



OPEN ACCESS

EDITED BY

Syed Afaq Ali Shah,
Edith Cowan University, Australia

REVIEWED BY

Sandeep Singh Sandha,
University of California, Los Angeles,
United States
Nebojsa Bacanin,
Singidunum University, Serbia

*CORRESPONDENCE

Philip Ehnert
✉ philip.ehnert@iits-consulting.de

RECEIVED 31 May 2023

ACCEPTED 01 March 2024

PUBLISHED 20 March 2024

CITATION

Ehnert P and Schröter J (2024) Key point generation as an instrument for generating core statements of a political debate on Twitter. *Front. Artif. Intell.* 7:1200949. doi: 10.3389/frai.2024.1200949

COPYRIGHT

© 2024 Ehnert and Schröter. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Key point generation as an instrument for generating core statements of a political debate on Twitter

Philip Ehnert^{1*} and Julian Schröter²

¹iits-consulting/ImpressSol GmbH, Department of Artificial Intelligence, Au in der Hallertau, Germany, ²FOM—Hochschule für Oekonomie und Management GmbH, Department of Business Informatics, Bonn, Germany

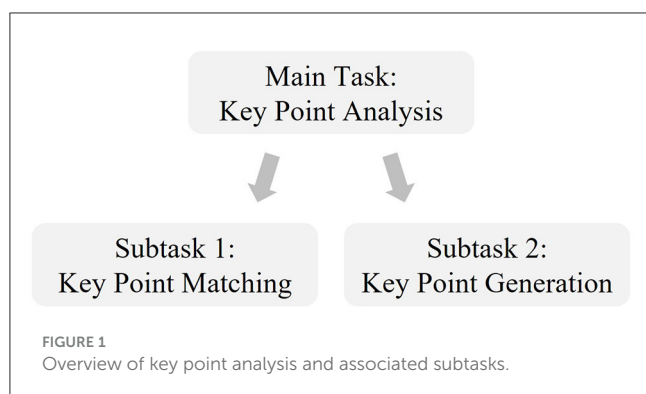
Identifying key statements in large volumes of short, user-generated texts is essential for decision-makers to quickly grasp their key content. To address this need, this research introduces a novel abstractive key point generation (KPG) approach applicable to unlabeled text corpora, using an unsupervised approach, a feature not yet seen in existing abstractive KPG methods. The proposed method uniquely combines topic modeling for unsupervised data space segmentation with abstractive summarization techniques to efficiently generate semantically representative key points from text collections. This is further enhanced by hyperparameter tuning to optimize both the topic modeling and abstractive summarization processes. The hyperparameter tuning of the topic modeling aims at making the cluster assignment more deterministic as the probabilistic nature of the process would otherwise lead to high variability in the output. The abstractive summarization process is optimized using a Davies-Bouldin Index specifically adapted to this use case, so that the generated key points more accurately reflect the characteristic properties of this cluster. In addition, our research recommends an automated evaluation that provides a quantitative complement to the traditional qualitative analysis of KPG. This method regards KPG as a specialized form of Multidocument summarization (MDS) and employs both word-based and word-embedding-based metrics for evaluation. These criteria allow for a comprehensive and nuanced analysis of the KPG output. Demonstrated through application to a political debate on Twitter, the versatility of this approach extends to various domains, such as product review analysis and survey evaluation. This research not only paves the way for innovative development in abstractive KPG methods but also sets a benchmark for their evaluation.

KEYWORDS

key point generation, topic modeling, abstractive summarization, hyperparameter tuning, semantic textual similarity

1 Introduction

The rapidly growing amount of user-generated content on platforms such as Twitter has created new opportunities and challenges in the field of political discourse analysis. Recognizing the potential of such interactions and the vast availability of textual data, previous research has underlined the importance of these engagements for data-driven political and economic decision-making (Zeng et al., 2010; Stieglitz and Dang-Xuan, 2013). However, the growing amount of data and interactions also brings challenges, in particular



the problem of information overload. In order to effectively address this problem and unlock the full potential of large amounts of text data, the use of big data analytics methods is increasingly recommended (He et al., 2017). This research seeks to tackle this problem by introducing an unsupervised abstractive KPG (Bar-Haim et al., 2020a) approach tailored to user-generated content on Twitter. The primary objective of key point analysis (KPA) is to identify and categorize core statements from large document collections of short, user-generated texts, thereby generating a comprehensive list of key points that reflect the dominant topics within the text collection. The result is a list of all identified key points, including the number of statements each key point represents to quantify its share of the debate.

The KPA task is divided into KPG and key point matching (KPM), as shown in Figure 1. KPG focusses on the identification and generation of key statements from a text corpus, while KPM involves the mapping of these statements back to the original corpus. Addressing the research question “How can an unsupervised abstractive key point generation method be developed and evaluated for application to unstructured, user-generated data in the context of political debates on Twitter?”, this study presents a novel approach to KPG and its evaluation. Unlike previous methods that rely on structured or specially annotated datasets, this research proposes an unsupervised KPG method that is tailored to the domain of political debate on Twitter. The first step is to prepare the Twitter data using a pre-processing pipeline to provide suitable data for a KPG. This is achieved by collecting tweets, applying pre-processing and topic modeling to identify high-level topics (see Figure 2). Our research innovatively integrates advanced topic modeling with hyperparameter tuning in the UMAP and HDBSCAN methods to achieve subtopic segmentation. Specifically, this tuning aims to make cluster assignment more deterministic, counteracting the inherent probabilistic variability of the process and ensuring more consistent output. In addition, we have developed a custom hyperparameter optimization for the abstractive summarization strategy. This strategy uses a modified Davies-Bouldin Index, specifically adapted for key point summarization, to improve the representativeness of the derived key points. Furthermore, we evaluated our KPG approach by considering it as a specialized form of MDS. In the future, this quantitative evaluation based on the word-based metric ROUGE (Lin, 2004) and the word-embedding-based metric BERTScore (Zhang T. et al., 2019) can be used to

compare the performance of different approaches of KPG, even when applied to unlabeled data.

The subject of KPG is a relatively new area of research, but it has strong overlaps with established research areas such as MDS, opinion, and argument summarization (Friedman et al., 2021). Whereas, MDS deals with the concise summarization of a collection of heterogeneous text documents, argument summarization deals with identifying and subsequently summarizing argumentative content in highly opinionated text documents.

A distinction is also made between extractive and abstractive summarization methods. Extractive procedures of the MDS focus on using, for example, lexical (Landauer et al., 1998; Mihalcea and Tarau, 2004), graph-based (Erkan and Radev, 2004), or transformer-based (Liu, 2019; Zhang X. et al., 2019) methods to use representative sentences of the input documents as components of the generated summary. Abstractive methods, in contrast, use transformer-based generative language models, such as BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020), to identify relevant content from various documents and summarize it in a concise, newly generated form.

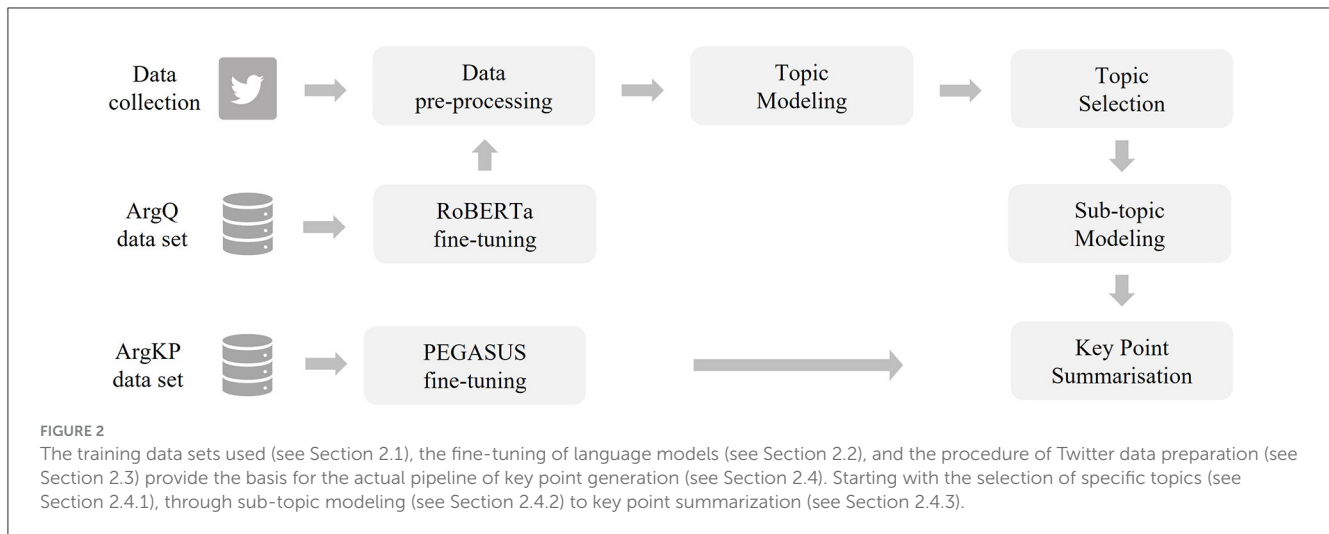
Recent advancements in the optimization of these complex abstractive models have seen the application of meta heuristic techniques for hyperparameter tuning. Such approaches, exemplified by studies such as Bacanin et al. (2022), harness the power of meta heuristics to efficiently navigate the vast hyperparameter space. These methods demonstrate potential in significantly improving model performance, particularly in tasks requiring fine-tuned control over multiple parameters, which is often the case in abstractive summarization methods.

Previous research in the field of MDS has mainly focussed on the processing of longer text documents, summarizing newspaper articles (Fabbri et al., 2019), Wikipedia articles (Ghalandari et al., 2020), and scientific literature (Lu et al., 2020).

However, due to the lack of publicly available training data, summarization of short user-generated texts is underrepresented. Within the research area of MDS, the practicability of extractive and abstractive methods was evaluated in the context of summarizing posts on the Reddit platform (Sotudeh et al., 2021), product reviews (Angelidis et al., 2021; Oved and Levy, 2021), and Twitter streaming data (Dusart et al., 2023).

Following the definition of the research task of KPA (Bar-Haim et al., 2020a), first extractive approaches were developed to generate key points (Bar-Haim et al., 2020b, 2021), identifying high quality statements using a RoBERTa language model (Devlin et al., 2019; Liu et al., 2019) specifically adapted to this task. Finally, following the ArgMining Workshop 2021, a corresponding training and evaluation data set (Friedman et al., 2021) was published, explicitly designed for the development of new KPA methods. Based on this data set, further extractive approaches of KPG have been developed, which solve the problem via a graph-based method (Alshomary et al., 2021) or via the selection of representative key point candidates with the help of a combination of the evaluation metric MoverScore (Zhao et al., 2019) and the maximal marginal relevance (MMR) based on word embeddings of the statements (Shirafuji et al., 2021).

Although abstractive methods have the advantage of summarizing different semantic aspects of similar statements in a concise key point, they have been underrepresented in



previous research. So far, only one approach has been developed to solve the KPG problem by using abstractive summarization with competitive performance (Friedman et al., 2021). For each statement in the text collection, key point candidates are generated using a language model (Zhang et al., 2020) that is specifically fine-tuned for this use case. Candidate key points are compared with expert annotated key points using the ROUGE evaluation metric, and only those with high scores are retained as final key points. Therefore, the approach described above is not applicable to unknown data.

2 Materials and methods

2.1 Data sets

2.1.1 Argument quality data set

The generation of high-quality key points requires a basis of statements of high argumentative quality. To ensure the argumentative quality of tweets, the distinction between non-argumentative and argumentative statements has proved to be effective (Bosc et al., 2016; Schaefer and Stede, 2021). In the course of this research, characteristics of non-argumentative and argumentative tweets were learned using a language model based on the freely available data set *ArgQ* (Gretz et al., 2020). This data set containing a total of 30,497 statements was generated with the help of annotators who were asked to write a pro and contra argument on a controversial discussion topic. The assessment of whether statements were argumentative or non-argumentative took the form of a binary annotation. For the quality of argumentation, two values are provided, the weighted average of all annotations and the MACE-P value (Habernal and Gurevych, 2016), which was also introduced to evaluate the argumentation quality of crowd annotations. The argument quality is given in the interval $[0, 1]$ for both values.

2.1.2 Key point summarization data set

To adapt language models to specific tasks, transfer learning using specialized training data sets has proven to be effective (Raffel

et al., 2020). We used the *ArgKP_2021* data set (Friedman et al., 2021), which was created specifically for the field of key point analysis. This data set is essentially a modified version of the *ArgQ* data set (Gretz et al., 2020). *ArgKP_2021* consists of a total of 7,238 statements on 31 topics and has been enriched with 276 key points through additional annotation by domain experts. Since each statement can be associated with several core statements, it consists of 27,519 rows with information on the topic affiliation, the stance toward the topic, the actual statement text, and the associated key point.

2.2 Model fine-tuning

2.2.1 Argument quality

In this research, the binary classification of non-argumentative and argumentative tweets is carried out via a fine-tuned RoBERTa-base model (Liu et al., 2019) using the Huggingface Transformers framework (Wolf et al., 2020). Following the approach of Bar-Haim et al. (2020a) to create a high-quality set of arguments, this research labeled statements from the *ArgQ* data set (Gretz et al., 2020) with a MACE-P argument quality value lower than 0.5 as non-argumentative. The binary labeled statements serve as training data for fine-tuning the model. The f_1 value, which reflects the ratio of precision and recall achieved during the evaluation, was 0.785. The detailed procedure of the fine-tuning and the evaluation can be found in the [Github repository](#) associated with this publication.

2.2.2 Key point summarization

This research uses an abstractive summarization technique to summarize similar statements. For this purpose, we use a PEGASUS language model since it can generate summaries of high quality with respect to the ROUGE score (Lin, 2004) achieved with only a few training samples for fine-tuning (Zhang et al., 2020). In addition to this property, we used a variant of the PEGASUS model pre-trained on the XSUM data set (Narayan et al., 2018). The XSUM data set consists of BBC articles and their highly condensed single-sentence summaries, making it ideal for generating key points. In

our research, the fine-tuning is carried out using the *ArgKP_2021* data set with statements from 25 of the 31 topics in the data set used for training. Participants of the ArgMining workshop used the concatenation of topic and statement as input and the corresponding key point as output to fine-tune the language model for the described abstractive approach (Friedman et al., 2021). In contrast, our research uses the concatenation of all statements explicitly assigned to a key point as input and the corresponding key point as output. This results in a smaller training data set of only 176 samples. In terms of the structure of input and output, this corresponds to a summary of similar statements to a key point in the context of a political debate on Twitter. Wolhandler et al. (2022) have found out that fine-tuning based on a single concatenated text rather than multiple source documents per training sample achieves a better performance. Details of the fine-tuning can be found in the [Github repository](#).

2.3 Twitter data preparation

2.3.1 Data collection

One of the aims of our research is to apply key point generation to the previously unaddressed domain of social media data. In order to collect a text corpus consisting of controversial topics and characteristics similar to the *ArgKP_2021* data set, we collected 258,184 statements of British parliamentarians active on Twitter between 3 September 2021 and 3 January 2022 using the Twitter API.¹ We took the account data of Members of Parliament (MPs) from a freely available source,² which contains, in addition to the account names, other metadata of interest for the analysis, such as party affiliation and constituency. We performed the iterative collection of Twitter data at 12-h intervals on an AWS EC2 instance using the *tweepy* library (see [Figure 3](#)). In addition to the tweet ID, timestamp, tweet text, and username, we added information on the party affiliation and constituency of each MP from the above source to the raw data and stored it in a MySQL database. The code, the account data of the MPs, and the structure of the SQL table are available in the [Github repository](#).

2.3.2 Data pre-processing

In a first step of the pre-processing pipeline, our research only considers tweets written between 2 September 2021 and 2 January 2022. We filtered out Retweets and duplicates, converted the text to lower case, and removed URLs and mentions. Neither did we remove stop words, nor did we apply stemming or lemmatization. These pre-processing steps do not add value to further processing using word embeddings (Schofield et al., 2017; Camacho-Collados and Pilehvar, 2018; Hickman et al., 2022) but may result in a loss of information. We excluded parties represented by <1,000 tweets. In a final step, this research uses a fine-tuned RoBERTa language model (see Section 2.2.1) to perform a binary classification into non-argumentative and argumentative tweets.

1 <https://developer.twitter.com/en>

2 <https://www.politics-social.com/>

2.3.3 Topic modeling

In a final step of Twitter data preparation, we used the BERTopic framework (Grootendorst, 2022) to segment the pre-selected and now pre-processed tweets into rough topics. Following the procedure of Grootendorst (2022), the tweets are transformed into word embeddings by using the SBERT transformer *all_mpnet_base_v2* (Reimers and Gurevych, 2019). In the second step, a dimensional reduction is performed using UMAP (McInnes et al., 2018) to prepare the input for the subsequent HDBSCAN clustering (Campello et al., 2013, 2015). Besides the minimum cluster size, set to 2% of the data set size, we used the standard parameters of the BERTopic framework. The clustering results were generated using the scripts of the [Github repository](#).

2.4 Key point generation

2.4.1 Topic selection

In order to perform key point generation on the Twitter data set created in chapter 2.3, we considered only one of the previously generated topics per run. According to Bar-Haim et al. (2020a), the aim of a key point analysis is to compare the arguments of two polarizing parties in the form of a list condensed into key points. Contrary to previous research (Bar-Haim et al., 2020a,b, 2021; Friedman et al., 2021), we did not determine the polarity of the two opposing parties by stance or sentiment on a topic but by party affiliation. The aim of the topic selection was to create a subset consisting of the statements of a political party on one of the latent topics identified in the course of topic modeling. Therefore, for all the topics listed in the [Table 2](#), key points can be generated for the political parties Conservative, Labor, Liberal Democrat, and Scottish National Party.

2.4.2 Sub-topic modeling

2.4.2.1 Objective

Our research followed the study of Reimers et al. (2019), who semantically segment the data space based on statement similarity. To achieve this segmentation, we used a procedure based on the BERTopic (Grootendorst, 2022) framework. This approach aims at identifying similar statements in an unsupervised manner, in order to summarize them in the subsequent step of key point summarization using an abstractive method. Furthermore, the BERTopic procedure makes it possible to quantify the statements by cluster affiliation. Although this does not meet the requirements of the original definition of key point matching (Bar-Haim et al., 2020a) in terms of granularity, it does provide an approximate overview of the distribution of semantically similar statements in a text collection. To ensure a certain stability of the topic modeling with respect to the adaptation to unknown data and data structures, our research complemented topic modeling with hyperparameter optimization. The code for the corresponding AzureML HyperDrive implementation is available in the [Github repository](#).

2.4.2.2 Word Embeddings

To determine the similarity of tweets for further processing, they are transformed into corresponding 768 dimensional vector

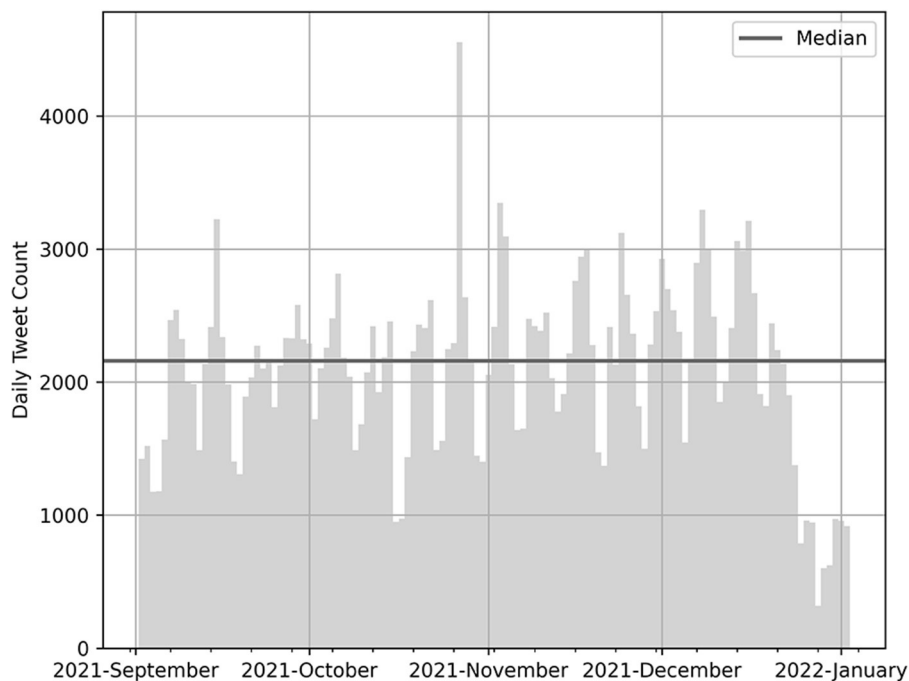


FIGURE 3
Number of daily tweets of British MPs retrieved via Twitter API between 2 September 2021 and 2 January 2022.

representations via SBERT transformer *all_mpnet_base_v2* (Reimers and Gurevych, 2019), a procedure that corresponds to (Grootendorst, 2022). In a recent benchmark study (Muennighoff et al., 2023), the clustering and semantic textual similarity performance of this small-scale transformer, pre-trained mainly on user-generated data, was demonstrated. Following Reimers et al. (2016), we used the cosine distance d_C (see Equation 1) as a distance measure for pairwise determination of the distance between two vectors \vec{a} and \vec{b} .

$$d_C(\vec{a}, \vec{b}) = 1 - \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|} \quad (1)$$

In the equation above, the symbols are defined as follows:

- $d_C(\vec{a}, \vec{b})$ is the cosine distance between vectors \vec{a} and \vec{b} ,
- \vec{a} and \vec{b} are vector representations of tweets,
- $\vec{a} \cdot \vec{b}$ denotes the dot product of vectors \vec{a} and \vec{b} ,
- $\|\vec{a}\|$ and $\|\vec{b}\|$ represent the Euclidean norm (magnitude) of vectors \vec{a} and \vec{b} , respectively.

2.4.2.3 Dimension reduction

The effectiveness of the UMAP dimension reduction method on the cluster quality of subsequent HDBSCAN clustering has recently been described (Allaoui et al., 2020). According to McInnes et al., the reduction of high-dimensional data to low-dimensional representations also had a positive effect on the computation time of the HDBSCAN clustering procedure (McInnes and Healy, 2017). When using UMAP, the adjustment of hyperparameters has a significant influence on the final low-dimensional mapping of

the data space. The most important parameters are the definition of the distance metric, the number of neighbors considered in the original data space, the number of desired target dimensions, and the height of the distance between the final low-dimensional representatives. To determine the distance of the data points in the original data space, we used the cosine distance (see Equation 1) because it is suitable for comparing high-dimensional data. UMAP allows greater control over the distance between the generated low-dimensional representatives than alternative dimension reduction methods such as t-SNE (van der Maaten and Hinton, 2008). We set the distance to 0, so that similar tweets had the highest possible density concentration in the target data space and are thus clearly separated for the subsequent clustering as described by Allaoui et al. (2020). In order to map relationships between data points in the original UMAP data space in a likelihood graph, the number of neighbors to be considered for the similarity calculation must also be specified. Since both the number of neighbors considered to generate the high-dimensional image and the desired number of target dimensions depend strongly on the nature of the data set being analyzed, we optimized these two parameters during hyperparameter tuning (see Section 2.4.2.5).

2.4.2.4 Clustering

Each generated key point should semantically cover as many homogeneous statements as possible to reflect a single aspect of a debate (Bar-Haim et al., 2021). Therefore, we used the density-based clustering method HDBSCAN (Campello et al., 2013) to identify density centers of semantically similar statements without prior knowledge of the number of relevant density centers. Data points within regions of lower density are automatically labeled

as noise (Campello et al., 2013). On the one hand, this has the advantage that only similar statements with a sufficient density concentration are considered for further key point summarization. On the other hand, it reduces the influence of the data noise often described in Twitter data (Barbosa and Feng, 2010; Derczynski et al., 2013) by excluding tweets that cannot be clearly classified.

The HDBSCAN clustering is mainly influenced by four parameter settings. A Euclidean distance metric (see Equation 2) is used to measure the distance between data points in the UMAP dimension-reduced data space.

$$d_E(\vec{a}, \vec{b}) = \sqrt{\sum_{i=1}^n (b_i - a_i)^2} \quad (2)$$

In the equation above, the symbols are defined as follows:

- $d_E(\vec{a}, \vec{b})$ represents the Euclidean distance between the points \vec{a} and \vec{b} ,
- \vec{a} and \vec{b} denote the position vectors of the two points in the n -dimensional UMAP reduced space, with components a_i and b_i , respectively,
- The summation $\sum_{i=1}^n (b_i - a_i)^2$ calculates the square of the differences between the corresponding components of points \vec{a} and \vec{b} ,
- n is the dimensionality of the space after dimension reduction via UMAP.

In order to understand the remaining parameter settings, the functionality of HDBSCAN is explained in more detail. First, the mutual reachability distance d_{mr} (Campello et al., 2013) is determined to assign data points to regions of low or high density. d_{mr} of the two data points a and b , the core distance $c_k(a)$ or $c_k(b)$, and the parameter k must be determined. The parameter k represents the number of neighboring data points that have to be reached starting from a point x before its core distance $c_k(x)$ can be determined. The core distance $c_k(x)$ corresponds to the radius which in turn corresponds to the Euclidean distance from x to the k -nearest neighbor. To provide a basis for classifying data points into low or high density, the mutual reachability distance d_{mr-k} is calculated dependent on the parameter k . The determination of $d_{mr-k}(a, b)$ (see Equation 3) of the two data points a and b is done by selecting the maximum value of the core distance $c_k(a)$, the core distance $c_k(b)$, and the Euclidean distance $d_E(a, b)$.

$$d_{mr-k}(a, b) = \max\{c_k(a), c_k(b), d_E(a, b)\} \quad (3)$$

In the equation above, the symbols are defined as follows:

- $d_{mr-k}(a, b)$ is the mutual reachability distance between data points a and b ,
- $c_k(x)$ denotes the core distance of a point x , defined as the Euclidean distance from x to its k -nearest neighbor,
- k is the parameter indicating the minimum number of neighbors required to consider a point's core distance,
- $d_E(a, b)$ represents the Euclidean distance between points a and b .

Prim's algorithm (Prim, 1957) is then used to generate a weighted graph in the form of a minimum spanning tree, which defines data points as vertices and weights edges between data points according to d_{mr-k} . By sorting the edges in ascending order based on their weight, a cluster hierarchy is created. The next step is to compress the fine-grained hierarchy using the min_{pts} parameter to determine the minimum cluster size. Finally, as described in the research study by Campello et al. (2013), the resulting clusters are extracted based on their cluster stability. In our research, we used the leaf method to extract all leaf nodes to subdivide the data space as finely as possible.

We set the distance metric and the clustering method as described. We set the parameter min_{pts} to one fiftieth of the size of the data set, so that only clusters representing a significant proportion of a debate are formed. In our research, only the parameter k , which determines the k -nearest neighbors to compute the core distance $c_k(x)$, was optimized by hyperparameter tuning.

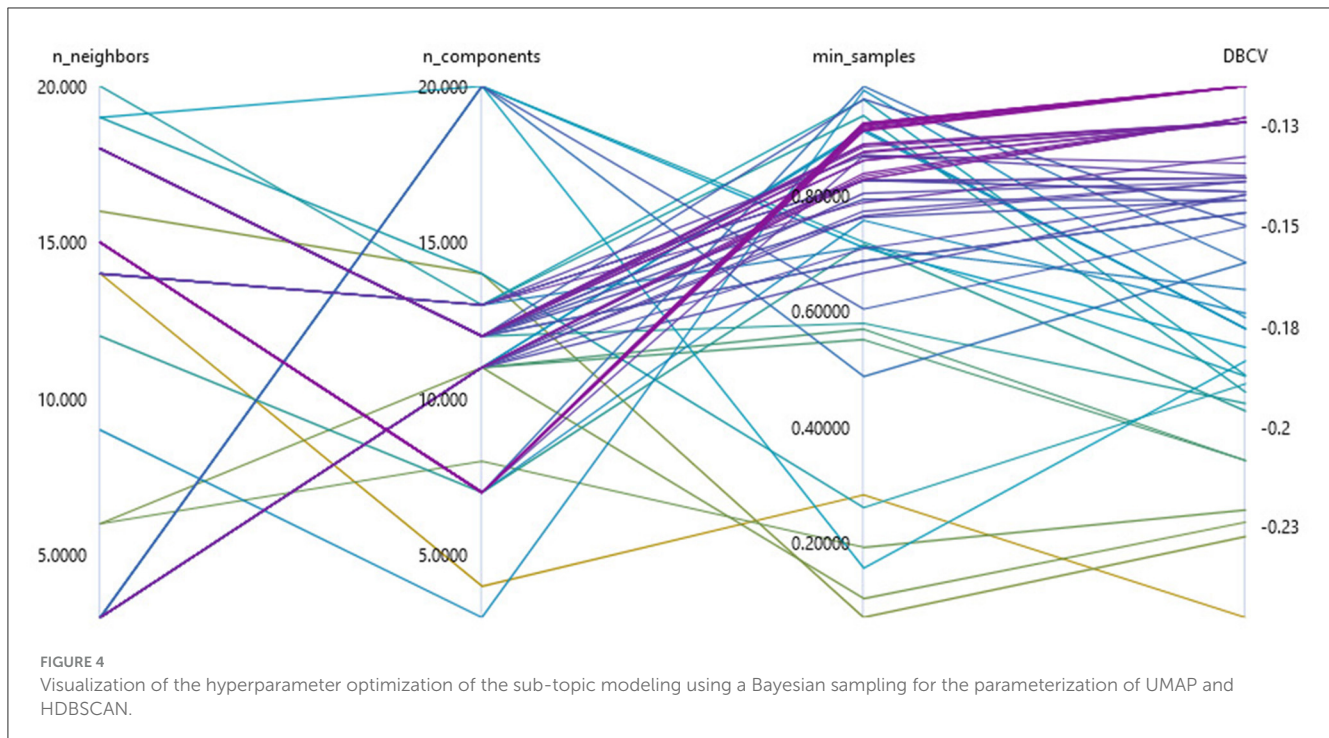
2.4.2.5 Hyperparameter tuning

Our research used hyperparameter optimization to stabilize the automatic detection of all relevant density centers of similar tweets through topic modeling. The definition of an optimal clustering strongly depends on the examined data characteristics and the actual use case (Hennig, 2015). Accordingly, the selection of parameters to be optimized, the sampling method, and the evaluation metric for determining the model quality must be individually adapted to the use case.

In our research, a total of three parameters were optimized to apply topic modeling to data with unknown structures (see Figure 4). For the UMAP method, we determined the optimal number of neighbors considered to create the high-dimensional image in the interval of [3,20] and the optimal number of target dimensions in the interval of [3,20]. To optimize the hyperparameters without prior knowledge of the structure of the data set under consideration, we used parameters that represent a multiplier of the data set size or the minimum cluster size. We determined the parameter k of the HDBSCAN clustering by the product of the minimum cluster size min_{pts} and the hyperparameter to be optimized k_f in the interval of [0.05,1]. We chose this procedure because the parameter k is limited to the minimum cluster size. We set the value of the minimum cluster size as described above, to one fiftieth of the data set size.

In our research, a sampling method based on a Bayesian optimization with a Gaussian process (Snoek et al., 2012) had been chosen so that the hyperparameters were optimized with regard to the required hardware resources and convergence. Specifically, we applied the HyperDrive implementation of Bayesian sampling within an AzureML cloud environment (Ranjit et al., 2019) using a *Standard_NC6S_v2* compute cluster with four compute nodes. The selection of parameters for a previously defined hyperparameter space is carried out by considering all evaluation results of previous procedures. This allows a probabilistic determination of which parameter configuration has the highest probability of optimizing model performance. Bayesian sampling achieves a significantly optimized parameter configuration with only a few samples (Wu et al., 2019).

In the course of Bayesian optimization, the quality of the model is determined by an evaluation metric. Density-based



clustering algorithms such as HDBSCAN, with relatively high noise, pose special requirements for a target metric (Palacio-Niño and Berzal, 2019). Evaluation metrics for unsupervised clustering such as the silhouette coefficient (Rousseeuw, 1987), the Calinski-Harabasz index (Caliński and Harabasz, 1974), or the Davies-Bouldin Index (Davies and Bouldin, 1979) determine the cluster quality from the ratio of intra-cluster dispersion and inter-cluster separation. However, dispersion and separation are determined by a distance metric and are therefore not suitable for evaluating density-based methods (Palacio-Niño and Berzal, 2019). They also do not handle noise adequately (Moulavi et al., 2014). Given the characteristics of density-based methods in the context of hyperparameter optimization, we used density-based clustering validation (DBCV) as the optimization metric (Moulavi et al., 2014). The DBCV index in the interval of $[-1, 1]$ also takes into account the influence of noise by considering all data points in the evaluation of the global cluster validity (Moulavi et al., 2014).

2.4.3 Key point summarization

2.4.3.1 Key Point Generation as abstractive summarization

Our research proposes a key point summarization once the data space has been segmented into semantic density centers. To generate a key point, we used abstractive methods to summarize the semantic content of identified sub-topics in a concise sentence. Previously, key points have often been generated using extractive summarization methods (Bar-Haim et al., 2020b, 2021), selecting highly representative and high-quality statements from the text collection itself. However, individual tweets rarely represent the entire content spectrum of their cluster, a fact partly explained

by the non-convex shape of density centers generated by sub-topic modeling. They also rarely meet formal requirements such as word length and structure. In contrast, abstractive summarization methods are able to generate a coherent summary without redundancies but with the appropriate information density and shape (Gupta and Gupta, 2019). To implement such an abstractive approach, we fine-tuned a PEGASUS (Zhang et al., 2020) language model, as described in Chapter 2.2.2.

2.4.3.2 Parameter setting

To generate concise key points, we set the maximum word length of the key points to 25. Our research determined the optimal minimum word length during hyperparameter tuning in an interval of $[8, 20]$. This ensured that key points have a sufficient number of words to represent the content of their cluster. Furthermore, we optimized the hyperparameter p in the interval $[0.2, 0.95]$, which represents a probability threshold for selecting potential candidates for generating the next word of a sequence during the top- p sampling (Holtzman et al., 2019) procedure. In each step of sequence generation, the smallest possible pool of potential word candidates is selected whose cumulative probability for a given vocabulary V exceeds the threshold value p . According to Holtzman et al. (2019), the pool of potential word candidates V_p for each generated word x in a sequence of length i is determined as follows:

$$\sum_{x \in V_p} P(x|x_1:i-1) \geq p \quad (4)$$

In the equation above, the symbols are defined as follows:

- p denotes the probability threshold for selecting potential word candidates during the top- p sampling process,

- V represents the entire vocabulary from which word candidates are chosen,
- V_p is the subset of V consisting of word candidates whose cumulative probability meets or exceeds the threshold p ,
- x refers to a potential word candidate within the sequence,
- $x_{1:i-1}$ denotes the sequence of words generated up to the current step $i - 1$,
- $P(x|x_{1:i-1})$ is the conditional probability of the candidate word x given the sequence of previously generated words $x_{1:i-1}$.

Furthermore, top-p sampling is combined with top-k sampling (Fan et al., 2018). The parameter k , which is optimized in the interval [10,100], indicates the maximum number of potential word candidates that will be considered for each step of sequence generation. This additional parameter prevents words with a very low probability from being considered as word candidates in the course of top-p sampling.

Since the maximum input sequence length of the language model is limited to 512 tokens, we filtered the most representative statements for each cluster. Therefore, we calculated the average cosine distance of each tweet to all other tweets in its cluster. In a final step, we concatenated the statements with the lowest average cosine distance up to a token length of 512.

2.4.3.3 Hyperparameter tuning

Automated evaluation of key point summarization is not trivial. Previous research has taken different approaches to evaluating the results of key point or opinion summarization: An automatic evaluation via the ROUGE value (Bražinskas et al., 2020a,b), a human evaluation (Suhara et al., 2020; Friedman et al., 2021) or by using a labeled ground truth data set and the obtained precision and coverage (Bar-Haim et al., 2020b, 2021). Since human evaluation in the course of hyperparameter tuning was excluded and since the aim of our research is an unsupervised application of key point generation to unknown data, two of the above options were already omitted. In order to perform an automated evaluation that does not assess the quality of the summaries at the word level, as the ROUGE metric (Lin, 2004) does, but also at the semantic similarity of the key points generated, this research performed an automated evaluation on the basis of word embeddings. As for the topic modeling, we used the SBERT Transformer *all_mpnet_base_v2* (Reimers and Gurevych, 2019) and the cosine distance to determine the semantic similarity. We evaluated automated key point summarization (exemplarily shown in Figure 5) using a modified version of the Davies-Bouldin Index (Davies and Bouldin, 1979) proposed in this research. This index, originally designed to assess cluster quality, is calculated from the ratio of intra-cluster dispersion and inter-cluster separation. We calculated the intra-cluster dispersion S_i (see Equation 5) from the average cosine distance d_C of all data points T_i of a cluster i to the cluster centroid A_i . Since a key point should semantically cover all data points in its cluster, we considered it to be the centroid of its cluster. A low cluster dispersion therefore suggests a key point that represents the semantics of its own cluster well.

$$S_i = \frac{1}{T_i} \sum_{j=1}^{T_i} d_C(X_j, A_i) \quad (5)$$

In the equation above, the symbols are defined as follows:

- S_i represents the intra-cluster dispersion for cluster i , indicating the average distance of all data points within the cluster to the cluster centroid,
- T_i denotes the total number of data points within cluster i ,
- $d_C(X_j, A_i)$ is the cosine distance between a data point X_j and the cluster centroid A_i ,
- X_j refers to the j -th data point within cluster i ,
- A_i is considered the centroid of cluster i , representing the key point that semantically covers all data points in its cluster.

We modified the calculation of cluster separation described by Davies and Bouldin (1979) and calculated the cluster separation M for each combination of generated key points (see Equation 6). We derived the cluster separation M_{ij} of the two clusters i and j by calculating the cosine distance d_C of the two cluster key points A_i and A_j . This is in line with the aim of generating a list of key points, which should not be redundant and should each represent only one aspect of a debate (Bar-Haim et al., 2021).

$$M_{ij} = d_C(A_i, A_j) \quad (6)$$

In the equation above, the symbols are defined as follows:

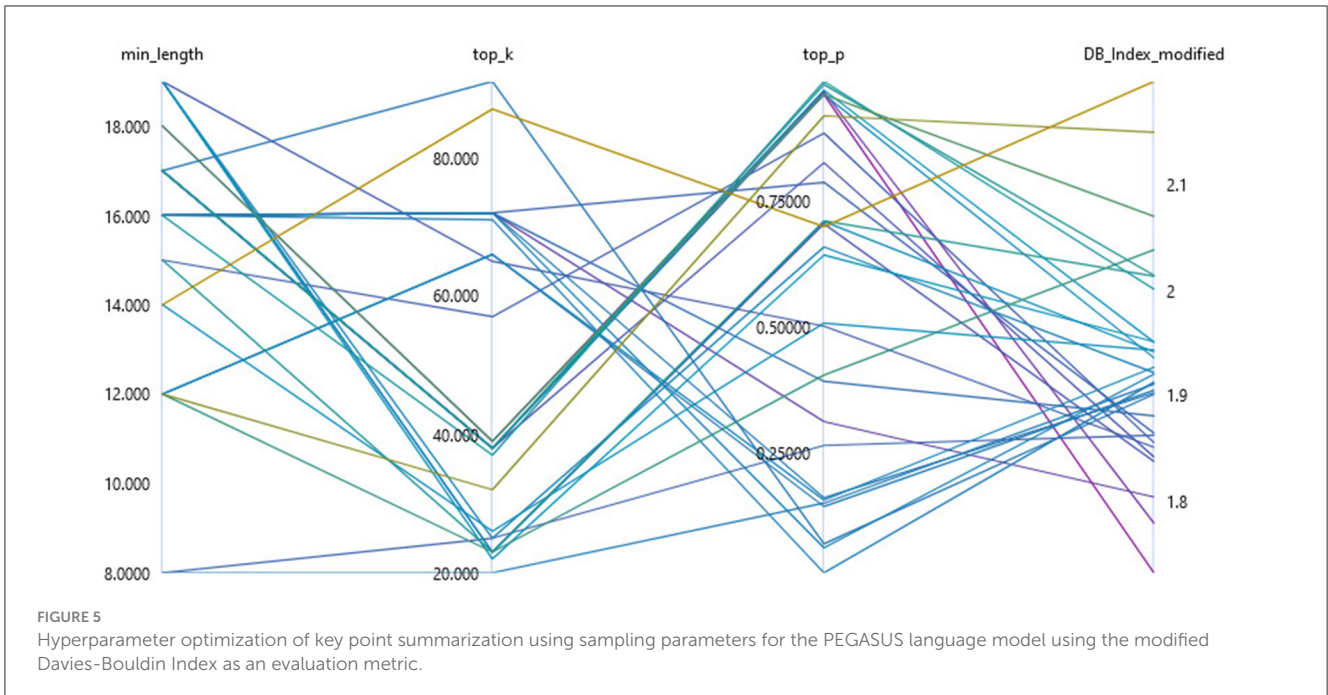
- M_{ij} denotes the separation between clusters i and j , intended to measure the dissimilarity between the clusters' key points,
- $d_C(A_i, A_j)$ represents the cosine distance between the key points A_i and A_j of clusters i and j , respectively,
- A_i and A_j are the centroids of clusters i and j , effectively serving as the key points that summarize the main concepts of their respective clusters.

Davies and Bouldin (1979) determine the relation R between cluster dispersion and cluster separation for each cluster combination (see Equation 7). The relation R_{ij} of the two clusters i and j is defined by the ratio of the sum of the cluster dispersion S_i and S_j to the corresponding cluster separation M_{ij} .

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \quad (7)$$

In the equation above, the symbols are defined as follows:

- R_{ij} quantifies the relation between the dispersion within clusters i and j and the separation between them, serving as a measure of clustering efficiency,
- S_i and S_j represent the intra-cluster dispersion for clusters i and j , respectively, indicating the average distance of data points within each cluster to their centroid,
- M_{ij} denotes the separation between clusters i and j , calculated as the cosine distance between their respective centroids (key points).



According to [Davies and Bouldin \(1979\)](#), the worst relation R_i is determined for each cluster i by taking the maximum value of all its cluster relations. Finally, the Davies-Bouldin Index \bar{R} is calculated by taking the average of the worst relations R_i of all the clusters (see [Equation 8](#)).

$$\bar{R} = \frac{1}{N} \sum_{i=1}^N R_i \quad (8)$$

In the equation above, the symbols are defined as follows:

- \bar{R} represents the Davies-Bouldin Index, which serves as a measure of the quality of clustering based on cluster dispersion and separation,
- R_i is the worst relation for cluster i , determined as the maximum relation among all cluster relations for that cluster,
- N denotes the total number of clusters in the clustering result.

2.5 Evaluating key point generation

To determine the quality of key point generation in an automated approach, our research considered the procedure as a multidocument summarization task. Quality of a multidocument summarization is assessed by comparing a generated summary with a reference summary of all documents. Since there is no generated summary available, we considered the concatenated list of all generated key points of a topic to be the generated summary. Similarly, we derived the reference summary from the concatenation of the key points of a topic, manually annotated by experts in the field.

In our research, the evaluation was based on the *ArgKP_2021* data set, which had already been used as a training data set for abstractive summarization (see [Section 2.2.2](#)). For the evaluation,

TABLE 1 Overview of each pre-processing step as well as the absolute number and relative proportion of the remaining tweets in the raw data set.

Pre-processing pipeline	
Processing step	Number tweets
Raw data extraction via Twitter API	258,184 (100%)
Delimitation of the survey period	257,149 (99.6%)
Filter retweets and duplicates	120,419 (46.6%)
Filter URLs and mentions	118,862 (46%)
Filter parties with <1,000 tweets	114,475 (44.3%)
Filter non-argumentative tweets	34,998 (13.6%)

we only considered six topics of the data set that had not been used in the model training. Since the topics are additionally divided into arguments with a positive or negative stance toward the topic, we included a total of twelve subsets in the evaluation.

We used the word-based metric ROUGE ([Lin, 2004](#)) and the word-embedding-based metric BERTScore ([Zhang T. et al., 2019](#)) as evaluation criteria. We chose ROUGE because it has become the standard metric for evaluating summaries ([Fabbri et al., 2021](#)). The lexical similarity is calculated in the course of the evaluation on the basis of unigrams (ROUGE-1) and bigrams (ROUGE-2).

To evaluate beyond the lexical dimension, we complemented the assessment with BERTScore. This metric computes a cosine similarity between the corresponding word embeddings for each pairwise token of the generated and reference summaries. Thus,

TABLE 2 Topics of the political debate on Twitter identified within the data set, including associated number of tweets and topic label.

Topic	Count	Topic label
0	3,413	social_care_tax_universal
1	2,634	great_local_work
2	2,445	climate_cop_green_energy
3	2,379	uk_people_trade
4	2,286	women_violence_police
5	2,143	covid_vaccine_booster
6	1,588	rules_MPs_tory_corruption
7	849	rail_transport_north_services
8	702	remembrance_armed_service
-1	16,559	Not assigned to a topic

the metric is able to detect semantic overlap between the reference and the generated summary even when there is no lexical overlap (Zhang T. et al., 2019). We chose “microsoft/deberta-xl-large-mnli” (He et al., 2021) as the underlying model for generating word embeddings as it has the highest correlation with human evaluation in terms of predicted similarity (Zhang T. et al., 2019). The BERTScore is given in the interval of [0, 1], where a higher value represents a higher precision, recall, or f1 score.

Our research carried out the evaluation per topic and stance of the evaluation data set. Only statements that can be explicitly assigned to a topic were taken into account. In order to perform the evaluation on subsets with an average number of 90 statements, we had to adjust the minimum cluster size parameter from one fiftieth to one twentieth of the data set under consideration. The code for the corresponding AzureML HyperDrive implementation of the evaluation is available in the [Github repository](#).

3 Results

3.1 Twitter data preparation

The first aim of our research was to prepare social media data for an exemplary application of key point generation to a political debate on Twitter. A total of 258,184 tweets from British MPs were obtained via the Twitter API between the 2 September 2021 and 3 January 2022. During the pre-processing pipeline (see Table 1), the number of tweets was reduced to 34,998. In particular, the two steps of filtering retweets and duplicates (257,149 → 120,419) and non-argumentative tweets (114,475 → 34,998) significantly reduced the data volume. To identify latent topics, we segmented the remaining data set into rough clusters using topic modeling (Grootendorst, 2022). This resulted in a total of nine topics with an average number of 2,049 tweets and 16,559 statements that were not assigned to any cluster (see Table 2).

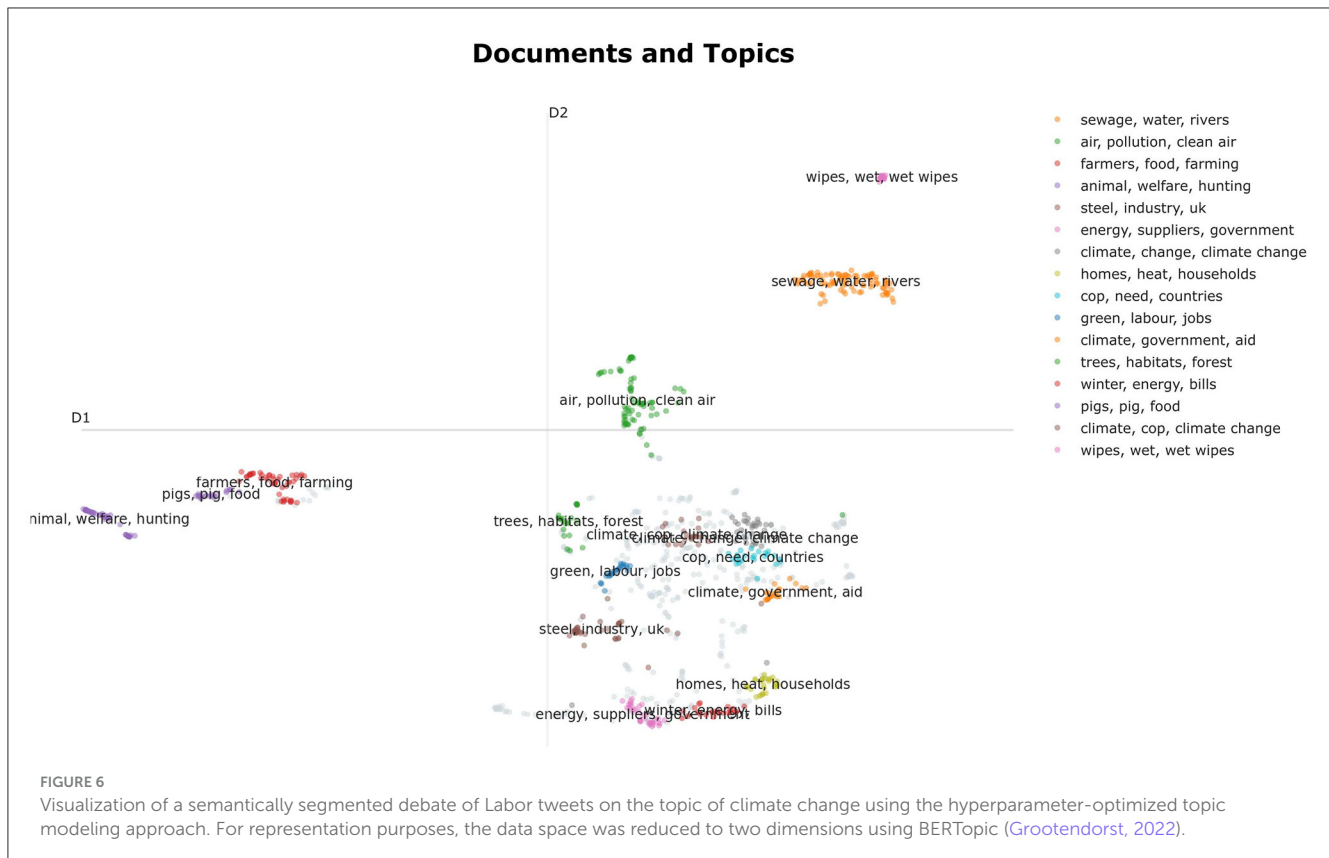
3.2 Sub-topic modeling

The next research objective was to develop a method to identify clusters of semantically similar tweets, as a basis for the subsequent key point summarization. For this purpose, we used a customized variant of BERTopic (Grootendorst, 2022), which we complemented by hyperparameter tuning. In the course of this procedure, we considered statements of one party on a specific topic. To perform topic modeling, we generated word embeddings of the statements using an SBERT transformer (Reimers and Gurevych, 2019). Then, we used UMAP (McInnes et al., 2018) to reduce the dimensions of the embeddings to prepare them for the subsequent HDBSCAN clustering (Campello et al., 2013) as suggested by Grootendorst (2022). We used the evaluation metric DBCV (Moulavi et al., 2014) to optimize relevant parameters of UMAP and HDBSCAN during hyperparameter tuning. We set parameters that depend on the size and structure of the data set relative to the size of the data set, so that the procedure could be automatically applied to unknown data without obvious limitations. As an example, the procedure of sub-topic modeling was applied to the Labor Party’s statements on the topic “climate_cop_green_energy”. The result of the sub-topic modeling was visualized in a dimensionally reduced representation using the BERTopic framework (see Figure 6).

3.3 Key point generation

The main goal of key point generation is to create a list of key points that semantically cover a party’s central statements on a given topic. Therefore, our research used the textual cluster contents of the sub-topic modeling results and concatenated them until the maximum input sequence length of the PEGASUS language model (Zhang et al., 2020) was reached. We prioritized statements based on their average cosine similarity to other statements in their cluster. We then summarized the concatenated statements using a PEGASUS language model that had been fine-tuned specifically for this use case. Our research evaluated KPG using a modified Davies-Bouldin Index (Davies and Bouldin, 1979), which calculates cluster dispersion and separation using the average distance to each key point rather than cluster centroids. We used this metric to optimize the minimum sequence length and the sampling parameters of the sequence generation in the course of hyperparameter tuning. Since each key point is supposed to represent the contents of its cluster, we considered the corresponding cluster size as an indicator of the key point’s share in the overall debate.

To assess the quality of the key point generation, we performed an automated evaluation using the *ArgKP_2021* data set. The approach taken in our research treats KPG as a summarization problem, where the concatenation of key points represents the summary 2.5. The quantitative results (see Table 3) show that the number of generated key points strongly deviates from the number of key points annotated by experts. Furthermore, the recall values for the evaluation metric



BERTScore are on average higher than for the corresponding precision values. Finally, the table shows that the ROUGE metric has a high variance compared to the corresponding BERTScore value.

We analyzed an exemplary result of a KPG on the topic of celibacy (see [Supplementary Table S1](#)) qualitatively. Overall, the key points that are generated largely reflect the semantics of the annotated key points. However, the generated key points generally show an increased number of words, occasional semantic redundancies, and partially uncovered content of the reference key points. In addition, a high number of statements are not assigned and the quantification of the key points differs greatly from the information provided by the experts. In order to analyze the adaptability to the domain of political debate on Twitter, we applied KPG to Conservative and Labor statements on the topic of “climate_cop_green_energy”, which roughly covers all topics related to climate change (see [Supplementary Tables S2, S3](#)). The generated key points largely meet the requirements of [Bar-Haim et al. \(2021\)](#) in terms of validity, informativeness and aspect focus. In some cases, however, the key points are formulated generically (see key point IDs in [Supplementary Tables S2, S3](#) Conservative: 2,9,12 Labor: 3,14), they do not contain complete sentences (Conservative: 8) or they do not provide all the essential information (Labor: 8,10). Furthermore, some key points contain non-informative fragments of sentences and are often formulated subjectively.

4 Discussion

4.1 Conclusion

In this research, we have successfully implemented an innovative unsupervised method for KPG, aligned with the key point analysis paradigm proposed by [Bar-Haim et al. \(2020a, 2021\)](#). Our approach, based on topic modeling and enhanced by hyperparameter tuning, not only segments the data space but also approximates the frequency of statements associated with each key point, filling a gap in previous KPG methodologies ([Bar-Haim et al., 2021](#)). Our decision to adopt an abstractive summarization approach, as opposed to traditional extractive methods ([Bar-Haim et al., 2020b, 2021](#); [Alshomary et al., 2021](#)), proved advantageous. It allowed for a broader semantic representation within key points, capturing different aspects of a subtopic more effectively. This was facilitated by optimizing the minimum sequence length during hyperparameter tuning, ensuring that key points more accurately reflected the content of the associated statements. Furthermore, we incorporated the Davies-Bouldin Index, modified to suit our research, into the hyperparameter tuning process. This procedure both optimizes the generated key points to better represent the semantic content of their cluster and at the same time differentiates them from other key points in terms of their semantic content, as exemplarily shown in [Figure 5](#).

Our research introduces an evaluation method for KPG that uses both ROUGE, a word-based metric, and BERTScore, a word-

TABLE 3 Results of the evaluation of key point generation based on the (ArgKP_2021) data set using the metrics ROUGE-1, ROUGE-2, and BERTScore (Precision, Recall, f1).

Topic	Stance	ROUGE		BERTScore			Number key points	
		1	2	P	R	f1	given	generated
The USA is a good country to live in	Positive	31.02	6.49	0.57	0.66	0.62	7	8
	Negative	16.67	2.8	0.54	0.63	0.58	7	11
Social media platforms should be regulated by the government	Positive	39.52	13.33	0.57	0.6	0.65	5	9
	Negative	37.13	14.55	0.57	0.66	0.61	5	10
Routine child vaccinations should be mandatory	Positive	31.72	16.78	0.62	0.69	0.66	5	9
	Negative	23.07	7.78	0.5	0.63	0.56	4	9
The vow of celibacy should be abandoned	Positive	22.95	6.67	0.63	0.73	0.67	5	7
	Negative	41.07	12.73	0.64	0.69	0.66	6	5
Assisted suicide should be a criminal offense	Positive	31.28	6.7	0.59	0.64	0.61	4	10
	Negative	29.71	12.72	0.56	0.69	0.62	6	12
Homeschooling should be banned	Positive	17.71	4.21	0.49	0.61	0.54	4	11
	Negative	43.14	11.88	0.61	0.68	0.64	6	8
Average:		30.42	9.72	0.57	0.66	0.61	5.33	9.08

embedding-based metric, by considering KPG as a multidocument summarization problem. This dual-metric approach allows for a more nuanced evaluation of the generated key points, capturing not only lexical similarity but also semantic coherence with the source documents. By framing KPG in the context of multidocument summarization, we acknowledge the complexity of distilling key information from multiple statements, and our evaluation method is designed to reflect both the precision of word choice and the depth of contextual understanding.

In conclusion, while our methodology was specifically designed to analyse data from the dynamic and often unstructured environment of social media, its theoretical foundation suggests potential applicability to a variety of contexts. The unsupervised principles underlying our approach, particularly in dealing with diverse and complex datasets through density-based clustering and subsequent hyperparameter tuning, may provide valuable insights for domains such as product reviews and customer surveys. However, it is important to note that the direct applicability of our methods to these domains remains to be fully explored and validated. The initial success in the social media context lays a promising foundation and encourages further exploration of the adaptability of our KPG techniques. This exploration not only addresses our initial research question but also invites future studies to rigorously test and expand the reach of our methodology and evaluation in various fields of key point analysis.

4.2 Limitations

4.2.1 Theoretical limitations

Apart from the advantages described above, our research has some theoretical limitations that need to be addressed. The quantitative evaluation revealed an overgeneration of key points compared to expert references, often leading to semantic redundancy, as reflected in the low precision of the BERTScore (see Table 3). This problem was less pronounced in denser semantic spaces such as political debates with more examples. In addition, the qualitative analysis showed that the key points generated sometimes did not match the criteria of Bar-Haim et al. (2021) (see Section 3.3). This suggests that treating key points as complete representations of clustered statements may be an oversimplification, especially given the observed semantic redundancy and deviation from established criteria.

Our research indirectly attempts to quantify the statements associated with a key point by approximating the cluster size, but this method deviates from the actual values 3.3. This discrepancy is due to the complexity of key point matching, where key points may not align exclusively with a single cluster. In addition, the use of HDBSCAN clustering often labels many statements as noise, making it difficult to accurately determine the relative prominence of a key points in political debates.

Our study also acknowledges the limitations of the ROUGE metric in evaluating abstractive approaches, as reflected in the

variance of ROUGE scores in our results (see Table 3) and confirmed by previous research as humans tend to formulate abstractive summaries that are mostly not equivalent in terms of word choice (Schluter, 2017). The fluctuations of the ROUGE value observed in Chapter 3.3 can be attributed to a property of ROUGE.

Another theoretical limitation arises from the inherent nature of our methodology, which includes the stochastic aspect of topic modeling and subsequent hyperparameter tuning. This process, when run repeatedly, can produce different results on the same dataset due to its optimization dynamics, adding another layer of complexity and uncertainty to our research approach.

4.2.2 Practical limitations

Our research has identified critical practical limitations that impact the effectiveness and broader applicability of our approach. A primary concern is the evaluation's credibility, which is compromised by our reliance on the scarcity of high-quality, rich datasets for short, opinionated statements. This limitation not only restricts the conclusiveness and generalizability of our findings but also severely hampers the fine-tuning of our language models. Such fine-tuning is crucial for effective key point generation. Consequently, this directly affects our ability to robustly validate the effectiveness of our topic modeling and hyperparameter tuning efforts, potentially compromising the accuracy and robustness of our results.

Second, the applicability of the KPG approach is somewhat limited, focussing primarily on social media data and political debates within the *ArgKP_2021* dataset. This specialization limits the broader applicability of our findings across different data types or domains.

Finally, the inherent limitations of the language models we use, in particular their constrained input sequence length, have an impact on our methodology. This constraint forces us to use only cluster representatives of a certain token length for abstractive summarization, which limits the quality of summarization in terms of coverage.

These practical limitations highlight the need for continuous refinement and development of our approach and methods.

4.3 Future work

Acknowledging these practical limitations not only informs the current state of our research but also provides a clear path for our future research in the field of KPG. A key part of this endeavor will be an extensive comparative analysis using different datasets such as *ArgKP_2021* (Bar-Haim et al., 2021), *Opinosis* (Ganesan et al., 2010), and *SPACE* (Angelidis et al., 2021). This study will not only benchmark our proposed KPG method against existing extractive methods (Bar-Haim et al., 2020b, 2021) but also explore its scalability and adaptability across different domains, data sizes, and types.

In parallel, we plan to enhance our evaluation techniques. While automated metrics such as BERTScore (Zhang T. et al.,

2019) have been valuable, incorporating comprehensive human scoring will provide deeper qualitative insights. This combination of automated and human scoring will ensure a more holistic assessment of the key points generated.

Addressing the challenge of semantic redundancy in KPG is also critical. We aim to develop advanced algorithms and refine techniques within frameworks such as BERTopic (Grootendorst, 2022) to minimize redundancy. This effort will also include exploring post-processing strategies to enforce semantic distinctiveness, ensuring the uniqueness and relevance of each key point. We also intend to conduct a detailed evaluation using crowd annotators, inspired by the methodology of Friedman et al. (2021). This approach will help us identify specific error patterns and understand user perspectives on the relevance and quality of key points.

The exploration of advanced language models is another key aspect of our future work. We will be experimenting with models such as BART (Lewis et al., 2020), T5 (Raffel et al., 2020), BLOOM (Scao et al., 2022), and their long form variants such as BigBird PEGASUS (Zaheer et al., 2020) and LongT5 (Guo et al., 2022). Our focus will be on evaluating how these models, with their enhanced input sequence capabilities, can improve the depth and quality of KPG.

An important consideration in our research will be the evaluation of computational efficiency, especially when using advanced sentence similarity methods such as BERTScore (Zhang T. et al., 2019) and MoverScore (Zhao et al., 2019). We aim to balance the trade-off between performance improvements and resource requirements, taking into account the practical implications for large-scale applications.

There will also be a focus on methodological innovation, in particular on optimizing the clustering process and experimenting with new hyperparameter tuning approaches. Inspired by recent advances described in Bacanin et al. (2022), we will explore the use of meta heuristic techniques for hyperparameter tuning. These techniques, known for efficiently navigating complex parameter spaces, show promise for improving model performance. This approach is especially pertinent for tasks like abstractive summarization, where precise control over multiple parameters is crucial.

Finally, we plan to expand the scope of our research to include multilingual and cross-cultural contexts. This extension will involve adapting the KPG methodology to different languages and cultural settings, thereby broadening its applicability and relevance.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and

institutional requirements. The social media data was accessed and analyzed using the Twitter API.

Author contributions

JS: Writing—review & editing, Supervision, Conceptualization, and Project administration. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

PE would like to thank his supervisor Prof. Dr. Rüdiger Buchkremer for the opportunity to write a study on this exciting topic and for his timely, challenging and empowering guidance throughout his research. PE would also like to thank Dr. Tim Delbrügger, Dr. Sindy Neumann, Dr. Dawid Kopetzki, and Elif Nallar Özkan of iits-consulting/ImpressSol GmbH for their confidence and assistance. This project would not have been possible without the support of his company and his university.

References

- Allaoui, M., Kherfi, M. L., and Cheriet, A. (2020). “Considerably improving clustering algorithms using umap dimensionality reduction technique: a comparative study,” in *Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings 9* (Berlin: Springer), 317–325.
- Alshomary, M., Gurcke, T., Syed, S., Heinisch, P., Spliethöver, M., Cimiano, P., et al. (2021). “Key point analysis via contrastive learning and extractive argument summarization,” in *Proceedings of the 8th Workshop on Argument Mining* (Stroudsburg, PA), 184–189.
- Angelidis, S., Amplayo, R. K., Suhara, Y., Wang, X., and Lapata, M. (2021). Extractive opinion summarization in quantized transformer spaces. *Transact. Assoc. Comp. Linguist.* 9, 277–293. doi: 10.1162/tacl_a_00366
- Bacantin, N., Zivkovic, M., Al-Turjman, F., Venkatchalam, K., Trojovský, P., Strumberger, L., et al. (2022). Hybridized sine cosine algorithm with convolutional neural networks dropout regularization application. *Sci. Rep.* 12:6302. doi: 10.1038/s41598-022-09744-2
- Barbosa, L., and Feng, J. (2010). “Robust sentiment detection on twitter from biased and noisy data,” in *Coling 2010: Posters*, A. K. Joshi, C. R. Huang, and D. Jurafsky (Stroudsburg, PA: Association for Computational Linguistics), 36–44.
- Bar-Haim, R., Eden, L., Friedman, R., Kantor, Y., Lahav, D., and Slonim, N. (2020a). “From arguments to key points: Towards automatic argument summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 4029–4039.
- Bar-Haim, R., Eden, L., Kantor, Y., Friedman, R., and Slonim, N. (2021). “Every bite is an experience: Key point analysis of business reviews,” 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)* (Stroudsburg, PA: Association for Computational Linguistics), 3376–3386.
- Bar-Haim, R., Kantor, Y., Eden, L., Friedman, R., Lahav, D., and Slonim, N. (2020b). “Quantitative argument summarization and beyond: Cross-domain key point analysis,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA: Association for Computational Linguistics), 39–49.
- Bosc, T., Cabrio, E., and Villata, S. (2016). “Tweeties squabbling: positive and negative results in applying argument mining on social media,” in *COMMA 2016*, 21–32.
- Bražinskas, A., Lapata, M., and Titov, I. (2020a). “Few-shot learning for opinion summarization,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA: Association for Computational Linguistics), 4119–4135.
- Bražinskas, A., Lapata, M., and Titov, I. (2020b). “Unsupervised opinion summarization as copycat-review generation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 5151–5169.
- Calinski, T., and Harabasz, J. (1974). A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* 3, 1–27. doi: 10.1080/03610927408827101
- Camacho-Collados, J., and Pilehvar, M. T. (2018). “On the role of text preprocessing in neural network architectures: an evaluation study on text categorization and sentiment analysis,” in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Stroudsburg, PA), 40–46.
- Campello, R. J., Moulavi, D., and Sander, J. (2013). “Density-based clustering based on hierarchical density estimates,” in *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14–17, 2013, Proceedings, Part II 17* (Berlin: Springer), 160–172.
- Campello, R. J., Moulavi, D., Zimek, A., and Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transact. Knowl. Discov. Data* 10, 1–51. doi: 10.1145/2733381
- Davies, D. L., and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1*, 224–227. doi: 10.1109/TPAMI.1979.4766909
- Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). “Twitter part-of-speech tagging for all: Overcoming sparse and noisy data,” in *Proceedings of the International Conference Recent Advances in Natural Language Processing Ranlp 2013* (Stroudsburg, PA: ACL), 198–206.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “Bert: pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of NAACL-HLT* (Stroudsburg, PA: ACL), 4171–4186.
- Dusart, A., Pinel-Sauvagnat, K., and Hubert, G. (2023). “Tsubert: how to sum up multiple years of reading in a few tweets. *ACM Transact. Inf. Syst.* 41, 1–33. doi: 10.1145/3581786
- Erkan, G., and Radev, D. R. (2004). Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* 22, 457–479. doi: 10.1613/jair.1523
- Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). Summeval: re-evaluating summarization evaluation. *Transact. Assoc. Comp. Linguist.* 9, 391–409. doi: 10.1162/tacl_a_00373
- Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. (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* (Stroudsburg, PA: ACL), 1074–1084.

Conflict of interest

PE was employed by iits-consulting/ImpressSol GmbH.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2024.1200949/full#supplementary-material>

- Fan, A., Lewis, M., and Dauphin, Y. (2018). "Hierarchical neural story generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Stroudsburg, PA: ACL), 889–898.
- Friedman, R., Dankin, L., Hou, Y., Aharonov, R., Katz, Y., and Slonim, N. (2021). "Overview of the 2021 key point analysis shared task," in *Proceedings of the 8th Workshop on Argument Mining* (Stroudsburg, PA: ACL), 154–164.
- Ganesan, K., Zhai, C., and Han, J. (2010). "Opinois: a graph based approach to abstractive summarization of highly redundant opinions," in *Proceedings of the 23rd International Conference on Computational Linguistics* (Beijing: Coling 2010 Organizing Committee), 340–348.
- Ghalandari, D. G., Hokamp, C., Glover, J., and Ifrim, G. (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* (Stroudsburg, PA: ACL), 1302–1308.
- Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., et al. (2020). "A large-scale dataset for argument quality ranking: Construction and analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence, volume 34*, 7805–7813. doi: 10.1609/aaai.v34i05.6285
- Grootendorst, M. (2022). Bertopic: Neural Topic Modeling With a Class-Based tf-idf Procedure. *arXiv [Preprint] arXiv:2203.05794*.
- Guo, M., Ainslie, J., Uthus, D. C., Ontanon, S., Ni, J., Sung, Y.-H., et al. (2022). "Longt5: Efficient text-to-text transformer for long sequences," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 724–736. doi: 10.18653/v1/2022.findings-naacl.55
- Gupta, S., and Gupta, S. K. (2019). Abstractive summarization: an overview of the state of the art. *Expert Syst. Appl.* 121, 49–65. doi: 10.1016/j.eswa.2018.12.011
- Habernal, I., and Gurevych, I. (2016). "Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Stroudsburg, PA: ACL), 1589–1599.
- He, P., Liu, X., Gao, J., and Chen, W. (2021). "Deberta: decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations* (Appleton, WI: ICLR).
- He, W., Wang, F.-K., and Akula, V. (2017). Managing extracted knowledge from big social media data for business decision making. *J. Knowl. Manag.* 21, 275–294. doi: 10.1108/JKM-07-2015-0296
- Hennig, C. (2015). What are the true clusters? *Pattern Recognit. Lett.* 64, 53–62. doi: 10.1016/j.patrec.2015.04.009
- Hickman, L., Thapa, S., Tay, L., Cao, M., and Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: review and recommendations. *Organ. Res. Methods* 25, 114–146. doi: 10.1177/1094428120971683
- Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). "The curious case of neural text degeneration," in *International Conference on Learning Representations* (Appleton, WI: ICLR).
- Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discour. Process* 25, 259–284. doi: 10.1080/01638539809545028
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (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* (Stroudsburg, PA: ACL), 7871–7880.
- Lin, C.-Y. (2004). "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, eds. M. F. Moens and S. Szpakowicz (Stroudsburg, PA: Association for Computational Linguistics), 74–81.
- Liu, Y. (2019). Fine-tune bert for extractive summarization. *arXiv [preprint]*. doi: 10.48550/arXiv.1903.10318
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: a robustly optimized bert pretraining approach. *arXiv [preprint]*. doi: 10.48550/arXiv.1907.11692
- Lu, Y., Dong, Y., and Charlin, L. (2020). "Multi-xscience: a large-scale dataset for extreme multi-document summarization of scientific articles," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Stroudsburg, PA: ACL), 8068–8074.
- McInnes, L., and Healy, J. (2017). "Accelerated hierarchical density based clustering," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (Piscataway, NJ: IEEE), 33–42.
- McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* 3:861. doi: 10.21105/joss.00861
- Mihalcea, R., and Tarau, P. (2004). "Textrank: bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA: ACL), 404–411.
- Moulavi, D., Jaskowiak, P. A., Campello, R. J., Zimek, A., and Sander, J. (2014). "Density-based clustering validation," in *Proceedings of the 2014 SIAM International Conference on Data Mining* (SIAM), 839–847.
- Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). "Mteb: massive text embedding benchmark," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, (Stroudsburg, PA: ACL), 2006–2029.
- Narayan, S., Cohen, S., and Lapata, M. (2018). "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," in *2018 Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA: Association for Computational Linguistics), 1797–1807.
- Oved, N., and Levy, R. (2021). "Pass: Perturb-and-select summarizer for product reviews," 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)*, (Stroudsburg, PA: ACL), 351–365.
- Palacio-Niño, J.-O., and Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. *arXiv [preprint]*. doi: 10.48550/arXiv.1905.05667
- Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell Syst. Tech. J.* 36, 1389–1401. doi: 10.1002/j.1538-7305.1957.tb01515.x
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 5485–5551.
- Ranjit, M. P., Ganapathy, G., Sridhar, K., and Arumugham, V. (2019). "Efficient deep learning hyperparameter tuning using cloud infrastructure: intelligent distributed hyperparameter tuning with bayesian optimization in the cloud," in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)* (Piscataway, NJ: IEEE), 520–522.
- Reimers, N., Beyer, P., and Gurevych, I. (2016). "Task-oriented intrinsic evaluation of semantic textual similarity," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (Osaka: COLING 2016 Organizing Committee), 87–96.
- Reimers, N., and Gurevych, I. (2019). "Sentence-bert: sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Stroudsburg, PA: ACL), 3982–3992.
- Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., and Gurevych, I. (2019). "Classification and clustering of arguments with contextualized word embeddings," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA: ACL), 567–578.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7
- Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., et al. (2022). Bloom: a 176b-parameter open-access multilingual language model. *arXiv [preprint]*. doi: 10.48550/arXiv.2211.05100
- Schaefer, R., and Stede, M. (2021). Argument mining on twitter: a survey. *Inf. Technol.* 63, 45–58. doi: 10.1515/itit-2020-0053
- Schluter, N. (2017). "The limits of automatic summarisation according to rouge," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 41–45.
- Schofield, A., Magnusson, M., and Mimno, D. (2017). "Pulling out the stops: Rethinking stopword removal for topic models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Stroudsburg, PA: ACL), 432–436.
- Shirafuji, D., Rzepka, R., and Araki, K. (2021). "Argument extraction for key point generation using mmr-based methods. *IEEE Access* 9, 103091–103109. doi: 10.1109/ACCESS.2021.3097976
- Snoek, J., Larochelle, H., and Adams, R. P. (2012). "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2* (Red Hook, NY: Curran Associates Inc.), 2951–2959.
- Sotudeh, S., Deilamsalehy, H., Dernoncourt, F., and Goharian, N. (2021). "Tldr9+: a large scale resource for extreme summarization of social media posts," in *Proceedings of the Third Workshop on New Frontiers in Summarization* (Stroudsburg, PA: ACL), 142–151.
- Stieglitz, S., and Dang-Xuan, L. (2013). Social media and political communication: a social media analytics framework. *Soc. Netw. Anal. Mining* 3, 1277–1291. doi: 10.1007/s13278-012-0079-3
- Suhara, Y., Wang, X., Angelidis, S., and Tan, W.-C. (2020). "OpinionDigest: a simple framework for opinion summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA: Association for Computational Linguistics), 5789–5798.
- van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605.

- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Stroudsburg, PA: Association for Computational Linguistics), 38–45.
- Wolhandler, R., Cattan, A., Ernst, O., and Dagan, I. (2022). How "Multi" is Multi-Document Summarization?. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, eds. Y. Goldberg, Z. Kozareva, and Y. Zhang (Stroudsburg PA: Association for Computational Linguistics), 5761–5769.
- Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., and Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on bayesian optimization. *J. Electron. Sci. Technol.* 17, 26–40.
- Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., et al. (2020). Big bird: transformers for longer sequences. *Adv. Neural Inf. Process. Syst.* 33, 17283–17297.
- Zeng, D., Chen, H., Lusch, R., and Li, S.-H. (2010). Social media analytics and intelligence. *IEEE Intell. Syst.* 25, 13–16. doi: 10.1109/MIS.2010.151
- Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning* (PMLR), 11328–11339.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). "Huggingface's transformers: state-of-the-art natural language processing," in *International Conference on Learning Representations* (Appleton, WI).
- Zhang, X., Wei, F., and Zhou, M. (2019). "Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Stroudsburg, PA: ACL), 5059–5069.
- Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. (2019). "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Stroudsburg, PA: ACL), 563–578.