annotation process should they feel uncomfortable at any point of time (it is worth noting that no annotator withdrew from the study). Similarly, our discussions related to the representation of dialect diversity in summaries are based solely on the summaries that were developed during this study and the summarization models that were adopted in our experiments. It remains to be seen whether these conclusions generalize to other social groups.

Acknowledgments

We would like to thank Camille Range, Vikram Nagapudi, and our annotators for their help in creating this dataset. We are grateful to the anonymous reviewers for their insightful suggestions, and to Cisco Research for partially supporting this research.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12489–12497.

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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. 2018. Fair and diverse dpp-based data summarization. In *International Conference on Machine Learning*, pages 716–725. PMLR.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *arXiv preprint arXiv:2106.14574*.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexis Dusart, Karen Pinel-Sauvagnat, and Gilles Hubert. 2021. Issumset: a tweet summarization dataset hidden in a trec track. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 665–671.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *arXiv preprint arXiv:2202.01711*.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

Ruifang He, Liangliang Zhao, and Huanyu Liu. 2020. Tweetsum: Event oriented social summarization dataset. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5731–5736.

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*, pages 507–517.

Neslihan Iskender, Oleg Vasilyev, Tim Polzehl, John Bohannon, and Sebastian Möller. 2021. Towards human-free automatic quality evaluation of german summarization. *arXiv preprint arXiv:2105.06027*.

Vijay Keswani and L Elisa Celis. 2021. Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, pages 3802–3814.

Misha Khalman, Yao Zhao, and Mohammad Saleh. 2021. ForumSum: A multi-speaker conversation summarization dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4592–4599, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Quanzhi Li and Qiong Zhang. 2020. Abstractive event summarization on twitter. In *Companion Proceedings of the Web Conference 2020*, pages 22–23.

Chen Lin, Zhichao Ouyang, Xiaoli Wang, Hui Li, and Zhenhua Huang. 2021a. Preserve integrity in real-time event summarization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–29.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021b. Csds: A fine-grained chinese dataset for customer service dialogue summarization.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. 2012. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–387.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond.

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).

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9408–9423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xinnuo Xu, Ondřej Dušek, Shashi Narayan, Verena Rieser, and Ioannis Konstas. 2021. Miranews: Dataset and benchmarks for multi-resource-assisted news summarization.

A Preprocessing Documents

Raw tweets can be notoriously noisy and challenging for summarization purposes[3]. Our preprocessing steps are as follows: (i) we considered a tweet as belonging to one of the three dialect groups if it had a dialect confidence score greater than 0.7 for a given dialect and less than 0.3 for all the other dialects, (ii) we removed any duplicate tweets, any mentions (i.e., @username), and any tweets shorter than seven tokens in length, and (iii) finally, since emojis may provide useful indicators during summary generation, we converted the Unicode emoji characters with their corresponding images.

After the initial preprocessing, we extracted a list of hashtags present in the remaining tweets with the goal of identifying the most frequent topics of at least thirty tweets per dialect. This considerably filtered down the set of tweets because of unbalanced distribution of dialect groups in the corpus, with significantly more *White* tweets than *AA* or *Hispanic* tweets. Finally, we found and settled on a set of twenty five topics that we hypothesize encompass tweets from a diverse set of users. These include: 49ers, Amazon, Beyonce, Chicago, Christmas, Eagles, Facebook, Flu, Graduation, Grammys, Iphone, Kobe, McDonalds, NBA, Netflix, NYC, Obama, Paris, Patriots, Seahawks, Superbowl, Thanksgiving, VMA, WWE, Xbox.

B Implementation Details

TEXTRANK model[4] was initiated with the word count set to 70 - which is the average number of tokens for 5 sentences in our dataset. BERT-EXT and LONGFORMER-EXT models were initiated from the extractive summarization model[5], with the number of output sentences set to 5. For the abstractive models BART, T5, and LED, we used pretrained model checkpoints BART_base[6], T5_base[7],and AllenAI LED_base_16384[8], respectively, with beam size set to 3 and minimum length of tokens set to 70. The model checkpoints were accessed from the HuggingFace library (Wolf et al., 2019) and further fine-tuned using the TWEETSUMM dataset (Feigenblat et al., 2021) which was chosen as it is one of the most similar tweet datasets to ours that was large enough to serve as a training set for fine-tuning purposes. For automatic clustering, we used k-means clustering with *tf-idf* vector representation, and set $k = 2$ for all our experiments after assessing it to generate reasonable results using silhouette coefficient scores[9]. The reference-based and reference-less evaluation metrics were computed using the SummEval toolkit[10] (Fabbri et al., 2021).

C Example System Summaries

Tables 7 and 8 present some system summaries as generated by an extractive model and an abstractive model, respectively.

D Representation Distribution

Figure 5 displays the violin plots for $\mathcal{R}(\mathcal{S})$ per dialect and approach for (a) BERT-Ext, and (b) LongFormer-Ext.

[3] And not to mention, potentially offensive. However, other than these preprocessing steps, we intentionally did not filter out any further tweets, neither automatically nor manually, in order to avoid inserting any biases. Given the nature of social media posts, it is possible that the dataset may thus unintentionally contain some offensive content. The annotators were carefully informed about the risks of participating in such a study.

[4] https://github.com/RaRe-Technologies/gensim

[5] https://pypi.org/project/bert-extractive-summarizer/

[6] https://huggingface.co/facebook/bart-base

[7] https://huggingface.co/t5-base

[8] https://huggingface.co/allenai/led-base-16384

[9] https://scikit-learn.org/stable/index.html

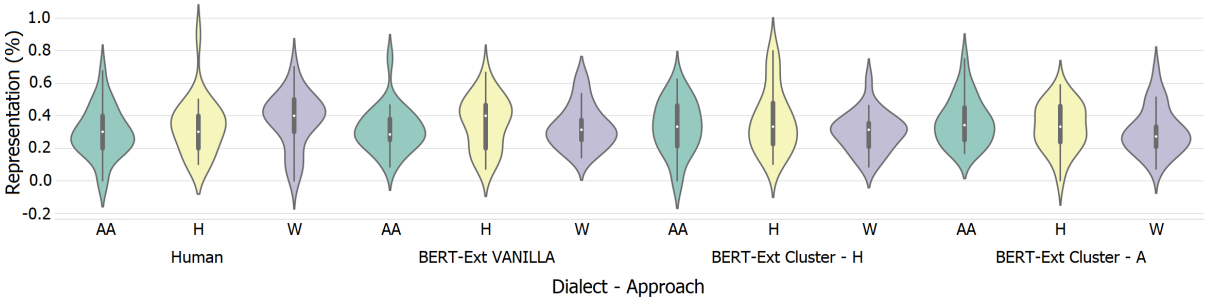
[10] https://github.com/Yale-LILY/SummEval

Approach	Summary
REFERENCE	Looks like I'm watching the sugar bowl game cz its halftime for the NBA. And the Lakers get back in the game by slowing it down, grinding on 'D', and going at the basket, Who knew? 0_0 #NBA. Aha best team in the NBA they by far better then your Celtics y'all lost last night btw to the grizzlies. NBA basketball n packing for the rest of my night. Miami comin out the east & Clippers comin out the west #NBA #FinalsPrediction.
VANILLA	Looks like I'm watching the sugar bowl game cz its halftime for the NBA. yea on NBA TV, they PLAYN n London" Oh Aite Thanks. Watching and NBA game live > Watching the play by play on an iPhone.. As I sit here lowkey buzzed playing NBA 2k13 thinking I'm actually in the tv screen. I love explaining the NBA to my mom & telling her which players are good & stuff like that.. Tim Duncan is always on my make believe NBA teams cause he's my favorite PF,. When girls actually know what they're talking about the NBA.
CLUSTER-H	When girls actually know what they're talking about the NBA. Any subliminal pics I post on IG will now be hash tagged #HNBAF. to comeback and do commentary on some Joker episodes? # West is actually good 1-8.. Tim Duncan is always on my make believe NBA teams cause he's my favorite PF,. NFL, NBA, and NHL were lockouts by owners.. Diandre doesn't understand there isn't really parity in NBA.
CLUSTER-A	The #Memphisgrizzlies just made the stupidest trade in NBA history trading Rudy gay at this point in the season #wow. NBA games are much better in person.. Looks like I'm watching the sugar bowl game cz its halftime for the NBA. Up watching NBA TV... He need to stick to coaching or being a GM or some shit.... @ The Black Mamba 81 point showcase in showing now on NBATV !!. Tim Duncan is always on my make believe NBA teams cause he's my favorite PF,.. he has a higher ceiling than granger but gotta hold off on calling him better, PG has never been top 5 in NBA in scoring. The Heat get every call to go their way...the NBA should be ashamed of this BS.

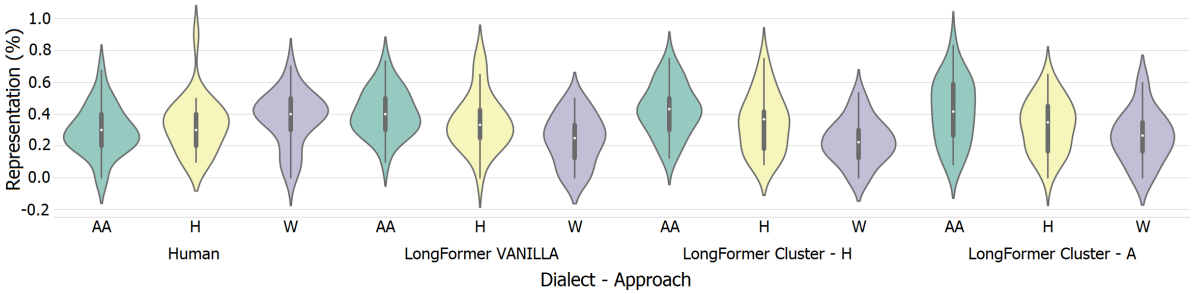
Table 7: An example of system summaries (along with human-generated reference summary) using BERT-Ext model

Approach	Summary
REFERENCE	It's an NBA game night. Many people are tuned in because it's the first game of the season. There are different reactions to the game because some think it's awesome and some think it's whack. It would greatly have to do with fans(the team they support). Viewers also gave opinions of different players they consider to be the best and Lebron is thought to be overhyped.
VANILLA	Looks like I'm watching the sugar bowl game and its halftime for the NBA.
CLUSTER-H	And the #Lakers get back in the game by slowing it down, grinding on 'D', and going at the basket. And the #Lakers get back in the game by slowing it down, grinding on 'D', and going at the basket. And the #Lakers get back in the game by slowing it down, grinding on 'D', and going at the basket.
CLUSTER-A	@ RobHall their are truly more lakers haters then Miami haters.. Who knew? he has a higher ceiling than granger but gotta hold off on calling him better, PG has never been top 5 in scoring. @robhall their are truly more lakers haters then Miami haters.. Who knew? he has a higher ceiling than granger but gotta hold off on calling him better, PG has never been top 5 in scoring. And the #Lakers get back in the game by slowing it down, grinding on 'D', and going at the basket. Looks like I'm watching the sugar bowl game and its halftime for the NBA

Table 8: An example of system summaries (along with human-generated reference summary) using LED model



(a)



(b)

Figure 5: *Violin plots of $\mathcal{R}(\mathcal{S})$ per dialect and approach, with (a) BERT-Ext, and (b) LongFormer-Ext models.* The values for $\mathcal{R}$ were determined across two runs and averaged. Of interest is how the models compare to the human-generated summaries. For BERT-Ext, the *White* dialect contains the consistently widest violins, indicating a more consistent $\mathcal{R}(\mathcal{S})$ average around $\sim$0.33. The violins for the *AA* and *Hispanic* dialects are skinnier, suggesting that they contain more outliers above and below the $\sim$0.33 mark. For LongFormer-Ext, once again the *White* dialect contains consistently wide violins, whereas the *AA* violins are wide for both VANILLA and CLUSTER-H approaches, but contain more outliers for the Cluster-A approach. The *Hispanic* dialect violins contain more outliers for both CLUSTER-H and CLUSTER-A approaches. More outliers is indicative of less consistent representation that deviates from the desired equal $\mathcal{R}$ value of 0.33 in the case of our equally represented groups.