Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 29, Num. 02 (2022) 11-34

RESEARCH ARTICLE

Building contrastive summaries of subjective text via opinion ranking

Construindo sumários constrastivos para textos subjetivos via ranqueametno de opinião

Raphael Rocha da Silva[1]* and Thiago Alexandre Salgueiro Pardo[1]

Abstract: This article investigates methods to automatically compare entities from opinionated text to help users to obtain important information from a large amount of data, a task known as "contrastive opinion summarization". The task aims at generating contrastive summaries that highlight differences between entities given opinionated text (written about each entity individually) where opinions have been previously identified. These summaries are made by selecting sentences from the input data. The core of the problem is to find out how to choose these more relevant sentences in an appropriate manner. The proposed method uses a heuristic that makes decisions according to the opinions found in the input text and to traits that a summary is expected to present. The evaluation is made by measuring three characteristics that contrastive summaries are expected to have: representativity (presence of opinions that are frequent in the input), contrastivity (presence of opinions that highlight differences between entities) and diversity (presence of different opinions to avoid redundancy). The novel method is compared to methods previously published and performs significantly better than them according to the measures used. The main contributions of this work are: a comparative analysis of methods of contrastive opinion summarization, the proposal of a systematic way to evaluate summaries, the development of a new method that performs better than others previously known and the creation of a dataset for the task.

Keywords: Summarization — Opinion mining — Evaluation

Resumo: Este artigo investiga métodos para comparar automaticamente entidades de textos opinativos para auxiliar usuários a obter informações importantes de uma grande quantidade de dados, uma tarefa conhecida como "sumarização contrastiva de opinião". A tarefa consiste em gerar resumos contrastivos que destacam as diferenças entre entidades a partir de textos opinativos (escritos sobre cada entidade individualmente) em que as opiniões foram previamente identificadas. Esses resumos são feitos selecionando-se sentenças dos dados de entrada, sendo que o cerne do problema é descobrir como escolher essas sentenças mais relevantes de maneira adequada. O método proposto usa uma heurística que toma decisões de acordo com as opiniões encontradas no texto de entrada e com as características que um resumo deve apresentar. A avaliação é feita medindo três características que se espera que os resumos contrastivos tenham: representatividade (ou seja, presença de opiniões frequentes na entrada), contrastividade (presença de opiniões que evidenciam diferenças entre entidades) e diversidade (presença de opiniões diferentes para evitar redundâncias). O novo método é comparado a métodos publicados anteriormente e tem um desempenho significativamente melhor do que eles de acordo com as medidas utilizadas. As principais contribuições deste trabalho são: uma análise comparativa de métodos de sumarização de opinião contrastiva, a proposta de uma forma sistemática de avaliação de resumos, o desenvolvimento de um novo método com desempenho melhor do que outros anteriormente conhecidos e a criação de um conjunto de dados para a tarefa.

Palavras-Chave: Sumarização — Mineração de opinião — Avaliação

[1] Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematical and Computer Sciences, University of São Paulo (USP), São Carlos - São Paulo, Brazil
*Corresponding author: raphsilva@alumni.usp.br

DOI: http://dx.doi.org/10.22456/2175-2745.118372 • **Received:** 14/09/2021 • **Accepted:** 13/04/2022

1. Introduction

People have access to large amounts of data of various kinds through the World Wide Web. It is therefore important to think of new ways to help people to explore the data and elutriate its contents to get the most relevant pieces of information. Ideally, users who need prompt information would get what they need by simply typing in some keywords or making questions in natural language (either written or spoken) instead of having to visit many different pages and mining relevant pieces themselves until they are able to put them together to answer the question they are trying to solve.

Many tasks can illustrate the use of automatic tools that help people to get quick information. Web search is one of them. Erst limited to finding documents containing keywords specified in a query, now they are able to answer questions in a very natural way. The questions can be worded as if they were directed to a human: '*what's the weather gonna be like this weekend?*'; '*how many dollars is eight hundred reais?*'. In some modern search services, instead of only showing a list of webpages that may help users to find what they need, the output of these questions are short sentences that actually answer them. This is done with techniques that allow to isolate, within the lots of data to which the tool has access, the piece of information that is really demanded in the given situation.

Finding relevant content in a set of data and presenting them in a friendly fashion for human consumption is a task called '**summarization**' [1]. This task has been investigated for many domains: news articles, web pages, biomedical documents and scientific articles, among others [2]. Many techniques have been explored: statistical methods, machine learning, lexical chains, clustering, graph-based methods, and swarm intelligence-based methods [2]. Text summarization was studied as early as in the 1950s and gained prominence in the 1990s [3]. Studies on the task are still simmering to this date, with several new ideas being published every year about new subtasks, new approaches and improvements of methods previously investigated.

More recently, efforts have been made towards the processing of subjective text. This type of text is demanded in several scenarios. For example, one may be interested in finding what people think of a certain government. This person can do queries on social media services and get thousands of excerpts containing opinions about the government, and then use tools to automatically analyze the data in order to understand what the most frequent opinions are, what topics are controversial, etc. This task is also a type of summarization. Since it is performed on opinionated text, it is called '**opinion summarization**' [4].

While it makes sense to process a subjective text written by a single author, the general goal of the methods discussed in this article is to understand global opinions about a subject. This requires collection and analysis of texts written by a sufficiently large number of people so that different points of views can be compared. Unlike facts, opinions are based on each person's experiences, habits, expectations, abilities and other personal traits. That is why it is important to hear many different people when it comes to subjective ideas.

Processing subjective text is demanded because this kind of text brings information that is not found in other types. [5] discusses a real example where a newly released model of phone was reported to bend after some days of use. Shortly after the phone was launched, many users published complaints on online forums and blogs. In this situation, three types of people may have the need to explore this information. First, a potential consumer who wants to buy the product and wants to figure if the problem is really serious. Second, a news writer who wants to make an article about what people have been saying about the fact. Third, an analyst from the phone manufacturer who needs to understand how the problem is affecting users and what to do to minimize it. In all three cases, as [5] highlights, due to the large amount of comments posted by users, it may be extremely hard and longstanding to find and explore the useful information in them. That is where automatic tools are demanded.

Opinion summarization can inherit some characteristics and techniques from regular summarization; indeed, it is even possible to try to summarize an opinionated text by using a summarizer designed for other domains. But processing subjective text has its singularities. The most prominent one is perhaps the fact that the ideas expressed should not come from a single source when one is trying to understand a certain entity from opinions about it. If the text being processed is objective (assuming it contains only real, trustworthy facts), one document is sufficient. For example, if someone wants to know what the shape of the Earth is, it is enough to find one respectful scientific work that provides the answer[1]. On the other hand, if someone wants to know whether or not the president is a good guy, it is certainly not enough to look up in only one source, no matter how respectful it is, because political views tend to be very subjective and biased according to each person's ideas and beliefs.

Subjective text processing usually needs to be quantitative, which means that it has to take into account which opinions are more frequent. When processing opinionated texts written by different people on a same subject, opinions with various points of argumentation will likely be found, some agreeing with one another, some disagreeing. Therefore, there may be the need to group opinions together based on how similar they are and then to identify which groups are larger, which indicates that the kind of opinion they contain occurs more frequently than others in the collected texts, meaning they can be considered more relevant (in most cases, at least).

Before processing opinionated text, the opinions in it need to be identified. In a simplified way, identifying an opinion means to figure out, in an opinionated sentence, what the target of the opinion is and whether the opinion is positive

[1] In practice, tools of information extraction are aware that mistakes can occur everywhere, so they use multiple sources whenever possible so that redundancy can compensate for the noise [6].

or negative. For example, in the sentence '*this is the best article I have ever read*', the target is 'article' and the opinion is positive. In Sentiment Analysis, the target of an opinion is called an 'aspect', and the value that can be positive or negative (sometimes neutral) is called 'polarity'. This kind of approach is called 'aspect-based sentiment analysis' [4].

When dealing with text written by regular Web users, it is usually necessary that the data is preprocessed to attenuate undesirable features like language errors. After that, the text may be preprocessed as a regular text would: it might need to be stemmed, tokenized and have its stopwords removed. After that, opinions can be identified. All this process is omitted in this article, which will deal only with the summarization methods given that opinions are already identified.

There are many methods for contrastive summarization available, but it is not known how they compare with each other since each work uses its own dataset and evaluation metrics. This work fills this gap by proposing a systematic way to evaluate the methods.

The general goal of this work is to use Sentiment Analysis and Summarization techniques to compare entities based on opinionated texts written about them (the next subsection defines the problem). Entities can be products, people, services, etc. The experiments made in this work were specifically performed on texts about consumer electronic products. The main contributions of this work are: it proposes a method for evaluating contrastive summaries, including metrics definition and creation of dataset; it compares previously published methods that have not yet been tested under standardized evaluation; it proposes a new method for contrastive summarization that is better than existing methods.

1.1 Problem definition

The problem investigated in this article is called 'contrastive opinion summarization' or simply '**contrastive summarization**'. Given a set of opinionated sentences about two entities (the **input set** or **source**), the goal is to generate a **contrastive summary**, which is a selection of sentences from the input set built in a way that aims at the comparison of the two entities in a fair manner by highlighting their differences according to the opinions found in the input set. Each input sentence is about only one entity (i.e., direct comparison between entities within sentences is not expected to occur).

The problem is defined here in the same way it appears in the work of [7], which is the only work found that deals with this exact problem. The name of the problem is also borrowed from that work. Other studies have been found that define similar problems:

- [8] develops a system that compares opinions about two or more entities in a quantitative way; it does not extract sentences from the input set to make the summary (like the present work), but counts the common opinions in the input set and displays them as a bar chart. They refer to the task as 'analyzing and comparing opinions'.

- [9] makes summaries that present not only the differences but the similarities between entities. The task is referred to as 'identifying comparative customer requirements'.

- [10] studies the task of finding divergent opinions about a single entity. Their aim is not at comparing entities, but at obtaining contrary points of view about the same topic. They call the task 'contrastive opinion summarization', the same name used in [7], even though it is a slightly different task.

1.2 Article's outline

This articles is organized as follows. **Section 2** brings the theoretical concepts that are important to understand this research effort. **Section 3** summarizes the main related work. **Section 4** presents an original method for contrastive summarization, while **Section 5** reports the experiments performed. Finally, **Section 6** analyses and discusses the achieved results.

2. Theoretical concepts

2.1 Sentiment Analysis

Sentiment Analysis is defined as the field of study that analyzes people's opinions towards entities and their attributes [4]. An **entity** is a 'product, service, topic, issue, person, organization, or event'. It can be defined as a pair $e : (T,W)$ where T is a hierarchy of parts and subparts of an entity e (with multiple levels, if necessary) and W is a set of attributes of e. A **part** is an element that composes an entity (or another part), e.g., 'keyboard' is a part of 'computer', and 'space bar' is a part of 'keyboard'. An **attribute** is a characteristic of an entity or one of its parts: 'weight' and 'design' are some attributes of 'computer'. In order to simplify the representation of entity, the hierarchy is reduced: 'parts of parts' are viewed simply as 'parts' and both parts and attributes are then called '**aspects**' [4]. Thus, 'keyboard', 'space bar', 'weight' and 'designs' are all aspects of the entity 'computer'. Even the entity as a whole can be considered an aspect.

Formally, an opinion is a quintuple $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, where e_i is an entity, a_{ij} is an aspect of e_i, s_{ijkl} is a sentiment about aspect a_{ij}, h_k is the opinion holder and t_l is the time when the opinion was expressed [4]. The **sentiment** is some information that tells what h_k thinks of a_{ij}.

In the present work, the definition of opinion will be simplified. There will be no need to know who the opinion holder is or at what time the opinion was given. The entity will be implicit. The sentiment will be represented by a **polarity**, which indicates whether the opinion is positive, negative or neutral. With those considerations, an **opinion** is a pair (a, p) where a is an aspect of a certain entity and p is a polarity attributed to a by someone. It is very important to remember, throughout the reading of this article, that the word **opinion** refers to the pair (a, p) that is an aspect (of an entity) and the polarity of that aspect; it does not refer to reviews or sentences.

The definition of opinion intends to extract from sentences only the information that is really needed for the task that will be performed. For example, if the sentence is '*The screen of this phone is too bright*', it is said that this sentence contains the opinion $(\text{screen}, -)$. The specific information about the cause for the negative polarity will not be considered by the summarization methods[2].

Even though the definition of opinion implies a simplification of the original information, this has the advantage of discarding unnecessary details. For example, even though '*NLP is always fun*' and '*I love NLP*' are very different sentences, they express basically the same feeling, so any of them would add about the same value to a summary if inserted into it, regardless of which one. These sentences are both seen as $(\text{NLP}, +)$ by the summarization algorithms of this work.

A pair formed by two opinions that disagree with each other will be called a **contrastive pair** in this work. In order to make a contrastive pair, two opinions must have the same aspect and opposite polarities. For example, from the sentences '*the screen is too bright*' and '*the image is perfect*', one can take the contrastive pair $((\text{screen}, -), (\text{screen}, +))$.

2.2 Opinion Summarization

Opinion summarization is the task of selecting and displaying the most relevant information of a set of opinionated texts [4]. The general goal is to help users to have an overview of a set of opinative texts so they can absorb the most common opinions without having to read the full dataset.

A frequent concern in opinion summarization works is that a summary has to be representative and diversified [11, 12, 9]. This can be assessed with measures called representativity and diversity. Different heuristics can be employed to estimate them.

Representativity measures how well the summary reflects in a fair way the information contained in the source. If in a set of opinions collected from several people half of the sentences speak positively of a product and half speak negatively, it is unfair that the summary contains only positive sentences; it is said that this summary is not representative. At the entity level, representativity also considers the topics mentioned: if 90% of people talk about the battery of a camera and only 5% of people talk about the lens, it is of greater value to include in the summary a sentence on the battery than one on the lens.

Diversity is a measure to indicate the amount of information contained in the summary. It is desirable that the summary contains as much information as possible in its limited space. Therefore, a summary that talks only about a single topic can be considered poor.

Diversity and representativity can be either allies or competitors of each other, that is, it is possible that one is automatically benefited when the other is maximized, or that one is impaired when the other is maximized. This depends mainly on the characteristics of the input set, because the input set, in turn, can be diverse or not. A source set that mentions each aspect approximately the same amount of times, if summarized in order to obtain a representative summary, immediately gives a diversified summary, because to be representative means to reflect in a fair way the content of the source, and since the source is diverse, the summary also is. However, if the source is not diverse (as in the case where 90% of people talk about a camera battery and 5% of people talk about the lens), a representative summary of it will not be either. In that case, a more diversified summary would include opinions about both battery and lens: for diversity, an opinion about the lens is as valuable as one about the battery. But this goes against representativity, where the most frequent opinions at source have greater value. So a trade-off between representativity and diversity has to be made in order to balance the characteristics of the summary.

2.3 Contrastive Opinion Summarization

The main task described in this paper is the contrastive opinion summarization, whose objective is to find and summarize the main information that allows to compare two or more entities given sets of opinion texts about each one of them [4, 10, 13, 9].

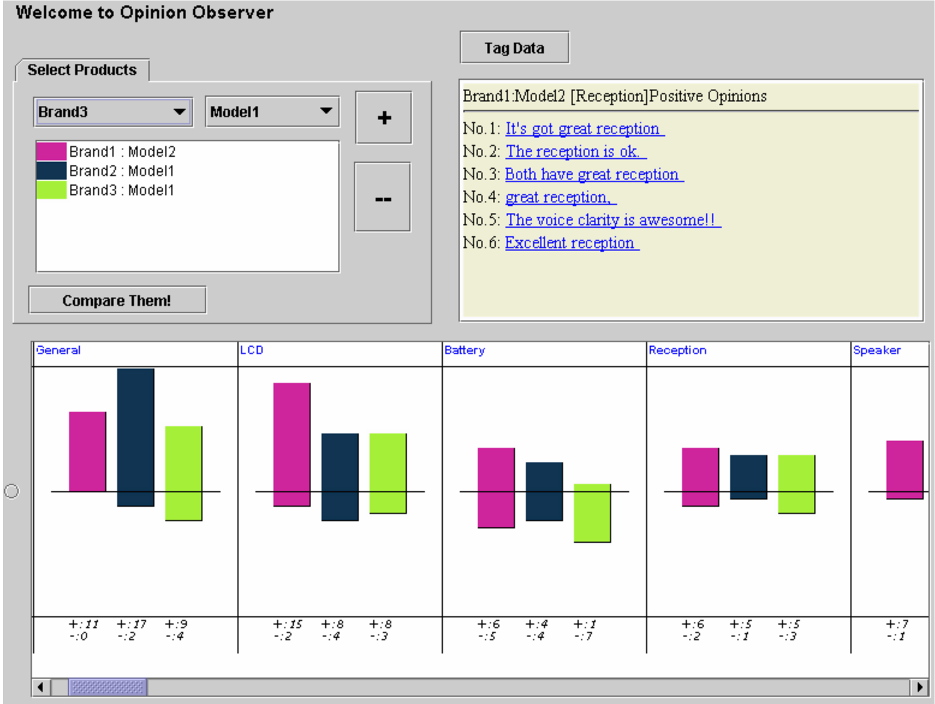
The earliest known work of contrastive analysis of opinion is the one of [8], which processes opinions about competing products. The main goal was to make a system that allows users to easily understand the strengths and weaknesses of each product. The proposed summary format is a bar chart that gives information about aspects of each product, shown in Figure 1. Since that work only separates positive and negative opinions from each aspect and counts them, it does not need to summarize the input set by selecting important parts, because opinions are only counted, and it is not necessary to choose the most relevant ones.

A work that uses summarization techniques in the contrastive analysis of opinion is [13], whose main objective is to extract information from entity pairs to generate a summary that highlights the differences between them. The most direct application suggested by the authors is the domain of consumer evaluations, where a potential buyer can use the summary to see differences between product reviews without having to read all the comments of the products.

In the previous section, the concepts of representativity and diversity were defined as desirable characteristics in an opinion summary. When it comes to *contrastive* opinion summaries, in addition to these two, another characteristic may be desirable: **comparability**, which measures how much the sentences contained in the summary talk about similar subjects of the entities being compared [9]. For example, if one is comparing two cameras, having a summary that talks only about the screen of one of them and about the battery of the other will not help comparing the products explicitly. Ideally, for each sentence on the battery of one of the cameras,

[2] Three detail levels are defined in [8]: in the first level, only the polarity of opinions are considered; in the second level, polarities and aspects define an opinion (like in the present work); the third level considers more refined details, like the actual adjectives used and specific issues expressed in the sentences.

Figure 1. Main screen of the system implemented by [8].



Brand1:Model2 [Reception]Positive Opinions

No.1: It's got great reception
No.2: The reception is ok.
No.3: Both have great reception
No.4: great reception,
No.5: The voice clarity is awesome!!
No.6: Excellent reception

there should be a sentence on the battery of the other.

A measure similar to the comparability is **contrastivity**, which, in addition to requiring sentences to deal with similar topics, requires them to have opposing polarities, that is, to present disagreeing opinions [10], forming contrastive pairs. In this paper, this measure will be more useful than comparability, since the focus here is to obtain divergent information about entities.

In this work, unlike any of the previous works mentioned, the contrastivity will be defined in a way that it rewards the presence of an opinion that is useful to compare entities even if that opinion can not form a contrastive pair with some other opinion in the summary about the other entity. This is useful in the following situation: suppose the entities being compared are e_1 and e_2; if many opinions of e_1 evaluate positively an aspect A but there is not any mention of A in the input set of e_2 (so no contrastive pair could be formed), having an opinion $(A,+)$ about e_1 in the summary helps understanding that that aspect is much more prominent in that entity. In this case, the evaluation metrics proposed here would give $(A,+)$ half the score it would give to a full contrastive pair $((A,+),(A,-))$ where one opinion is about each entity.

Some works [13, 9] describe the contrastive opinion summarization as a combinatorial optimization problem: given two sets E_1 and E_2, each containing opinative sentences about a different entity, a function L is defined that estimates how good a summary $R \subseteq (E_1 \cup E_2)$ is. After choosing a size limit t, the problem is to solve $\arg\max L(R)$ with $|R| \leq t$. This is an NP-hard problem, so the most effective practice is to use a greedy algorithm to find suboptimal solutions [14]. A large part of the study in those works is devoted to setting an appropriate L.

2.4 Summary format

Properly choosing a set of information to be part of a summary is just one of the concerns of contrastive summarization. Another concern is the presentation of the summary: once one has chosen the information that will compose the output, how can the system present it to the user? The presentation can be textual or graphic. There is a work that uses bar graphs to quantitatively display the opinions found [8]. When it comes to works that opt for a textual format, it is common to use a structured model, such as a table [10, 9]; an exception is [13], which does not describe a specific format for the output, implying that the sentences of each entity are simply concatenated as plain text.

The summary generated in [9] is aligned: the output is viewed as a pair of summaries, one for each entity; both summaries are the same size, and the n-th sentence of one is paired with the n-th sentence of the other. Paired sentences speak about the same topic of the two entities, and may indicate a difference or similarity between them (they may have the same polarity or opposite polarities). This summary is shown in Table 1. In [10], something similar is done, but this time it is required that pairs have opposite polarities, and all sentences are about a single entity. An example of summary made in that work is in Table 2.

[15] presents a study that explores different presentation formats. To make the summaries, they used argumentative texts on the debate 'Is global climate change man-made?'. One of the formats presented in the work is illustrated in Figure 2. It was the best evaluated format according to the

Table 1. Example of summary made by [9].

	#	Pair of sentences
Positive vs. positive	1	Battery life is more than 24 h with moderate use. Battery life is excellent as well.
	2	Battery life is great and camera takes good pictures. Battery life is good, as expected with a GSM phone.
Negative vs. negative	1	But not much better battery life than my old phone. The battery life was relavively poor.
	2	When I first bought the old Lumia 521 it had about the same battery life. Sometimes I would even have to remove the battery and put it back in before I could get the phone to turn back on.
Negative vs. positive	1	The battery life is not better than my old phone. Battery life is excellent as well.
	2	Battery life, call quality, apps, all of the things are not better. Battery life is good, as expected with a GSM phone.
Positive vs. negative	1	Overall, a phone with good battery life. On occasion the phone will not turn on and you have to take the battery out and put it back in to get it to respond, battery life is horrible as well.
	2	Does have very good battery life. The battery had not a longer life than I expected.

Table 2. Example of summary made by [10].

No	Positive	Negative
1	oh ... and file transfers are fast & easy .	you need the software to actually transfer files
2	i noticed that the micro adjustment knob and collet are well made and work well too.	the adjustment knob seemed ok, but when lowering the router, i have to practically pull it down while turning the knob.
3	the navigation is nice enough , but scrolling and searching through thousands of tracks , hundreds of albums or artists , or even dozens of genres is not conducive to save driving .	difficult navigation - i wo n't necessarily say " difficult ," but i do n't enjoy the scrollwheel to navigate .
4	i imagine if i left my player untouched (no backlight) it could play for considerably more than 12 hours at a low volume level.	there are 2 things that need fixing first is the battery life. it will run for 6 hrs without problems with medium usage of the buttons.

authors' experiments. It is formed by a bar chart containing the various topics of argumentation; each bar has size proportional to the amount of times the corresponding topic has been mentioned, and its position indicates the amount of comments that agree or disagree with the debate question. For example, for the topic 'CO2', there are 38 comments that mention it in texts written by advocates (i.e., ones who claim that climate change is man-made) and 14 people who think the opposite. Coupled to the bar chart is a table that shows the comments for each topic. Other formats evaluated by the authors were the paired table and the simple list.

In the present work, the focus is solely to choose relevant information to make the summary. Thus, it will not deal with any graphical techniques to exhibit the summary, but it will suggest a new format of summary that can be further investigated in future work.

3. Related work

This section summarizes four lines of research that are important in the research of contrastive summarization of opinions. Works are organized from the least to the most similar compared to the present work.

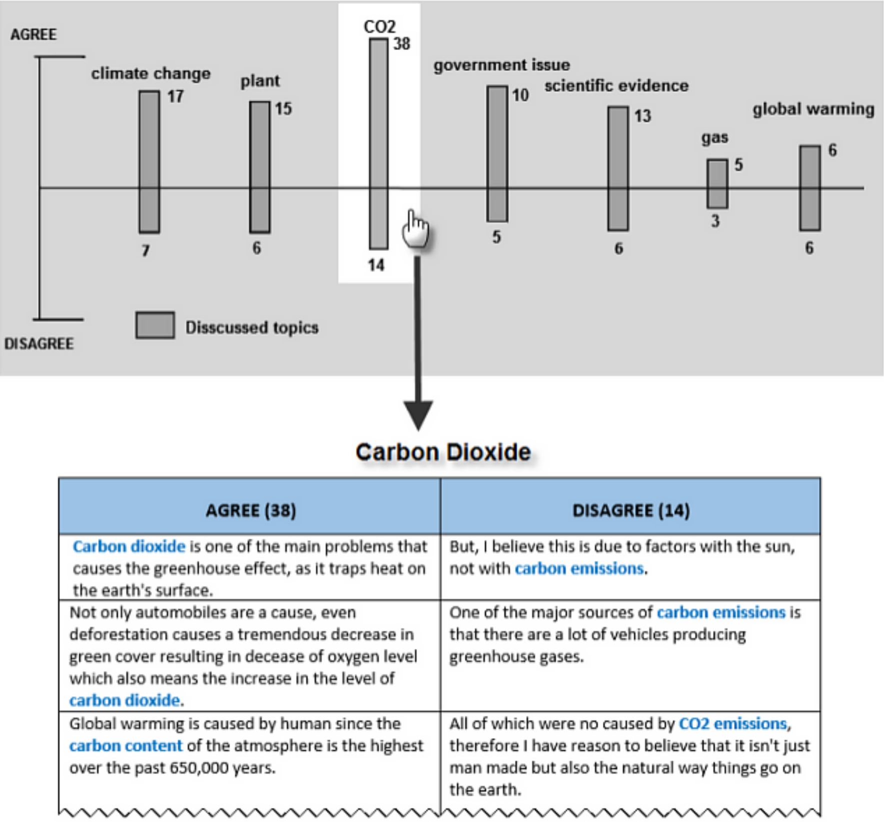
[8] has the same final goal as the present work: automatically comparing entities from opinionated texts. However, its problem definition has a very different format than the one that this paper deals with: instead of selecting opinionated sentences to build a textual summary, its system counts the opinions found and displays the numbers in a graphical format. This work is summarized here to illustrate a different kind of summary that can be used for the same purpose as the ones studied in this paper.

Figure 1 shows charts generated by the system of [8]. This chart is considered to be a summary; it can be said to be a quantitative summary since it only shows numbers about the information from the source rather than excerpts of the input set like the problem defined in the present work.

Figure 1 is a screenshot of the main window of the system when it is performing a comparison of three products. Products can be selected at the top left part of the window. At the bottom, charts exhibit the quantity of positive and negative opinions for each aspect of the products: the distance from the horizontal line to the top end of a bar is proportional to the amount of positive opinions (that talk about the product indicated by that bar, and about the aspect indicated at the top left of each chart); the distance from the horizontal line to the bottom end of a bar is proportional to the amount of negative opinions. The number of opinions is indicated below each bar. If a user clicks on a bar, the system shows (at the top right board) all the sentences that have been counted to generate that bar.

In the work of [10], the main objective is to receive opinionated texts of a single entity and identify divergent opinions about it. In the output, as shown in Table 2, opinions are placed by pairs in the summaries, where a pair is composed of

Figure 2. The best contrastive summary format according to [15].



Carbon Dioxide

AGREE (38)	DISAGREE (14)
Carbon dioxide is one of the main problems that causes the greenhouse effect, as it traps heat on the earth's surface.	But, I believe this is due to factors with the sun, not with carbon emissions.
Not only automobiles are a cause, even deforestation causes a tremendous decrease in green cover resulting in decease of oxygen level which also means the increase in the level of carbon dioxide.	One of the major sources of carbon emissions is that there are a lot of vehicles producing greenhouse gases.
Global warming is caused by human since the carbon content of the atmosphere is the highest over the past 650,000 years.	All of which were no caused by CO2 emissions, therefore I have reason to believe that it isn't just man made but also the natural way things go on the earth.

two opinions that disagree on a certain topic about the entity.

[10] handles the task as an optimization problem based on heuristic functions that estimate representativity and contrastivity of a summary. To solve the problem, two greedy algorithms are proposed.

In one of the proposed algorithms, clustering is done to identify similar sentences, which supposedly are about the same subject. Then a sentence from each cluster is chosen to be inserted into the summary. Thus, it is guaranteed that the summary will be representative (since each cluster will be represented by a sentence) and diversified (because the sentences of one cluster are supposed to be very different from those of another). Two strategies are tested to choose the sentence that will represent each cluster in the summary. At the end, pairs of contrastive sentences are formed by identifying, among the chosen sentences, which of them disagree with each other. This algorithm is called **R-First** (representativity first) because it prioritizes representativity.

In another algorithm proposed by [10], the first step is the formation of contrastive pairs from the input sentences. These pairs are ranked from the most contrastive to the least contrastive, according to some heuristic function that estimates the contrastivity of sentence pairs. Then, these pairs are selected one by one to fill in the summary, starting at the top of the rank until it reaches the size limit of the summary. However, instead of simply selecting the top-ranked elements, the selection also considers the elements previously selected to avoid redundancy and to maximize the representativity of the summary; some top sentences can be dismissed for being unnecessary when topics they mention have already been covered in the summary by sentences previously inserted. This algorithm is called **C-First** (contrastivity first) because it prioritizes contrastivity.

[9] does a job to identify comparable information between products for opinative texts. The process outline is: given opinionated texts about two products, select pairs of sentences (one about each product) in a way that the two sentences of the pair are representative and comparable. Thus, the extracted sentences are expected to explore similar points (either with opposing or confluent opinions) about similar features of the two products. For example, for a camera, a feature may be the battery, and a point of argument related to the battery may be the duration; so, when comparing two cameras, it is valuable to find two sentences (one about each) where both talk about the duration of the battery, regardless of whether they agree or disagree. As Table 1 shows, the summaries are made of four parts, one for each combination of polarities of the two entities. Sentences are paired inside each group.

The method developed by [9] considers that a summary must be representative, comparative and diversified. These metrics are defined at the sentence level: similar sentences about an entity are said to be representative of one another; similar sentences about competing entities are said to be comparative; sentences with low similarity about an entity are said to be diversified.

To measure how similar two sentences are, [9] considers topics: the more topics sentences have in common, the more similar they are. A topic can be any point of argumentation about an entity or an aspect.

To solve this (optimization) problem, one would need to find the summary that has the largest values of representativity, comparability and diversity (while also considering their relative importance). However, the researchers inform that it is not feasible because it would require a brute force solution that needs an absurdly large amount of time to compute. Because of this, they propose three algorithms where each one looks only at a single measure:

- **R-First**: considers only representativity;
- **C-First**: considers only comparability;
- **D-First**: considers only diversity.

In each one of the three algorithms, they compute, for every available sentence, how much value it aggregates to the summary according to the correspondent measure. The sentences are ranked according to this value. At the end, the top sentences are inserted into the summary.

[7] developed a non-contrastive opinion summarizer by using statistical methods that consider the likelihood of a certain aspect having a certain polarity in the input set. This work is used by [13], where the authors develop a contrastive summarizer that generates summaries that highlight differences between two entities from two sets of data, each containing opinion texts about each of the entities.

In the work developed in [13], the task of summarizing opinion is seen as the resolution of the equation

$$R = \arg\max_{|R|\leq k} L(R) \tag{1}$$

where L is a function that assigns a score to the summary R and k is the summary size limit. The score is higher the more amount of desirable information the summary contains. The problem of opinion summarization is finding an ideal way to calculate L, that assigns good scores for good summaries and bad scores for bad summaries. After this function is defined, the problem can be solved with optimization techniques.

The model proposed in [7] uses a probabilistic approach to find a summary where each aspect appears with polarities that reflect its polarities in the input set: they find the mean and standard deviation of the polarity of each aspect and build the summary so that the polarity distribution for each aspect in the summary is as close as possible to that observed in the input set, assuming that the polarities of each aspect respect the normal distribution.

The model in [7] is adapted in [13] to solve the problem of contrastive summarization. Starting from the opinion summarizer described in [7], the authors produce a method of generating contrastive summaries by modifying the objective function L, which initially is only capable to generate summaries of a single entity.

The contrastive summary can be seen as a pair of summaries R_1 and R_2, one for each entity. Three different strategies for generating the summaries R_1 and R_2 are tested:

1. Generate the two summaries independently, without modifying the scoring function;
2. Modify the L function so that it increases the score of summary pairs if they diverge from each other;
3. Modify the L function so that it increases the score of each table if it differs from the source set relative to the opposite party.

The third strategy was perceived as the most advantageous in the evaluations made by the authors.

This section has listed the three methods that were replicated in this paper: [13], [10] and [9]. Besides these, another work has been shown for illustrative purposes: [10]. The three works chosen have in common that they allow to select, within a set of opinionated texts, the most relevant sentences that allow to contrast points of view. Nevertheless, the problem definitions of the three works differ: only [13] defines the problem as it is studied in this paper. For this reason, the other two works were adapted (in the implementations made in the present paper) so that they all have the same format.

The methods used in the three works are quite different: one uses a statistical approach [13], one uses clustering [10], and another uses ranking with similarity measures [9]; one considers the aspects of opinions [13], another considers every lexical item of the sentence [10], and another separates opinions by aspect before summarization [9]. These differences are interesting because they indicate that the methods chosen have diversified strategies.

This article compares the methods to highlight the main strengths and weaknesses of each. In addition, a new method is proposed. It is explained in details in the next section.

4. Method

The proposed method considers an opinion to be a pair formed by an aspect and a polarity. The method works by first identifying the most relevant opinions of a source based on their frequency there, and then selecting sentences that contain the relevant opinions.

4.1 Basic definitions

For this method, **opinion** is defined as a tuple (a, p) where a is an **aspect** (main topic in the opinion) and p is the polarity assigned to the aspect[3]. The **polarity** will be represented by a number in the interval $[-1,+1]$ with $p<0$ if the opinion about the aspect is negative, $p>0$ if positive. When enough, the polarity will be indicated by a sign: $+$ if the opinion is positive and $-$ if the opinion is negative. Neutral opinions will be ignored because we believe that it is convenient to consider only 'strong' opinions, that is, opinions that are forth or against something.

The function $OP(s)$ is defined to extract opinions from a sentence s. This function has as output a set that contains all the opinions identified in s.

Example 1: Considering the sentences $s_1 =$ '*the screen is clear and has good resolution, but the battery doesn't last*' and $s_2 =$ '*the battery lasts enough but takes time to charge, speaker is awesome*', one possible[4] way to extract opinions from them is (consider that "scr" stands for "screen", "bat" for "battery" and "aud" for "audio"):

$$OP(s_1) = \{(\text{scr},+),(\text{bat},-)\}$$
$$OP(s_2) = \{(\text{bat},+),(\text{bat},-),(\text{aud},+)\}$$

The function OP is also defined over a *set* of sentences. If $S=\{s_i\}_{i=1}^{n}$ is a set of sentences, $OP(S)$ is the multiset made of all opinions found in the sentences of S. Each opinion appears in $OP(S)$ the same amount of times it appears in the sentences of S.

$$OP(S) = \bigcup_{i=1}^{n} OP(s_i)$$

Example: Considering the sentences s_1 and s_2 from Example 1. If $S=\{s_1,s_2\}$, then:

$$OP(S) = OP(s_1)\cup OP(s_2) = \\ \{(\text{scr},+),(\text{bat},-),(\text{bat},+),(\text{bat},-),(\text{aud},+)\}$$

Being:

e_1 : an entity;

e_2 : another entity, to be compared with e_1;

E_1 : a set of opinionated sentences about e_1;

E_2 : a set of opinionated sentences about e_2;

O_1 : the multiset of opinions contained in E_1 (defined by $OP(E_1)$);

O_2 : the multiset of opinions contained in E_2 (defined by $OP(E_2)$);

R_1 : a subset of E_1;

R_2 : a subset of E_2.

The task of **contrastive summarization of opinions** is to find the sets R_1 and R_2 given E_1 and E_2 in a way that $R=(R_1,R_2)$ is a summary that allows to compare both entities considering their main differences. The set R_1 must have relevant opinions of O_1, and R_2 of O_2. Each summary R_1 and R_2 will be called a **side** of the summary R.

A **contrastive pair** is a set of opinions (o_1,o_2) so that o_1 and o_2 both have the same aspect and opposite polarities. That is, if $o_1=(a_1,p_1)$ and $o_2=(a_2,p_2)$, then o_1 and o_2 can make a contrastive pair (o_1,o_2) if $a_1=a_2$ and $p_1\times p_2<0$.

An opinion is **opposite** to the opinion o if it has the same aspect of o and polarity opposite to that of o; it is, therefore, an opinion that could make a contrastive pair with o.

If two sentences s_1 and s_2 are such that o_1 is in s_1 and o_2 is in s_2 and (o_1,o_2) is a contrastive pair, the sentences $s1$ and $s2$ are said to (be able to) represent the contrastive pair (o_1,o_2).

4.2 Selection of opinions

The algorithm to select relevant opinions starts by identifying contrastive pairs that can be formed from O_1 and O_2: these are the pairs (o_1,o_2) where $o_1\in O_1$ and $o_2\in O_2$. The label $C(O_1,O_2)$ will be used to denote the set that encompasses all these pairs.

Example: If

$O_1 = \{(\text{A},-),(\text{A},+),(\text{B},-),(\text{B},-),(\text{B},+),(\text{C},+),(\text{D},-)\}$;

$O_2 = \{(\text{A},+),(\text{A},+),(\text{B},+),(\text{B},+),(\text{B},+),(\text{C},+),(\text{D},-)\}$,

then

$$C(O_1,O_2) = \{((\text{A},-),(\text{A},+)),((\text{B},-),(\text{B},+))\}.$$

The idea is to use the set $C(O_1,O_2)$ to choose the sentences that will go into the summary: for every contrastive pair (o_1,o_2) of C, one sentence of E_1 is chosen so that it contains the opinion o_1 and one sentence of E_2 is chosen so that it contains the opinion o_2. These two sentences are inserted into the summary. This strategy only indicates which *opinions* will be in the summary, but the summary is not built with opinions, but with sentences that contain opinions; for a given opinion, there may be more than one choice of sentence that contains it.

An ideal summary would contain all the contrastive pairs of $C(O_1,O_2)$ (that is, $C(OP(R_1),OP(R_2)=C(O_1,O_2))$, but the size limit imposed to the summary can hinder this to happen. So, there should be a way to rank the itens of C so that relevant opinions are prioritized, and less relevant opinions are inserted into the summary only if there is space left.

[3] To the interested reader, aspect and polarity identification may be performed by different methods, from using sentiment lexicons to advanced machine learning-based solutions. We suggest consulting [4], for an overview of the subject, and, specifically for Portuguese processing, the approaches in [16].

[4] There might be other solutions because opinion identification is by itself a subjective task. Besides, the specification of which elements are considered aspects can vary in different cases.

4.3 Ranking of opinions

To decide which opinions are the most important to be included in the summary, the contrastive pairs of the set C defined in the previous section must be ranked, that is, placed in order of relevance. Opinions that favor the representativeness of the abstract are considered relevant. This will be set by a score assigned to each opinion. Two strategies are proposed: one that scores each contrastive pair in a unified way, and one that splits the opinions in the pair and scores each one independently. They will be described in the next subsections; for now, the ranking method will be defined devoid of that function.

Scoring an opinion means to give it a value that expresses how much it deserves to be chosen for a summary. The function that assigns a score to an opinion o will be called $L(o)$. After scoring all the opinions of C, a priority queue is set up to guide the construction of the summary, where the highest rated opinions are placed first.

By labeling the set of contrastive pairs generically as $C = \{(o_1^i, o_2^i)\}_{i=1}^n$, two sets are defined: C_1, which contains the opinions from C that belong to O_1, and C_2, which contains the opinions from C that belong to O_2:

$$C_1 = \{o_1 \mid \exists(o_1, o_2) \in C\}$$

$$C_2 = \{o_2 \mid \exists(o_1, o_2) \in C\}$$

⌜Example: If

$$C(O_1, O_2) = \{((\mathrm{A}, -), (\mathrm{A}, +)), ((\mathrm{B}, -), (\mathrm{B}, +)), ((\mathrm{D}, +), (\mathrm{D}, -))\},$$

then

$$C_1 = \{(\mathrm{A}, -), (\mathrm{B}, -), (\mathrm{D}, +)\}$$

$$C_2 = \{(\mathrm{A}, +), (\mathrm{B}, +), (\mathrm{D}, -)\}$$

⌟

By using the score function $L(o)$ (yet to be defined), one can obtain the priority queues Q_1^c for the entity e_1 and Q_2^c for the entity e_2:

$$Q_1^c = (o_1, o_2, ..., o_n), o_i \in C_1, L(o_i) \leq L(o_{i-1}) \forall i > 1$$

$$Q_2^c = (o_1, o_2, ..., o_n), o_i \in C_2, L(o_i) \leq L(o_{i-1}) \forall i > 1$$

This definition simply shapes priority queues as two tuples (one for each entity) where an opinion is always positioned to the left of the opinions that score lower than it.

Once the priority queues are formed, the comparative summary is generated independently for each entity. First, the queue is inquired to find out which opinion is the most relevant; a sentence is retrieved from the dataset so that the most relevant opinion is covered in the sentence; every opinion that is present in the chosen sentence is placed at the end of the queue (to avoid redundancy in the summary, which may occur only if there is space left); the procedure repeats, now with the modified queue.

To make the side R_i of entity e_i given the set of sentences E_i and the priority queue Q_i^c:

1. The summary R_i is initialized as an empty set.
2. The first element of Q_i^c is labeled o_i^c.
3. A sentence s of E_i that contains the opinion o_i^c is inserted into the summary, as long as s fits in the summary and is not in the summary yet.
4. For each opinion o_j^s in $OP(s)$, if $o_j^s \in Q_i^c$, then o_j^s is removed from Q_i^c and added as the last element of Q_i^c.
5. Step 2 is revisited and the procedure repeats while a new sentence can be added to the summary.

The following sections describe the calculation of opinion scores with two strategies: the combined scoring (which scores the opinions of a contrastive pair together) and the independent scoring (which separates the opinions of a contrastive pair to score each one individually).

4.3.1 Combined scoring

In combined scoring, opinions that belong to the same pair of set C receive the same score. That is, considering the pair $(o_1, o_2) \in C$, the computing of $L(o_1)$ and that of $L(o_2)$ are made with a function $L(o_1, o_2)$ that considers both elements of the pair simultaneously. This section shows how to calculate $L(o_1, o_2)$.

Let: c_1 be the amount of elements of O_1 equal to o_1 and c_2 the amount of elements of O_2 equal to o_2; f_1 be the relative frequency of elements equal to o_1 in O_1 ($f_1 = \frac{c_1}{|O_1|}$) and f_2 the relative frequency of elements equal to o_2 in O_2. The following heuristics were tested for the calculation of L:

1. $L(o_1, o_2) = \min(c_1, c_2)$: the maximum quantity of contrastive pairs equal to (o_1, o_2) that can be made with sentences of E_1 and E_2 without repetition;
2. $L(o_1, o_2) = c_1 \times c_2$: the number of combinations of sentences from E_1 and E_2 that can make a contrastive pair (o_1, o_2);
3. $L(o_1, o_2) = \frac{1}{2}(f_1 + f_2)$: the average frequency of each opinion of the pair in its own input set;
4. $L(o_1, o_2) = \max(f_1, f_2)$: the frequency of the most frequent opinion of the pair in its own input set;
5. $L(o_1, o_2) = r$, where r is a random number: no actual criteria, so the priority queue is arbitrary.

Initial tests showed that using either $L(o_1, o_2) = c_1 \times c_2$ or $L(o_1, o_2) = \min(c_1, c_2)$ yields better results, especially when the size limit of the summary is small (and only a few opinions from the beginning of the queue are able to get in). Globally, no difference was perceived between these two choices; therefore, it was arbitrarily decided that $L(o_1,o_2) = c_1 \times c_2$ will always be used.

Example: Consider the sets of opinions

$$\begin{aligned} O_1 = \{&(\mathrm{A},-),(\mathrm{A},+),(\mathrm{B},-),(\mathrm{B},-),(\mathrm{B},+),\\ &(\mathrm{C},+),(\mathrm{D},+)\}\\ O_2 = \{&(\mathrm{A},+),(\mathrm{A},+),(\mathrm{B},+),(\mathrm{B},+),(\mathrm{B},+),\\ &(\mathrm{C},+),(\mathrm{D},-)\} \end{aligned}$$

that lead to the formation of contrastive pairs

$$\begin{aligned} p_A &= ((\mathrm{A},-),(\mathrm{A},+))\\ p_B &= ((\mathrm{B},-),(\mathrm{B},+))\\ p_D &= ((\mathrm{D},+),(\mathrm{D},-)) \end{aligned}$$

so that C is

$$C(O_1,O_2) = \{p_A, p_B, p_D\}.$$

Denoting by $c_1(i)$ the absolute frequency of i in O_1 and $c_2(i)$ the absolute frequency of i in O_2, one can obtain, with the use of combined scoring (defined above), for the elements of C:

$$\begin{aligned} L(p_A) &= c_1((\mathrm{A},-)) \times c_2((\mathrm{A},+)) = 1 \times 2 = 2\\ L(p_B) &= c_1((\mathrm{B},-)) \times c_2((\mathrm{B},+)) = 2 \times 3 = 6\\ L(p_D) &= c_1((\mathrm{D},+)) \times c_2((\mathrm{D},-)) = 1 \times 1 = 1 \end{aligned}$$

and, from these results, the priority queues are set up as

$$\begin{aligned} Q_1^c &= ((\mathrm{B},-),(\mathrm{A},-),(\mathrm{D},+))\\ Q_2^c &= ((\mathrm{B},+),(\mathrm{A},+),(\mathrm{D},-)). \end{aligned}$$

This indicates that B is the most relevant aspect of the source, and opinions about B will be the first ones to be chosen for the summary.

4.3.2 Independent scoring

Now, another strategy will be proposed to score opinions. In this strategy (unlike the one described in the previous section), each opinion receives a score that is independent of its contrastive pair.

Consider $o \in O$, where O is any of the two sets of opinions ($O \in \{O_1, O_2\}$). Call c the amount of elements of O that are equal to o. The score L will be computed by simply assigning $L(o) = c$. In other words, the score of each opinion is equal to the absolute frequency of that opinion in its input set.

Example: Consider the sets of opinions

$$\begin{aligned} O_1 = \{&(\mathrm{A},-),(\mathrm{A},+),(\mathrm{B},-),(\mathrm{B},-),(\mathrm{B},+),\\ &(\mathrm{C},+),(\mathrm{D},+)\}\\ O_2 = \{&(\mathrm{A},+),(\mathrm{A},+),(\mathrm{B},+),(\mathrm{B},+),(\mathrm{B},+),\\ &(\mathrm{C},+),(\mathrm{D},-)\} \end{aligned}$$

that lead to the following formation of contrastive pairs:

$$\begin{aligned} C(O_1,O_2) = \{&((\mathrm{A},-),(\mathrm{A},+)),((\mathrm{B},-),(\mathrm{B},+)),\\ &((\mathrm{D},+),(\mathrm{D},-))\}. \end{aligned}$$

Denoting by $c_1(i)$ the absolute frequency of i in O_1 and $c_2(i)$ the absolute frequency of i in O_2, one can obtain, with the use of independent scoring, for the elements of C_1:

$$\begin{aligned} L((\mathrm{A},-)) &= c_1((\mathrm{A},-)) = 1\\ L((\mathrm{B},-)) &= c_1((\mathrm{B},-)) = 2\\ L((\mathrm{D},+)) &= c_1((\mathrm{D},+)) = 1 \end{aligned}$$

and for the elements of C_2:

$$\begin{aligned} L((\mathrm{A},+)) &= c_2((\mathrm{A},+)) = 2\\ L((\mathrm{B},+)) &= c_2((\mathrm{B},+)) = 3\\ L((\mathrm{D},-)) &= c_2((\mathrm{D},-)) = 1 \end{aligned}$$

and, from these results, the priority queues are set up as

$$\begin{aligned} Q_1^c &= ((\mathrm{B},-),(\mathrm{D},+),(\mathrm{A},-))\\ Q_2^c &= ((\mathrm{B},+),(\mathrm{A},+),(\mathrm{D},-)) \end{aligned}$$

The draw between $(\mathrm{D},+)$ and $(\mathrm{A},-)$ is arbitrarily resolved during the selection of sentences.

4.4 Representativity maximization

The ranking of contrastive pairs (as previously established) aims to maximize the representativity of the summary since it favors the most frequent opinions. However, this strategy still seems to be very focused on contrastivity. In fact, only opinions that have the possibility of forming contrastive pairs are selected for the summary. If an opinion is very frequent in one of the source sets but there is no opposite opinion in the source set of the competing entity, it has no chance of being in the summary.

This section proposes a strategy that gives more value to frequent opinions of each input set regardless of the contrastive pairs that can be formed. To do so, other priority queues will be made (and they will be used along with the queues Q_1^c and Q_2^c previously defined) that consider only occurrences of each opinion in its own set. This new queues will be named Q_1^r and Q_2^r:

$$Q_1^r = (o_1, o_2, ..., o_{n_1}), o_i \in O_1, L(o_i) \leq L(o_{i-1}) \forall i \neq 1$$

$$Q_2^r = (o_1, o_2, ..., o_{n_2}), o_i \in O_2, L(o_i) \leq L(o_{i-1}) \forall i \neq 1$$

This definition resembles that of Q_1^c and Q_2^c, with the distinction that the new queues consider all opinions of O_1 and O_2, and not only those in C. The scoring function $L(o)$ is simply defined as the absolute frequency of o in its source set.

Then, in order to generate each side R_i of the summary ($i \in \{1,2\}$), two queues, Q_i^c and Q_i^r, are used alternately: first, Q_i^c is queried to find out which element this queue prioritizes; this element is added to the summary; this element is removed from the beginning of the queue and put at the end of the queue (then, eventually, if all elements in the queue have been already represented in the summary and there still is space left, this element can be chosen again); this element is also moved to the end in the other queue, Q_i^r (because it is already represented in the summary and there is no need to repeat it, unless after all elements in the queue have been chosen once); the query is alternated to Q_i^r, when the first element of this queue is selected and the procedure repeats.

The algorithm for the generation of the summary R_i of entity e_i given the set of sentences E_i and the priority queues Q_i^c and Q_i^r is now changed to:

1. Initialize the summary R_i as an empty set.

2. Label o_1^c the first element of Q_i^c.

3. Insert into the summary a sentence s_c of E_i that contains the opinion o_i^c and that fits in the summary and that is not in the summary yet (if such a sentence exists).

4. For each opinion o_j^s of $OP(s_c)$, if $o_j^s \in Q_i^c$, remove o_j^s from Q_i^c and add o_j^s to the end of Q_i^c.

5. Label o_1^r the first element of Q_i^r.

6. Insert in the summary a sentence s_r of E_i that contains the opinion o_1^r and that fits in the summary and that is not in the summary yet (if such a sentence exists).

7. For each opinion o_j^s of $OP(s_r)$, if $o_j^s \in Q_1^r$, remove o_j^s from Q_1^r and add o_j^s to the end of Q_1^r.

8. Go to step 2 and repeat the procedure while a new sentence can be added.

4.5 Strategies

Previous sections (4.3.1 and 4.3.2) presented two ways to compute the score used to rank the list of possible contrastive pairs: the combined scoring and the independent scoring. Each of these choices conceives a different ranking method. For each of these choices, one may or may not use the additional priority queue to maximize representativity (as described in Section 4.4). So there are four combinations of strategies to be tested. The strategies will be refered with numbers as listed in Table 3.

Table 3. Strategies used for the ranking of opinions.

strategy	use of Q_r	computation of L
1	no	combined
2	no	independent
3	yes	combined
4	yes	independent

4.6 Improvement

In an attempt to obtain a more informative summary with less irrelevant text, a way of prioritizing sentences according to their number of words was tested. With the assumption that too short sentences are of little use because they are too generic and too long sentences tend to ramble and contain matters other than opinion of interest, it was established that sentences whose number of words (not counting stopwords) as close as possible to 5 would be prioritized. This criterion was used as a tiebreaker in cases where there is more than one sentence in the source set containing the opinion of interest to be added to the summary. The ideal sentence size was chosen heuristically to find sentences short enough to make better use of the limited space of the summary, but long enough to be informative. The improvement will be called **ranking+**.

5. Experiments

This section reports the experiments performed in this work, including the dataset used, the evaluation method and the achieved results.

5.1 Dataset

The dataset made in this work consists of opinions about four products: two cameras and two smartphones. The comments were extracted from Buscapé[5], which is a Brazilian website for product search. A total of 542 comments were gotten, 450 about smartphones and 102 about cameras. The opinions about smartphones compose the dataset labeled D1, and the opinions about cameras compose the dataset D2. Other datasets were artificially created based on these two in order to get diversified scenarios:

- The set **D1** contains all the opinions about two smartphones.

- The set **D2** contains all the opinions about two cameras.

- The set **D3** was made by taking D1 and deleting some sentences so as to balance the amount of positive and negative sentences for each entity. Both D1 and D2 have a lot more positive than negative sentences (see Table 6). This new set will be used to simulate a case where there is strong controversy in the opinions about the entities.

- The set **D4** was made by taking D1 and arbitrarily deleting sentences so that one of the entities has significantly more text than the other. This set aims to assess the method in the case where one of the entities has more prominence than the other.

- The set **D5** was made by selecting the half of D2 composed by the most recent comments (i.e., the last comments published by users).

[5] ⟨www.buscape.com.br⟩

- The set **D6** was made by selecting the half of D2 composed by the least recent comments (i.e., the first comments published by users).

- The set **D7** was made by selecting only the sentences of D2 that contain opinions about the four most relevant (i.e., most frequent) aspects of D2. Its purpose is to simulate entities that have few aspects.

- The set **D8** was made by selecting arbitrary sentences of D1.

Comments were automatically divided into sentences. Table 4 counts the sentences and words of each subset and introduces the names of the sets.

Two people identified the opinions contained in the dataset. Each sentence was analyzed by a single collaborator; cases that generated doubt for some of the collaborators were discussed by the two in order to clarify eventual misinterpretations. Collaborators were instructed to identify all the opinions contained in each sentence. Each opinion is identified by an aspect and a polarity.

As it is not the focus of this work, there was no concern for deepening in difficult annotation cases (which were rare) or with redundant annotations for calculation of agreement. Single annotation strategy was already adopted by several other works in the area of Sentiment Analysis and shows to be a viable approach (see, e.g., [17], [18] and [19]).

It is interesting to notice that the evaluations that will be made will use the opinions manually identified as reference; therefore, even if there are errors (or eventual ambiguities) in the opinion identification, these errors will not affect the confidence of the evaluation, since they would be present both in the execution of the algorithm and in the evaluation: the errors would be between the sentences and the identification of opinion, but the methods studied do not look for the sentences, only for the opinions identified, and therefore these errors would not even be perceived by the methods. In addition, it would be possible to make a set of data without sentences by means of the automatic generation of random opinions to test the methods; this was not done because randomly generating opinions do not necessarily simulate occurrences of opinions in a set of natural text.

For the identification of polarity, collaborators were given three options:

- **positive**: if the opinion reflects a good, desirable thing about the aspect it mentions;

- **negative**: if the opinion reflects a bad, undesirable thing about the aspect it mentions;

- **neutral**: if the opinion is neither positive or negative.

To identify the aspects, collaborators were given a list of aspects and their definitions, which was based on [20] and adapted to this task. Exceptionally, the following cases that do not contain aspects should be identified:

- **generic**: if the sentence contains only opinions about the product as a whole and mentions no specific characteristic of it;

- **alien**: if the sentence contains opinions about entities other than the product of interest, such as the manufacturer and the shipping service.

Table 5 shows examples of some opinions identified in the sentences. Sentences were translated from Portuguese. Table 6 shows how many opinions of each polarity were identified for each product. The table also shows the number of different aspects identified for each entity.

5.2 Metrics

To evaluate the summaries obtained, the manual labeling of the dataset was used as reference. With this labeling, one can compare the opinions contained in a summary generated with those contained in the source according to a classification of opinions considered ideal by humans.

To better assess the identification of specific aspects, the opinions marked as generic by the human annotators were not considered in the evaluation. The opinions identified as aliens were also disregarded, since they are undesirable in the summary because they do not contribute with information about the product.

For the evaluation, three measures were defined: one that evaluates representativity, one that evaluates the contrastivity and one that evaluates diversity.

The **representativity** is the percentage of opinions of the input set that are represented in its summary. For each opinion of the input set, if there is any opinion of the summary that is equal to it (same aspect and same polarity), then that opinion is said to be represented in the summary. Let S be the summary generated from the source set E. Let c be the number of opinions of E that are represented in S: for each opinion of E, if there is any opinion in S with the same aspect and same polarity of it, then it is represented in S. The representativity of S is defined as

$$Pr(S) = \frac{c}{|E|}$$

The **contrastivity** considers contrastive pairs that can be formed from the two input sets that are the summarization target. A contrastive pair (o_1, o_2) containing the opinions $o_i \in E_1$ and $o_2 \in E_2$ can be made from the two input sets E_1 and E_2 if the aspect of o_1 is equal to the aspect in o_2 and the polarity of o_1 is opposite to that of o_2. Call C the set of all possible contrastive pairs that can be formed from E_1 and E_2 (without repetitions[6]), as defined in Section 4.2. In order to evaluate the summaries S_1 and S_2 (generated, respectively, from E_1 and E_2), let c_1 be the amount of pairs $(o_1, o_2) \in C$ such that $o_1 \in S_1$ and c_2 be the amount of pairs $(o_1, o_2) \in C$

[6] It would be possible to consider the repetitions to value the most frequent contrastive pairs, however it is already the role of the representativity measure to verify if the most frequent opinions are in the summary.

Table 4. Dataset size overview.

type	name		entity	sentences	words
phone	**D1**	D1a	Motorola Moto G5 Plus	269	3767
		D1b	Galaxy S7	253	3462
camera	**D2**	D2a	Canon EOS Rebel T5	68	1108
		D2b	Canon PowerShot SX520 HS	52	594
phone	**D3**	D3a	*(subset of D1a)*	150	1948
		D3b	*(subset of D1b)*	109	1508
phone	**D4**	D4a	*(subset of D1a)*	43	518
		D4b	*(copy of D1b)*	253	3462
camera	**D5**	D5a	*(subset of D2a)*	39	686
		D5b	*(subset of D2b)*	30	277
camera	**D6**	D6a	*(subset of D2a)*	29	422
		D6b	*(subset of D2b)*	22	317
camera	**D7**	D7a	*(subset of D2a)*	31	636
		D7b	*(subset of D2b)*	25	261
phone	**D8**	D8a	*(subset of D1a)*	39	572
		D8b	*(subset of D1b)*	32	284

Table 5. Examples of aspect and polarity identification.

sentence	opinions
Very fast!	performance +
Gorgeous.	design +
Great phone.	*generic* +
Don’t buy it!	*generic* −
Good but could be best.	*generic* +
Good speed to play videos and good memory storage.	performance + storage +
The screen doesn’t have the best color contrast (I found I prefer super AMOLED screens), but the sharpness is unbeatable (to read texts, for example)	screen − screen +
Very good battery, stays alive after more than one day without recharging in my use experience, excellent display with vivid colors, besides a very good processor that optimizes a lot its everyday usage.	battery + screen + performance +
For games undoubtedly it runs anything and with a great sized screen, for me it has the best cost-benefit of the market.	performance + screen + price +
Fair Device, but little memory for the functionalities it has, the camera and the screen are bellow expectations for this category and the battery duration falls short, because if one day you have a longer call or decide to use the TV, it would quickly drain.	*generic* • storage − camera − screen − battery −
The battery is just fair, because a day with longer conversation calls would drain it entirely.	battery •
However, to my frustration, the product did not arrive on time.	*alien* −
I bought this product last month.	
When it comes to that, the phone doesn’t disappoint.	

Table 6. Numerical overview of identified opinions.

set	aspects	opinions	positive	negative
D1a	15	337	247	90
D1b	14	331	242	89
D2a	13	55	44	11
D2b	15	51	43	8
D3a	11	158	103	55
D3b	10	113	52	61
D4a	13	52	41	11
D4b	14	331	242	89
D5a	12	37	27	10
D5b	10	29	26	3
D6a	8	18	17	1
D6b	11	22	17	5
D7a	4	39	33	6
D7b	4	26	22	4
D8a	12	57	47	10
D8b	12	38	24	14

such that $o_2 \in S_2$. The contrastivity of summary $S = (S_1, S_2)$ is defined as

$$Pc(S) = \frac{\frac{1}{2}(c_1 + c_2)}{|C|}$$

The **diversity** is the amount of different opinions contained in the summary in relation to the amount of different opinions contained in the corresponding source set. Consider the summary S generated from the input set E. Let C_R be the set of all reviews contained in S (without repetitions) and C_E the set of all reviews contained in E. The diversity of S is

$$Pd(R) = \frac{|C_R|}{|C_E|}$$

The evaluation is made considering only the aspect and the polarity of each opinion (according to the manual labeling); if two opinions have the same aspect and the same polarity, they are considered to be the same opinion regardless of whether the sentences are different.

To evaluate the representativity of the whole comparative summary, considering the two entities, the simple mean of the percentage of representativity calculated for each side of the summary is used. The same happens with the diversity.

Example: Suppose the opinions in the input sets are

$$E_1 = \{(\text{scr},+),(\text{scr},+),(\text{scr},-),(\text{bat},+),(\text{design},+)\},$$
$$E_2 = \{(\text{scr},+),(\text{scr},-),(\text{scr},-),(\text{bat},-),(\text{design},+)\}.$$

The set of all possible contrastive pairs that can be formed from E_1 and E_2 is

$$C = \{((\text{scr},+),(\text{scr},-)),((\text{scr},-),\text{scr},+)), \\ ((\text{bat},+),(\text{bat},-))\}$$

If the summaries are

$$R_1 = \{(\text{scr},+),(\text{bat},+)\},$$
$$R_2 = \{(\text{scr},-),(\text{design},+)\},$$

the contrastivity is:

$$Pc(R_1,R_2) = \frac{\frac{1}{2}(c_1 + c_2)}{|C|} = \frac{\frac{1}{2}(2+1)}{3} = 50\%.$$

This means that half of the pairs of C are represented in R_1 and R_2; indeed, of the three pairs of C, there is one fully represented $(((\text{scr},+),(\text{scr},-))$, because $(\text{scr},+) \in R_1$ e $(\text{scr},-) \in R_2)$ and one partially represented $(((\text{bat},+),(\text{bat},-))$, because $(\text{bat},+) \in R_1$ but $(\text{bat},-) \notin R_2)$.

The representativity score is the average of the representativity of R_1 and R_2:

$$Pr(R_1,R_2) = \frac{1}{2}\left(\frac{q_1}{|E_1|} + \frac{q_2}{|E_2|}\right) = \frac{1}{2}\left(\frac{3}{5} + \frac{3}{5}\right) = 60\%.$$

The diversity of R_1 is 50%:

$$Pd(R_1) = \frac{|\{(\text{scr},+),(\text{bat},+)\}|}{|\{(\text{scr},+),(\text{scr},-),(\text{bat},+),(\text{design},+)\}|} = \frac{2}{4}.$$

The **score of a summary** is defined as the harmonic mean of the scores Pr, Pc and Pd. The harmonic mean was chosen because, of the three Pythagorean means, it is the one that most emphasizes low values within a set, and it is considered that summaries that have any of the three measures too low are bad, even if the other measures are high.

$$H(R) = 3 \times \left((Pr(R))^{-1} + (Pc(R_1,R_2))^{-1} + (Pd(R))^{-1} \right)^{-1}$$

5.3 Evaluation

The best version of each described method was chosen according to preliminary experiments. The following lists the versions and the names that will refer to them.

- **Statistic**: the method of [13] in the variation called Strategy 3, which considers the contrast between each side of the summary and the source set of the opposite entity;
- **Clustering**: method of [10] in the C-First version, which prioritizes the contrastivity of the table of contents, adapted to summarize two entities instead of one as in the original work;
- **Similarity**: method of [9] in the D-First strategy, which prioritizes the diversity of the summary;
- **Ranking**: the original method described in Section 4 in the version called Strategy 3, which scores sentences according to contrastive pairs and considers representativity as well as contrastivity;

- **Ranking+**: the original method described in Section 4 in a variation of Strategy 3 that prioritizes sentences according to their number of words.

It is valid to question whether any choice of sentences from the source set is enough to obtain a contrastive summary, without specific algorithms for summarization. Therefore, a naive method that forms summaries through the selection of sentences by chance will also be evaluated. This method will be called the '**random** method'.

To eliminate the effect of the order of the sentences in the dataset, that order was shuffled before each execution. Since this makes the results non-deterministic, each test case was executed 100 times, and the results reported in this article are the average of the values found for the executions after discarding the 10 worst and 10 best results. This strategy of discarding the minimum and maximum results (inspired on other works in the literature, as [21] and [22]) allows removing outliers and finding more realistic evaluation results.

We show two tables with the results of the content evaluation for summaries generated with size restriction of 100 words for each entity (all measures are in a scale from 0 to 1, but they will be shown in percentage for better visualization). In the tables, the best result for each column is underlined and the worst is wavy underlined; we considered tied values with a difference of up to two points. Table 7 shows the results for the four largest datasets. The representativity, contrastivity and diversity are indicated with the labels **R**, **C**, **D**, and the overall score (harmonic mean of those three values) is indicated by the letter **H**. Table 8 shows the overall score of each test case and the average of each test score: the next to the last column shows the arithmetic mean of the scores of each method and the last one shows the average disregarding the best and the worst scores to enhance occurrence of median cases. The Student's T-test calculated for the scores of the ranking method compared to those of the statistic method is 0.015; for the ranking+ method compared to the ranking method it is 0.003.

The random method is clearly a failure because its scores are far below the other methods'. The adaptation of [10] (clustering) seems fruitless, having been competitive only for datasets D3 and D7, which are the most restrained. The method of [9] (similarity) performed poorly in all cases, except in the D2 set, where it scored much better than the clustering method and was among the three methods that scored best in that set. The method of [13] (statistic) proved to be superior to the clustering method, with a huge difference in four cases, but found a setback in the set D8, which left it with an average not far from its preceding in the rank. The novel method was better than the previous ones in all cases, with the difference (in relation to the previous method in the rank) being stronger in sets D3, D7 and D8 and almost null in the less balanced set, D4. The method obtained an average of 15 percentage points more than the previous one according to this evaluation. This difference is mainly due to the improvement in representativity, which is noticeable in the first three sets. The improvement of the novel method was able to increase 8 more points and was best placed for all datasets.

Table 9 shows the details of the evaluation of the methods, including the scores obtained for each of the three criteria. The table also shows the standard deviations within each test set (which consists of 100 runs, excluding the 10 best and 10 worst). Figure 3 shows the average of the scores and the mean of the standard deviations of each method. Small standard deviations indicate that the method generates very similar summaries regardless of the order of the sentences in the source set.

5.4 Human perception

To evaluate the human perception of the usefulness of the summaries, 7 people were invited to evaluate them. Each method was evaluated 13 times for the output obtained for different datasets. Each person was instructed to rate the summary between -3 and 3 according to how much they think the summary helps to understand the differences between the products. Table 10 shows the scores obtained. Figure 4 shows the mean and standard deviation of the scores after converting them to a scale from 0 to 100. The results of this experiment suggest that people do not perceive differences between the implemented methods, despite the huge difference found in the metrics. This probably occurred because, unlike the metrics, the volunteers did not have access to the source set so they could not compare the summaries with their inputs. Asking people to read the input set would not be a good idea because the datasets are too long and people do not have accurate memories like computers.

5.5 Outputs

This section shows examples of summaries (translated from Portuguese) obtained with the original method and its enhanced version. The summaries had a size limit set to 100 words for each entity. The summary in Table 12 contains more sentences than the one in Table 11 because the enhanced method tries to choose short sentences to better use the space (nevertheless avoiding too short sentences that could be not very informative).

6. Results and discussion

This section presents considerations about the elements presented throughout this text.

6.1 Regarding the dataset

We sought to use datasets with diversified characteristics for each test case. The main characteristics in which variability was sought are: absolute size of the dataset, size of the dataset in relation to the size of the other set of the same pair, relative frequency of each polarity and number of aspects (the latter two being obtained by means of manual identification). In the search for this diversification, we manipulated the datasets to create two pairs of artificial sets (which nevertheless simulate

Table 7. Evaluation of the methods for summaries of size 100.

	D1				D2				D3				D4			
method	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
random	47	27	27	**31**	48	24	30	**31**	45	40	42	**42**	51	28	31	**34**
clustering	64	42	35	**44**	48	44	34	**41**	51	56	43	**50**	52	41	36	**42**
similarity	24	39	37	**32**	73	67	60	**66**	37	50	50	**45**	39	46	48	**43**
statistic	64	49	47	**52**	69	70	53	**63**	52	75	55	**59**	72	50	52	**56**
ranking	78	52	46	**56**	79	64	59	**66**	79	74	63	**71**	79	53	50	**57**
ranking+	90	64	58	**68**	92	72	65	**75**	92	90	84	**89**	85	56	54	**62**

Table 8. Overall evaluation of the methods for summaries of size 100.

method	D1	D2	D3	D4	D5	D6	D7	D8	average	
random	31	31	42	34	45	43	62	48	42	**41**
clustering	44	41	50	42	41	58	73	48	50	**47**
similarity	32	66	45	43	60	65	63	72	56	**57**
statistic	52	63	59	56	66	83	56	39	59	**59**
ranking	56	66	71	57	79	92	94	81	75	**74**
ranking+	68	75	89	62	81	95	100	85	82	**82**

Table 9. Detailed evaluation of methods for summaries of size 100.

	D1				D2				D3				D4			
method	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
random	47 ±6	27 ±4	27 ±4	**31** **±4**	48 ±7	24 ±6	30 ±4	**31** **±5**	45 ±6	40 ±8	42 ±5	**42** **±5**	51 ±8	28 ±6	31 ±4	**34** **±5**
clustering	64 ±2	42 ±2	35 ±2	**44** **±2**	48 ±0	44 ±0	34 ±0	**41** **±0**	51 ±5	56 ±0	43 ±3	**50** **±3**	52 ±2	41 ±2	36 ±3	**42** **±3**
similarity	24 ±3	39 ±2	37 ±2	**32** **±2**	73 ±0	67 ±0	60 ±0	**66** **±0**	37 ±7	50 ±6	50 ±5	**45** **±6**	39 ±10	46 ±3	48 ±2	**43** **±5**
statistic	64 ±1	49 ±2	47 ±2	**52** **±1**	69 ±8	70 ±3	53 ±3	**63** **±4**	52 ±0	75 ±0	55 ±0	**59** **±0**	72 ±0	50 ±0	52 ±0	**56** **±0**
ranking	78 ±3	52 ±4	46 ±3	**56** **±4**	79 ±7	64 ±3	59 ±4	**66** **±4**	79 ±3	74 ±6	63 ±4	**71** **±4**	79 ±3	53 ±5	50 ±5	**57** **±5**
ranking+	90 ±1	64 ±2	58 ±2	**68** **±1**	92 ±0	72 ±1	65 ±1	**75** **±1**	92 ±1	90 ±3	84 ±2	**89** **±2**	85 ±1	56 ±1	54 ±1	**62** **±1**
	D5				D6				D7				D8			
	R	C	D	H	R	C	D	H	R	C	D	H	R	C	D	H
random	57 ±6	39 ±14	46 ±5	**45** **±7**	49 ±11	42 ±13	44 ±8	**43** **±10**	79 ±8	53 ±10	62 ±9	**62** **±9**	61 ±7	42 ±7	45 ±6	**48** **±6**
clustering	34 ±3	50 ±0	43 ±2	**41** **±2**	51 ±2	100 ±2	45 ±1	**58** **±2**	78 ±1	69 ±6	71 ±4	**73** **±4**	56 ±3	48 ±4	41 ±4	**48** **±4**
similarity	66 ±4	53 ±8	62 ±2	**60** **±5**	57 ±0	100 ±0	53 ±0	**65** **±0**	57 ±7	62 ±0	72 ±4	**63** **±4**	72 ±0	71 ±2	73 ±1	**72** **±1**
statistic	66 ±2	75 ±0	59 ±1	**66** **±1**	81 ±1	91 ±8	78 ±2	**83** **±3**	60 ±0	62 ±0	50 ±0	**56** **±0**	43 ±0	41 ±0	34 ±0	**39** **±0**
ranking	81 ±2	100 ±0	64 ±3	**79** **±2**	92 ±2	100 ±0	86 ±4	**92** **±2**	96 ±5	91 ±7	94 ±5	**94** **±5**	89 ±3	78 ±6	76 ±6	**81** **±5**
ranking+	87 ±1	100 ±0	65 ±2	**81** **±1**	97 ±1	100 ±0	88 ±2	**95** **±1**	100 ±0	100 ±0	100 ±0	**100** **±0**	93 ±1	84 ±1	80 ±3	**85** **±1**

Figure 3. Score of methods, according to the bold column of Table 8.

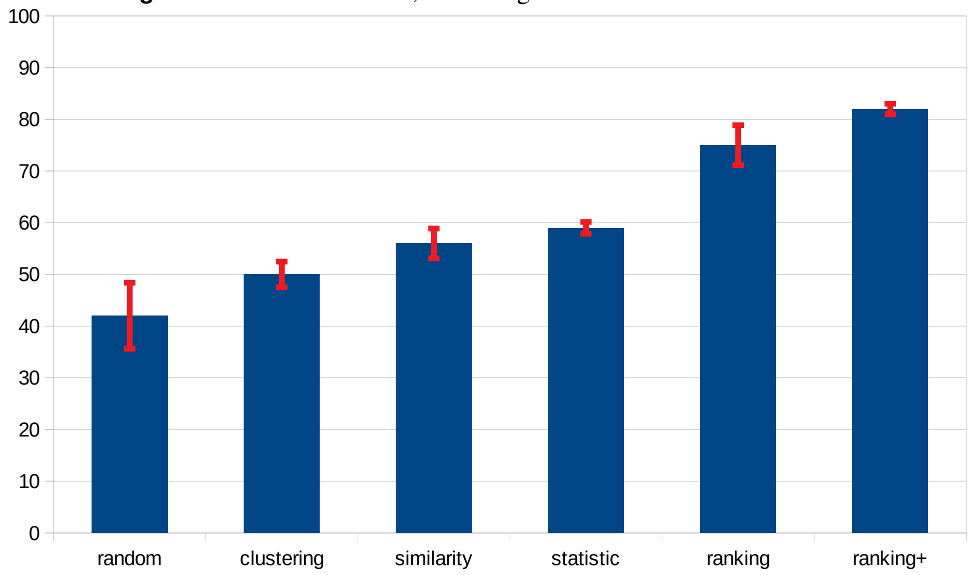
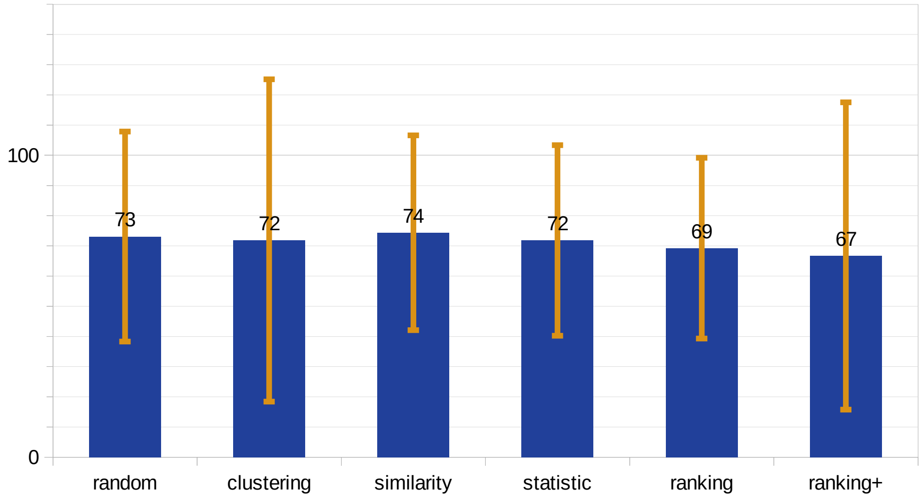


Figure 4. Average of scores assigned by volunteers to the summaries.



scenarios that may be real), thus doubling the number of test cases.

It was of great interest to do diversified tests, in which each entity pair has different characteristics of the other as to the quantity of opinions and the proportion between positive and negative opinions (as shown in Table 6); in addition, the ratio between the size of an entity's dataset relative to the opposing entity is different between entity pairs (see Table 4). The fact that the method performed well for the four pairs of entities evidences the coherence of the theory and the correctness of the method, showing that it is robust and does not destabilize with these variations. Searching for new datasets that were diversified and manually identifying their views would be very laborious, so it was convenient to do the manipulation to generate new sets from the existing ones.

6.2 On the evaluation

The metrics proposed for evaluation were useful for evaluating the composition of the summary: the percentage of representativeness indicates whether the important opinions appear in the summary; the percentage of contrastivity indicates whether the summary presents the differences between the two entities; the percentage of diversity indicates whether a summary shows different opinions among them, avoiding redundancy.

The proposed evaluation is normalized by the expectations about the source sets. For example, it is seen that the evaluation of the tests in D4 were much more satisfactory than those of D1. This occurred because D4 is a subset of D1. When you decrease a set of data, expectations about a summary of it also decrease (for example, there are fewer contrastive pairs to be formed). Besides, since the word limit of the summary is maintained, the compression of D4 is smaller than that of D1, which also contributes so that the evaluation of D4 is better

Table 10. Scores assigned by volunteers to the summaries.

		method					
set	volunteer	random	clustering	similarity	statistic	ranking	ranking+
D1	a	3	3	2	3	1	3
	b	3	0	2	1	2	1
	f	1	2	0	0	1	−2
D2	e	1	2	2	1	2	2
D3	d	1	−2	1	2	0	−1
	e	1	0	2	2	1	1
D4	d	2	1	2	1	2	3
	g	0	2	0	2	0	−1
D6	c	1	3	2	0	1	2
	f	0	3	2	1	1	1
D7	a	3	−1	3	2	3	2
	b	1	2	0	0	0	1
	g	1	2	1	2	1	1
	average	**1.4**	**1.3**	**1.5**	**1.3**	**1.2**	**1.0**
	standard deviation	1.0	1.6	1.0	0.9	0.9	1.5
	mode	1	2	2	2	1	1
	median	1	2	2	1	1	1

than that of D1.

Although not measured, the informativeness of the summary (amount of information from the source texts that is preserved in the summary [1]) has been respected thanks to some decisions made in this work. For example, in order to maximize the three evaluation measures defined, it would always be possible to put in the summary the smallest possible sentence. In fact, this would save a lot of space and increase the number of opinions in the summary, which would make it easier for metrics to reach a high value. However, that would make the summary have poor explanatory sentences like '*bad camera*' and '*liked the screen*', and a summary with only sentences like that would be uninformative. We then chose to select sentences at random to take either long sentences (which tend to be more descriptive) or short sentences (which save space and make reading easier). Also to improve informativeness, short sentences that do not contain a specific aspect have been eliminated. In addition to this, other justifications mitigate the lack of informativeness calculation:

- Representativity already measures, to a certain extent, the informativeness of the summary (albeit in a simplified form, considering only aspects and polarities).
- One way to measure informativeness is by manually generating summaries: volunteers take the source text and make summaries they deem ideal [23]. This task would be very complicated for contrastive summaries (compared to traditional summaries), as it would require an analysis of many possibilities, which humans are bad at doing.

The evaluation proposed here is by no means the only way to gauge the quality of summaries. To accept the results found here, one must agree with the evaluation criteria adopted. Someone may set different quality criteria for contrastive summaries than those presented here and, by repeating the experiments, find different results. This depends on what is considered an ideal contrastive summary.

6.3 Concerning the new method

According to the definition of the problem and the proposed evaluation, the new method presented proved to be efficient in all tests and successfully surpassed other previously published methods. Its four variations had similar performances. Even with similar results, each variation has its own strengths, which will be analyzed in Section 6.3.1.

The fact that the method is non-deterministic may bring some inconvenience in execution. How this affects results is explained in Section 6.3.2. There will be ways to mitigate possible problems that this may cause.

The method had a better evaluation than the competing methods in practically all tests, but each set of data showed a different score gain: some improved greatly, others practically remained the same as the method that obtained the second best evaluation. This will be analyzed in Section 6.3.3.

Because it did not use costly computational resources and because it generated each side of the summary independently, the proposed method had an advantage over runtime compared to other methods. Section 6.4 will report the efficiency.

Although there was a satisfactory performance for the expectations of this work, there are several actions (some very simple, others more laborious) that can be considered in an eventual implementation for real use of the method to improve the results. Some suggestions are made in Section 6.5.

Table 11. Summary obtained with the ranking method for the dataset D1.

Phone A	Phone B
Good cost-benefit. The device is great. Its performance is excellent especially for its price but it could be a little cheaper. I believe that for more picky users would not pay off. I think the battery life and camera quality were excellent points, but the main thing is that the eight processors do their job very well! But in some situations it crashes. Beautiful. Very good android device and good dispatch very fast, only the camera is not all that is advertised and disappointed me. Beautiful design, good processor, long battery life, easy to use functions. Device with problem.	It could be more cost effective. Let's see how the battery will be after some time. Everything good. Product meets expectations, no crashes, cool camera. It takes amazing pictures, has a great screen and has a perfect finish. Anyway, this phone disappointed compared to other devices I've had. Great phone with battery life of almost 1 day. At the moment I can say that the battery does not last as specified in the product. Fine, but fragile. But perfect design, the perfect screen, a great camera, super fast digital player, and the freedom to set up the device any way you want...

6.3.1 Strategies

Some differences among the four variations of the method that were observed are worthy to be commented.

Independent Score The use of independent scoring strategy scores each opinion according only to its frequency in its source set. Thus, the most frequent opinions of each group always have higher priority (unlike the conjugated score, which considers the two elements of each contrastive pair to estimate the relevance of both opinions). Therefore, this approach is expected to maximize the representativeness of the abstract.

Conjugated score Using the conjugated score, the priority queues for each entity are ordered so that the *n*-th element of the e_1 queue can form a contrastive pair with the *n*-th element of e_2 (because both opinions of each contrastive pair receive the same score and consequently take the same place in the ranking). Thus, it is expected that the *n*-th sentence on one side of the summary has an opposite opinion to the *n*-th sentence on the opposite side, which may favor reading the summary to understand the differences between the entities. This can be seen on Table 13 (where the strategy 1 of the method was used, since it considers only contrastivity, which enhances this effect). Matching is not perfect because of sentences that contain more than one opinion.

Representative ranking The use of an extra priority queue to maximize representativity helped to generate summaries that better reflect the frequent views of source sets, and also solved a crucial problem: without it, the generated abstracts would bring only opinions that could form contrastive pairs between entities, but there may be a very small number of such pairs. In the extreme case where no contrastive pair can be formed, the generated summary would be empty; if there is only one possible contrastive pair, the summary would repeat opinions (that occur in different sentences) until the size limit is reached (or until there are no more sentences that can be added without repetition), since each priority queue would contain only one opinion. With the representativity queue, the summary would contain the most frequent opinions of each source set independently of the possibility of forming contrastive pairs. Not using a representative queue would be feasible only in the case where a totally comparative summary is desired, with interest only in opinions that oppose the two entities. In the tests done in this work, this problem was not observed because coincidentally there were always a sufficient number of contrastive pairs between the entities.

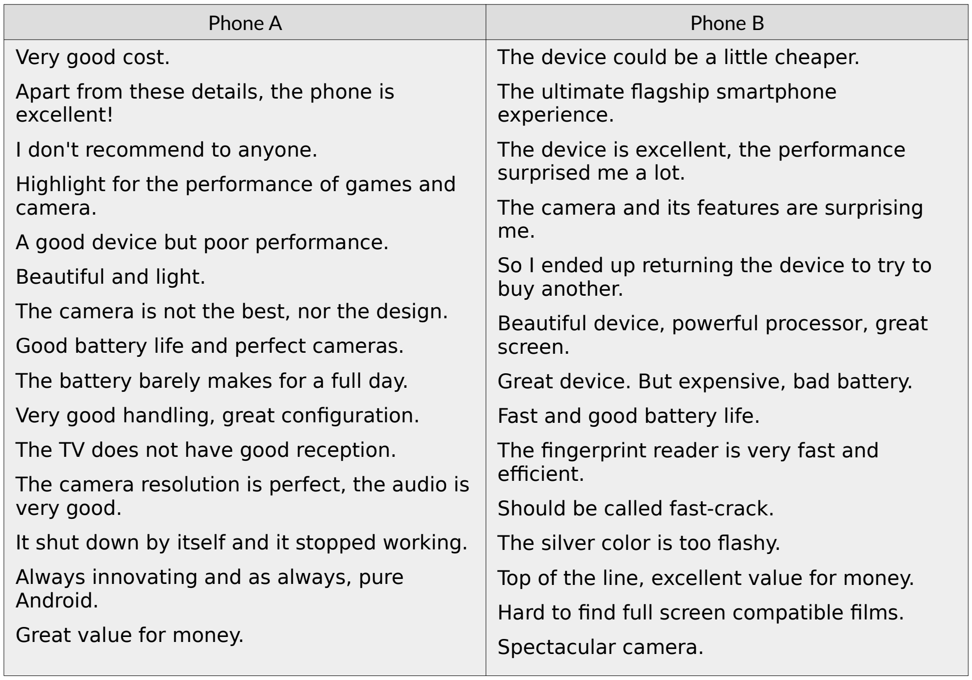
6.3.2 Non determinism

The summarization algorithm is entirely based on the opinions identified in each sentence: it indicates, at each iteration, an opinion that is desirable to have in the summary so that a later step chooses a sentence that contains that opinion to be inserted in the summary. Often, there is more than one sentence in the source set containing the indicated opinion. In this case, a sentence is arbitrarily chosen. The fact that the choice is random, in addition to making the output non-deterministic, affects the performance of the method (according to the used evaluation) mainly for reasons related to two characteristics of the selected sentence:

- Sentence size: it is possible that the tied sentences have a large variety of sizes. Choosing a very large sentence rather than a small one means leaving less space in the summary for the iterations to come. The strategy proposed in Section 4.6 can soften the non-determinism caused by this.

- Other opinions in the sentence: it is possible that the chosen sentence contains opinions other than those indicated by the algorithm. When more than one sentence can be chosen (because they contain the opinion of in-

Table 12. Summary obtained with the ranking+ method for the dataset D1.



Phone A	Phone B
Very good cost.	The device could be a little cheaper.
Apart from these details, the phone is excellent!	The ultimate flagship smartphone experience.
I don't recommend to anyone.	The device is excellent, the performance surprised me a lot.
Highlight for the performance of games and camera.	The camera and its features are surprising me.
A good device but poor performance.	So I ended up returning the device to try to buy another.
Beautiful and light.	Beautiful device, powerful processor, great screen.
The camera is not the best, nor the design.	Great device. But expensive, bad battery.
Good battery life and perfect cameras.	Fast and good battery life.
The battery barely makes for a full day.	The fingerprint reader is very fast and efficient.
Very good handling, great configuration.	Should be called fast-crack.
The TV does not have good reception.	The silver color is too flashy.
The camera resolution is perfect, the audio is very good.	Top of the line, excellent value for money.
It shut down by itself and it stopped working.	Hard to find full screen compatible films.
Always innovating and as always, pure Android.	Spectacular camera.
Great value for money.	

terest), being the choice random, it is not possible to determine which of these intrusive opinions will enter the summary, since different sentences may have different opinions. This case will be discussed in Section 6.5.1.

6.3.3 Comparison with other methods

Observing the performance (Table 9) of the method proposed here compared to [13], which was the second best evaluated, it was found that the novel method overcame it with a significant difference in all cases, except for the set D4 (where both obtained the score 59), which is the most unbalanced set in terms of the number of opinions of each entity. This anomaly can have two explanations:

- [13]'s method may be more tolerant to unbalanced datasets, and the method proposed here performs better in more balanced sets. Another evidence of this would be the fact that the novel method has given significant leverage of the results for the D3 set, which is the most balanced on the occurrence of polarities.
- [13]'s method may have succeeded in getting the close to best possible solution for this case, and it is impossible to significantly improve it.

Tests made with a brute force algorithm discarded the second hypothesis, as they were able to find a summary with a score of 75 for the same case (far higher than the score of 56 achieved before). This fact also proves that all methods are far from perfect, and that there is room for creating new algorithms that perform much better. However, the original method can find summaries with score 75 sometimes; the fact is that it is non-deterministic, and most solutions found have lower scores. In any case, an algorithm that tests all the possibilities would be ideal to find the best solution according to the proposed scoring method, notwithstanding the setback of it being pathless due to the slowness of the operation.

6.4 Regarding efficiency

An implementation for practical use would raise concern about the efficiency of implementation, especially with regard to runtime. In the experiments done on a common personal computer, no case was observed that took more than 0.1 second for the algorithm to complete its execution.

The algorithm proposed here has the advantage of confronting the two input sets only once (to form the contrastive pairs), unlike others (such as the statistic method) where the optimization must be done with an algorithm that at each new insertion in the summary must check all possible pairs that can be formed by the two sets. In addition, the algorithm proposed here does not use computationally costly features like the probability distributions used by the statistic method or the similarity clustering used by the clustering method. Just as the ranking method is the similarity method: it also selects only the most valuable sentences for the summary, but the similarity method does a lot of simplification (to avoid a more rigorous optimization problem), which leaves it with humble results.

Table 14 shows the runtime to generate a summary for the six largest datasets using each of the methods. Time is indicated in seconds and was obtained from the average of 100 runs of each test case on a standard personal computer.

The random method is of course the fastest, as it does not use any algorithm to select appropriate sentences, but instead

Table 13. Summary obtained with the first strategy of the improved ranking method for set D1, with size limit 75 and ideal sentence size equal to 5.

Phone A	Phone B
Very good product, great value for money. A good device that does not underperform. The camera is not the best, nor the design. Good battery life and perfect cameras. Very good, only battery and design disappoint. There is no option to delete a single outgoing or incoming call. Quality phone, beautiful and with a super durable battery. The camera resolution is perfect, the audio is very good. Device with problem.	The problem for me is the battery, cost and not being dual chip. It is very fast, in my use it never crashed. The camera is great for high speed photos. Very fast! Doesn’t crash! Battery lasts a long time! The speed of the fingerprint reader is also a plus point. It's fast, very beautiful, functional, great camera. The silver on the back of the device gets dirty easily . Great product, product made to last. Good image.

Table 14. Runtime (in seconds) of the methods.

method	D1	D2	D3	D4	D5	D6	total
random	0,04	0,00	0,01	0,02	0,00	0,00	**0,07**
clustering	13,35	0,02	1,72	0,50	0,01	0,00	**15,6**
similarity	0,15	0,01	0,02	0,07	0,00	0,00	**0,25**
statistic	40,46	1,78	9,10	4,54	0,47	0,21	**56,56**
ranking	0,07	0,01	0,02	0,03	0,00	0,01	**0,14**
ranking+	0,07	0,01	0,01	0,03	0,00	0,01	**0,13**

performs a draw to choose them. Because it is a seemingly useless method, it is irrelevant to compare its execution time with those of other methods.

6.5 Future improvements

Finally, we suggest in this section some modifications that could be made to improve the quality of the method. They remain as future work.

6.5.1 Extra opinions in the sentences

Ideally, only the opinions chosen by the algorithm would be included in the summary, but some sentences have more than one opinion, which may favor or impair performance if they are inserted in the summary, as exemplified by the following cases:

- If all the opinions contained in a chosen sentence would be chosen in the following steps of the algorithm, it reaches its goal for the current iteration and (coincidentally) for the next ones, eventually saving space for having found a single sentence that brings several opinions of interest.
- If some opinion contained in the sentence has already entered the summary, redundancy occurs because it is not desirable to include repeated opinions (unless there is enough space in the summary after all opinions of interest have been included).
- If any opinion contained in the sentence would never be indicated by the algorithm, there is the presence of irrelevant information in the summary.

The observation of these points allows to formulate improvements to the method, so that the tie-breaker is made considering all the opinions contained in a candidate sentence, assigning a higher score to the sentences that contain opinions that favor the method.

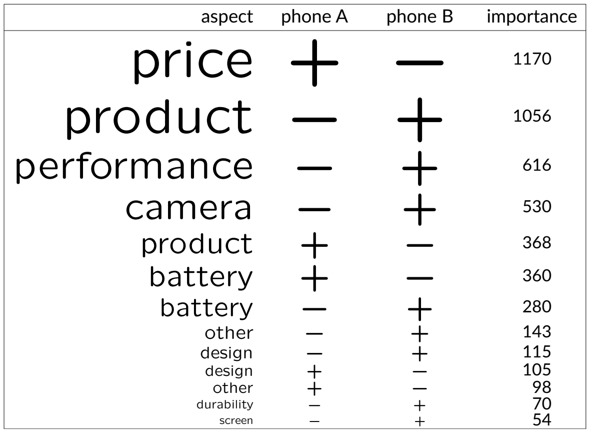
6.5.2 Quantitative summary

The summary generated by the methods concatenates the text segments that are considered the most relevant ones. From the user’s point of view, it is obscure to know which opinions presented in the summary are the most relevant and most frequent of the source set, since:

- Some sentences in the summary may also contain irrelevant information (i.e., information not desirable to be included in the summary, but was inserted because the sentence has some other opinion of interest).
- Some sentences may have been selected only for the possibility of forming contrastive pairs, without necessarily containing frequent opinions.
- The summary may contain redundant information (when space is left in the summary and there are no diversified sentence options to insert), but the fact that an opinion appears repeated in the summary does not necessarily reflect the frequency of such opinion in the source set.
- The summary does not indicate how often the opinions are.

We propose here a contrastive summary format that displays statistics about the contrastive pairs detected in the source set. Figure 5 shows a real example made for the D1 set. It shows all contrastive pairs detected, ordered by importance[7], where importance is estimated by the scoring function defined in Section 4.3.1. To emphasize the most frequent opinions, the font size is roughly proportional to the importance of each pair.

Figure 5. Quantitative summary based on statistics of contrastive pairs of dataset D1.



aspect	phone A	phone B	importance
price	+	−	1170
product	−	+	1056
performance	−	+	616
camera	−	+	530
product	+	−	368
battery	+	−	360
battery	−	+	280
other	−	+	143
design	−	+	115
design	+	−	105
other	+	−	98
durability	−	+	70
screen	−	+	54

In addition to allowing the most important contrastive pairs to be easily identified, this format allows a general assessment of each entity: one can look at the columns containing the polarity indications in order to find out which of the two have the most positive points and how frequent these good points are.

It is suggested to use a quantitative summary as defined here along with an extractive summary so that the user has both an overview of the opinions and a sample of those that are deemed the most relevant and informative.

Author contributions

Raphael Rocha da Silva has worked on the investigation and experiments, while Thiago Alexandre Salgueiro Pardo was his advisor. Both the authors have contributed to the writing of the paper.

Acknowledgments

We are grateful to Otávio Augusto Ferreira Sousa for helping building the dataset used in this project and to São Paulo Research Foundation (grant #2017/12236-0) and USP Research Office (PRP N. 668) for the financial support. We also acknowledge that this initiative is part of the Center for Artificial Intelligence (C4AI) of the University of São Paulo, sponsored by IBM and FAPESP (grant #2019/07665-4).

[7] Because it is information that requires theoretical knowledge of the method to be interpreted, it is not recommended to display the importance value to the user (which, for didactic purposes only, was included in Figure 5), but only use it to sort the opinions.

References

[1] MANI, I. *Automatic Summarization*. [S.l.]: John Benjamins Publishing, 2001.

[2] NAZARI, N.; MAHDAVI, M. A. A survey on automatic text summarization. *Journal of AI and Data Mining*, v. 7, n. 1, p. 121 – 135, 2019.

[3] NENKOVA, A.; MASKEY, S.; LIU, Y. Automatic summarization. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (HLT '11), p. 3:1–3:86.

[4] LIU, B. *Sentiment Analysis and Opinion Mining*. [S.l.]: Morgan and Claypool Publishers, 2012.

[5] HOQUE, E.; CARENINI, G. Multiconvis: A visual text analytics system for exploring a collection of online conversations. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. New York, NY, USA: ACM, 2016. (IUI '16), p. 96–107.

[6] ZHANG, C. et al. Big data versus the crowd: Looking for relationships in all the right places. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (ACL '12), p. 825–834.

[7] LERMAN, K.; BLAIR-GOLDENSOHN, S.; MCDONALD, R. Sentiment summarization: Evaluating and learning user preferences. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (EACL '09), p. 514–522.

[8] LIU, B.; HU, M.; CHENG, J. Opinion observer: Analyzing and comparing opinions on the web. In: *Proceedings of the 14th International Conference on World Wide Web*. New York, NY, USA: ACM, 2005. (WWW '05), p. 342–351.

[9] JIN, J.; JI, P.; GU, R. Identifying comparative customer requirements from product online reviews for competitor analysis. *Eng. Appl. Artif. Intell.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 49, n. C, p. 61–73, mar. 2016.

[10] 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*. New York, NY, USA: ACM, 2009. (CIKM '09), p. 385–394.

[11] XU, X.; MENG, T.; CHENG, X. Aspect-based extractive summarization of online reviews. In: *Proceedings of the 2011 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2011. (SAC '11), p. 968–975.

[12] WANG, D.; ZHU, S.; LI, T. Sumview: A web-based engine for summarizing product reviews and customer opinions. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 40, n. 1, p. 27–33, jan. 2013.

[13] LERMAN, K.; MCDONALD, R. Contrastive summarization: An experiment with consumer reviews. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. (NAACL-Short '09), p. 113–116.

[14] WANG, D. et al. Comparative document summarization via discriminative sentence selection. *ACM Trans. Knowl. Discov. Data*, ACM, New York, NY, USA, v. 6, n. 3, p. 12:1–12:18, out. 2012.

[15] SANCHAN, N.; BONTCHEVA, K.; AKER, A. Understanding man Preferences for Summary Designs in Online Debates Domain. *Polibits*, scielomx, p. 79 – 85, 12 2016.

[16] MACHADO, M. et al. Learning rules for automatic identification of implicit aspects in portuguese. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. [S.l.: s.n.], 2021. p. 82–91.

[17] MOZETIC, I.; GRCAR, M.; SMAILOVIC, J. Multilingual twitter sentiment classification: The role of human annotators. *PLOS ONE*, v. 11, 02 2016.

[18] LITMAN, D.; FORBES, K.; SILLIMAN, S. Towards emotion prediction in spoken tutoring dialogues. In: *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*. [S.l.: s.n.], 2003. p. 52–54.

[19] CORTIZ, D. et al. A weakly supervised dataset of fine-grained emotions in portuguese. In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, RS, Brasil: SBC, 2021. p. 73–81.

[20] VARGAS, F. A.; PARDO, T. A. S. Aspect clustering methods for sentiment analysis. In: *Proceedings of the International Conference on Computational Processing of the Portuguese Language*. [S.l.: s.n.], 2018. p. 365–374.

[21] PIASECZNY, W. *Adaptive document discovery for vertical search engines*. Dissertação (Mestrado) — Simon Fraser University, 2020.

[22] SARPE, I.; VANDIN, F. Presto: Simple and scalable sampling techniques for the rigorous approximation of temporal motif counts. In: *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. [S.l.: s.n.], 2021. p. 145–153.

[23] CONDORI, R. E. L.; PARDO, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. *Expert Systems with Applications*, v. 78, p. 124–134, 2017.