International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00434-4

REVIEW

Contrastive text summarization: a survey

Thomas Ströhle[1] · Ricardo Campos[2,3,4] · Adam Jatowt[1]

Received: 15 April 2023 / Accepted: 19 July 2023

Abstract
In our data-flooded age, an enormous amount of redundant, but also disparate textual data is collected on a daily basis on a wide variety of topics. Much of this information refers to documents related to the same theme, that is, different versions of the same document, or different documents discussing the same topic. Being aware of such differences turns out to be an important aspect for those who want to perform a comparative task. However, as documents increase in size and volume, keeping up-to-date, detecting, and summarizing relevant changes between different documents or versions of it becomes unfeasible. This motivates the rise of the contrastive or comparative summarization task, which attempts to summarize the text of different documents related to the same topic in a way that highlights the relevant differences between them. Our research aims to provide a systematic literature review on contrastive or comparative summarization, highlighting the different methods, data sets, metrics, and applications. Overall, we found that contrastive summarization is most commonly used in controversial news articles, controversial opinions or sentiments on a topic, and reviews of a product. Despite the great interest in the topic, we note that standard data sets, as well as a competitive task dedicated to this topic, are yet to come to be proposed, eventually impeding the emergence of new methods. Moreover, the great breakthrough of using deep learning-based language models for abstract summaries in contrastive summarization is still missing.

Keywords Contrastive summarization · Comparative summarization · Text summarization · Change analysis · Systematic literature review · Survey

1 Introduction

In our highly digitized world, an enormous amount of information in the form of text is collected on a daily basis. Especially, user-generated data are increasing enormously on the Internet through user reviews, news, blogs, social networks, idea contests, etc. This growth in text data has taken on exponential proportions, with large amounts of redundant information being collected [19].

Due to this overload of data, it is becoming increasingly difficult to pick out relevant information and obtain an overview of a topic. It takes a lot of time and cognitive effort to read and understand all the available content, making it increasingly impossible. People are faced with a lot of irrelevant, noisy and redundant content, so summarizing textual data is becoming more and more important [5] as a way to create a short summary of a single document or a set of documents while maintaining the sense of the content. Although automatic text summarization has always been an important application of natural language processing, it is now becoming more relevant due to the improved effectiveness of pre-trained deep natural language models [16].

Despite these advances, text summarization generally focuses on creating a single generic summary that covers the most common points of one or multiple documents, likely addressing the greatest similarities, as opposed to the greatest differences [17]. In practice, however, multiple redundant ideas, solutions, or different opinions are now produced with user-generated data on a large-scale basis, which motivates the identification of the most important differences as an

✉ Thomas Ströhle
thomas.stroehle@uibk.ac.at

✉ Ricardo Campos
ricardo.campos@ubi.pt

✉ Adam Jatowt
adam.jatowt@uibk.ac.at

[1] University of Innsbruck, Innsbruck, Austria

[2] University of Beira Interior, Covilhã, Portugal

[3] INESC TEC, Porto, Portugal

[4] Ci2 - Smart Cities Research Center - Polytechnic Institute of Tomar, Tomar, Portugal

Published online: 09 August 2023

alternative or complement to common text summarization approaches. This is especially relevant when the content is diverse and varied, making it difficult for humans to effectively process and understand all the distinct information available. By automating the process of comparing and summarizing text, one can possibly extract relevant information more efficiently and gain diverse insights from large amounts of data. For example, user comments on two different accommodations in one tourist destination can be comparatively summarized [25]. Furthermore, the different opinions of several political parties can be summarized comparatively [6, 44]. Therefore, it is of great importance, in addition to detecting the key similarities, to identify the differences and reduce redundancies in the text summaries.

This particular task of natural language processing can be found in the literature under the term *contrastive summarization* or *comparative summarization* [23, 25, 34, 48]. Unlike generic summarization, query-oriented summarization, and update summarization [51], contrastive summarization addresses the problem of generating a summary that highlights the differences between multiple texts.

In this paper, we conduct a research overview on the topic of *Contrastive/Comparative summarization*, which, despite its high importance and many promising applications, no standardized evaluation benchmark has yet been developed [16].

Furthermore, our paper is intended to highlight possibilities and promising research avenues that can still be explored and considered. In particular, our review is structured as follows. Section 2 defines and explains the problem and describes, in detail, the systematic literature conducted in this paper. Section 3 presents the main data sets and evaluation criteria used for contrastive summarization. Section 4 presents and compares the given approaches, methods, and techniques. All presented papers are then classified according to their methods used, while trends and future research directions are also identified. Section 5 introduces the most important applications of contrastive summarization. We will show that the main focus of contrastive summarization is usually to find different opinions on a certain topic, different reviews of a product, or differences in newspapers. Section 6 concludes our research article by highlighting our findings and further research opportunities.

2 Problem definition and literature selection

Contrastive summarization was first mentioned in [34] as the joint creation of summaries for two entities to highlight their differences. In their study, product reviews of 56 electronic products with about 70 ratings per product were used to collectively create two summaries that highlight the differences between two products.

Three years later, [56] narrowed the definition of Comparative summarization to an extractive summarization approach (*Comparative extractive document summarization*) where the task is to summarize differences between comparable groups of documents. In contrast to [34, 56] comparatively summarizes more than two groups of documents, using a method that extracts the most discriminatory sentences for each group.

More recently, [25] presented this problem as *Comparative opinion summarization*. In this work, the authors aim to produce two contrastive summaries and a common summary from two different candidate sets of reviews. As an example, several reviews of one hotel are compared with reviews of another hotel. The two contrastive summaries stem from the differences of the respective reviews, while the common summary represents the intersection of the two. Liu et al. [40], on the other hand, and in contrast to [25, 34], defines one-to-many comparative summarization task, where one document is compared to many others.

Our understanding is that contrastive or comparative summarization is an attempt to summarize the text from different documents related to the same theme in such a way that the relevant differences in the text become highlighted. This results in summaries that either describe the differences between different versions of the same document (e.g., the same document with its different versions) or the differences between different documents which discuss the same subject or are comparable in a certain way (e.g., the differences between two political party programmes).

Having this defined, we adopted a systematic literature research that uses the following logical expression as a means to obtain relevant literature to the discussed theme[1]:

$$(\text{comparative} \vee \text{contrastive} \vee \text{compare} \vee \text{contrast} \vee \text{contrasting} \vee \text{comparing}) \wedge (\text{summarization} \vee \text{summarization} \vee \text{summarizing} \vee \text{summarising} \vee \text{summarize} \vee \text{summarise} \vee \text{summaries})$$

To perform the search, we used the dblp computer science bibliography primarily as a search basis and verified it with ACL Anthology and Google Scholar. The different combinations of search queries yielded 138 publications as of April 12, 2023. We then conducted a curation process that involved reading all titles and abstracts. By doing this, we found that some publications include a comparative study or a comparative evaluation or are only about contrastive learning. A related topic of Timeline Summarization was also found along with some non-English language publications. The removal of these publications resulted in a final set of 26 relevant publications, all listed in Table 1. The number

[1] The search took into account that Summarization is written with a "z" in American English and with an "s" in British English.

of citations was determined using Google Scholar[2], and the *Method* column is intended to provide a brief overview of the approach used.

Table 1 shows that in addition to [34], who were the first to address the problem, [30, 44, 56] are the most cited publications, with more than 50 citations. Along with this, we also show the method used in each research paper. A previous survey of this area has been conducted in the past by [32], however, in their work, the authors only focused on Opinion summarization and surveyed only a small subset of 6 publications [20, 30, 34, 43, 44, 53]. Our work extends this survey by including in the analyses a larger number of publications on the topic of Contrastive/Comparative summarization and covers also ones published after 2019.

Before delving into the details of each research work, we proceed by setting up the data sets and the evaluation metrics that are often used in the community for evaluation purposes.

3 Data sets and evaluations metrics

Although several research works have been proposed over the years, the lack of standard data sets and/or a competition task dedicated to this topic has limited a comparison of the different contrastive summarization methods proposed so far and makes it difficult to thoroughly understand their strengths and weaknesses, ultimately impeding the emergence of new methods. Despite this, a few data sets have been used and/or adapted, compare Table 2. However, none has established the ground to become a reference data set.

The efforts done by the research community have also led to the emergence of different metrics used to evaluate the proposed methods. The *Overall Responsiveness* scale, a manual human-based Likert scale rating, is one such metric. In this metric, raters assign an overall score for each summarization based, for example, on content as well as readability [24].

Another measure is the *Comparative Aspect Recall* scale that is used to measure the effectiveness of comparative extraction, defined as the number of human-agreed comparative aspects in the summary [24].

The *Aspect Coverage*, on the other hand, appears as an alternative, by measuring the number of unique aspects collected in the summary divided by the number of unique aspects labeled by human annotators [30].

Also, the *precision* and *recall* metrics can be adapted to the contrastive realm by counting the number of sentences in the automatically generated contrastive summary and the manually generated one. The *F-Measure* combines both measures. A formal representation of each of the metrics is given below. Let a and m be the number of sentences in the automatically generated contrastive summary and the manually written contrastive summary, respectively; let c be the number of human-agreed correct comparative sentences in the automatically generated summary, then the precision, recall, and F-measure are given as follows [23]:

$$\text{Precision} = \frac{c}{a},$$
$$\text{Recall} = \frac{c}{m},$$
$$\text{F} = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}.$$

Higher precision means that more correct contrastive sentences are retrieved in the automatically generated contrastive summary, and high recall means that more correct contrastive sentences are retrieved in the summary when compared to all those that were manually labeled as relevant in the reference summaries. The F-measure averages both values, becoming larger when both measures are balanced.

Another standardized procedure for evaluating automatically generated contrastive summaries is the *ROUGE Metric* toolkit. All ROUGE metrics count the number of units overlapping between the candidate summaries and the reference summaries [38]:

- ROUGE-N measures the overlap of n-grams between a candidate summarization and a reference summarization, that is, ROUGE-1 measures the overlap of unigrams and ROUGE-2 measures the overlap of bigrams.
- ROUGE-L measures the longest common subsequence between a candidate summary and a reference summary.
- ROUGE-W is an improvement of ROUGE-L that weights the measure of longest common subsequence with word order.
- ROUGE-SU is a combination of Skip-bigram and unigram-based measure.

The main disadvantage of ROUGE metrics, as well as precision, recall, and F-measure, is that they are more suitable for extractive contrastive summarization, since similar synonym words cannot be considered [42]. Summaries that are semantically very similar, but which may not use common words, will then receive a low score. With abstractive summarization, it is not mandatory that the same content is described with the same word phrases, i.e., a small overlap of words to the comparison summary is possible, despite the content being well captured. Thus, in contrast to standardized scoring metrics that rely on word overlap, human-based measures such as Overall Responsiveness can be used for abstractive summaries.

Another development that has only been used by [25] in the contrastive summarization literature is the use of semantic similarities using pre-trained language models to evaluate

[2] The citation count lookup was made on April 12, 2023.

Table 1 Selected literature

Title	Author(s)	Method	Year	Citations
Contrastive summarization: an experiment with consumer reviews	Lerman and McDonald	Optimization problem	2009	80
Generating comparative summaries of contradictory opinions in text	Kim and Zhai	Optimization problem	2009	154
Summarizing contrastive viewpoints in opinionated text	Paul et al	Topic modeling, graph-based model	2010	179
Comparative news summarization using linear programming	Huang et al	Optimization problem	2011	39
Lightweight contrastive summarization for news comment mining	Raveendran and Clarke	Statistical approach	2012	5
Query sensitive comparative summarization of search results using concept based segmentation	Chitra et al	Statistical approach	2012	6
Comparative document summarization via discriminative sentence selection	Wang et al	Optimization Problem	2012	94
Generating comparative summaries from reviews	Sipos and Joachims	Optimization problem	2013	34
Research on contrastive viewpoint summarization for opinionated texts	Liang et al	Topic modeling, graph-based model	2013	1
Topic models for comparative summarization	Campr and Jezek	Topic modeling	2013	10
Comparative summarization via latent Dirichlet allocation	Campr and Jezek	Topic modeling	2013	7
Concise comparative summaries (CCS) of large text corpora with a human experiment	Jia et al	Machine learning approach	2014	22
Comparative news summarization using concept-based optimization	Huang et al	Optimization problem	2014	23
Contrastive max–sum opinion summarization	Özsoy and Çakici	Optimization problem	2014	6
Summarizing contrastive themes via hierarchical non-parametric processes	Ren and de Rijke	Topic modeling	2015	28
Expert-guided contrastive opinion summarization for controversial issues	Guo et al	Topic modeling, machine learning approach	2015	15
Exploring differential topic models for comparative summarization of scientific papers	He et al	Topic modeling	2016	16
Comparative document summarization via classification	Bista et al	Machine learning approach	2018	5
Comparative summarization of rich media collections	Bista	Machine learning approach, neural network approach (pre-trained word embeddings)	2019	2
A survey on contrastive opinion summarization	Lavanya and Parvathavarthini	Survey	2019	1
Context-sensitive contrastive feature-based opinion summarization of online reviews	Lavanya and Parvathavarthini	Topic modeling	2020	0
Context aware contrastive opinion summarization	Lavanya and Parvathavarthini	Neural network approach (LSTM)	2020	0
Comparative opinion summarization via collaborative decoding	Iso et al	Neural network approach (transformer)	2021	4

Table 1 continued

Title	Author(s)	Method	Year	Citations
Comparative graph-based Summarization of scientific papers Guided	Chen et al	graph-based model, neural network approach (pre-trained BERT)	2022	0
One-to-many comparative summarization for patents	Liu et al	Graph-based model, optimization problem, neural network approach (transformer)	2022	0
Building contrastive summaries of subjective text via opinion ranking	Rocha da Silva and Salgueiro Pardo	Optimization problem, statistical approach	2022	0

generated contrastive summaries [41]. Iso et al. [25] compute semantic distances using a pre-trained BERT model.

In the following, we describe in detail each of the methods used by different research works considered in this study to perform a contrastive summarization.

4 Contrastive summarization methods

As with automatic text summarization, contrastive summarization can be divided into the following approaches: *Extractive*, *Abstractive*, or *Hybrid*. In the extractive approach, the important sentences are selected from the input text and used to create the summary. In the abstractive approach, a summary is created using words and phrases that can be different from the original text sentences. Finally, the hybrid approach combines both extractive and abstractive methods. In the context of Contrastive summarization works, all 26 publications studied in this survey, except [25], follow an extractive approach. Extractive text summarization approaches can be further subdivided according to their methods that will guide us through the rest of this chapter.

We refer to optimization-based, topic-based, statistical, machine learning-based, neural network-based, and graph-based methods. Most publications use a combination of these methods to achieve the best possible result, which we now present in more detail.

4.1 Optimization problems

The extractive summarization task can be defined as an optimization problem, where sentences are transformed into vectors with many statistical or linguistic features and are selected through multiobjective optimization functions.

Lerman and McDonald [34]'s work is based on an optimization problem called the Sentiment Aspect Match model, which attempts to summarize contrastiveness based on the sentiment scores of different user reviews of electronic products with a pre-specified length constraint. To evaluate the contrasting summaries, the authors asked 55 online reviewers to find differences between two contrasting summaries and to identify the usefulness of these differences, giving three ratings for each summary on a four-star Likert scale. The results show that about 80% of the raters were able to find at least two points of contrast in the summaries generated by the contrastive summarization, compared to 40% for the summaries generated by the simple Sentiment Aspect Match model, meaning that the contrastive summarization clearly outperforms the single product summary.

In addition, [52] use sentiment scores in their optimization framework, which highlight differences between entities in an opinionated text and satisfy the following three characteristics: Representativeness (presence of opinions that are common in the input), contrastiveness (presence of opinions that highlight differences between entities) and diversity (presence of different opinions to avoid redundancy). A specially developed score (from 0 to 100) for representativeness, contrast and diversity is used to validate their contrastive summaries of consumer reviews. It is compared with [30, 34], whose method achieves an overall score (harmonic mean of all three measures over all data sets) of 82, compared with 59 [30, 34], respectively.

Instead of sentiment scores, [30] apply word overlaps and semantic word matches to define content similarity and contrastive similarity within the context of optimization problems. They use existing product review summaries from [22] and two human reviewers to identify representative pairs of contrastive sentences from this data. The results of their experiments show that the proposed methods are effective in producing contrastive opinion summaries, with the contrastivity-first (0.54 precision, 0.80 aspect coverage) approach performing better than the representativeness-first (0.50 precision, 0.74 aspect coverage) approach.

In [23], a linear optimization problem with term frequency and inverse document frequency is used, which optimizes between the comparativeness within the summary and the representativeness of the topics. Singular sentences with the most different vocabulary from all sentences in the other documents are selected, without considering synonyms. They collected their own data set consisting of articles of differ-

Table 2 Data sets, metrics and applications

Research article(s)	Data description	Metric(s)	Application(s)
Lerman and McDonald [34]	Compiled data set of consumer electronics reviews for 56 electronics products, with an average of 70 reviews per product from a variety of sources, including CNET, Epinions, and PriceGrabber. Data covers 15 categories of electronics, including MP3 players, digital cameras, laptops, GPS systems, and more	Four-star Likert scale from 0 (very useful) to 3 (not useful)	User-generated reviews
Kim and Zhai [30]	14 tagged product review data sets from [22] containing Amazon reviews are scored by two human evaluators to identify representative contrastive sentence pairs	Precision, Aspect Coverage	User-generated reviews
Paul et al. [44]	The Gallup® telephone survey on the 2010 US health care bill [29], which is a set of 948 verbatim responses with an even mix of the two viewpoints (45% for and 48% against), and the Bitterlemons corpus, a collection of 594 editorials on the Israel-Palestine conflict with 312 articles by Israeli authors and 282 articles by Palestinian authors [39]	ROUGE metric	User-generated opinions and arguments
Huang et al. [23]	Special data set of ten related news articles for each topic (Haiti Earthquake vs. Chile Earthquake, Chile Mining Accident vs. New Zealand Mining Accident, Iraq Withdrawal vs. Afghanistan Withdrawal, Apple iPad 2 vs. BlackBerry Playbook, 2006 FIFA World Cup vs. 2010 FIFA World Cup) using Google News search engine and a manually written comparative summary for each pair of topics	Precision, Recall, F-measure and ROUGE metric	Newspaper articles
Raveendran and Clarke [48]	No data set	No validation	Newspaper articles and comments
Chitra et al. [11]	A collection of 200 Web documents from the Internet related to educational institutions, algorithms, banking, and household items	Overall responsiveness via the Five-Star Likert scale	Web pages
Wang et al. [56]	Blog entries without comments collected by NEC's internal blog crawler in 2005 and 2006, and 279 abstracts of computer science research papers	ROUGE metric	User-generated opinions from blogs and scientific articles
Sipos and Joachims [53]	They used scraped reviews from the Amazon website for eight tablets (from different manufacturers) and decomposed them into sentences based on punctuation to automatically match pairs of snippets describing the opinions of reviewers on different features of two products	Manual validation	User-generated reviews
Liang et al. [37]	The Gallup® telephone survey [29], compare [44]	ROUGE metric	User-generated opinions and arguments

Table 2 continued

Research article(s)	Data description	Metric(s)	Application(s)
Campr and Jezek [7, 8]	A specially created data set with data from the TAC 2011 conference containing 100 news articles, divided into 10 topics with 10 articles each, is used	ROUGE metric	Newspaper articles
Jia et al. [27]	A set of 2009 New York Times International section articles for 15 different countries (China, Iran, Iraq, Afghanistan, Israel, Pakistan, Russia, France, India, Germany, Japan, Mexico, South Korea, Egypt, and Turkey)	Five-star Likert scale	Newspaper articles
Huang et al. [24]	Same data set as [23]	ROUGE metric, Comparative Aspect Recall, Overall Responsiveness	Newspaper articles
Özsoy and Çakici [43]	The same English data set as [30] and an own scraped Turkish data set about 1400 user reviews and ratings for 31 movies from *beyazperde.com*	Precision, Aspect coverage	User-generated reviews
Ren and de Rijke [49]	The Gallup® telephone survey [29], compare [44]	ROUGE metric	User-generated opinions and arguments
Guo et al. [20]	Own data set on the topic of gay marriage by collecting expert opinion data (procon.org) and Twitter data, providing controversy and various arguments	Precision, Coverage	User-generated opinions and arguments, Twitter data
He et al. [21]	35 papers with 6,636 sentences from the ACL Anthology Searchbench	ROUGE metric	Scientific articles
Bista [1], Bista et al. [2]	News articles on three controversial topics (Beef Ban—controversy over the slaughter and sale of beef for religious reasons, Gun Control—restrictions on carrying, and Capital Punishment—use of the death penalty) that appeared between June 2017 and July 2018 to comparatively summarize news articles from different time periods to determine what changed about the topic between the summarization periods	Human evaluation experiment	Newspaper articles
Lavanya and Parvathavarthini [31]	Uses the same product review data set as in [30] and the benchmark car data set [18] with 42,230 reviews for about 140–250 car models	ROUGE metric	User-generated reviews
Lavanya and Parvathavarthini [33]	The SemEval 2014 Task 4 restaurant reviews data set [46], which consists of 1,125,457 documents, is reduced to a sample of 50,000 documents for validation	ROUGE metric	User-generated reviews
Iso et al. [25]	A special comparative opinion summarization corpus containing human-written contrastive and common summaries for 48 pairs of entities that are sampled from the TripAdvisor corpus [57]	ROUGE metric, BERT score	User-generated reviews

Table 2 continued

Research article(s)	Data description	Metric(s)	Application(s)
Chen et al. [10]	A corpus of scientific summarization based on comparative citations is built using citations as a guide, and the DUC2006 and DUC2007 data sets	ROUGE metric	Scientific articles (using the citation linking structure)
Liu et al. [40]	A special comparative patent summarization data set of real-world patentability of approximately 28,000 infringement analysis reports on various topics provided by patent law firms and patent search engines	Precision, Recall, F-measure	Patents (using the citation linking structure)
Rocha da Silva and Salgueiro Pardo [52]	Opinions on four products, two cameras, and two smartphones, which are extracted from a Brazilian website (www.buscape.com.br) with a total of 542 comments	A self-developed rating (from 0 to 100) for representativeness, contrast and diversity	User-generated reviews and opinions

ent newspapers on comparable topics and used precision, recall, F-measure, and ROUGE metric toolkits to validate their comparative summarization approach. In summary, their linear programming-based comparative model (0.36 precision, 0.42 recall, 0.39 F-measure, 0.43 ROUGE-1), which focuses on both comparability and representativeness at the same time, achieves better performance in both comparison extraction and summarization than the self-defined non-comparative model (0.24 precision, 0.26 recall, 0.25 F-measure, 0.40 ROUGE-1) and the co-ranking model (0.31 precision, 0.29 recall, 0.29 F-measure, 0.43 ROUGE-1).

Similarly, [56] use a combinatorial optimization problem designed with the use of document-sentence representation: Sentences are considered as features, and the challenge of selecting discriminant sentences is formulated as a sentence-based feature selection problem. The combinatorial optimization problem is estimated with a multivariate normal generative model and the developed sequential selection method. Unlike [23, 30, 34], the features are no longer based on single words, but on entire sentences, taking into account the sentence–document and sentence–sentence relationships. The authors used blog entries without comments and abstracts from computer science research papers as a data set to create groups from the data and identify differences between them. The ROUGE metric is used to validate their three types of contrastive summarization with human-generated summaries. Their approach yielded a ROUGE-1 of 0.53 in the research papers, compared to seven other approaches that fall within the ROUGE-1 range of 0.21–0.32. The experiments demonstrate the effectiveness of their proposed method, which benefits from the document-sentence representation.

In the work of [53], a vector of features with bag-of-words and tf-idf scores is used to formulate an optimization problem that generates short and comparative summaries based on product reviews. The authors manually labeled their data for supervised learning and validation. In summary, their approach selects pairs of review snippets in such a way as to produce a summary comparison of the product that outperforms a naive self-defined baseline solution by more than 20%.

Huang et al. [24] employ a concept-based optimization approach, which uses cross-topic pairs of semantically related concepts as evidence of comparability and topic-related concepts as evidence of representativeness. The similarities of the tf-idf measure based on whole sentences and sparse one-hot encodings are used in the optimization. The authors use the same data set as [23] and rate each summary with a Five-Star Likert scale for both content and readability/fluency. In addition, their contrastive summarization approach is validated using the ROUGE metric. The authors show that comparative analysis of related news topics is useful in many applications and that with the use of linear programming, their model (3.5 Comparative Aspect Recall, 3.4 Overall Responsiveness, 0.27 ROUGE-2 in English data set) outperforms the self-defined non-comparative model (1.9 Comparative Aspect Recall, 2.6 Overall Responsiveness, 0.22 ROUGE-2) and the co-ranking model (2.3 Comparative Aspect Recall, 2.9 Overall Responsiveness, 0.22 ROUGE-2) in comparative extraction and summarization.

Finally, in [43] a max sum optimization problem is formulated. Word-level differences are determined by cosine similarity with tf-idf and are used in summed form for whole sentences and documents. This approach attempts to determine a list of pairs of the most representative sentences

related to a given aspect, where each pair contains a positive and a negative sentence that have contrasting meanings, e.g., "*The design is really well done.*" as a positive sentence and "*But my biggest criticism is still the extremely ugly design.*" as a negative sentence. The same English data set as [30] and a self-created Turkish data set on user reviews are used for validation and comparison with [30]. Their approach obtained 0.65 precision and 0.89 aspect coverage using term frequency and $\lambda = 0.80$ versus 0.50 precision and 0.74 aspect coverage [30]'s representativeness-first approach and 0.54 precision and 0.80 aspect coverage [30]'s contrastivity-first approach. They obtained better results and observed that using cosine similarity with term frequency for computations performed better than using tf-idf in their max sum optimization.

4.2 Topic models

The use of topic models as a means is also a widely adopted approach by the research community. In general, works that implement such a technique aim to identify the topic of a document more precisely before selecting the desired sentences based on these topics.

One of the most popular methods in this regard is the Topic-Aspect Model, an extension of Latent Dirichlet Allocation (LDA). It is a Bayesian mixed model that jointly discovers topics and aspects. In [44], the Topic-Aspect Model is used as the first step to build multiple topics and extract viewpoints in combination with a Comparative LexRank, an adaptation of the PageRank algorithm, to contrast these viewpoints. The Gallup® telephone survey on the 2010 US health care bill [29] and the Bitterlemons corpus [39] are used as pre-existing sources in their contrastive summarization approach. They used the ROUGE metric to validate their contrastive summaries and compared it to a standard LexRank summarization approach and to [34]'s approach. Their approach yielded 0.43 ROUGE-1 without stop words compared to 0.36 for the standard LexRank summarization approach and 0.35 for the [34] approach. The results show that their method outperforms both comparison approaches, and thus, their approach can produce more informative summaries of viewpoints in opinionated texts.

In addition to the Topic-Aspect Model, [37] also use various graph-based centrality scoring approaches, namely the basic LexRank, Comparative LexRank, Topic-sensitive tf-idf LexRank, Topic-sensitive tf-idf & Comparative LexRank, and Biased & Comparative LexRank, to rank the centrality of each sentence. The sentences with the highest centrality values for each topic are selected to form the contrastive summaries. They, like [44], use the same Gallup® telephone survey [29] to build contrasting viewpoints and summaries and use the ROUGE metric for validation. Empirical experiments show that all proposed methods have similar ROUGE-1 precision values (ranging from 0.08 to 0.11) and can effectively perform the summarization task, especially the proposed topic-sensitive tf-idf & Comparative LexRank method could be used for both multi-topic summarization and contrastive viewpoint summarization with high performance.

In [7, 8], Latent Semantic Analysis and Latent Dirichlet Allocation are used to determine the different topics of the documents. These topics are then compared, and the main differences are used to select the most important different sentences to build the contrastive summaries. A specially created data set with data from the TAC 2011 conference containing news articles is used in combination with the ROUGE metric for evaluation. The average scores for both algorithms—LSA and LDA—show comparable results (ROUGE-1 recall and precision for LDA just over 0.35, ROUGE-1 recall just under 0.35 and precision over 0.4 for LSA), with LDA providing better recall results, but LSA providing better precision.

Ren and de Rijke [49] present a three-step approach: A hierarchical sentiment Latent Dirichlet Allocation is used to model contrastive topics, which are filtered in a structured determinantal point process to the most diverse topics and used in an iterative optimization algorithm that selects sentences with explicit consideration of contrast, relevance, and diversity to form the contrastive summary. The approach is compared with several other topic models, namely the Topic-Aspect Model [44], the Sentiment-topic Model [35], the Latent Dirichlet Allocation [3], and the hierarchical Latent Dirichlet Allocation in combination with summarization approaches such as LexRank and clustering-based sentence ranking. Like [44], they use the same Gallup® telephone survey [29] and add extracted news articles from the New York Times to build contrastive summaries across topics. The ROUGE metric is used to validate their three-step approach and shows that the contrastive summaries produced, which meet the three main criteria of contrast, diversity, and relevance, demonstrate the effectiveness of the proposed method through significant improvements over the three manually annotated data sets. Their approach yields a ROUGE-1 of 0.4 for the Healthcare Corpus compared to 0.4 for the Topic Aspect Model and 0.31 for the Sentiment-Topic Model.

A semi-supervised Probabilistic Latent Semantic Analysis model is used in combination with a sentence selection strategy that uses a contrastive similarity measure that indicates how well two sentences with opposing opinions match [20]. Taking into account prior information from experts, the topic model groups the arguments. Guo et al. [20] created their own data set on controversial topics and used the precision and coverage measure to evaluate their contrastive summarization approach and compare it to the ones of [30, 43]. Their model (precision of 0.6 and coverage of 0.67) outperforms [30]'s approach (precision of 0.2 and coverage of

0.17) and the one of [43] (precision of 0.2 and coverage of 0.33) in terms of precision and coverage.

The focus of [21] is a differential topic model dTM-SAGE with a sentence scoring method that measures the discriminative power of sentences to summarize the differences between groups of documents. The topic model is used to obtain deviations in group-specific word distributions to indicate how words are used differently in different document groups from a background word distribution. They used 35 papers with 6636 sentences from the ACL Anthology Searchbench for their contrastive summarization approach, which is validated with the ROUGE metric. In their evaluation, they significantly outperform generic baseline summarization approaches (ROUGE-1 0.42) such as the centroid-based method [47] (ROUGE-1 0.23), the graph-based method LexPageRank [47] (ROUGE-1 0.25) and the MMR-based method [9] (ROUGE-1 0.28), as well as two contrastive summarization approaches, [56] (ROUGE-1 0.31) and [51] (ROUGE-1 0.32).

Lavanya and Parvathavarthini [31] use a context-sensitive PLSA model with initial linguistic rules based on dependency relationships to extract context-feature-opinion phrases and then, automatically cluster the extracted context-feature-opinion phrases into contrastive summaries. In their work, the same product review data set as in [30] and the benchmark car data set [18] are used to extract context-feature-opinion phrases and automatically group them into contrasting arguments. The ROUGE metric is used as a validation metric to compare their approach (ROUGE-1 0.33) with [20] (ROUGE-1 0.31), which achieved similar results.

4.3 Statistical approaches

Purely statistical approaches compute statistical and/or linguistic characteristics and their weights for sentences or words, and then, select the most important words or phrases based on these characteristics. In these scoring algorithms, additional attempts are made to explicitly extract orthogonal sentences to represent the most discussed items.

References [11, 48] both use a statistical sentence weight calculation to compute a comparative summary, where [48] use Kullback–Leibler divergence and a bag-of-words model to quickly isolate interesting opinions and provide analyst feedback on how users generally feel about a given topic. [48] apply their summarization approach to the content and commentary of news stories, but do not provide a detailed description of the data set or any indication of validation of the method, nor have they made comparisons with other methods.

Chitra et al. [11], on the other hand, generate comparative summaries from a set of URLs using the HTML DOM tree structure of these web pages and using feature keywords to score sentences. A collection of 200 web documents related to educational institutions, algorithms, banking, and household items is collected to contrastively summarize these web pages. They use a five-star Likert scale to measure the overall responsiveness of their contrastive website summaries and show that their system reduces the time and effort required for the user to browse different websites to compare information.

4.4 Machine learning approaches

The advent of machine learning and neural networks has also reached contrastive text summarization.

For example, [27] use a sparse predictive classification approach that automatically labels text units for a given topic, pre-processes the possible summarizing phrases and phrase counts, and sparsely selects a contrastive phrase list of interest using Lasso and L^1 penalized logistic regression on automatic labels. Each contrastive summarization approach is applied to the set of articles in the New York Times. They compared their four feature selection methods in a crossed and randomized experiment in which non-experts read both the original documents and their summaries and rated the quality and relevance of the results using a five-star Likert scale. Based on their human experiment, they concluded that features, such as Lasso or tf-idf, selected using a sparse prediction framework, can generate informative summaries of keywords for topics of interest.

On the other hand, [1, 2] define extractive contrastive summarization as a binary classification problem using the maximum mean discrepancy in combination with a gradient optimization method. In this process, for given groups, the algorithm learns to select sentences that represent each group, but also to highlight differences between groups. References [1] and [2] use news articles on controversial topics to comparatively summarize news articles from different time periods to determine what changed about the topic between summarization periods. In a human evaluation setting, where crowd workers are given some contrasting summary articles from two groups and asked to classify them into one of two groups, the comparative summarization is then evaluated with this resulting accuracy value. They found that the gradient optimization summaries were 7% more accurate in classification than discrete optimization. References [1] and [2] use not only machine learning approaches but also pre-trained GLOVE vectors to represent documents in their comparative summarization through the binary classification method. In recent years, there has been a trend towards using or fine-tuning pre-trained language models, which we present in the next chapter.

4.5 Deep learning and transformer language models

A breakthrough of neural networks in text summarization was achieved by the invention of word embeddings and transformer language models. Word embeddings such as [45] but also [12] are representations of words and phrases that are mapped to a numerically dense vector that captures their semantic meaning and context. Transformer, which is a neural network architecture introduced in 2017 by [54], has led to significant improvements in tasks such as machine translation, language modeling, and summarization. Many well-known language models such as BERT [12], but also GPT-3 [4] are based on this architecture.

Lavanya and Parvathavarthini [33] train a long short-term memory on pre-trained Word2Vec embeddings with attention mechanisms such as feature attention, opinion attention, and context attention to automatically build context-sensitive contrastive summaries. Context-sensitive sentiment classification using a soft-max classifier is used to identify and present contrasting summaries from a given set of positive and negative summaries of two entities. They used the SemEval 2014 Task 4 restaurant reviews data set [46] to train the context-sensitive contrastive opinion summarization model. The precision measure and the ROUGE metric are used to validate their approach, and the experimental results show that the proposed model achieves better or similar performance (ROUGE-1 0.36) than the baseline models, such as [58] (ROUGE-1 0.36) and [50] (ROUGE-1 0.32).

A Word2Vec model is trained from scratch on a large patent data set to determine the vector representations of [40]'s patent vocabulary, and [10] use a pre-trained BERT model to represent the scientific vocabulary. Both approaches are described in more detail in the graph-based subsection.

None of the above methods, however, uses the language models for abstractive summarization, only for extractive summarization. On the other hand, the COCOSUM [25] comparative summary system is a fine-tuned language model that produces abstractive contrastive summaries. This work clearly distinguishes itself from the others, as it highlights not only what is different but at the same time what is common, and it uses an abstractive approach. Iso et al. [25] created a comparative opinion summarization corpus containing human-written contrastive and common summaries of hotel reviews. The ROUGE metric is used for the validation of the contrastive summary of hotel reviews, and the experimental results on their created benchmark data set show that COCOSUM can produce higher-quality contrastive and common summaries than two extractive and four abstractive opinion summarization models. All variants of their method have a ROUGE-1 of more than 0.4 versus the range of 0.32–0.37 for the baseline methods and a BERT score of more than 0.29 versus a range of 0.2–0.24.

4.6 Graph-based methods

In graph-based models, documents are represented as graphs based on their sentences. After a long period of obscurity [37, 44], they have recently experienced a somewhat renaissance with pre-trained language models. Two works stand out here, [10, 40] both have developed graph-based comparative summarization methods using pre-trained language models as a basis.

Liu et al. [40] use the vector representation of patent vocabulary given by the Word2Vec model as features in a multi-relational graph to generate contrastive summaries. They created a data set of real-word-linked patents to create contrasting patent summaries. Precision, recall and F-measure are used to validate their approach (Results for the mechanical engineering data set: 0.68 precision versus range 0.10–0.23 for baseline approaches, 0.59 recall versus 0.44–1.00, 0.63 F-measure versus 0.16–0.35), with experimental results and detailed analysis in the case study showing the effectiveness of the proposed framework.

Similarly, [10] use a pre-trained BERT model to generate comparative summaries using a comparative graph-based summarization method that uses citations as a guide. The ROUGE metric is used for the evaluation on a specially created corpus of scientific articles with linked citations. Experiments show that their method (ROUGE-1 0.49) outperforms nine single and multi-document summary methods (ROUGE-1 ranges from 0.23 to 0.43) in their own corpus and also performs well in DUC2006 (ROUGE-1 of 0.40 against a range of 0.32 to 0.41) and DUC2007 (ROUGE-1 of 0.41 against a range of 0.33 to 0.43).

In addition to plain text, both approaches [10, 40] use linking data, such as those found in patents or scientific publications, to connect and contrast content. The next section describes the applications used in contrastive summarization approaches.

5 Applications on contrastive summarization

Contrastive summarization is used in many areas where different and controversial contributions to a particular topic or problem are collected and need to be distinguished. This motivated us to investigate which applications are considered and aimed for in contrastive summarization, and by that inspire future applications.

Overall, we found that Contrastive Summarization is used for opinions or sentiments on an issue, reviews, content in blogs or tweets, websites, patents, news articles, or even scientific research articles. References [1, 2, 7, 8, 23, 24, 27, 48, 49] used their approach to analyze newspaper articles on controversial topics. In addition to newspaper articles, [20, 25, 30, 31, 33, 34, 37, 43, 44, 48, 52, 53] used user-generated

data as an application of their method. They tried to contrast and summarize opinions, comments, arguments, and reviews on controversial topics or products. References [20, 56] crawl their user-generated data from blogs and Twitter to contrast and summarize controversy and various arguments.

Contrastive summarization can also make use of more complex data structures, such as those found in patents, web pages, or scientific articles. Chitra et al. [11] use the HTML DOM tree structure to summarize multiple web pages comparatively. Liu et al. [40] utilize the underlying network resulting from the link data structure of patents, and [10] use the citation structure of scientific articles in addition to the link structure of patents. Without using the citation link structure, [21] also uses scientific articles as an application of their approach.

6 Conclusion and outlook for further research

Our survey shows that the problem of contrastive or comparative summarization has been known for a long time and has attracted a great deal of interest from the research community. As in the evolution of language models, where rule-based systems were replaced by statistical ones, followed by a neural revolution [28], contrastive summarization systems also faced a similar evolution: In the early days of this research, contrastive summarization was always considered as an optimization problem, where representativeness, contrastivity, and diversity were the constraints for extractive sentence selection. Due to the statistical revolution of generative statistical models such as LDA, topic generation was very much in vogue at that time. Bayesian methods and expectation maximization were also used to detect contrastive themes in a text. Meanwhile, these have been replaced by machine learning approaches, in particular, deep learning methods, with pre-trained language models increasingly forming the basis of contrastive summarization. On the basis of this, another trend has emerged that involves the use of additional complex data structures linking the documents. Patents, scientific articles, or websites are connected by citations or hyperlinks in addition to raw text, enabling the use of graph-based methods. However, the development of contrastive summarization methods must take into account the genre of the text, as different genres pose different challenges. For example, patents and scientific articles tend to be very technical and formal, while tweets and comments on forums often adopt a colloquial nature. This is something acknowledged by [55] who recognize in their work that different techniques may be applied depending on the type of text.

Several efforts have also been made by researchers for evaluation purposes. However, the lack of standard data sets and the fact that the task of contrastive summarization differs on the basis of its application have led to the emergence of a multitude of data sets, most of which were created by researchers themselves. To further popularize the task and provide grounds for fair method comparison, competitions with standardized data sets are expected to emerge.

Standardized measures such as accuracy, recall, F-measure, and ROUGE, in addition to human-based Likert scales, were also used to evaluate contrastive summarization systems. These, however, are more appropriate for extractive approaches. In the future, human-based metrics or semantic distances using pre-trained language models [25] should be used when evaluating systems that stem from abstractive approaches as means to correctly evaluate sentences that have the same semantic meaning but are described by synonyms. There is a need for evaluation metrics that correctly evaluate sentences that have the same semantic meaning, but are described by synonyms.

Finally, there is a body of research work that incorporates temporal aspects in contrastive summarization, which was not covered in this survey. These are either research works that compare content published at different time periods [13, 14, 26] or that conduct comparative timeline summarization[3] [15]. Temporal comparative summarization is challenging and poses unique challenges as it needs to consider other issues such as event ordering and coreference, general/background context change of different times, vocabulary semantic drift, fragmentary or missing documents especially when considering distant time periods, or even different OCR result quality. This is something researchers should look at in the future.

Parallel to this, we envision that in the near future, modern pre-trained language models, which have shown strong effectiveness in abstractive summarization benchmarks [36], will be extended to contrastive summarization, which has always been considered, with the exception of [25], as an extractive approach.

Author Contributions The authors certify that they have read the journal policies and submit this manuscript in accordance with those policies, and that authorship is justified as follows: Thomas Ströhle, Ricardo Campos and Adam Jatowt conceived and designed the survey. Thomas Ströhle collected all scientific publications and analyzed all methods, data sets and metrics used. Thomas Ströhle and Ricardo Campos wrote all parts of the survey. Adam Jatowt wrote parts of the introduction and the conclusion of the survey and provided feedback on the remaining parts.

Funding Open access funding provided by University of Innsbruck and Medical University of Innsbruck. Ricardo Campos was financed by National Funds through the FCT - Fundação para a Ciência e a

[3] We note here that, actually, even a standard, hence, non-contrastive (unlike [15]) timeline summarization problem could also be to some degree regarded as a variant of comparative summarization task since summary content selection at different time points may be affected by summaries generated at other time points.

Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC.

Data Availability The authors confirm that the data supporting the findings of this study are available in the article.

Declarations

Conflict of interest The authors have no conflicts of interest as defined by Springer, or other interests that could be perceived as influencing the results and/or discussion reported in this paper.

References

1. Bista, U.: Comparative summarisation of rich media collections. In: Culpepper, J.S., Moffat, A., Bennett, P.N., et al. (eds.) Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11–15, 2019, pp. 812–813. ACM (2019). https://doi.org/10.1145/3289600.3291603
2. Bista, U., Mathews, A. P., Shin, M., et al.: Comparative document summarisation via classification. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27–February 1, 2019, pp. 20–28. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.330120
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). https://doi.org/10.5555/944919.944937
4. Brown, T.B., Mann, B., Ryder, N., et al.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20. Curran Associates Inc., Red Hook (2020). https://doi.org/10.5555/3495724.3495883
5. Campos, R., Pasquali, A., Jatowt, A., et al.: Automatic generation of timelines for past-web events. In: The Past Web, pp. 225–242. Springer, Berlin (2021). https://doi.org/10.1007/978-3-030-63291-5_18
6. Campos, R., Jatowt, A., Jorge, A.: Text mining and visualization of political party programmes using keyword extraction methods: the case of Portuguese legislative elections. In: Lecture Notes in Computer Science. Proceedings of the iConference'23, Barcelona, Spain. March 27–30 (2023). https://doi.org/10.1007/978-3-031-28035-1_24
7. Campr, M., Jezek, K.: Comparative summarization via latent dirichlet allocation. In: Snásel, V., Richta, K., Pokorný, J. (eds.) Proceedings of the Dateso 2013 Annual International Workshop on DAtabases, TExts, Specifications and Objects, Pisek, Czech Republic, April 17, 2013, CEUR Workshop Proceedings, vol. 971, pp. 80–86. CEUR-WS.org (2013a). http://ceur-ws.org/Vol-971/poster11.pdf
8. Campr, M., Jezek, K.: Topic models for comparative summarization. In: Habernal, I., Matousek, V. (eds.) Text, Speech, and Dialogue—16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1–5, 2013. Proceedings, Lecture Notes in Computer Science, vol. 8082, pp. 568–574. Springer, Berlin (2013b). https://doi.org/10.1007/978-3-642-40585-3_71
9. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335–336 (1998). https://doi.org/10.1145/290941.291025
10. Chen, J., Cai, C., Jiang, X., et al.: Comparative graph-based summarization of scientific papers guided by comparative citations. In: Calzolari, N., Huang, C., Kim, H., et al. (eds.) Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12–17, 2022. International Committee on Computational Linguistics, pp. 5978–5988 (2022). https://aclanthology.org/2022.coling-1.522
11. Chitra, P., Baskaran, R., Sarukesi, K.: Query sensitive comparative summarization of search results using concept based segmentation. CoRR. arXiv:1201.2304 (2012)
12. Devlin, J., Chang, M. W., Lee, K., et al.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019). https://doi.org/10.18653/v1/N19-1423
13. Duan, Y., Jatowt, A.: Across-time comparative summarization of news articles. In: Culpepper, J.S., Moffat, A., Bennett, P.N., et al. (eds.) Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11–15, 2019, pp. 735–743. ACM (2019). https://doi.org/10.1145/3289600.3291008
14. Duan, Y., Jatowt, A., Tanaka, K.: Discovering latent threads in entity histories. Data Sci. Eng. **4**(4), 336–351 (2019). https://doi.org/10.1007/s41019-019-00108-x
15. Duan, Y., Jatowt, A., Yoshikawa, M.: Comparative timeline summarization via dynamic affinity-preserving random walk. In: Giacomo, G.D., Catalá, A., Dilkina, B., et al. (eds.) CAI 2020—24th European Conference on Artificial Intelligence, 29 August–8 September 2020, Santiago de Compostela, Spain, August 29–September 8, 2020—Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 1778–1785. IOS Press (2020). https://doi.org/10.3233/FAIA200292
16. El-Kassas, W.S., Salama, C.R., Rafea, A.A., et al.: Automatic text summarization: a comprehensive survey. Expert Syst. Appl. **165**, 113679 (2021). https://doi.org/10.1016/j.eswa.2020.113679
17. Ermakova, L., Cossu, J.V., Mothe, J.: A survey on evaluation of summarization methods. Inf. Process. Manag. **56**(5), 1794–1814 (2019). https://doi.org/10.1016/j.ipm.2019.04.001
18. Ganesan, K., Zhai, C.: Opinion-based entity ranking. Inf. Retr. **15**(2), 116–150 (2012). https://doi.org/10.1007/s10791-011-9174-8
19. Garg, A., Popli, R., Sarao, B.: Growth of digitization and its impact on big data analytics. In: IOP Conference Series: Materials Science and Engineering, p. 012083. IOP Publishing (2021). https://doi.org/10.1088/1757-899X/1022/1/012083
20. Guo, J., Lu, Y., Mori, T., et al.: Expert-guided contrastive opinion summarization for controversial issues. In: Gangemi, A., Leonardi, S., Panconesi, A. (eds.) Proceedings of the 24th International Con-

ference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18–22, 2015—Companion Volume, pp. 1105–1110. ACM (2015). https://doi.org/10.1145/2740908.2743038
21. He, L., Li, W., Zhuge, H.: Exploring differential topic models for comparative summarization of scientific papers. In: Calzolari, N., Matsumoto, Y., Prasad, R. (eds.) COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11–16, 2016, Osaka, Japan, pp. 1028–1038. ACL (2016). https://aclanthology.org/C16-1098/
22. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Kim, W., Kohavi, R., Gehrke, J., et al. (eds.) Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22–25, 2004, pp. 168–177. ACM (2004). https://doi.org/10.1145/1014052.1014073
23. Huang, X., Wan, X., Xiao, J.: Comparative news summarization using linear programming. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19–24 June, 2011, Portland, Oregon, USA—Short Papers, pp. 648–653. The Association for Computer Linguistics (2011). https://aclanthology.org/P11-2114/
24. Huang, X., Wan, X., Xiao, J.: Comparative news summarization using concept-based optimization. Knowl. Inf. Syst. **38**(3), 691–716 (2014). https://doi.org/10.1007/s10115-012-0604-8
25. Iso, H., Wang, X., Angelidis, S., et al.: Comparative opinion summarization via collaborative decoding. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22–27, 2022, pp. 3307–3324. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-acl.261
26. Jatowt, A., Bron, M.: HistoryComparator: Interactive across-time comparison in document archives. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. The COLING 2016 Organizing Committee, Osaka, Japan, pp. 84–88 (2016). https://aclanthology.org/C16-2018
27. Jia, J., Miratrix, L., Yu, B., et al.: Concise comparative summaries (CCS) of large text corpora with a human experiment. Ann. Appl. Stat. **8**(1), 499–529 (2014). https://doi.org/10.1214/13-AOAS698
28. Johri, P., Khatri, S.K., Al-Taani, A.T., et al.: Natural language processing: history, evolution, application, and future work. In: Abraham, A., Castillo, O., Virmani, D. (eds.) Proceedings of 3rd International Conference on Computing Informatics and Networks, pp. 365–375. Springer Singapore, Singapore (2021). https://www.springerprofessional.de/en/proceedings-of-3rd-international-conference-on-computing-informa/18963732
29. Jones, J.M.: In U.S., 45% favor, 48% oppose Obama healthcare plan. Gallup Poll, March 9. https://news.gallup.com/poll/126521/favor-oppose-obama-healthcare-plan.aspx (2010)
30. Kim, H.D., Zhai, C.: Generating comparative summaries of contradictory opinions in text. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM'09. Association for Computing Machinery, New York, pp. 385–394 (2009). https://doi.org/10.1145/1645953.1646004
31. Lavanya, S., Parvathavarthini, B.: Context-sensitive contrastive feature-based opinion summarisation of online reviews. Int. J. Enterp. Netw. Manag. **11**(2), 144–163 (2020). https://doi.org/10.1504/IJENM.2020.106309
32. Lavanya, S.K., Parvathavarthini, B.: A survey on contrastive opinion summarisation. Int. J. Reason. Based Intell. Syst. **11**(2), 141–150 (2019). https://doi.org/10.1504/IJRIS.2019.10021326
33. Lavanya, S.K., Parvathavarthini, B.: Context aware contrastive opinion summarization. In: Chandrabose, A., Furbach, U., Ghosh, A., et al. (eds.) Computational Intelligence in Data Science—Third IFIP TC 12 International Conference, ICCIDS 2020, Chennai, India, February 20–22, 2020, Revised Selected Papers, IFIP Advances in Information and Communication Technology, vol. 578, pp. 16–29. Springer, Berlin (2020b). https://doi.org/10.1007/978-3-030-63467-4_2
34. Lerman, K., McDonald, R.T.: Contrastive summarization: An experiment with consumer reviews. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31–June 5, 2009, Boulder, Colorado, USA, Short Papers, pp. 113–116. The Association for Computational Linguistics (2009). https://aclanthology.org/N09-2029/
35. Li, F., Han, C., Huang, M., et al.: Structure-aware review mining and summarization. In: Huang, C., Jurafsky, D. (eds.) COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23–27 August 2010, Beijing, China, pp. 653–661. Tsinghua University Press (2010). https://aclanthology.org/C10-1074/
36. Li, H., Einolghozati, A., Iyer, S., et al.: EASE: extractive-abstractive summarization with explanations. CoRR arXiv:2105.06982 (2021)
37. Liang, X., Qu, Y., Ma, G.: Research on contrastive viewpoint summarization for opinionated texts. J. Interconnect. Netw. **14**(3) (2013). https://doi.org/10.1142/S0219265913600037
38. Lin, C., Hovy, E.H.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Hearst, M.A., Ostendorf, M. (eds.) Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27–June 1, 2003. The Association for Computational Linguistics (2003). https://aclanthology.org/N03-1020/
39. Lin, W.H., Wilson, T., Wiebe, J., et al.: Which side are you on? Identifying perspectives at the document and sentence levels. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), pp. 109–116. Association for Computational Linguistics, New York City (2006). https://aclanthology.org/W06-2915
40. Liu, Z., Zhang, J., Qin, T., et al.: One-to-many comparative summarization for patents. Scientometrics **127**(4), 1969–1993 (2022). https://doi.org/10.1007/s11192-022-04307-8
41. Lymperaiou, M., Manoliadis, G., Menis-Mastromichalakis, O., et al.: Towards explainable evaluation of language models on the semantic similarity of visual concepts. In: Calzolari, N., Huang, C., Kim, H., et al. (eds.) Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12–17, 2022. International Committee on Computational Linguistics, pp. 3639–3658 (2022). https://aclanthology.org/2022.coling-1.321
42. Moratanch, N., Chitrakala, S.: A survey on extractive text summarization. In: 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1–6. IEEE (2017). https://doi.org/10.1109/ICCCSP.2017.7944061
43. Özsoy, M.G., Çakici, R.: Contrastive max–sum opinion summarization. In: Jaafar, A., Ali, N.M., Noah, S.A.M., et al. (eds.) Information Retrieval Technology—10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3–5, 2014. Proceedings, Lecture Notes in Computer Science, vol. 8870, pp. 256–267. Springer, Berlin (2014). https://doi.org/10.1007/978-3-319-12844-3_22
44. Paul, M.J., Zhai, C., Girju, R.: Summarizing contrastive viewpoints in opinionated text. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9–11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 66–76. ACL (2010). https://aclanthology.org/D10-1007/
45. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W.

(eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1532–1543. ACL (2014). https://doi.org/10.3115/v1/d14-1162

46. Pontiki, M., Galanis, D., Pavlopoulos, J., et al.: SemEval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35. Association for Computational Linguistics, Dublin, Ireland (2014). https://doi.org/10.3115/v1/S14-2004
47. Radev, D.R., Jing, H., Sty, M., et al.: Centroid-based summarization of multiple documents. Inf. Process. Manag. **40**(6), 919–938 (2004). https://doi.org/10.1016/j.ipm.2003.10.006
48. Raveendran, G., Clarke, C.L.A.: Lightweight contrastive summarization for news comment mining. In: Hersh, W.R., Callan, J., Maarek, Y., et al (eds.) The 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'12, Portland, OR, USA, August 12–16, 2012, pp. 1103–1104. ACM (2012). https://doi.org/10.1145/2348283.2348490
49. Ren, Z., de Rijke, M.: Summarizing contrastive themes via hierarchical non-parametric processes. In: Baeza-Yates, R., Lalmas, M., Moffat, A., et al. (eds.) Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9–13, 2015, pp. 93–102. ACM (2015). https://doi.org/10.1145/2766462.2767713
50. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. In: Barzilay, R., Kan, M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30–August 4, Volume 1: Long Papers, pp. 1073–1083. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1099
51. Shen, C., Li, T.: Multi-document summarization via the minimum dominating set. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 984–992 (2010). https://doi.org/10.5555/1873781.1873892
52. Rocha da Silva, R., Salgueiro Pardo, T.A.: Building contrastive summaries of subjective text via opinion ranking. Revista de Informática Teórica e Aplicada **29**(2), 11–34 (2022). https://doi.org/10.22456/2175-2745.118372
53. Sipos, R., Joachims, T.: Generating comparative summaries from reviews. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, CIKM'13, pp. 1853–1856. Association for Computing Machinery, New York (2013). https://doi.org/10.1145/2505515.2507879
54. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. (2017). https://doi.org/10.5555/3295222.3295349
55. Vodolazova, T., Lloret, E., Muñoz, R., et al.: Extractive text summarization: can we use the same techniques for any text? In: Métais, E., Meziane, F., Saraee, M., et al. (eds.) Natural Language Processing and Information Systems—18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19–21, 2013. Proceedings, Lecture Notes in Computer Science, vol. 7934, pp. 164–175. Springer, Berlin (2013). https://doi.org/10.1007/978-3-642-38824-8_14
56. Wang, D., Zhu, S., Li, T., et al.: Comparative document summarization via discriminative sentence selection. ACM Trans. Knowl. Discov. Data **6**(3), 12:1-12:18 (2012). https://doi.org/10.1145/2362383.2362386
57. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data: a rating regression approach. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 783–792 (2010). https://doi.org/10.1145/1835804.1835903
58. Yang, M., Qu, Q., Shen, Y., et al.: Aspect and sentiment aware abstractive review summarization. In: Bender, E.M., Derczynski, L., Isabelle, P. (eds.) Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018, pp. 1110–1120. Association for Computational Linguistics (2018). https://aclanthology.org/C18-1095/