Analysis of techniques for automatic summarization of hotel opinions

Paulo César M. Sousa[1], **Márcio de Souza Dias**[1], **Sérgio Francisco da Silva**[1]

[1]Instituto de Biotecnologia – Universidade Federal de Catalão (UFCAT)
Catalão – GO – Brazil

paulocmsousa99@gmail.com, {marciodias,sergio}@ufcat.edu.br

Abstract. *This paper presents a comparison of different techniques aimed at automatic summarization of textual content found in hotel reviews. Extractive techniques that generate an aspect-based summary as well as techniques that generate a general summary are analysed. The reviews themselves were extracted from a novel corpus comprising data collected from the TripAdvisor platform, focusing on hotels from different regions of Brazil. All automatic summaries were evaluated using the ROUGE set of metrics against summaries created by human annotators. The results revealed some key limitations within ROUGE when used on shorter, informal documents, as well as variations in the effectiveness of different techniques in addressing specific aspects of summarization.*

1. Introduction

Automatic text summarization is a task that focuses on generating a summary from one or more information sources, from where it extracts the content that is most relevant to the topic of interest, presenting that information in a condensed form which is tailored to the user's needs [13]. That content can vary according to the topic and the user's interest, so it is possible to generate more than one summary from the same information source.

To ensure the effectiveness of automatic summarization, it is crucial to employ a highly representative corpus that closely aligns with the data to be summarized. With the corpus in hand, the task of text summarization can begin, which can be done by either an extractive approach, or an abstractive one [14]. Extractive summarization involves extracting the most informative and relevant sentences from the source text without modifications of any kind. In contrast, abstractive summarization aims to enhance clarity and readability by rewriting sentences of the final summary. The resulting summary can be categorized into three distinct types: indicative, comprising essential topics only; informative, encompassing the most crucial information as a self-contained text; and critical, incorporating critiques of the generated content [3].

The importance of automatic opinion summarization has been steadily increasing due to the escalating volume of opinions posted on various websites in recent times. According to [4], reviews and opinions found on travel and booking sites play a vital role in users' decision-making processes within these same platforms. This significance is expected to grow further as the number of users and available data online continues to increase in upcoming years. Summarizing hotel opinions holds crucial importance for two primary reasons: 1) enhancing the efficiency of opinion verification by hotel managers, leading to improvements, and 2) simplifying the user's experience by providing them with a concise summary of the most informative opinions regarding one or more desired hotels based on their needs.

This work focuses on studying different algorithms and summarization models, with the objective of analyzing how each technique handles the task of summarizing hotel opinions. The techniques were employed on a corpus of hotel opinions obtained from TripAdvisor[1]. Subsequently, an evaluation was conducted using ROUGE [12] to compare the automatically generated summaries with "Gold Standard" summaries (generated by human annotators) on the mentioned corpus.

The remainder of this article is organized as follows: Section 2 presents a selection of the most relevant related works; Section 3 provides details about the corpus and the process of human summary generation; Section 4 covers the summarization techniques employed; Section 5 describes the achieved results; and finally, Section 6 presents the conclusion derived from the experiments.

2. Related Works

Opinion summarization is a relativelly new field, but one that already contains relevant studies exploring many different approaches. Important and relevant works related to this topic will be discussed in the following paragraphs.

Condori [3] developed aspect-based techniques for extractive and abstractive opinion summarization in Brazilian Portuguese, as well as a method to compare their generated summaries. Comparisons were conducted using multiple metrics, including informativity, utility, linguistic quality, and readability. Human reviewers and the ROUGE package were utilized for evaluation. Two corpora, ReLi [8] (containing reviews related to literary works) and Buscapé [9] (focused on electronic products), were employed for generating summaries. The findings indicated that Condori's proposed algorithms generally outperformed other models, including those proposed in [10].

Raut and Londhe [15] mined and summarized hotel opinions using supervised machine learning. They developed a technique based on SentiWordNet [6] to classify opinions as positive or negative and extract them based on this classification method. The task of text summarization was integrated into a framework for opinion mining, retrieval, and summarization. Using a corpus composed of TripAdvisor opinions, the systems proposed by the authors achieved over 87% accuracy in classifying opinion polarities with the Naive Bayes model.

Akhtar et al. [1] focused on studying and classifying the types of information found in opinions. Latent Dirichlet Allocation (LDA) was employed to uncover hidden information, while sentiment analysis was used to determine the polarity of each sentence. The summarization process relied on results returned by polarity classification, selecting the most informative sentences of each identified aspect. The resulting summary included up to three sentences of both polarized (positive and negative), as well as neutral nature for each aspect.

Freires and Holanda [7] developed SumOpinions as a method to summarize opinions extractively on various tourist attractions present in booking websites such as TripAdvisor. The method uses topics and unsupervised probabilistic modeling. Results obtained

[1] https://www.tripadvisor.com.br

by SumOpinions are shown to be better than those from K-Medoids on Topic Coverage from [11], indicating that SumOpinions is better at selecting informative sentences that cover more of the base set of opinions.

Our approach differs from related works in three key aspects: (i) We evaluate various techniques using the same corpus, including both general summary and aspect-based approaches; (ii) We assess the effectiveness of the ROUGE metrics in handling informal and opinionated content, despite being originally designed to evaluate formal and factual documents; (iii) We provide a large corpus consisting exclusively of hotel opinions in Brazilian Portuguese.

3. Brazilian TripAdvisor Corpus

The corpus used in the experiments is a new corpus created from scratch and is comprised entirely by data collected from TripAdvisor, including 413 hotels and 826,436 opinions. The average number of opinions per hotel was approximately 2,001. This corpus consists of a total of 65,715,668 tokens in opinion content and 2,999,528 tokens in titles, resulting in a combined total of 68,715,196 tokens. Hotels were selected from various regions of Brazil, without any regional filtering. Opinions are written in Brazilian Portuguese and categorized as positive (4 and 5 stars), negative (1 and 2 stars), or neutral (3 stars) based on a 1 to 5 scale as that's what TripAdvisor uses.

The process of generating general summaries relied on guidelines and methodologies employed in the Document Understanding Conference (DUC)[2] iterations of 2001, 2002, and 2003. For aspect-based summaries, an extractive approach similar to the one described in [3] was utilized. This methodology was chosen for its alignment with the experimented techniques.

Five hotels were used for summary generation, where an hotel was considered its own topic and each one of the four annotators were assigned to generate one general summary, and one aspect-based summary with fifty sentences for each one of those five topics. Fifty opinions were sampled from every hotel and given to the annotators from which they would extract said fifty sentences.

Given that some techniques employed aspect-based approaches, certain characteristics were taken into consideration: (i) Aspects should be categorized and ranked based on their importance, which was determined by their frequency in the opinions; (ii) ideally, each aspect should comprise ten sentences, with an equal distribution of five positive and five negative sentences. However, this distribution was not mandatory, as cases with a limited number of positive or negative sentences could occur. This requirement of ten sentences was done in order to make aspect-based summaries contain a similar amount of sentences compared to general summaries. As for the recommendation to balance positive and negative sentences, that is because one of the techniques (Opizer-E) tries to present positive and negative sentences for each aspect as much as it can, disregarding even sentence informativeness in some cases.

4. Summarization Methods

This section briefly describes the summarization methods that were implemented and experimented with in this research. More details about each method can be found in the

[2] https://duc.nist.gov

cited references.

K-Medoids: this technique, based on one of the approaches from [11], uses K-Medoids clustering to select the top-k sentences for the final summary. The dissimilarity between sentences is calculated using Euclidean Distance on TF-IDF-based opinion matrices. Partitioning Around Medoids (PAM) is used to find the k-medoids.

Tadano: Tadano, Shimada and Endo [16] developed a method that extracts opinion characteristics by TF-IDF values, polarity bias and intra-cluster mentions, with K-Means clustering with Euclidean Distance and Lloyd's Algorithm being used as well. TF-IDF values and sentence mentions are used to measure cluster importance in order to identify representative sentences. Aspect qualifiers (a set of markings that indicate if a given aspect is being rated positively or negatively) are also considered.

Opizer-E: developed by Condori [3], it generates general summaries by extracting sentences related to the most recurrent aspects of opinions. It employs a two-stage process: sentence clustering and sentence ranking. Unlike traditional clustering methods, Opizer-E groups sentences based on aspects and polarities. Euclidean Distance is used to calculate the dissimilarity between aspects and qualifiers. Sentence ranking considers a sentence's position and proximity of aspects and qualifiers present on it. Sentences are extracted per aspect and polarity, with only the n most important sentences for each aspect and polarity being included in the summary.

LexRank: introduced by Erkan and Radev [5], is a graph-based summarization technique that assigns salience to sentences based on their lexical centrality within a document. Sentences are represented as vertices and sentence similarities as edges. LexRank calculates the similarity between sentence pairs and determines the centrality of each sentence based on its similarity to others. It uses a Bag-of-Words model and a modified IDF equation to compute sentence similarities. By applying a threshold, sentences with similarity coefficients below a certain value (0.25 in this case) are removed.

MMR: Maximal Marginal Relevance (MMR), presented by Carbonell and Goldstein [2], is a summarization method that aims to balance the relevance and diversity of retrieved information. MMR uses a greedy algorithm to select relevant sentences from a corpus based on a given query. The selection criterion considers both relevance to the query and similarity to previously selected subsets.

5. Experiments and Results

All summaries generated by the previously described techniques were evaluated using ROUGE-1, ROUGE-2, and ROUGE-L, a set of metrics that measures the matching of unigrams, bigrams, and longest sub-sequence, respectively, between the automatic summaries and the summaries created by annotators. Each ROUGE metric is composed of three sub-measures: Recall, Precision and F-measure [12]. From all experimented techniques, MMR, LexRank, and K-Medoids are applied to generate general summaries, while Opizer-E and Tadano are applied to generate aspect-based summaries. Results are separated by annotator (see Tables 1, 2, 3 and 4), and the best scores for each hotel are highlighted.

To ensure consistency with annotated summaries, it was decided that all automatic summarizers would generate summaries consisting of approximately 50 sentences each.

For aspect-based summaries, the techniques were required to generate a summary that included the five most important aspects based on their frequency in the source opinions, with 10 informative sentences for each aspect. Opizer-E, which aims to balance positive and negative polarities, was configured to choose five sentences of each polarity for each aspect.

Analyzing the results, some patterns emerge. Regardless of annotator or technique, all results show rather low scores. This can be attributed to the subjective nature of opinion summarization, where sentence selection can vary significantly among annotators. Unlike journalistic or scientific documents, opinions tend to have shorter summaries, which hampers ROUGE's performance that usually benefits from longer summaries with more content. Consequently, there is a significant variance in results between hotels and annotator summaries, indicating that ROUGE struggles to handle short documents effectively. This is logical considering that ROUGE was originally designed for evaluating larger documents. The subjective content and informal writing style of opinions also contribute to the high variance in results.

Notably, Tadano's approach outperformed Opizer-E in aspect-based summarization for all annotators. Tadano's method focuses on extracting aspects and informative sentences without considering polarity, while Opizer-E aims to balance positives and negatives; this difference benefits Tadano's results, as the annotators were instructed to prioritize sentence selection based on importance in case polarity balancing wasn't possible.

As for general summaries, K-Medoids and LexRank consistently outperform MMR, demonstrating competitive performance and better sentence selection. This result suggests that MMR, designed for ranking webpages based on topic diversity and coverage, is not optimized for handling redundant information in text summarization, which is often used as an indicator of importance on other techniques.

Overall, ROUGE does not appear to be a suitable evaluation metric for opinion summaries due to its consistently low scores with too much variance, which can also seen on [3], where the highest score achieved was 0.393 on a corpus with far less words and number of opinions compared to ours. These challenges are not limited to extractive techniques, but also extend to abstractive summaries according to [17], on which ROUGE's performance is shown to be similarly weak across the board, failing to differentiate between accurate summaries and inaccurate ones. ROUGE scores are below average even when evaluating summaries in a multitude of configurations and generation setups, which implies that not even cutting edge techniques based on machine learning will be able to achieve decent scores with opinion summaries, as the issue lies in how ROUGE evaluates those, and not in how they are generated in the first place.

Those difficulties faced by ROUGE can be attributed to the shorter nature of opinion summaries, as well as the informality, grammatical errors, and inconsistencies found in source opinions. These limitations of ROUGE raise important questions regarding what could be considered a good way of evaluating opinion summaries, given that ROUGE evaluates word and sentence co-occurrence between summaries.

Annotator #1										
		ROUGE-1			ROUGE-2			ROUGE-L		
		Recall	Precision	F-measure	Recall	Precision	F-measure	Recall	Precision	F-measure
General Summaries										
K-Medoids	Hotel 1	**0.2011**	**0.2107**	**0.1902**	**0.0901**	**0.0966**	**0.0910**	**0.1864**	**0.1943**	**0.1780**
	Hotel 2	0.0889	0.0988	0.0848	0.0137	0.0179	0.0149	0.0738	0.0821	0.0695
	Hotel 3	**0.1176**	**0.1200**	**0.1080**	**0.0221**	**0.0235**	**0.0228**	**0.1030**	**0.1047**	**0.0940**
	Hotel 4	**0.0751**	0.0829	**0.0710**	**0.0150**	**0.0158**	**0.0154**	**0.0691**	0.0738	**0.0644**
	Hotel 5	**0.1700**	**0.1711**	**0.1630**	**0.0665**	**0.0694**	**0.0677**	**0.1521**	**0.1535**	**0.1465**
LexRank	Hotel 1	0.0674	0.1184	0.0798	0.0056	0.0092	0.0065	0.0615	0.1114	0.0735
	Hotel 2	**0.1070**	**0.1444**	**0.1142**	**0.0383**	**0.0434**	**0.0388**	**0.0948**	**0.1260**	**0.1011**
	Hotel 3	0.0612	0.0941	0.0683	0.0074	0.0123	0.0088	0.0566	0.0874	0.0632
	Hotel 4	0.0562	0.1067	0.0646	0.0118	0.0218	0.0127	0.0533	**0.1043**	0.0620
	Hotel 5	0.0932	0.1605	0.0997	0.0286	0.0378	0.0309	0.0866	0.1528	0.0933
MMR	Hotel 1	0.0685	0.0762	0.0646	0.0014	0.0006	0.0009	0.0593	0.0680	0.0564
	Hotel 2	0.0558	0.1184	0.0606	0	0	0	0.0503	0.1020	0.0537
	Hotel 3	0.0484	0.0521	0.0413	0.004	0.0011	0.0017	0.0445	0.0497	0.0383
	Hotel 4	0.0665	**0.0884**	0.0676	0.0033	0.0033	0.0031	0.0607	0.0821	0.0621
	Hotel 5	0.0630	0.0775	0.0612	0.0035	0.0036	0.0035	0.0482	0.0581	0.0463
Aspect-based Summaries										
Opizer-E	Hotel 1	0.0447	0.0594	0.0487	0	0	0	0.0438	0.0581	0.0477
	Hotel 2	0.0560	0.0644	0.0531	**0.0188**	**0.0215**	**0.0195**	0.0521	0.0582	0.0490
	Hotel 3	0.0721	0.0626	0.0609	**0.0115**	**0.0064**	**0.0074**	0.0677	0.0579	0.0563
	Hotel 4	0.0599	0.0775	0.0610	**0.0206**	**0.0224**	**0.0211**	0.0585	0.0769	0.0601
	Hotel 5	0.0436	0.0602	0.0425	0	0	0	0.0403	0.0574	0.0397
Tadano	Hotel 1	**0.1152**	**0.1296**	**0.1145**	**0.0224**	**0.0265**	**0.0220**	**0.1123**	**0.1247**	**0.1110**
	Hotel 2	**0.0865**	**0.0960**	**0.0838**	0.0004	0.0030	0.0007	**0.0846**	**0.0940**	**0.0818**
	Hotel 3	**0.0868**	**0.1121**	**0.0923**	0.0008	0.0008	0.0008	**0.0813**	**0.1060**	**0.0866**
	Hotel 4	**0.0846**	**0.0990**	**0.0857**	0.0181	0.0181	0.0181	**0.0820**	**0.0976**	**0.0839**
	Hotel 5	**0.1109**	**0.1413**	**0.1165**	**0.0222**	**0.0318**	**0.0252**	**0.1070**	**0.1344**	**0.1121**

Table 1. ROUGE's results for Annotator nº 1.

Annotator #2										
		ROUGE-1			ROUGE-2			ROUGE-L		
		Recall	Precision	F-measure	Recall	Precision	F-measure	Recall	Precision	F-measure
General Summaries										
K-Medoids	Hotel 1	**0.0961**	0.0852	0.0811	0.0165	0.0147	0.0142	**0.0783**	0.0740	0.0687
	Hotel 2	0.1034	0.1020	0.0895	0.0084	0.0075	0.0076	0.0891	0.0871	0.0761
	Hotel 3	**0.0820**	**0.0869**	**0.0752**	0.0026	0.0058	0.0031	**0.0701**	0.0734	**0.0638**
	Hotel 4	**0.0893**	0.0807	0.0676	0.0015	0.0021	0.0015	0.0768	0.0659	0.0562
	Hotel 5	**0.1236**	**0.1083**	**0.1071**	**0.0229**	**0.0174**	**0.0193**	**0.1035**	0.0862	**0.0874**
LexRank	Hotel 1	0.0906	**0.1457**	**0.1030**	**0.0171**	**0.0187**	**0.0173**	0.0781	**0.1287**	**0.0889**
	Hotel 2	**0.1116**	**0.1669**	**0.1186**	**0.0189**	**0.0346**	**0.0217**	**0.0997**	**0.1485**	**0.1049**
	Hotel 3	0.0683	0.0799	0.0650	0.0053	0.0027	0.0036	0.0616	0.0721	0.0582
	Hotel 4	0.0830	**0.0855**	**0.0731**	**0.0051**	**0.0035**	**0.0042**	**0.0769**	**0.0767**	**0.0665**
	Hotel 5	0.0940	0.1086	0.0877	0.0128	0.0124	0.0113	0.0840	**0.0995**	0.0790
MMR	Hotel 1	0.0718	0.0813	0.0612	0.0055	0.0035	0.0042	0.0597	0.0697	0.0519
	Hotel 2	0.0554	0.0991	0.0625	0.0040	0.0172	0.0063	0.0448	0.0824	0.0513
	Hotel 3	0.0708	0.0845	0.0652	**0.02**	**0.02**	**0.0199**	0.0637	**0.0745**	0.0594
	Hotel 4	0.0683	0.0762	0.0590	0.0004	0.0005	0.0004	0.0606	0.0701	0.0532
	Hotel 5	0.0654	0.0594	0.0543	0.0050	0.0043	0.0042	0.0582	0.0531	0.0479
Aspect-based Summaries										
Opizer-E	Hotel 1	0.0763	0.0793	0.0710	**0.0157**	**0.0060**	**0.0086**	0.0706	0.0756	0.0667
	Hotel 2	0.0508	0.0500	0.0443	0.0038	0.0026	0.0031	0.0482	0.0457	0.0416
	Hotel 3	0.0759	0.0602	0.0626	0.0150	0.0068	0.0086	0.0691	0.0546	0.0566
	Hotel 4	0.0824	0.0696	0.0660	**0.0296**	**0.0122**	**0.0154**	0.0798	0.0663	0.0632
Tadano	Hotel 5	0.0918	0.0720	0.0652	**0.0125**	0.0013	0.0024	0.0918	0.0720	0.0652
	Hotel 1	**0.0976**	**0.1055**	**0.0943**	0.0051	0.0012	0.0020	**0.0923**	**0.1022**	**0.0902**
	Hotel 2	**0.0978**	**0.1275**	**0.0986**	**0.0181**	**0.0181**	**0.0181**	**0.0907**	**0.1220**	**0.0931**
	Hotel 3	**0.1165**	**0.1403**	**0.1160**	**0.0246**	**0.0228**	**0.0231**	**0.1140**	**0.1343**	**0.1128**
	Hotel 4	**0.0936**	**0.1134**	**0.0902**	0.0018	0.0022	0.0020	**0.0921**	**0.1127**	**0.0893**
	Hotel 5	**0.1063**	**0.1234**	**0.0995**	0.0029	**0.0042**	**0.0034**	**0.1034**	**0.1176**	**0.0959**

Table 2. ROUGE's results for Annotator nº 2.

Annotator #3										
		ROUGE-1			ROUGE-2			ROUGE-L		
		Recall	Precision	F-measure	Recall	Precision	F-measure	Recall	Precision	F-measure
General Summaries										
K-Medoids	Hotel 1	0.1097	0.1086	0.0942	0.0033	0.0043	0.0034	0.0912	0.0866	0.0767
	Hotel 2	0.1211	0.0916	0.0897	0.0254	0.0213	0.0212	0.1115	0.0804	0.0800
	Hotel 3	**0.0847**	0.0790	0.0681	0.0005	0.0029	0.0009	**0.0700**	0.0697	0.0581
	Hotel 4	**0.0942**	0.0720	**0.0740**	**0.0180**	**0.0161**	**0.0168**	**0.0822**	0.0625	0.0649
	Hotel 5	**0.1251**	0.1062	**0.1076**	**0.0247**	**0.0222**	**0.0225**	**0.1103**	0.0931	**0.0949**
LexRank	Hotel 1	**0.1253**	**0.1601**	**0.1229**	**0.0540**	**0.0518**	**0.0504**	**0.1192**	**0.1561**	**0.1185**
	Hotel 2	**0.1357**	**0.1681**	**0.1345**	**0.0389**	**0.0373**	**0.0372**	**0.1295**	**0.1585**	**0.1276**
	Hotel 3	0.0699	**0.0917**	**0.0692**	**0.0067**	**0.0095**	**0.0074**	0.0647	**0.0864**	**0.0644**
	Hotel 4	0.0729	**0.0811**	0.0680	0.0075	0.0063	0.0064	0.0707	**0.0763**	**0.0650**
	Hotel 5	0.0987	**0.1410**	0.0993	0.0137	0.0180	0.0137	0.0886	**0.1315**	0.0907
MMR	Hotel 1	0.0870	0.0812	0.0755	0.0199	0.0214	0.0205	0.0811	0.0783	0.0719
	Hotel 2	0.0583	0.1165	0.0616	0.0029	0.0126	0.0043	0.0540	0.1057	0.0560
	Hotel 3	0.0707	0.0665	0.0589	0.0033	0.0005	0.0009	0.0586	0.0528	0.0476
	Hotel 4	0.0591	0.0621	0.0533	0.0022	0.001	0.0013	0.0520	0.0546	0.0470
	Hotel 5	0.0568	0.0520	0.0454	0.0064	0.0042	0.0048	0.0477	0.0458	0.0388
Aspect-based Summaries										
Opizer-E	Hotel 1	**0.1046**	0.1176	**0.1016**	**0.0597**	**0.0518**	**0.0533**	**0.1020**	0.1072	**0.0977**
	Hotel 2	0.0691	0.0876	0.0691	**0.0219**	**0.0221**	**0.0202**	0.0636	0.0832	0.0643
	Hotel 3	**0.0518**	0.0551	0.0459	**0.0049**	0.0024	**0.0031**	**0.0489**	0.0521	0.0431
	Hotel 4	0.0375	0.0387	0.0336	0.0020	0.0012	0.0015	0.0349	0.0377	0.0322
	Hotel 5	0.0433	0.0629	0.0462	0.0026	0.0053	0.0034	0.0401	0.0596	0.0431
Tadano	Hotel 1	0.0963	**0.1213**	0.1007	0.0029	0.0051	0.0036	0.0927	**0.1166**	0.0973
	Hotel 2	**0.0909**	**0.0955**	**0.0826**	0.0036	0.0022	0.0027	**0.0883**	**0.0910**	**0.0795**
	Hotel 3	0.0500	**0.0808**	**0.0560**	0.0006	0.0024	0.0010	0.0476	**0.0738**	**0.0527**
	Hotel 4	**0.0742**	**0.0961**	**0.0797**	**0.0193**	**0.0194**	**0.0193**	**0.0734**	**0.0938**	**0.0785**
	Hotel 5	**0.1048**	**0.1297**	**0.1102**	**0.0375**	**0.0391**	**0.0379**	**0.1022**	**0.1260**	**0.1071**

Table 3. ROUGE's results for Annotator nº 3.

Annotator #4										
		ROUGE-1			ROUGE-2			ROUGE-L		
		Recall	Precision	F-measure	Recall	Precision	F-measure	Recall	Precision	F-measure
General Summaries										
K-Medoids	Hotel 1	**0.1581**	**0.1946**	**0.1646**	**0.0721**	**0.0741**	**0.0727**	**0.1421**	**0.1750**	**0.1485**
	Hotel 2	**0.1686**	**0.2106**	**0.1785**	**0.0869**	**0.0921**	**0.0888**	**0.1537**	**0.1948**	**0.1642**
	Hotel 3	**0.1030**	0.1446	**0.1072**	**0.0189**	**0.0258**	**0.0208**	**0.0880**	0.1245	**0.0926**
	Hotel 4	**0.1279**	**0.1494**	**0.1314**	**0.0564**	**0.0575**	**0.0569**	**0.1152**	**0.1351**	**0.1187**
	Hotel 5	**0.1410**	0.1799	**0.1483**	**0.0608**	**0.0638**	**0.0622**	**0.1281**	0.1652	**0.1358**
LexRank	Hotel 1	0.0643	0.1909	0.0910	0.0035	0.0103	0.0052	0.0573	0.1722	0.0815
	Hotel 2	0.0831	0.1818	0.1049	0.0123	0.0288	0.0153	0.0739	0.1600	0.0927
	Hotel 3	0.0695	**0.1460**	0.0876	0.0079	0.0165	0.0103	0.0623	**0.1322**	0.0785
	Hotel 4	0.0550	0.1127	0.0705	0.0065	0.0095	0.0076	0.0516	0.1042	0.0658
	Hotel 5	0.0892	**0.1925**	0.1076	0.0231	0.0257	0.0235	0.0784	**0.1736**	0.0954
MMR	Hotel 1	0.0620	0.0939	0.0693	0.0013	0.0015	0.0014	0.0536	0.0826	0.0598
	Hotel 2	0.0610	0.1501	0.0752	0.0143	0.0433	0.0183	0.0557	0.1438	0.0698
	Hotel 3	0.0712	0.1102	0.0753	0.0063	0.0123	0.0072	0.0554	0.0907	0.0598
	Hotel 4	0.0774	0.1216	0.0819	0.0211	0.0247	0.0218	0.0673	0.1108	0.0726
	Hotel 5	0.0679	0.1141	0.0738	0.0032	0.0018	0.0023	0.0514	0.0932	0.0568
Aspect-based Summaries										
Opizer-E	Hotel 1	**0.0755**	**0.1121**	**0.0703**	**0.0240**	**0.0382**	**0.0259**	**0.0726**	**0.1052**	**0.0670**
	Hotel 2	**0.0706**	0.0724	0.0428	0.0021	0.0076	0.0032	**0.0669**	0.0624	0.0377
	Hotel 3	**0.0773**	0.0751	**0.0570**	0.0004	0.0025	0.0008	**0.0701**	0.0672	**0.0497**
	Hotel 4	0.0237	0.0380	0.0274	0	0	0	0.0218	0.0357	0.0253
	Hotel 5	0.0556	0.0623	0.0435	**0.0026**	0.0008	0.0013	0.0556	0.0623	0.0435
Tadano	Hotel 1	0.0487	0.0783	0.0491	0.0019	0.0083	0.0028	0.0480	0.0738	0.0479
	Hotel 2	0.0591	**0.1252**	**0.0701**	**0.0058**	**0.0131**	**0.0076**	0.0506	**0.1136**	**0.0608**
	Hotel 3	0.0472	**0.0903**	0.0529	**0.0046**	**0.0074**	**0.0055**	0.0416	**0.0795**	0.0461
	Hotel 4	**0.0446**	**0.0718**	**0.0501**	**0.0194**	**0.0208**	**0.0198**	**0.0438**	**0.0708**	**0.0492**
	Hotel 5	**0.0650**	**0.1130**	**0.0700**	0.0014	**0.0032**	**0.0019**	**0.0610**	**0.1058**	**0.0650**

Table 4. ROUGE's results for Annotator nº 4.

6. Conclusion

This work compares some automatic summarization techniques applied to a newly created corpus designed for summarizing hotel opinions. The study evaluates not only the effectiveness of the techniques themselves but also examines how the ROUGE metrics perform when evaluating opinion summaries, which is a relatively unexplored area, particularly in Brazilian Portuguese. As ROUGE is a widely used metric for evaluating automated summaries, it is crucial to analyze its performance on subjective opinionated content, which may include slang, grammatical errors, and other aspects that can impair sentence extraction and evaluation.

Our results pointed out that ROUGE appears to be dependent on the textual structure of the summaries it is evaluating, and also has difficulty evaluating opinative texts, which are naturally short and informal. Those characteristics could be seen on the generally low scores, and in the case where Tadano's summarizer scored consistently higher than Opizer-E. It can be seen that when it comes to general summaries, K-Medoids and LexRank score higher than MMR on the vast majority of cases, pointing out that redundancy is more desirable for opinion summaries, and not topic diversity and coverage.

It was also shown that utilizing ROUGE to evaluate shorter, opinative content is not recommended, And as previously cited, [17] also arrived at a similar conclusion but focusing on abstractive summaries on a myriad of summary configurations and techniques. From both studies, it is possible to conclude that ROUGE has a difficult time evaluating opinative documents of all kinds, and not only those that are extractive or abstractive in nature. This stems from those documents often shorter and informal contents which contrast the type document ROUGE was initially intended to evaluate.

The source code for all techniques and results, as well as our corpus, can be found on Github. They are on separate repositories to facilitate cloning and checking the techniques [3] and the corpus [4] separately from each other.

References

[1] Akhtar, N., Zubair, N., Kumar, A., Ahmad, T.: Aspect based sentiment oriented summarization of hotel reviews. Procedia computer science **115**, 563–571 (2017)

[2] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 335–336 (1998)

[3] Condori, R.E.L.: Sumarização automática de opiniões baseada em aspectos. Ph.D. thesis, Universidade de São Paulo (2014)

[4] Cortez, M.C.A., Mondo, T.S.: Comentários on-line: formação de expectativa e decisão de compra de consumidores hoteleiros. Rosa dos Ventos **10**(1), 119–136 (2018)

[5] Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research **22**, 457–479 (2004)

[6] Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (2006)

[7] Freires Junior, J.H.: Sumopinions: Sumarização automática de opiniões sobre pontos turísticos. Repositório Institucional UFC (2018), `https://repositorio.ufc.br/handle/riufc/52200`

[8] Freitas, C., Motta, E., Milidiú, R., Cesar, J.: Vampiro que brilha... rá! desafios na anotação de opinião em um corpus de resenhas de livros. In: XI Encontro de Linguística de Corpus (ELC 2012) (2012)

[3] https://github.com/AShiningRay/ExtractiveSum-Comparison

[4] https://github.com/AShiningRay/Corpus-ExtractiveSum-Comparison

[9] Hartmann, N.and Avanço, L., Balage Filho, P.P., Duran, M.S., Nunes, M.D.G.V., Pardo, T.A.S., Aluísio, S.M., et al.: A large corpus of product reviews in portuguese: Tackling out-of-vocabulary words. In: LREC. pp. 3865–3871 (2014)

[10] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177 (2004)

[11] Hu, Y.H., Chen, Y.L., Chou, H.L.: Opinion mining from online hotel reviews–a text summarization approach. Information Processing & Management **53**(2), 436–449 (2017)

[12] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)

[13] Mani, I.: Automatic summarization, vol. 3. John Benjamins Publishing (2001)

[14] Nenkova, A., McKeown, K.: Automatic summarization. Now Publishers Inc (2011)

[15] Raut, V.B., Londhe, D.: Opinion mining and summarization of hotel reviews. In: 2014 International Conference on Computational Intelligence and Communication Networks. pp. 556–559. IEEE (2014)

[16] Tadano, R., Shimada, K., Endo, T.: Multi-aspects review summarization based on identification of important opinions and their similarity. In: Proceedings of the 24th Pacific Asia conference on language, information and computation. pp. 685–692 (2010)

[17] Tay, W., Joshi, A., Zhang, X.J., Karimi, S., Wan, S.: Red-faced rouge: Examining the suitability of rouge for opinion summary evaluation. In: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association. pp. 52–60 (2019)