UNIVERSIDADE FEDERAL DE ITAJUBÁ - UNIFEI
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO/POSCOMP
ENGENHARIA DA COMPUTAÇÃO

Sumarizador de Avaliações Usando TextRank e Modelagem de Tópicos.

Fernando Hideki Takenaka

Itajubá, 25 de outubro de 2023

UNIVERSIDADE FEDERAL DE ITAJUBÁ - UNIFEI
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO/POSCOMP
ENGENHARIA DA COMPUTAÇÃO

Fernando Hideki Takenaka

Sumarizador de Avaliações Usando TextRank e Modelagem de Tópicos.

Dissertação submetida ao Programa de Pós-Graduação em Ciência E Tecnologia Da Computação como parte dos requisitos para obtenção do Título de Mestre em Ciência E Tecnologia Da Computação.

Área de Concentração: Inteligência Artificial

Orientador: Prof. Dr. Laércio Augusto Baldochi Junior

25 de outubro de 2023

Itajubá

UNIVERSIDADE FEDERAL DE ITAJUBÁ - UNIFEI
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E TECNOLOGIA DA COMPUTAÇÃO/POSCOMP
ENGENHARIA DA COMPUTAÇÃO

Sumarizador de Avaliações Usando TextRank e Modelagem de Tópicos.

Fernando Hideki Takenaka

Banca Examinadora:
Prof. Dra. Isabela Neves Drummond
Prof. Dr. Rafael Duarte Coelho dos Santos

Itajubá

Fernando Hideki Takenaka
Sumarizador de Avaliações Usando TextRank e Modelagem de Tópicos/ Fernando Hideki Takenaka. – Itajubá, 25 de outubro de 2023-
67 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Laércio Augusto Baldochi Junior

Dissertação (Mestrado)
Universidade Federal de Itajubá - UNIFEI
Programa de pós-graduação em Ciência E Tecnologia Da Computação/POSCOMP
Engenharia da Computação, 25 de outubro de 2023.

1. PROCESSAMENTO DE LINGUAGEM NATURA. 2. SUMARIZAÇÃO. I. Laércio Augusto Baldochi Junior. II. Universidade Federal de Itajubá.

CDU 07:181:009.3

Fernando Hideki Takenaka

Sumarizador de Avaliações Usando TextRank e Modelagem de Tópicos

Dissertação submetida ao Programa de Pós-Graduação em Ciência E Tecnologia Da Computação como parte dos requisitos para obtenção do Título de Mestre em Ciência E Tecnologia Da Computação.

Prof. Dr. Laércio Augusto Baldochi Junior
Orientador

Prof. Dra. Isabela Neves Drummond

Prof. Dr. Rafael Duarte Coelho dos Santos

Itajubá
25 de outubro de 2023

Agradecimentos

Agradeço a Deus por me presentar com esta oportunidade, agradeço à minha família por me apoiar durante esta fase da minha vida e ao meu orientador pelo direcionamento e instrução durante este projeto.

"Toda decisão acertada é proveniente de experiência. E toda experiência é proveniente de uma decisão não acertada."

(Albert Einstein)

Resumo

Na última década a Internet mudou o modo como as pessoas trabalham, fazem compras e se socializam. Essas mudanças resultaram em um aumento no Conteúdo Gerado pelos Usuários (CGU) como, por exemplo: avaliações, notas, artigos e vídeos. Os CGUs possuem informações relevantes para a tomada de decisão, especialmente no que se refere à aquisição de bens e serviços. Entretanto, o grande volume e dispersão deste conteúdo torna difícil a obtenção de informações relevantes. Neste contexto, a sumarização de textos é apresentada como um modo de tornar este conteúdo mais acessível às pessoas.

Um dado sumário A pode ser considerado melhor que um outro sumário B se o primeiro for mais curto que o segundo com o mesmo conteúdo, ou quando mesmo sendo mais longo, possui mais informações relevantes. Analisando a literatura disponível, foi constatado que é possível produzir sumários de melhor qualidade do que aqueles que correspondem ao estado da arte em sumarização de textos. Neste trabalho, apresentamos um sumarizador automático multilingual que combina e expande os algoritmos Latent Dirichlet Allocation (LDA) e TextRank. Em comparação com o estado da arte, este trabalho gerou sumários melhores em termos de tamanho e conteúdo.

Palavras-chaves: PROCESSAMENTO DE LINGUAGEM NATURAL, TEXTRANK, MODELAGEM DE TOPICOS, SUMARIZAÇÃO.

Abstract

Over the past decade, the Internet has changed the way people work, shop and socialize. Those changes resulted in the increase of User Generated Content (UGC) such as: ratings, reviews, wikis, and videos. UCG contains relevant information for decision-making, especially with regard to the acquisition of goods and services. However, the large volume and dispersion of this content makes it difficult to obtain relevant information. Text summarization appears as a way to make this content more accessible to people.

A summary A can be considered better than another B when A is shorter than B while maintaining the same content relevance, or when A, despite being longer, presents more relevant content. Analyzing the literature, we observed that it is possible to produce better quality summaries than those produced by algorithms that correspond to the state of the art in text summarization. We present a multilingual automatic text summarizer that combines and extends the algorithms Latent Dirichlet Allocation (LDA) and TextRank. Our approach, when compared to the state of the art, generates better text summaries in terms of size and content relevance.

Key-words: NATURAL LANGUAGE PROCESSING. TEXTRANK. TOPIC MODELING. SUMMARIZATION

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CGU	*Conteúdo Gerado pelos Usuários*	16
ETTM	*Enriched Two-tiered topic model*	57
IA	*Inteligência Artificial*	19
LDA	*Latent Dirichlet Allocation*	10
LSA	*Latent Semantic Analysis*	26
NLTK	*Natural Language Toolkit*	26
NMF	*Non Negative Matrix Factorization*	26
PLN	*Processamento da Linguagem Natural*	16
PLSA	*Probabilistic Latent Semantic Analysis*	26
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*	54
STA	*Sumarização Textual Automática*	10
TF-IDF	*Term Frequency Inverse Document Frequency*	57
TTM	*Two-tiered topic model*	57

Sumário

1 Introdução

Durante a última década, a tecnologia da Internet alterou consideravelmente o cotidiano das pessoas. É cada dia mais comum que produtos e serviços sejam oferecidos por meios eletrônicos e plataformas Web. Além disso, muitas empresas começaram a ofertar seus produtos pela Internet, o que permite alcançar um número maior de pessoas. Essas mudanças deram origem à Web 2.0, que permitiu um aumento expressivo na quantidade de CGU (*Conteúdo Gerado pelos Usuários*) na Internet (Ramadhan *et. al.*, 2020) [1].

Os CGUs são aqueles conteúdos originários de pessoas comuns que, voluntariamente, contribuem com dados, informações ou mídias que são apresentados a outros com conteúdo útil ou com fins de entretenimento (Krumm *et. al.*, 2008) [2]. Uma pesquisa realizada por Yan *et. al.* em 2016 [3], aponta que 91% dos participantes responderam que usam avaliações online, blogs e outras formas de CGU na decisão de adquirir um produto ou serviço, e, 46% dos participantes disseram que tais avaliações influenciam em suas decisões.

Isso demonstra que as avaliações são um conteúdo importante para os negócios e para os usuários finais. Porém, quanto maior o número de avaliações que o usuário tiver que analisar, mais tempo e recursos serão necessários para essa tarefa. Portanto, um meio de condensar as informações relevantes em um curto texto é necessário, o que pode ser alcançado com um sumário.

O dicionário de Oxford (2023) [4] define um sumário como: “um texto curto que contém apenas os principais pontos de algo, sem incluir detalhes”. Essa pode ser a solução para o problema da leitura e análise das avaliações dos usuários, esse trabalho pode ser executado por outra pessoa ou pelo próprio usuário. Entretanto, com o constante crescimento no número de avaliações, os recursos humanos e o tempo necessários para executar a sumarização das avaliações também aumentam. Eventualmente, o custo se torna muito alto, inviabilizando a execução desta tarefa. Neste contexto, o uso dos computadores oferece um meio de automatizar esta tarefa.

A principal solução apresentada pela ciência da computação para resolver o problema da sumarização é o uso do PLN (*Processamento da Linguagem Natural*), uma área de pesquisa que busca permitir que computadores consigam extrair significado de textos e documentos em linguagem humana.

Desde a definição de STA, em 1958, por Hans Peter Luhn [5], várias abordagens foram propostas para o problema da sumarização de textos. Um dos maiores obstáculos nesta área de pesquisa é o fato de que os computadores não possuem conhecimento humano, o que inclui a linguagem e o processamento das palavras que dão sentido a elas. A

ausência dessas características fazem desta tarefa algo desafiador e complexo.

Uma das abordagens mais populares para a geração de sumários é o TextRank, uma variação do PageRank criado por Sergey Brin e Lawrence Page [6]. A abordagem do TextRank permite que ela seja bastante abrangente, tornando possível aplicá-lo em uma ampla gama de áreas que atuem com textos. Porém, as *stopwords*, palavras que são comuns em qualquer idioma (Zaware *et. al.*, 2021) [7], tornam a escolha das frases problemática, pois o TextRank não é capaz de atribuir valores apropriados para cada palavra, impactando na qualidade do sumário.

Uma proposta para resolver o problema da escolha de palavras para o Textrank consiste em substituir os termos genéricos por termos que possuam relevância ao tópico abordado, ou desejado, no documento (ou conjunto de documentos). Neste contexto, o Latent Dirichlet Allocation (LDA), uma técnica de modelagem de tópicos, oferece um meio de realizar a seleção de palavras-chave para este fim. A ideia básica do LDA é que os documentos são representados como misturas aleatórias sobre tópicos latentes, onde um tópico é caracterizado por uma distribuição sobre palavras (Jelodar *et. al.*, 2019)[8].

O presente trabalho propõe uma abordagem para o problema do STA aplicando uma combinação do TextRank com o LDA para criar um programa multilingual capaz de gerar sumários focados em avaliações.

1.1 Objetivos

O objetivo geral deste trabalho é a criação de um sistema capaz de gerar sumários textuais de forma automática e que aceite fontes escritas em múltiplos idiomas. Para alcançar este objetivo, é necessário definir as técnicas que serão utilizadas e qual ferramenta para identificação de idioma e tradução das frases será utilizada.

A fim de se alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

1. Desenvolver um STA capaz de aceitar múltiplos arquivos, escritos em diversos idiomas, como fonte para sumarização;
2. Validar o STA desenvolvido em (1) com bases disponíveis publicamente;
3. Revisar o estado da arte dos STAs extrativos;
4. Comparar o STA desenvolvido em (1) com outros trabalhos disponíveis na literatura.

1.2 Metodologia

Para alcançar o primeiro objetivo específico foi desenvolvida uma funcionalidade para identificar o idioma das frases extraídas e, caso necessário, traduzi-las. Além disso, foi necessário mesclar a técnica LDA com a técnica TextRank, estendendo suas funcionalidades de modo que um complemente o outro na tarefa de gerar sumários.

Uma vez desenvolvido o sistema, foram escolhidas duas bases de dados para realizar dois estudo de casos. Os resultados obtidos mostraram a eficiência do sistema proposto em relação às tecnologias usadas na composição da abordagem proposta. Desta forma, o objetivo específico 3 foi atingido.

Para alcançar o quarto objetivo específico, foi realizada uma busca sobre o estado-da-arte dos STAs extrativos usando a ferramenta de busca do Google™ para pesquisa acadêmica, o Google Scholar.

Por fim, as características da abordagem proposta foram comparadas com as técnicas estado-da-arte encontradas no objetivo 4, apontando as vantagens e desvantagems de cada uma. Os resultados mostram que a abordagem proposta possui vários pontos positivos quando comparado com o estado-da-arte. Deste modo, o objetivo específico 5 foi atingido.

1.3 Organização do trabalho

Este trabalho encontra-se dividido em 6 Capítulos. O Capítulo 2 apresenta uma revisão teórica dos conceitos usados.

O Capítulo 3 discute o TextRank, o LDA e a abordagem proposta, iniciando pela explicação do funcionamento e estrutura do TextRank, seguindo para o LDA e por fim para o sistema proposto.

No Capítulo 4 são apresentados os experimentos realizados para validação do sistema proposto e os resultados obtidos em cada experimento são analisados.

No Capítulo 5 são discutidos os trabalhos relacionados, apontando as vantagens e desvantagens de cada abordagem em comparação com o sistema desenvolvido.

Por fim, no Capítulo 6 são apresentadas as conclusões sobre a dissertação e os trabalhos futuros.

2 Revisão teórica

Neste Capítulo são explicados em detalhes os conceitos usados neste trabalho e como ele foram aplicados. Este Capítulo possui três seções: na primeira é explicado o que é a sumarização textual automática e suas classificações, na segunda seção são explicados os algoritmos TextRank e PageRank. Na última seção é explicado o que é a modelagem de tópicos e algumas de suas técnicas mais comuns.

2.1 Sumarização Textual Automática

A STA é uma das tarefas mais desafiadoras do PLN e da IA (*Inteligência Artificial*) em geral. Sua pesquisa iniciou-se em 1958 com o trabalho de Lunh [5] onde um sistema capaz de gerar sumários automatizados de revistas e artigos técnicos foi apresentado. Desde então, vários trabalhos propuseram diferentes soluções para resolver este problema.

De acordo com El-Kassas *et. al.* , 2021 [9], os STAs podem ser classificados em diversas categorias, dependendo das características que são consideradas. Dentre as várias existentes, as divisões mais aceitas são apresentadas a seguir.

2.1.1 Com base na quantidade de entradas

Nesta divisão, Chatterjee *et. al.* (2012) [10] divide os STAs em duas subcategorias: individual ou múltiplo. Como o nome indica, na subcategoria individual cada sumário é gerado a partir de um único documento, não sendo permitida a inserção de mais documentos para a geração deste sumário, enquanto na subcategoria múltiplo, uma coleção de documentos são requisitados para se extrair o sumário (Ahmad, 2017) [11]. Gupta e Siddiqui [12](2012), apontam ainda que sumarizadores da categoría múltiplo são mais complexos que os da categoria individual e que possuem mais obstáculos como por exemplo: redundância, abrangência, características temporais, taxa de compressão etc. Tais pontos são reforçados por outros trabalhos ([13], [14], [15]), indicando que, apesar da quantidade significativa de avanços, ainda há muito a ser explorado.

2.1.2 Com base no tipo de extração

El-Kassas *et. al.* (2021) [9] considera que esta divisão pode ser subcategorizada em 3(três) subcategorias: Extrativa, Abstrativa e Híbrida. Neelima Bhatia e Arunima Jaiswal(2016)[16] definem como sumarização extrativa aquela que gera o sumário por meio da extração de frases presentes no(s) documento(s) original(is), também definem

como sumarização abstrativa aquela na qual alguns dos pontos não se encontram no(s) documento(s) original(is).

Tandel *et. al.* [17] (2016), afirmam que soluções extrativas são mais simples e rápidas que as soluções abstrativas. Por outro lado, Moratanch e Chitrakala [18](2017) argumentam que as soluções extrativas possuem diversas falhas, incluindo a ausência de semântica, coesão e a presença de redundância. Em contra partida, as soluções abstrativas em troca de gerarem sumários melhores que se aproximam daqueles escritos manualmente, pecam por palavras fora do vocabulário, repetição e saliência, problemas notórios a serem resolvidos (Gui *et. al.*, 2019)[19]. Por fim, as soluções híbridas tentam unir as soluções anteriores para criar um sistema com uma performace geral melhor (Wang *et. al.*, 2017)[20] ao custo de gerar sumários com qualidade inferior ao abstrativo devido ao fato da função abstrativa deste sistema depender da parte extraída pela parte extrativa.

2.1.3 Com base na natureza da resposta

Gong and Liu[21] (2001) separam esta divisão em 2(duas) subcategorias: genérica e baseada em consulta, explicando que sumarizadores genéricos provêm um olhar superficial do conteúdo dos documentos sem se importar com palavras-chave, enquanto os sumarizadores baseados em consulta apresentam os conteúdos que são mais próximos e relacionados a aqueles usados na consulta inicial.

2.1.4 Com base no idioma

Esta divisão possui 3 classificações, de acordo com Gambhir e Gupta (2016)[22]: mono-idiomático, multi-idiomas e cruzado. A categoria mono-idiomática se caracteriza pelo fato do idioma dos textos-fonte e do sumário gerado serem iguais, por exemplo, português brasileiro. Os sumarizadores classificados como multi-idiomas podem possuir texto-fontes com diversos idiomas (exemplo: espanhol e inglês), entretanto o sumário gerado deve, obrigatoriamente ser escrito em um dos idiomas usados pelos textos-fonte. Por fim, os sumários de idiomas cruzados geram sumários em um idioma que não está entre aqueles usados nos textos-fonte, por exemplo: Os textos-fonte estão escritos em chinês e o sumário em inglês.(Wang *et. al.*, 2022) [23].

2.1.5 Com base no algoritmo de sumarização

Nesta divisão os STA são distribuídos em duas categorias: supervisionados e não-supervisionados, de acordo com o tipo de aprendizagem. Na categoria não-supervisionada o resultado é desconhecido pelo usuário, apenas os dados de entrada estão disponíveis, enquanto que, na categoria supervisionada, existem variáveis de entrada, variáveis de saída e o algoritmo é treinado de modo que ele aprenda a função que mapeará as entradas e saídas

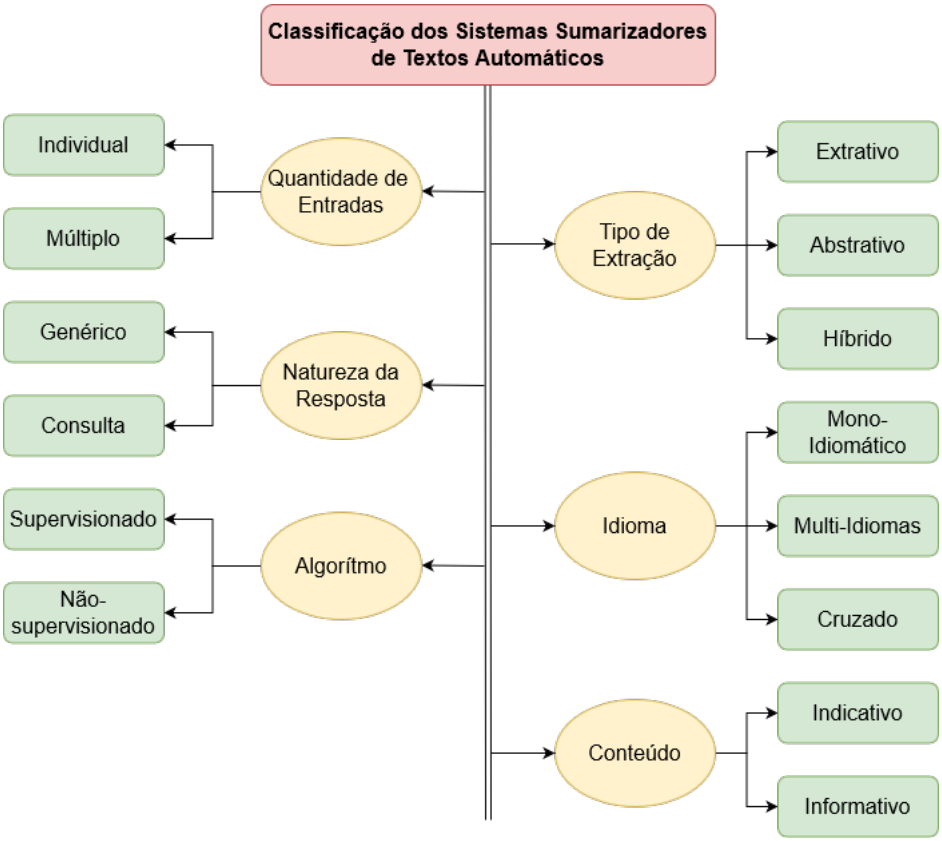

Figura 1 – Classificação dos STA

(Charitha *et. al.*, 2018) [24]. Os sistemas que usam o método supervisionado apresentam excelentes resultados para a sumarização automática de textos, em contra partida, exige que se possua uma base de dados categorizados abundande bem como a disponibilidade de mão-de-obra humana suficiente para operar o sistema, o que o torna caro e difícil de criar. No caso dos sistemas não-supervisionados, são exploradas as características estatísticas e linguísticas para a geração do sumário a partir do texto-fonte (Mohd e Jan e Shah)[25].

2.1.6 Com base no conteúdo

De acordo com Bhat *et. al.*[26] (2018), esta divisão pode ser categorizada em: indicativa e informativa. Hovy e Lin (1998) [27] informam que um sumarizador indicativo provê ou fornece um sumário com apenas indicativos dos temas e assuntos abordados no documento original, sem incluir seu conteúdo. Já os sumarizadores informativos refletem, em parte, o conteúdo original, permitindo que alguém descreva partes do que existe no texto-fonte.

Com base nesses critérios estabelecidos nas seções anteriores, é possível usar a Figura 1 para ilustrar as várias divisões e classificações dos STA.

O presente trabalho implementa um sumarizador de texto automático classificado como múltiplo, extrativo, baseado em consulta, multi-idioma, não-supervisionado e indi-

cativo, de acordo com as definições apresentadas nesta seção. Na Seção 2.2 é apresentado o algoritmo TextRank e um pouco de sua origem.

2.2 TextRank

A fim de explicar o algoritmo TextRank se faz necessário explicar o algoritmo em que ele se baseia: o PageRank. O PageRank é um algoritmo criado por Sergey Brin e Lawrence Page no ano de 1998 [6] e é considerado uma medida objetiva da importância da citação correspondente à ideia subjetiva de importância das pessoas (Kim e Lee, 2002)[28], e apesar do PageRank ter sido originalmente projetado para grafos Web, seus conceitos e definições funcionam bem para qualquer grafo (Chung, 2014)[29].

De acordo com Ishii e Suzuki [30] (2018), o PageRank tem recebido grande interesse, especialmente quando o contexto se refere a redes complexas, como na medição de centralidade efetiva. Esta afirmação é sustentada pela quantidade de artigos e trabalhos que utilizam este algoritmo e pela quantidade de varições do mesmo que tais obras científicas apresentam, incluindo, mas não se limitando ao: TextRank[31], weighted PageRank[32] e ArticleRank[33].

2.2.1 O PageRank

O algoritmo PageRank pode ser dividido em duas principais partes, a construção do grafo de citações (Cheang *et. al.*,2014) [34], seguido pelo cálculo da qualidade de cada página utilizando o princípio base do PageRank em que se assume que a classificação das páginas web é decidida com base em suas arestas de saída, arestas que saem da sua página para outras páginas e arestas de entrada, arestas essas que fazem o percurso contrário, saindo das outras páginas em direção à sua (Patel e Patel, 2015) [35].

Sen e Chaudhary (2017) [36] descrevem a fórmula usada pelo algoritmo do PageRank para classificar as páginas na Equação 2.1:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v))}{N_v} \tag{2.1}$$

Onde: u representa a página

$PR(u)$ e $PR(v)$ representam as classificações as páginas u e v, respectivamente

$B(u)$ é o conjunto de páginas que apontam para u

N_v se refere ao número de arestas saindo da página v

c é o fator de normalização

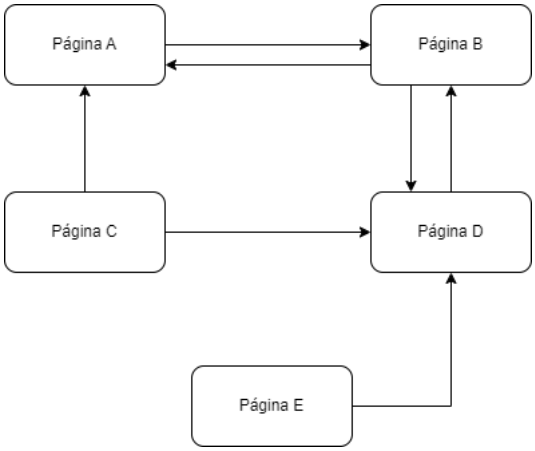

Figura 2 – Exemplo de grafo para o PageRank com arestas de entrada e saída

Considerando a Equação 2.1, é possível inferir que se uma página possui arestas importantes em direção a ela, então as arestas que saem dela em diração às outras também devem ser consideradas como importantes (Chowdhary *et. al.*, 2019).[37]

A fim de ilustrar como este algoritmo funciona, é usado como exemplo um grafo com 5(cinco) páginas distribuídas e ligadas de acordo com a Figura 2:

Aplicando a lógica do PageRank na Figura 2, é possível calcular o valor de cada página resolvendo as Equações: 2.2, 2.3, 2.4, 2.5, 2.6:

$$PR(A) = c * (\frac{PR(B)}{N_B} + \frac{PR(C)}{N_C}) \tag{2.2}$$

$$PR(B) = c * (\frac{PR(A)}{N_A} + \frac{PR(D)}{N_D}) \tag{2.3}$$

$$PR(C) = c * 0 \tag{2.4}$$

$$PR(D) = c(\frac{PR(B)}{N_B} + \frac{PR(C)}{N_C} + \frac{PR(E)}{N_E}) \tag{2.5}$$

$$PR(E) = c * 0 \tag{2.6}$$

Para o valor inicial das páginas, é utilizado o valor proposto por Sergey Brin e Lawrence Page. Tal valor deve ser entre 0 e 1 e deve ser o mesmo para todas as páginas [6], portanto, com 5 páginas, cada página possui o valor inicial de 0.20. Resolvendo as equações com o valor proposto, temos as Equações 2.7, 2.8, 2.9, 2.10, 2.11:

$$PR(A) = c * (\frac{0.20}{2} + \frac{0.20}{2}) = 0.20 * c \tag{2.7}$$

$$PR(B) = c * (\frac{0.20}{1} + \frac{0.20}{1}) = 0.40 * c \tag{2.8}$$

$$PR(C) = c * 0 = 0 \tag{2.9}$$

$$PR(D) = c(\frac{0.20}{2} + \frac{0.20}{2} + \frac{0.20}{1}) = 0.40 * c \tag{2.10}$$

$$PR(E) = c * 0 = 0 \tag{2.11}$$

Como resultado, o PageRank classifica as páginas B e D como sendo as mais relevantes, a página A como de alguma relevância e as páginas C e E como as menos relevantes. Uma vez explicado este conceito, é possível explicar o funcionamento do algoritmo TextRank.

2.2.2 O TextRank

O TextRank é um algoritmo criado como uma expansão do PageRank e foi proposto por Mihalcea e Tarau em 2004 [38]. No contexto do TextRank, as páginas são substituídas por palavras que serão equivalentes aos nós do grafo (Li e Zhao, 2016) [39]. Complementando a informação anterior, Zhang *et. al.* (2018) [40], dizem que uma aresta é colocada entre duas palavras se elas aparecerem dentro do contexto uma da outra, o que é definido como janela de palavras.

A fim de se construir este grafo, Petasis e Karkaletsis (2016)[41] dizem que as seguinte etapas devem ser seguidas:

- O texto do documento deve ser *tokenizado* em palavras e frases. Em outras palavras, o texto deve ser separado em palavras ou frases únicas.
- O texto deve então ser convertido em um grafo no qual cada vértice é representado por uma frase ou palavra.
- Conexões ou arestas entre as frases ou palavras são estabelecidas, baseado em uma relação de similaridade. O cálculo do valor de "similaridade" entre dois vértices conectados é considerado o peso da aresta.
- O algoritmo de classificação é aplicado ao grafo, atribuindo a cada vértice uma pontuação.
- As frases ou palavras são ordenadas em ordem decrescente de sua pontuação e aquelas melhores classificadas, ou seja com a menor pontuação, são selecionadas para serem incluídas no sumário.

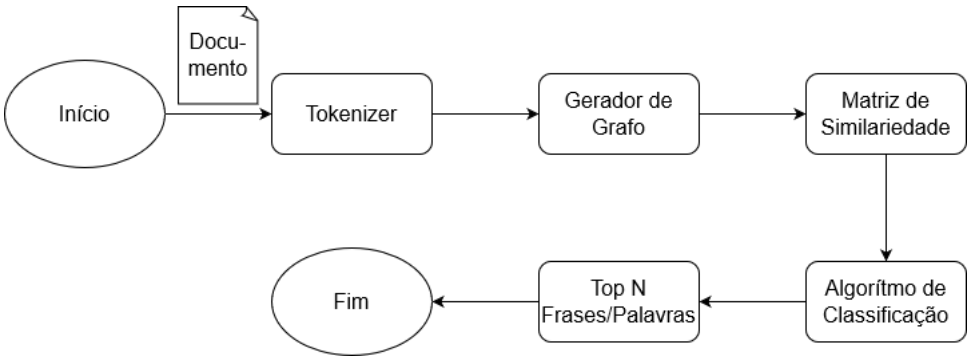

Figura 3 – O algoritmo TextRank

Em relação ao TextRank, a similaridade é definida por Mihalcea and Tarau (2004) [38] de acordo com a Equação 2.12:

$$Similarity(S_i, S_j) = \frac{|W_k|W_k \epsilon S_i \& W_k \epsilon S_j|}{log(|S_i|) + log(|S_j|)} \tag{2.12}$$

Onde:

Cada frase S_i é representada por um conjunto de N_i de palavras que aparecem na frase: $S_i = W_1^i, W_2^i, ..., W_{N_i}^i$.

S_i *e* S_j correspondem às frases.

W_k correspondem às palavras.

Porém, Fakhrezi *et. al.* (2021) [42], propõem um método diferente para se avaliar a similaridade, nomeado cosseno de similaridade. O cosseno de similaridade computa o quão similar 2 dados vetores são por meio da relação direta do número de palavras que elas possuem em comum. Esta equação é descrita pela Equação:

$$Cos(X, Y) = \frac{X^t.Y}{\sqrt{X^t.X} \times \sqrt{Y^t.Y}}, \ \ with \ \ X^t.Y = \sum_{n}^{i=1} x_i y_i \tag{2.13}$$

Em outras palavras, a Equação 2.13 atribuirá uma pontuação mais alta, quanto mais palavras idênticas forem encontradas em ambos os textos. O valor do cosseno de similaridade é muito útil porque permite a remoção de frases que são muito similares do sumário, possibilitando outras frases com valores menores de similaridade, e portanto diferente das demais, sejam escolhidas.

O processo utilizado pelo TextRank para encontrar as frases, palavras ou documentos mais relevantes é ilustrado na Figura 3

Por fim, a literatura registra uma grande quantidade de obras aplicando, monitorando e comparando o TextRank com suas variantes e customizações, mas ainda há muitas possibilidades para serem exploradas. A fim de se desenvolver o sistema proposto

no presente trabalho, é utilizada uma versão modificada do TextRank para separar e classificar as avaliações em ordem de relevância, utilizando os tópicos obtidos pelo LDA como parâmetro.

2.3 Modelagem de tópico

A fim de explicar o termo modelagem de tópicos, se faz necessário uma breve explicação sobre o Processamento dd Linguagem Natural(PLN). O PLN é definido por Chowdhary[43] como um conjunto de técnicas computacionais para a automatização da análise e representação da linguagem humana.

Em 2017, Sun *et. al.*[44] apresentam em seu trabalho várias técnicas de PLN para o processamento de textos e as ferramentas disponíveis para esta tarefa, como por exemplo: *Natural Language Toolkit* (NLTK), *OpenNLP*, *Gensim* e *The Language Technology Platform*.

Seguindo para 2019, Jelodar *et. al.*[8] explica que a modelagem de tópicos não é uma técnica, mas sim um grupo de poderosas técnicas inteligentes aplicadas amplamente no processamento da linguagem natural para a descoberta de tópicos e mineração de semântica em documentos não-ordenados. Algumas das técnicas mais comuns são: LDA, LSA (*Latent Semantic Analysis*), NMF (*Non Negative Matrix Factorization*) e PLSA (*Probabilistic Latent Semantic Analysis*).

2.3.1 Processamento de Linguagem Natural

Chopra *et. al.*(2013)[45] afirmam que a história do PLN se inicia em 1950 com Alan Turing por meio da publicação de seu trabalho chamado *Machine and Intelligence* (Máquinas e inteligência, em tradução livre) na qual anunciou o *imitation game* (jogo da imitação, em tradução livre), hoje conhecido como o "Teste de Turing".

O termo PLN é definido por Chowdhary, 2020,[43] como uma área de pesquisa acadêmica que compreende uma variedade de técnicas computacionais para representação e análise automática de linguagens humanas. Wu *et. al.*, 2022, [46] complementa explicando que o PLN é um subconjunto da IA, cuja relação entre ambas é de uma via de mão dupla, onde os avanços da IA impactam no desenvolvimento do PLN e vice-versa.

Dentre as diversas ferramentas de software disponíveis para se trabalhar com o PLN, Lauriola, Lavelli e Aiolli, 2022,[47] listam e fornecem uma breve explicação delas. Algumas delas são:

- NLTK (*Natural Language Toolkit*): De acordo com a descrição original, encontrada

na plataforma NLTK [1] [48], NLTK é uma plataforma para auxiliar o desenvolvimento de programas em Python, focado na atuação com dados com linguagem humana. Possui interfaces de fácil uso com mais de 50 diferentes recursos, como o WordNet, mais uma ampla gama de bibliotecas para processamento de textos para classificação, tokenização, lematização, marcação, entre outros, além de um fórum de discussão ativo.

- Gensim: De acordo com a descrição original, encontrada na plataforma Gensim [2] [49], Gensim é uma biblioteca Python para a modelagem de tópicos, indexação e recuperação por similaridade e outras funcionalidades. Focado em PLN e Recuperação de informações, a biblioteca possui implementações eficientes de diversos algoritmos populares, por exemplo: LSA, LDA, Random Projections (RP) e Hierarchical Dirichlet Process (HDP).

- SpaCy: Uma biblioteca Python para executar tarefas relacionadas ao PLN. Esta biblioteca foi concebida especificamente para criar sistemas industriais complexos, e opera sem complicações com TensorFlow, PyTorch, scikit-learn, Gensim e o restante do ecossistema IA do Python. Inclui diversas funcionalidades como, por exemplo, a tokenização, segmentação de frases e agrupamento dependente[50].

Uma vez definido e explicado o que é o PLN e apresentada algumas ferramentas, é possível definir e explicar o que é a modelagem de tópicos.

2.3.2 Técnicas de Modelagem de Tópicos

A origem da modelagem de tópicos pode ter datada de 1990 por meio do trabalho de Deerwester *et. al.*, intitulado *Indexing by latent semantic analysis*[51], na qual foi desenvolvido um modelo chamado de *Latent Semantic Indexing*.

A modelagem de tópicos é um termo criado em 2001 por Blei *et. al.*, quando estes propuseram o método LDA [52]. Churchill e Sing, 2021[53], dividem a história da modelagem de tópicos em fases: Modelagem de tópicos iniciais (1990-2006), Modelagem de tópicos temporais e online (2006-2011) e Modelagem de tópicos moderna (2011-Presente). Cada fase trouxe um avanço significativo para a área, de acordo com Churchill e Sing, 2021[53] as contribuições foram as seguintes: a primeira trouxe a criação da área de estudo, a segunda permitiu que se trabalhasse com modelagens dinâmicas e a mais recente com a criação do Word2Vec e outros modelos de incorporação de palavras, bem como a incorporação de novas técnicas para o processamento de linguagem natural nas modelagens de tópicos.

[1] https://www.nltk.org/
[2] https://radimrehurek.com/gensim/intro.html

Dentre os diversos métodos usados para a modelagem de tópicos existentes, os mais conhecidos são:

2.3.2.1 Latent Dirichlet Allocation

O LDA é um modelo de gerador probabilístico de um *corpus*. A ideia base deste modelo é que os documentos podem ser representados por um conjunto aleatório de tópicos latentes, no qual o tópico é caracterizado pela sua distribuição sobre as palavras (Mustakim *et. al.*, 2021)[54].

Sua primeira aparição, que também definiu o termo modelagem de tópicos, foi em 2001 por Blei *et. al.* [52], quando os autores propuseram o método. Desde então o método tornou-se popular e diversos trabalhos propuseram mudanças ao algoritmo original para atender diferentes fins, como, por exemplo: Author-Topic model[55], Dynamic Topic Model[56] e Labeled LDA [57].

2.3.2.2 Latent Semantic Analysis

A ideia fundamental da técnica do LSA é que o significado de cada passagem no texto (documento) é relacionado a padrões de presença ou ausência de certas palavras, enquanto a coleção de documentos (corpus) é modelado como um sistema de equações simultâneas que podem determinar a similaridade do significado das palavras e documentos uma das outras (Evangelopoulos *et. al.*, 2012) [58].

Sua primeira aplicação foi no trabalho de Deerwester *et. al.* [51], os mesmos autores que iniciaram a modelagem de tópicos, que inclusive gerou uma patente para este método (ver nota[3], atualmente expirado) este modelo é chamado de *Latent Semantic Indexing*. Mesmo sendo uma das técnicas mais antigas na modelagem de tópicos, muitos trabalhos ainda são realizados com esta técnica, por exemplo: o DeleSmell [59], engenharia aeroespacial [60] e avaliação de redações ([61], [62]).

2.3.2.3 Non-Negative Matrix Factorization

O NMF é um método novo para a fatorização de matrizes, no qual permite lidar com dados em larga escala. Proposto inicialmente por Lee, em 1999,[63] como um método para fatoração de matrizes que permite trabalhar com dados em larga escala. Ganhou destaque entre os pesquisadores de todo o mundo por introduzir uma restrição de não-negativo que tornou sua aplicação mais ampla e com melhor interpretabilidade (Gan *et. al.*, 2021)[64].

Como exemplos de aplicações deste método Berry*et. al.* (2007) exploram a mineração de texto e análise espectral de dados[65] e Zhang (2012), explora a aplicação deste

[3] http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=4839853

modelo e suas variações para a mineração textual, processamento de imagens, bioinformática, entre outros[66].

2.3.2.4 Probabilistic Latent Semantic Analysis

Esta é uma técnica para redução de dimensionalidade em mineração de texto com base no *bag of words* (saco de palavras, em tradução literal) para detecção semântica da coocorrência de termos utilizando um *framework* probabilístico em um corpus[67].

Sua primeira aplicação foi no trabalho de Hofmann, mesmo autor da técnica, onde propos o modelo Aspect [68], que assume que cada palavra é gerada a partir de um único tópico e que diferentes palavras de um documento podem ser geradas a partir de diferentes tópicos. Cada documento é então representado como uma lista de proporções de cada mistura de componentes, o transformando em uma distribuição de problabilidade com um conjunto fixo de tópicos.

Para este trabalho, foi escolhida como ferramenta a biblioteca Gensim, disponível em Python, para implementar o LDA. Esta técnica para a modelagem de tópicos considera cada documento como uma coleção de tópicos em uma certa proporção e que cada tópico corresponde a uma coleção de palavras-chaves, novamente, em uma certa proporção.

3 Solução proposta

Com o objetivo de criar um sumarizador automático de textos multilingual, este trabalho combina o TextRank com a modelagem de tópicos, mais especificamente o LDA.

Neste trabalho foram utilizadas 2 bases de dados disponíveis em domínio público como fonte de dados para uma das experiências de extração das avaliações de usuário a serem sumarizados. Estas bases de dados são chamadas *Airbnb Ratings*[1] e Opinosis[2] onde é possível comparar os sumários obtidos com os sumários escritos por humanos.

O sistema proposto usa o LDA a fim de extrair os tópicos que serão usados com palavras-chave para o TextRank, este realiza a classificação das avaliações por ordem de relevância, mais especificamente as palavras dos tópicos selecionados servem como parâmetro para a classificação dos resultados mais relevantes.

3.1 O algoritmo

O algoritmo utilizado pelo sistema proposto aplica o TextRank e a biblioteca Gensim-LDA com o intuito de classificar as avaliações por ordem de importância e para extrair os principais tópicos das avaliações.

Este trabalho é iniciado com a apresentação do componente TextRank, usado como uma das bases para comparações com o sistema proposto, explicando as modificações realizadas no algoritmo, o pré-processamento dos dados e o modo de seleção das frases. Posteriormente, o mesmo procedimento é realizado no componente LDA, usando a mesma fonte e especificações utilizadas no componente TextRank. Finalmente, é realizado o mesmo para o sistema proposto.

Deve ser levado em consideração que os componentes possuem algumas adaptações que permitem o processamento das avaliações. Essas adaptações permitem que os componentes realizem as seguintes tarefas:

1. Detectar o idioma da avaliação.
2. Realizar a tradução da avaliação para o Inglês, caso o idioma detectado pelo processo anterior seja diferente do Inglês e o usuário opte pelo uso destas avaliações.

[1] https://www.kaggle.com/datasets/samyukthamurali/airbnb-ratings-dataset?select=airbnb-reviews.csv
[2] https://github.com/kavgan/opinosis-summarization

3. Pré-processamento das avaliações, removendo emoticons e emojis utilizados com frequência no idioma Inglês. A lista de emoticons a ser removida foi obtida por meio do sítio eletrônico Wikipedia[3].

4. Separar a avaliação em frases que serão avaliadas de forma independente, onde cada frase é considerada como uma avaliação.

Para fins de simplificação, este trabalho se refere às adaptações de detecção de idioma e de tradução da avaliação como *pré-processamento linguístico*.

Essas adaptações são parte do pré-processamento das avaliações, permitindo que o programa traduza as avaliações escritas em outros idiomas pelos usuários para o Inglês ao mesmo tempo em que evita que se tente traduzir avaliações que já se encontrem em Inglês.

3.1.1 O componente TextRank

Esta implementação do algoritmo TextRank é uma versão adaptada do PageRank desenvolvida pela Google™. É importante destacar que o passo da detectção de idioma é aplicado para todas as avaliações, independente da escolha do usuário. A justificativa para essa decisão se deve ao fato de que se o usuário optar pelo descarte de avaliações que não tenham sido escritas em Inglês, o sistema deve identificar se a avaliação a ser considerada está ou não em Inglês a fim de decidir se a avaliação será utilizada ou descartada para a geração do sumário. Por outro lado, caso o usuário opte pelo uso das avaliações em outros idiomas, essa atua como um filtro, enviando para tradução apenas as avaliações que não foram escritas em Inglês, evitando o consumo desnecesário de recursos e reduzindo o tempo total de execução.

Nesta versão do TextRank a aplicação recebe o documento e o usuário decide se deseja usar ou descartar as avaliações em outros idiomas como parâmetro, então é realizada a separação das avaliações em frases que são traduzidas ou descartadas de acordo com a escolha do usuário.

O próximo passo é o pré-processamento das frases, na qual os emoticons e emojis são removidos uma vez que estes não possuem valor para a geração do sumário. As frases, depois de pré-processadas são então usadas pelo TextRank como parâmetros para gerar a matriz em conjunto das *stopwords*. Wang and Hu, 2021, [69] apresentam o termo *stopwords* como: "palavras comumente utilizadas mas que possuem pouco ou nenhum significado prático". Para fins de testes foram utilizadas como *stopwords* as palavras contidas na biblioteca *nltk.corpus*.

[3] https://en.wikipedia.org/wiki/List_of_emoticons

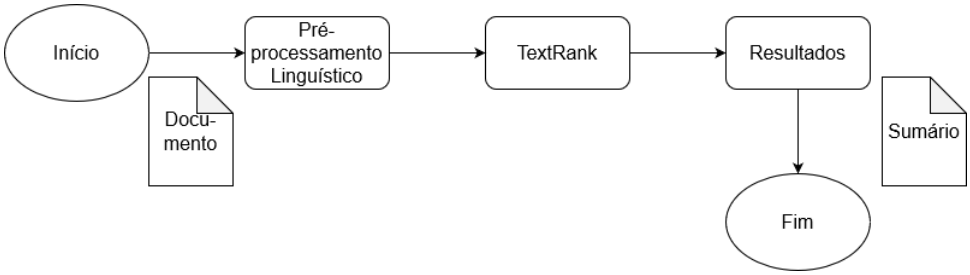

Figura 4 – Diagrama da solução TextRank modificada

A fim de avaliar a similaridade entre as frases, foi utilizada a função cosseno chamada de *cosine_distance* presente na biblioteca *nltk.cluster.util*. Esta função recebe como parâmetro de entrada dois vetores com valores numéricos e retorna um valor numérico entre 0(zero) e 1(um), como descrito pela Equação 3.1.

$$Cosine\_distance(X, Y) = 1 - Cos(X, Y) \tag{3.1}$$

Onde $Cos(X, Y)$ é descrita pela Equação 2.13. Uma vez calculados os valores de cada frase, o TextRank seleciona as frases com a maior pontuação para gerar o sumário.

O fluxo de execução desta versão modificada do TextRank é ilustrada pela Figura 4, onde o módulo de preprocessamento linguístico foi adicionado antes da execução do TextRank.

Após a execução desta versão do TextRank utilizando as amostras previamente selecionadas, os sumários resultantes foram coletados para serem usados como base de comparação. Este processo é descrito em maiores detalhes no Capítulo 4.

A próxima etapa foi realizar o mesmo procedimento para o LDA. Os detalhes da implementação deste algoritmo são descritos a seguir.

3.1.2 O componente Latent Dirichlet Allocation

O componente LDA explora a biblioteca Python chamada *Gensim*. Assim como aplicado no componente TextRank, foi implementado o processamento linguístico, e toda avaliação é verificada para identificação do idioma pelas mesmas razões descritas na Seção 3.1.1.

O LDA propriamente dito não é capaz de gerar sumários por padrão uma vez que seu propósito principal é a geração de tópicos com base no(s) texto(s) a ele fornecidos. Com o intuito de gerar sumários usando o LDA é necessário que sejam realizadas extensões no algoritmo que permitam a seleção de frases e a união destes em um sumário.

Na adaptação do LDA implementada, foram adicionadas as seguintes funcionalidades:

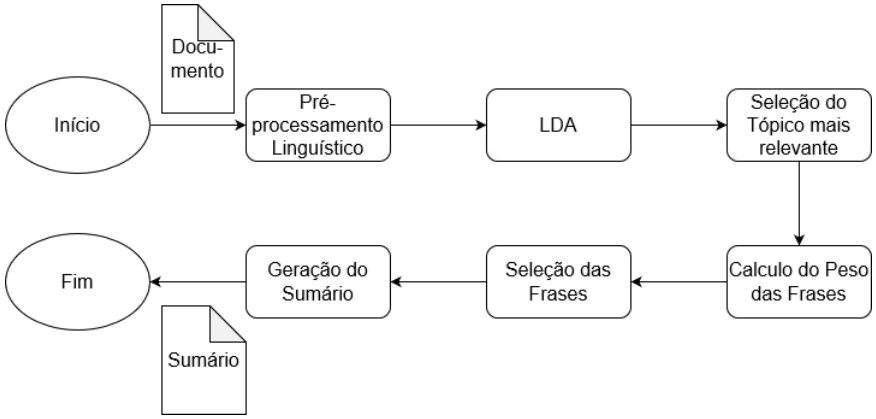

Figura 5 – Diagrama da solução LDA modificada

- Geração de uma matriz de pesos para encontrar os tópicos dominantes e ordená-los por ordem de dominância;
- Divisão do documento em frases, descartando aquelas que são muito curtas. Para esta implementação, foram consideradas como muito curtas frases com menos de 10 (dez) palavras (ver nota [4]);
- Cálculo do valor distribuído dos tópicos para cada frase com base no tópico dominante;
- Filtragem de frases com alto valor de similaridade usando a equação de cosseno de similaridade descrita na Equação 3.1;
- Seleção de frases com as maiores pontuações no tópico mais dominante.

O funcionamento desta versão modificada do LDA é ilustrado pela Figura 5, onde a extensão de preprocessamento linguístico foi adicionada antes da execução do próprio LDA, seguido pela seleção de frases baseada nos tópicos mais relevantes e seus termos.

De modo similar ao feito com o TextRank, o LDA foi executado com as mesmas amostras previamente selecionadas e os sumários resultantes foram coletados para serem avaliados e usados como base de comparação, os resultados obtidos são apresentados no Capítulo 4.

Por fim, ambos os algoritmos são usados para desenvolver o sistema proposto, esta junção é explicada na Seção 3.1.3.

3.1.3 Solução proposta

A implementação original do TextRank é capaz de gerar sumários curtos, mas por usar palavras genéricas, os sumários gerados por este método tendem a não escolher

[4] Este valor mínimo de palavras foi baseado nos valores usados nos trabalhos de: Alshboul e Odat [70]; Sun *et. al.* [71]; Peng e Shen [72]; Murkas *et. al.* [73]; Hasan *et. al.* [74] e Shen *et. al.* [75]

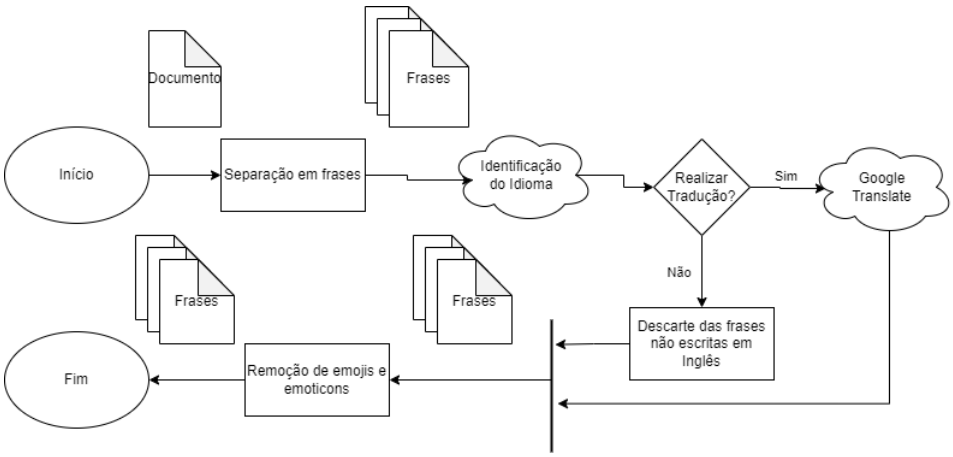

Figura 6 – Diagrama do componente de pré-processamento linguístico

frases que possuem relevância. Por outro lado, o LDA é capaz de agrupar as palavras por tópicos e assim selecionar as palavras que, possivelmente, possuem maior relevância para aquele tópico, mas sua implementação original não permite a seleção de frases para gerar sumários. Além disso, pelo fato do LDA depender muito dos tópicos e das palavras-chave escolhidas, frases que não possuam as palavras-chaves escolhidas são excluídas da seleção, mesmo que elas sejam relevantes.

Além disso, ao atuar com opiniões dos usuários, existe a possibilidade de que essas opiniões não estejam escritas em um idioma na qual o avaliador conheça, tornando a sumarização mais difícil, devido a necessidade de se traduzir os comentários antes que possam ser usados para a escrita de um sumário.

A pricipal contribuição deste trabalho é a criação de um sumarizador extrativo multilingual facilmente configurável, permitindo ao usuário selecionar características do componente LDA e o número de frases a serem selecionadas para a geração do(s) sumário(s).

Esta seção, portanto, apresenta as modificações e extensões realizadas para a criação da solução proposta e inicia com a explicação do pré-processamento linguístico, mencionado no início deste Capítulo.

3.1.3.1 Pré-processamento

O pré-processamento inicia com a separação da avaliação ou texto em frases. As frases obtidas são então tratadas para remover emoticons e emojis que podem estar presentes. Esta etapa, permite que o sistema selecione os trechos das avaliações que são mais relevantes para a geração do sumário.

Em seguida as frases são avaliadas para identificar o idioma na qual foram escritas usando o Google Translate™ para a detecção online do idioma e para a tradução do mesmo, caso necessário e desejado. Este processo é ilustrado na Figura 6.

Dependendo da opção do usuário, as frases podem ser agrupadas de acordo com a fonte de origem, caso o usuário opte por tratar cada fonte como um tópico diferente. É possível também agrupar todas as frases sem aplicar esta divisão, caso o usuário opte por tratar todas as fontes como relativo ao mesmo tópico. É importante notar que esta escolha afeta o(s) sumário(s) gerado(s), pois ao agrupar as frases pela fonte, um sumário é gerado para cada fonte existente, caso escolha tratar as fontes como relativo ao mesmo tópico, apenas um sumário é gerado.

Por fim, após esta etapa, as frases são consumidas pelo LDA, onde os tópicos e palavras-chave de cada documento ou do conjunto de documentos, de acordo com a escolha feita previamente pelo usuário, são selecionados.

3.1.3.2 Modelagem de tópicos - LDA

Após o pré-processamento da fonte, as frases são processadas pelo LDA para geração dos tópicos e para agrupar as palavras-chave nestes tópicos. Este procedimento é executado para cada grupo de frases gerado pela etapa de pré-processamento, em outras palavras, se na etapa de pré-processamento foram gerados 10 conjuntos de frases, um para cada fonte fornecida, então esta etapa gera 10 conjuntos de tópicos e suas palavras-chave, um para cada fonte fornecida. No caso dessas fontes serem tratadas como uma única fonte, apenas um conjuntos de tópicos e suas palavras-chave é gerado.

O comportamento do LDA é inalterado em sua maior parte, apenas foi adicionada a classificação de dominância de tópicos e a seleção dos termos mais influentes para o tópico mais dominante.

A criação de tópicos e a separação das palavras-chave nos tópicos criados segue o conceito original da técnica: "Cada documento exibe uma mistura de tópicos latentes em que cada tópico é caracterizado por uma distribuição sobre as palavras"(Bastani, Namavari e Shaffer, 2019) [76].

A classificação de dominância de tópicos é, como o nome sugere, a classificação dos tópicos gerados por ordem decrescente de qual tópico é, possivelmente, o mais adequado para o documento. Esta classificação é realizada por meio de uma função já disponibilizada na biblioteca *Gensim*. Com base nesta classificação, a extensão do LDA seleciona os termos que são usados pelo TextRank para seu saco de palavras.

É importante ressaltar que a seleção dos termos exerce um papel muito importante no sistema, pois os termos selecionados nesta etapa são usados como parâmetros de classificação para a porção do sistema que corresponte ao TextRank modificado.

3.1.3.3 Classificação e seleção de frases - TextRank

Na extensão do TextRank utilizada neste sistema, os termos selecionados pelo LDA na etapa anterior são considerados como parâmetros de entrada para o saco de palavras do TextRank. Já as frases agrupadas pela etapa de pré-processamento são usadas para geração do sumário.

Uma característica a ser destacada nesta extensão do TextRank é o fato de que, originalmente, as palavras que compõem o saco de palavras são usadas como palavras a serem desconsideradas, conhecidas como *stopwords*, conforme apresentado na Seção 3. Neste trabalho as palavras são consideradas como não importantes, exceto por aquelas que compõem o saco de palavras. Neste contexto, as *stopwords* que, originalmente, compõem o saco de palavras, são susbstituídas pelo conjunto de palavras selecionados pelo LDA. Além disso, as palavras presentes na frase avaliada que não existem neste novo conjunto recebem o valor 0, e as que existem o valor 1, comportamento este que difere do TextRank original, onde as palavras que não existem no saco de palavras recebem o valor 1, e as que existem o valor 0.

A próxima etapa é a criação e população da matriz de similaridade, matriz essa que, de modo numérico, indica o quão similar é um par X, Y de frases, sendo 0 indicativo de que este par não possui similaridades e 1 indicando que as frases são idênticas. Para calcular os valores que preenchem cada par de coluna e linha da matriz de similaridade é realizado o seguinte procedimento: Cada par X, Y de frases é usado para calcular a distância de cosseno entre elas por meio da fórmula transcrita na Equação 2.13. Os resultados obtidos são, então, usados para popular a matriz de similaridade, na qual cada par de linha e coluna corresponde a um par X, Y.

Uma vez gerada a matriz de similaridade, esta é usada como entrada para gerar o grafo de similaridade que, por sua vez, é usado para gerar um vetor de pontuações de cada frase. Esse vetor é ordenado pela sua pontuação em ordem decrescente, as frases com maior pontuação são aquelas que menos possuem similaridade com as outras frases presentes e as mais repetitivas recebem as pontuações menores.

Por fim, as frases que obtiveram as maiores pontuações são selecionadas para compor o sumário. O fluxo de operação do sistema é ilustrado na Figura 7.

Novamente, são utilizadas as mesmas amostras usadas na geração dos sumários do TextRank e do LDA, para gerar novos sumários que foram coletados e comparados com aqueles obtidos por essas duas aplicações. A discussão dos resultados obtidos são apresentado no Capítulo 4.

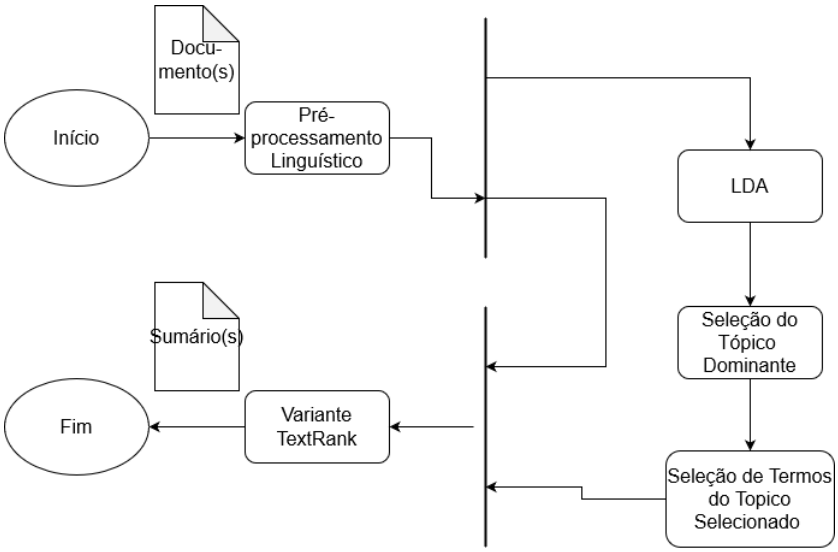

Figura 7 – Diagrama do sistema proposto

4 Avaliação

Neste capítulo são apresentados dois experimentos que foram realizados com o objetivo de cobrir a abordagem proposta. Para isso, foram usadas duas bases de dados: *Airbnb Ratings*[1] e Opinosis[2]

Para o primeiro experimento foi utilizada a base *Airbnb Database* e foi iniciada pela análise e apresentação das características do banco, seguindo para a explicação dos procedimentos realizados, dos resultados obtidos, e, por fim, as conclusões obtidas neste experimento.

Para o seguindo experimento foi utilizada a base *Opinosis* e, similar ao primeiro experimento, foi iniciada pela análise e apresentação das características do banco, seguindo para a explicação dos procedimentos realizados, dos resultados obtidos, e, por fim, as conclusões obtidas neste experimento.

4.1 Experimento 1 - Airbnb

Nesta seção são apresentados os resultados obtidos com cada solução, iniciando pela variante do TextRank, seguido pela variante do LDA e, por fim, o sistema proposto.

4.1.1 Analisando o banco de dados

O primeiro passo deste experimento, foi a aquisição e a extração do banco de dados. Em seguida foi aberto o arquivo, para se gerar amostras que foram usadas com textos de entrada e analisar as colunas do banco. O nome das colunas e suas respectivas descrições são apresentadas na Tabela 2

Foram encolhidos 4 valores de *listing_id* de forma aleatória a fim de se criar 4 conjunto de amostras, essas amostras possuem todas as avaliações para aquele *listing_id*. Após uma análise mais detalhada foi possível notar que as amostras possuem vários idiomas, como, por exemplo: Koreano, Russo, Inglês, Português (Brasil) e Chinês.

Este trabalho não visa gerar sumários com múltipos idiomas, portanto, o idioma Inglês foi selecionado como o idioma em que o sumário é escrito. Com o intuito de decidir qual opção seguir, foi decidida a execução de simulações com ambas as possibilidades, com e sem tradução, e comparamos os resultados. A ferramenta escolhida para tratar da tradução de avaliações que não foram escritas em Inglês foi o Google ™Translate, um

[1] https://www.kaggle.com/datasets/samyukthamurali/airbnb-ratings-dataset?select=airbnb-reviews.csv

[2] https://github.com/kavgan/opinosis-summarization

Nome	Descrição
listing_id	listing_id é o identificador da propriedade
id	id é o identificador da avaliação
date	date é a data em que a avaliação foi submetida
reviewer_id	reviewer_id se refere ao identificador do avaliador
reviewer_name	reviewer_name é o nome do avaliador
comments	comments é o próprio comentário

Tabela 2 – Cabeçalho do banco *Airbnb Ratings*

serviço de tradução multilingual desenvolvido pela empresa Google™para tradução de textos, documentos e páginas Web.

Ao analisar os resultados, alguns pontos devem ser considerados, tanto do lado da tradução, quanto do lado do descarte.

Quando usada a tradução automática, também conhecida como tradução de máquina:

1. A tradução pode não ser fiél à avaliação original. Isto ocorre devido ao fato de que expressões, vícios linguísticos e figuras de linguagem podem não ser traduzidos corretamente em outro idioma. Um exemplo seria a frase *break a leg* em inglês que, em tradução livre para o Português brasileiro seria *Quebre a perna*, mas que seria o equivalente a desejar boa sorte à um artista;

2. Traduções imprecisas ou errôneas podem influenciar negativamente na qualidade do sumário;

3. Outros fatores podem interferir na qualidade da tradução, a exemplo: erros de escrita, abreviações(BO, Ca.), emojis(e.g. XD, XP, :]).

É importante destacar que ao usar o serviço de terceiros, existem mais dois pontos a serem considerados: a disponibilidade do serviço e a variabilidade dos resultados. A disponibilidade do serviço se refere a fatores que podem resultar na impossibilidade do uso do serviço. (falta de conexão com o servidor, indisponibilidade do serviço, seja por descontinuidade ou por problemas técnicos), já a variabilidade dos resultados se refere a um certo grau de imprevisibilidade no resultado da tradução uma vez que o sistema não possui controle sobre como a tradução é realizada.

Se por um lado a tradução não garante a precisão do sumário, por outro o descarte também não o garante, pois ao realizar o descarte de avaliações escritas em outros idiomas

existe a possibilidade de se descartar informações relevantes ao tópico. Portanto, assim como uma tradução imprecisa, o descarte de boas avaliações pelo fato de serem escritas em outro idioma pode prejudicar a qualidade do sumário no final do processo.

Tendo em consideração todos os pontos apresentados, foi decidido que o sistema permita o usuário escolher entre o recurso multilingual ou usar apenas as avaliações escritas em Inglês.

Conforme explicado na Seção 3, todas as técnicas implementaram o pré-processamento linguístico e o aplicaram para todas as avaliações, de modo que todas as avaliações são traduzidas para o Inglês sem serem descartadas. Devido ao fato de cada amostra se referir a uma determinada localidade, usá-los em conjunto para gerar um único sumário composto pela união desses documentos não resultará em um sumário significativo. Portanto, para cada amostra foi gerado um sumário desta amostra para cada técnica. Para fins comparativos, todos os sumários gerados possuem 5 frases selecionadas por cada técnica. Os sumários gerados são apresentados nas Tabelas 3 e 4.

Em uma primeira avaliação nota-se que o TextRank e o LDA geram, respectivamente, o sumário mais curto e o mais longo dos 3 sistemas. Por outro lado, o sistema proposto gerou sumários que são geralmente mais curtos que o LDA porém mais longos que o TextRank o que em primeira análise pode causar dúvidas sobre a qualidade dos sumários gerados por esse novo sistema. Mas, como explicado no Capítulo 1, um sumário maior não resulta necessariamente em maior qualidade e vice-versa. Mesmo assim, o comprimento de um sumário é considerado como parte da avaliação de sua qualidade, para compará-los é necessário uma avaliação mais crítica sobre cada resultado descrito nas Tabelas 3 e 4 para julgar se as frases selecionadas são relevantes ou se não o são.

Para organizar as comparações entre os métodos, cada amostra será avaliada em uma subseção, na qual os resultados obtidos por cada método é comparado para a amostra em questão. Por último, uma conclusão geral é apresentada em uma outra Seção.

4.1.2 Amostra 1

A partir dos resultados obtidos para a Amostra 1 é notória a diferença em comprimento entre os sumários gerados. O TextRank gerou o sumário mais curto dentre os três sistemas. Porém, as frases selecionadas não possuem tanto valor quanto aquelas selecionadas pelos outros métodos, pois apresentam conteúdo com informações pouco relevantes ou que possuem conteúdo similar, por exemplo: “Thank you alex” ou “Alex was a nice host”, que não fornece informações relevantes ou que é similar à primeira frase “Alex is a great host...”.

Por outro lado, o LDA selecionou frases muito longas, que possuem uma quantidade abundante de informações sobre o proprietário e sobre o local. Em contra-partida, possui

Amostra	TextRank	LDA	Sistema Proposto
Amostra 1	Alex is a great host, the apartment is very clean . Alex was a nice host . I had really good time in alex apartment and i felt like in my parents house, alex is very nice host and apartment is lovely, also very important is excellent location. Great apartment and alex is a really nice guy. Thank you alex	I had really good time in alex apartment and i felt like in my parents house, alex is very nice host and apartment is lovely, also very important is excellent location. The appartment is really very clean, there is enough place for your belongings, a small but beautiful balcony, and on the top of the house you have an incredible view over the city. It was really nice to be there and it helped for our few days staying to be more efficient and very very good: . The entire house is very clean, the room is very nice with space for clothes and a big bed. Alex is a very nice host, he gave us a little tour and presented the city from the terrace of his building and then gave us all the advice we needed when we asked for recommendations or information.	Alex house's is a really nice flat well situated too. Alex is a friendly guy who like to give advice and talk about this city he really loves so that is a very nice meeting with him. Alex's appartment is a very nice place to stay in barcelona. The flat was very modern, nice and clean, lots of light in the room and we even had a small balcony which was just beautiful. The flat is absolutely pristine and bedroom was nice and comfortable.
Amostra 2	We had a great stay at patricia and jose's- the apartment was clean, modern with comfy beds and nice linen and towels . Patricia greeted us at the apartment to walk us through all it has to offer. Jose and patricia kept in touch with us before our arrival, and jose met us at the apartment with the keys and a short tour. Patricia met us at the apartment and was very professional and helpful, explaining how everything worked. Our stay in their apartment was perfect.	The apartment is gorgeous, nicely located, has everything one would need for a short/ long stay in bnc . Patricia is the best and left us with a nice book of information and suggestions and was so helpful at check in. The apartment had everything you could need, was very clean and the location was great. Patricia was warm and friendly, and gave us a detailed tour of the apartment with a friendly self-made guide to all of the nearby restaurants, metro stations, neighborhoods etc. We left a phone in the apartment the morning of check-out and patricia dropped everything to come and unlock the door for us and make sure we still caught our flight.	Nice apartement clean on a good location and very nice people: . Nice apartment, very clean and spacious. Nice location, clean and tidy apartement very helpful and nice people. I will mention it is about a 30 minute train / bus ride away from most attractions so it is not an easy hop in / hop out during the day kind of place but the neighborhood is really nice and itś nice to be outside of everything at night. The place was perfect for us, walking distance to the beach and near local transportation that works perfectly in barcelona.

Tabela 3 – Experimental output - Part 1

Amostra	TextRank	LDA	Sistema Proposto
Amostra 3	Federico made us a map and roberto gave us the location on the mobile. Federico was very friendly and gave us some great recommendations. Federico is a great host i recommend Federico was a really nice and helpful host, he does everything to make our stay pleasant. We could not meet federico but he left us with roberto as host and the truth that was very nice and kind.	Federico was a great host, the area in which the floor is very accessible and safe and the floor is wide and comfortable. I slept very well and there is everything you might need to make your stay comfortable in a clean environment, well equipped kitchen, etc. The room and the house in general are clean, organized and very comfortable. He provided lots of tips on where to visit in the city. Roberto who lives in the apartement and cleans the place is a warm, friendly and very funny guy.	Federico made us a map and roberto gave us the location on the mobile. New and pleasant mattresses, and if not the location of the apartment is top metro station at 100 meters and a few stations from the center. The location of this apartment is a major asset, very well located a few hundred meters from the rocafort metro station in connection with all lines covering the city and the possibility of completing other routes with the bus. Roberto who lives in the apartement and cleans the place is a warm, friendly and very funny guy. He taught us the whole apartment and made us feel at home.
Amostra 4	But very good location and great host Really nice flat and sergio was really nice. Xavier is good and nice for this journey, thank very much We had a really great time in barcelona and xavierś and claudiaś place was excellent. A friend of xavier .	Everything went very well, xavier was very attentive at all times. They took time and dedication to explain the operation of everything in the department as well as show me on the maps to visit, where to go and eat. We were made to feel very welcome by claudia who took the time to tell us the nice places to go out on barcelona. During our stay, the apartment was quiet and fairly clean exceptions made of the kitchen that lacks maintenance and equipment to be able to use it properly. Location of the house is brilliant - you have metro, beach, supermarket just within minutes of walking.	They took time and dedication to explain the operation of everything in the department as well as show me on the maps to visit, where to go and eat. The location is great, thereś a metro station one block away, and many great restaurants short walks away. Location is perfect: 10 min from the beach, 1 min to the metro station. He did everything to make our stay perfect, it was extremely touching. Would gladly stay here again the next time i am in barcelona.

Tabela 4 – Experimental output - Part 2

a desvantagem de tornar o sumário mais longo. A tendência deste método em selecionar frases longas pode resultar em frases com pouca relevância sendo escolhidas para compor o sumário. Como exemplo do problema citado, é possível usar a seguinte frase: “It was really nice to be there and it helped for our few days staying to be more efficient and very very good: .”. Note que nesta frase, apesar de ser longa, ela pouco contribui para o sumário uma vez que não possui informações relevantes nem sobre o proprietário, nem

sobre o local.

Em comparação, o sistema proposto, selecionou frases relevantes, que contribuem mais para a geração de um sumário melhor que aquele gerado pelo TextRank, e as frases escolhidas são mais curtas que aquelas selecionadas pelo LDA. Apesar da terceira frase escolhida acrescentar pouco valor ao sumário, similar ao LDA, o sistema ainda foi capaz de gerar um sumário tão relevante quanto ele com frases mais curtas.

4.1.3 Amostra 2

Avaliando os resultados obtidos para a Amostra 2 nota-se, novamente, que o TextRank gerou o sumário mais curto dentre os 3, porém o resultado obtido mostra frases com pouca relevância para o sumário, similar ao que aconteceu com resultado apresentado na Seção 4.1.2. Como exemplo de frases com pouco valor para o sumário é possível apontar: “Our stay in their apartment was perfect.” e “Patricia greeted us at the apartment to walk us through all it has to offer”. Note que em ambas as frases, mesmo que sejam relacionados ao local, não fornecem informações relevantes para o leitor, ou seja, informações que justifiquem a boa avaliação.

O LDA, por outro lado, selecionou frases longas, que contém mais informações relevantes. Por exemplo a quarta frase fornece informações não existentes no sumário gerado pelo TextRank, que são as frases que tratam dos restaurantes e estações de metrô próximos à propriedade. Porém, assim como apontado na Seção 4.1.2, a escolha de frases longas pode resultar em frases que não são relevantes. Exemplificando esse ponto, a quinta frase é: “We left a phone in the apartment the morning of check-out and patricia dropped everything to come and unlock the door for us and make sure we still caught our flight”. Esta frase, embora indique uma característica positiva da proprietária, não agrega valor para o sumário a ser lido pelo leitor.

O sistema proposto aborda melhor o problema da falta de relevância nas frases selecionadas pelo TextRank por meio da inclusão de informações relevantes, como por exemplo, transportes e atrações próximas. Nesta amostra é possivel argumentar que o sumário gerado pelo sistema proposto é melhor que o apresentado pelo LDA, sendo mais curto e com informações que podem ser consideradas tão úteis quanto. Entretanto, ela ainda apresenta algumas falhas na seleção de frases, a primeira frase apresentada na Tabela 3, poderia ser removida sem afetar a qualidade do sumário gerado.

4.1.4 Amostra 3

Na Amostra 3 o TextRank gerou um sumário que possui apenas informações pertinente aos proprietários Frederico e Roberto. O conteúdo das frases é repetitivo de tal modo que apenas a quarta frase se torna relevante enquanto todas as outras poderiam ser

ignoradas sem prejuízo à qualidade do sumário gerado.

Nesta amostra o LDA apresenta um sumário que é menor que o gerado pelo sistema proposto, o que não ocorreu nos sumários das amostras anteriores. Além disso, se compararmos os sumários gerados pelo LDA e pelo TextRank, se torna notória a diferença na quantidade de informações presentes em ambas, o LDA possui informações não apenas sobre os proprietários como também sobre a propriedade como mobília e limpeza do local.

Nesta amostra ocorreu o único caso em que o sistema resultou em um sumário mais longo que o LDA, mesmo assim o sumário gerado é muito mais rico em informações que o sumário gerado pelo TextRank. Ainda mais, é possível argumentar que o sumário gerado pelo sistema proposto possui mais informações que o LDA, posto que além das informações presentes naquela estarem presentes nesta, há informações sobre estações de metrô próximas ao local, informação esta que não consta em nenhum dos outros dois sumários.

4.1.5 Amostra 4

Na última amostra utilizada, o TextRank gerou o sumário mais curto dentre todas as amostras para este método, em sua maioria as frases selecionadas elogiam os proprietários ou o *flat* sem expor mais detalhes. Além de que, a quinta frase, não contribui em qualquer aspecto para a qualidade do sumário, pelo contrário, pode-se dizer que a frase "A friend of xavier ." piora a qualidade do sumário, pois não se sabe o contexto da qual esta frase foi extraída.

Para esta amostra, o método LDA produziu um sumário com várias informações relevantes, como por exemplo: locais para se visitar, a existência de comércio próximo e da existência de transporte próximo ao local. Um ponto que chamou a atenção é apontado na quarta frase selecionada: "...the kitchen that lacks maintenance and equipment to be able to use it properly.", essa frase em específico chama a atenção pois é a única ocorrência de uma frase que cita uma característica negativa da propriedade, tal fato não ocorreu em nenhum outro sumário analisado.

Nesta amostra, o sistema proposto falha em incluir o aspecto negativo explicado no sumário gerado pelo LDA, além de que, a quinta frase selecionada pouco contribui para o sumário. Mesmo assim, o sumário gerado possui relevância, por trazer informações ao leitor sobre a existência de comércio, transporte e sobre o proprietário.

4.1.6 Sumário dos resultados

Após a análise individual dos sumários apresentados nas Tabelas 3 e 4, é possível extrair diversas informações sobre os pontos positivos e negativos de cada método, que são discutidos nesta seção.

Em quesito de qualidade de sumários gerados, o TextRank teve o pior desempenho entre os 3 métodos uma vez que ele não é capaz de atribuir de maneira apropriada valores para cada palavra presente no saco de palavras (tradução livre do termo original *Bag of Words*). Isso impacta diretamente na escolha de frases, resultando na escolha de frases que não possuem grande relevância. Este problema pode ser constatado principalmente nos sumários gerados para as amostras 3 e 4 correspondentes às Seções 4.1.4 e 4.1.5, respectivamente, onde diversas informações relevantes não são incluídas no sumário gerado. Esses problemas foram apontados em trabalhos anteriores como o de Fang *et. al.* [77] e Wang *et. al.*[20]. Como resultado, apesar dos sumários gerados pelo TextRank serem mais curtos, eles possuem como desvantagem o fato de geralmente serem menos ricos de informações.

Os sumários gerados pelo LDA, por outro lado possuem a tendência de selecionar frases mais longas, o que resulta em sumários mais longos. Esta tendência pode ser confirmada em todos os sumários, com exceção daquele gerado na Seção 4.1.4. El-Kassas *et. al.* (2021), [9] afirmam que soluções que geram sumários usando a abordagem baseada em tópicos são afetados pelos tópicos escolhidos, esta afirmação é complementada por Wang e Ma (2013) [78], explicando que frases que não possuem as maiores pontuações não serão selecionadas para compor o sumário, mesmo em casos em que a frase seja relacionada ao tópico principal. Estas características, juntamente com o lógica implementada, podem explicar o motivo do LDA selecionar as frases mais longas, uma vez que frases mais longas possuem uma probabilidade maior de conter os termos que compõem os tópicos, explicando o motivo dos sumários serem mais longos. Em resumo, o LDA gera sumários mais longos, porém mais ricos em informação.

O sistema proposto substitui o conteúdo genérico do saco de palavras usado pelo TextRank por termos selecionados pelo LDA. Esta modificação aprimora a seleção de frases ao fornecer ao algoritmo de classificação quais termos devem ser considerados relevantes para o sumário ao mesmo tempo que evita frases longas. Contudo, os problemas apontados por El-Kassas *et. al.* (2021) [9] no que diz respeito a influência dos termos selecionados permanece, pois se a frase não possuir determinado termo, ela não será incluída no sumário. Essas características podem ser observadas nos resultados obtidos nas Seções 4.1.2, 4.1.3, 4.1.4 e 4.1.5, onde os sumários gerados com as amostras 1, 2 e 4 retiveram informações relevantes ao tópico ao mesmo tempo sendo mais curtos que aqueles gerados pelo LDA. Já a desvantagem pode ser notada no sumário produzido usando a amostra 3.

De forma geral, apesar de indicar que pode ser aprimorada, a solução proposta demonstra que é capaz de gerar sumários com maior qualidade que o TextRank e, por vezes, até que o LDA, ao mesmo tempo que consegue, geralmente, ser menor que o último.

A pontuação com a metodologia Rouge não é possível com esta base de dados devido a ausência de sumários humanos oficiais para serem usados como base de compa-

ração.

4.2 Experimento 2 - Opinosis

Nesta seção são apresentados os resultados obtidos com a solução proposta, iniciando pela avaliação da base e da apresentação de suas características, seguindo para os resultados obtidos e, por fim, apresentando as conclusões.

4.2.1 Avaliação do banco de dados

Este experimento foi iniciado com a obtenção e extração do base de dados Opinosis, de modo similar ao realizado com o banco de dados *Airbnb Database*. Após avaliarmos seu conteúdo, foram escolhidos 4 tópicos para servirem como documentos de entrada para o sistema.

Este banco, diferentemente daquele apresentado na Seção 4.1 é composto majoritariamente por frases em inglês, existindo poucas palavras em outros idiomas. O banco é formado por diversas frases agrupadas em documentos por tópico, e possui, além das avaliações uma pasta chamada *summaries-gold*, pasta que possui sumários gerados por humanos para cada tópico.

Novamente, este trabalho não visa gerar sumários para idiomas que não sejam o inglês, portanto, os sumários gerados foram todos escritos em inglês. Devido ao fato do banco ser composto quase em sua totalidade de frases em inglês, foi escolhida a sumarização sem tradução.

Após a leitura da documentação do Opinosis e a avaliação de seu conteúdo, alguns pontos chamaram atenção, são eles:

- Devido a segmentação imperfeita das frases, pode haver frases incompletas. Isso pode comprometer a qualidade do sumário gerado.
- A existência de sumários feitos por humanos, os quais podem ser usados como base para averiguar a qualidade do sumário gerado. Tais sumários são chamados de “sumários dourados” (*summaries gold*, no original) na fonte. Cada tópico possui uma média de 4 “sumários dourados”.
- Por ser uma base em inglês, não houve oportunidades para aplicar a função de tradução, diferente do primeiro experimento.

Assim como no primeiro experimento, as avaliações foram pré-processadas na etapa de processamento linguístico a fim de se remover símbolos que não são considerados relevantes, como emojis.

É importante destacar sobre o item de segmentação de frases, que isto é uma caractarística da própria fonte na qual a base foi extraída, não sendo possível, para este trabalho, encontrar as avaliações originais. Isto pode resultar em sumários menos precisos, pois a ausência de informações pode comprometer o sentido do restante da frase, bem como resultar na ausência de informações relevantes para a geração do sumário.

Isso é ainda mais relevante quando se trata de sumarizadores extrativos, pois esse tipo de sumarizador depende das palavras e frases existentes na(s) fonte(s), uma vez que os sumários gerados são frases selecionadas pelo sumarizador que foram extraídas da(s) fonte(s).

Diferentemente da primeira base de dados, esta não possui uma estrutura padrão, e pode ser descrita como um conjunto de frases agrupadas em um arquivo texto, na qual cada linha possui apenas uma frase. Os sumários dourados são apresentados nas tabelas das Seções 4.2.2, 4.2.3 e 4.2.4 como "Sumário 1", "Sumário 2", "Sumário 3", "Sumário 4", "Sumário 5" (quando houver).

Tendo estes pontos em consideração, os resultados obtidos são apresentados nas Seções 4.2.2, 4.2.3, 4.2.4 e as conclusões na Seção 4.2.5.

4.2.2 Resultados - TextRank

Nesta parte do experimento, o componente TextRank foi executado com a base de dados, e os resultados foram coletados para serem comparados com os sumários escritos por humanos.

Para testar o componente TextRank, foi usado a base de palavras *stopwords* para compor o conjunto de palavras menos relevantes.

Um ponto importante em relação à diferença entre esta aplicação do TextRank e o componente TextRank usado nesta abordagem é o fato de que as *stopwords* são usadas como palavras de interesse e não como filtro. Outro detalhe é que estes experimentos foram extraídas as 2 frases consideradas mais relevantes pelo algoritmo, a fim de se comparar as frases extraídas com as frases dos sumários.

As Tabelas 5, 6, 7 e 8 exibem os resultados obtidos e os "sumários dourados" para as duração da bateria de um netbook, os botões de um Amazon Kindle, a localização do hotel Holliday e a tela de um netbook.

O resultado obtido para o tópico da Tabela 5 resultou em um sumário extremamente curto, mas não fornece estimativas sobre a duração da bateria, se limitando a dizer apenas que a duração da bateria é longa.

Para o tópico da Tabela 6, a primeira frase é confusa, sendo difícil extrair informações dela, já a segunda frase pode ser considerada parcialmente equivalente à segunda

TextRank	Sumário 1	Sumário 2	Sumário 3	Sumário 4	Sumário 5
the reason i went with asus is the long battery life without the huge bump in the back , a fast processer, easy memory upgrade, good looks and great price . the image is nice and crisp and the longevity of the battery is a definite plus .	The battery life is longer then 5 hours. But due to the battery charger this may decrease or not work at all.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.	battery-life is fantastic and good. The 6 hours battery life is great.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.

Tabela 5 – Sumário do Opinosis para o TextRank - Duração da bateria de um netbook

TextRank	Sumário 1	Sumário 2	Sumário 3	Sumário 4
sym button depressed at same time , inside, edge buttons , these have saved me from accidentally pressing buttons many times .	It is not user friendly and the buttons are not easily pressed.	The buttons are well placed, but can be hard to press down.	Magical five way button. Next page button on both side of kindle. No reset button.	New buttons are easy to use and effective. No more accidental button presses. Buttons make navigation easy.

Tabela 6 – Sumário do Opinosis para o TextRank - Botões do Amazon Kindle

TextRank	Sumário 1	Sumário 2	Sumário 3	Sumário 4
nice location by glouchester tube stop . outstanding location, even an internet cafe nearby .	Location is excellent, very close to the Glouchester Rd. tube stop.	The locations are excellent. Nice and close to where you want to be.	The location is excellent. The hotel is very convenient to shopping, sightseeing, and restaurants. It is located just minutes from the tube stations.	Excellent location. Near the tube station.

Tabela 7 – Sumário do Opinosis para o TextRank - Localização do hotel Holliday em Londres

frase do sumário 4 sobre evitar acionamentos acidentais dos botões.

Para o tópico referente a localização do hotel Holliday, mostrado na Tabela 7, temos um desempenho satisfatório, pois as frases condizem com a informação dos sumários sobre a presença do ponto de transporte, além de indicar a presença de outros estabelecimentos próximos ao local.

Para o último tópico, a tela do netbook, foi analisado os sumários e comparado com os resultados do sumário do TextRank. Os sumários são unânimes quanto ao fato de que a tela é pequena, três deles elogiam o brilho da tela e dois deles relatam que a tela apresenta erros com certa frequência. Por outro lado, o sumário gerado cita que a tela não vira completamente para trás e que, com a tela maximixada, a tela é grande o bastante para trabalhar e ler documentos. Uma curiosidade é que nesta amostra o sumário gerado

TextRank	Sumário 1	Sumário 2	Sumário 3	Sumário 4
one of the downside i have noticed is that the netbook is heavy near the hinge because that's where the battery is located and if you want to put it on your lap while sitting, you have to put your hand on the bottom chassis so that the netbook doesn't tip over, and the screen doesn't bend all the way backwards which can be annoy, otherwise this is a really nice netbook . with the view maximized, the screen is large enough to work on documents and read online research materials .	Screen is clear and bright. However, the screen is smaller than most screens. Blue screen experienced often.	Although the screen is small, it's sharp, bright and readable.	The screen is sharp, bright but small sized.	Screen quality is up to par. A common bug is the screen going black and the computer still runs. The screen can seem small and extremely reflective.

Tabela 8 – Sumário do Opinosis para o TextRank- Tela do netbook

foi maior que o gerado pelo LDA.

4.2.3 Resultados - LDA

Nesta parte do experimento, de modo similar ao feito com o TextRank, o componente LDA foi executado com a base de dados, e os resultados foram coletados para serem comparados com os sumários escritos por humanos.

Diferentemente do TextRank, o LDA não possui um saco de palavras para decidir sobre as palavras a serem desconsideradas. Em seu lugar existem uma série de parâmetros que modificam o comportamento da aplicação. Para fins de comparação, este teste foi realizado usando configurações similares àquelas usadas pela solução proposta.

De modo similar ao realizado com o TextRank, optou-se por gerar sumários com 2 frases escolhidas pelo LDA para se comparar com os "sumários dourados".

As Tabelas 9, 10, 11 e 12 exibem os resultados obtidos e os "sumários dourados" para serem comparados.

O resultado obtido por este método para o tópico da Tabela 9 resultou em um sumário mais longo que aquele gerado pelo TextRank na Tabela 5. Neste sumário é fornecida a informação que a bateria do netbook possui uma boa duração chegando a 9 horas, diferente das 5 ou 6 horas apresentados pelos "sumários dourados", mas que não o invalida.

Seguindo para o tópico da Tabela 10, as frases indicam que os botões são bem posicionados, permitindo o uso do Kindle com facilidade. Porém, como indicado na Seção 4.1.6, as frases escolhidas são consideravelmente longas, se comparadas com os "sumários dourados" pela proporção de informações contidas.

LDA	Sumário 1	Sumário 2	Sumário 3	Sumário 4	Sumário 5
it won't give you all the functionality of a full, sized laptop, and the small screen may take some getting used to, but no regular laptop gives you this kind of battery life, and most of the functions absent here aren't used by most people most of the time anyway, so if the tradeoff is fewer features for greatly, increased battery life, i'll take the latter . i have not done an actual test of the full battery life but from what iv seen in general use it has around 9 hours while running wifi and other applications with around half screen brightness .	The battery life is longer then 5 hours. But due to the battery charger this may decrease or not work at all.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.	battery-life is fantastic and good. The 6 hours battery life is great.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.

Tabela 9 – Sumário do Opinosis para o LDA - Duração da bateria de um netbook

LDA	Sumário 1	Sumário 2	Sumário 3	Sumário 4
i started reading with the default size without my glasses and noticed i was squinting a bit, so changed to one size larger with a couple button clicks and it was much easier without feeling like the print was too big and took up too much of the page . i don't need two hands to hold the book open when i'm on the train, or whenever, just one hand gives you good access to all the buttons even while you are holding the device .	It is not user friendly and the buttons are not easily pressed.	The buttons are well placed, but can be hard to press down.	Magical five way button. Next page button on both side of kindle. No reset button.	New buttons are easy to use and effective. No more accidental button presses. Buttons make navigation easy.

Tabela 10 – Sumário do Opinosis para o LDA - Botões do Amazon Kindle

LDA	Sumário 1	Sumário 2	Sumário 3	Sumário 4
the breakfast we had 1 morning was to say the least quite incredible there was so much variety, there was also no complaint in this area with the staff they were right on the ball helpfulthis hotel while being high rise had some low moments for us as we felt some areas were really lacking but the hotels location is superb and staff like peter did try and make your stay worth while . great location , a couple of minutes walk from gloucester rd underground station, and shops, restaurants, internet cafe etc	Location is excellent, very close to the Glouchester Rd. tube stop.	The locations are excellent. Nice and close to where you want to be.	The location is excellent. The hotel is very convenient to shopping, sightseeing, and restaurants. It is located just minutes from the tube stations.	Excellent location. Near the tube station.

Tabela 11 – Sumário do Opinosis para o LDA - Localização do hotel Holliday em Londres

Para o tópico referente à localização do hotel Holliday, Tabela 11, o resultado foi um sumário com informações relevantes, mas que poderia ser reduzida para a segunda frase apenas sem comprometer a quantidade de informações contida. Para este caso, é possível dizer que, o sumário gerado pelo TextRank, transcrito na Tabela 7 possui uma qualidade melhor que o LDA, sendo mais curto e próximo do que existe nos "sumários

LDA	Sumário 1	Sumário 2	Sumário 3	Sumário 4
some of the asus special on screen controls, update and expand features built into the machine have not been all that useful and some border on being advertisements for buying asus stuff . it won't give you all the functionality of a full, sized laptop, and the small screen may take some getting used to, but no regular laptop gives you this kind of battery life, and most of the functions absent here aren't used by most people most of the time anyway, so if the tradeoff is fewer features for greatly, increased battery life, i'll take the latter .	Screen is clear and bright. However, the screen is smaller than most screens. Blue screen experienced often.	Although the screen is small, it's sharp, bright and readable.	The screen is sharp, bright but small sized.	Screen quality is up to par. A common bug is the screen going black and the computer still runs. The screen can seem small and extremely reflective.

Tabela 12 – Sumário do Opinosis para o LDA - Tela do netbook

dourados".

Por fim, em relação ao último tópico, a tela do netbook, o sumário gerado deixa muito a desejar, pois, apesar de possuir a informação de que a tela é pequena como citam os "sumários dourados", praticamente todo o restante do sumário não se refere ao tópico escolhido. Deste modo, o sumário gerado é pouco eficiente se comparado com o gerado pelo TextRank para o mesmo tópico.

4.2.4 Resultados - Solução Proposta

Nesta subseção são apresentos os resultados obtidos para a solução proposta. Os resultados são comparados com os sumários escritos por humanos, assim como foi realizado com o TextRank e o LDA.

Para gerar os sumários, a solução executa o LDA para selecionar as palavras mais relevantes para o TextRank, que realiza a distribuição de pesos para as frases. Neste experimento, o sistema usa as 2 frases consideradas mais relevantes para compor o sumário.

As Tabelas 13, 14, 15 e 16 exibem os resultados obtidos e os "sumários dourados" para serem analisados e comparados.

Proposta	Sumário 1	Sumário 2	Sumário 3	Sumário 4	Sumário 5
The battery life is unbelievable, even while using usb to transfer files, watching movies and surfing the internet with screen brightness all the way up . I have not done an actual test of the full battery life but from what iv seen in general use it has around 9 hours while running wifi and other applications with around half screen brightness .	The battery life is longer then 5 hours. But due to the battery charger this may decrease or not work at all.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.	battery-life is fantastic and good. The 6 hours battery life is great.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.	Battery lasts about 5 hours. Time is shorter when running many drives or using bright backlight.

Tabela 13 – Sumário do Opinosis para a Proposta - Duração da bateria de um netbook

Proposta	Sumário 1	Sumário 2	Sumário 3	Sumário 4
This means that when you handle the kindle 2 by the edges, the pages don't change even if you grab by the buttons . They seem to be in the perfect location, i can grip the k2 easily and do not have to worry about pressing a button on accident but when i need to change the page they are right where i want them .	It is not user friendly and the buttons are not easily pressed.	The buttons are well placed, but can be hard to press down.	Magical five way button. Next page button on both side of kindle. No reset button.	New buttons are easy to use and effective. No more accidental button presses. Buttons make navigation easy.

Tabela 14 – Sumário do Opinosis para a Proposta - Botões do Amazon Kindle

Proposta	Sumário 1	Sumário 2	Sumário 3	Sumário 4
The location in south kensington is excellent as the gloucester road tube station is just around the corner and is served by 3 main lines, , district, circle and piccadilly . Excellent location next to gloucester road tube station that has 3 train lines making getting around london really easy .	Location is excellent, very close to the Glouchester Rd. tube stop.	The locations are excellent. Nice and close to where you want to be.	The location is excellent. The hotel is very convenient to shopping, sightseeing, and restaurants. It is located just minutes from the tube stations.	Excellent location. Near the tube station.

Tabela 15 – Sumário do Opinosis para a Proposta - Localização do hotel Holliday em Londres

Proposta	Sumário 1	Sumário 2	Sumário 3	Sumário 4
But even with that, and the glossy screen, fingerprint magnet, limited screen tilt, tinny speakers, and intermittent odd glitches, i still think it's a great value and would buy it again today if i were on the market for a netbook . Hard to see full web pages on small screen, resolution inadequate .	Screen is clear and bright. However, the screen is smaller than most screens. Blue screen experienced often.	Although the screen is small, it's sharp, bright and readable.	The screen is sharp, bright but small sized.	Screen quality is up to par. A common bug is the screen going black and the computer still runs. The screen can seem small and extremely reflective.

Tabela 16 – Sumário do Opinosis para a Proposta - Tela do netbook

O resultado obtido por este método para o tópico da Tabela 13 resultou em um sumário mais longo que aquele gerado pelo TextRank, mas mais curto que aquele gerado pelo LDA. Neste sumário, assim como no gerado pelo LDA, é fornecida a informação que a bateria do netbook possui uma boa duração chegando a 9 horas, diferente das 5 ou 6 horas apresentados pelos "sumários dourados", mas que não o invalida.

Para o tópico dos botões do Amazon Kindle, Tabela 14, o sumário gerado possui informações sobre como os botões são mais difíceis de se acionar por acidente e sobre o bom posicionamento dos botões. Assim, o sumário gerado é satisfatório por conter informações aplicáveis ao tópico tratado e estar alinhado parcialmente com o conteúdo dos "sumários dourados".

Seguindo para o tópico referente à localização do hotel Holliday, Tabela 15, o

sumário gerado cita o ponto de transporte, assim como os "sumários dourados", e como isso facilita a locomoção pela cidade. Novamente, nesta situação o sumário do TextRank é mais próximo aos sumários humanos que a proposta pelo fato de ser mais curto e com as mesmas informações.

Por fim, avaliando o resultado obtido para o tópico da Tabela 16, temos a informação sobre a resolução da tela e a reclamação em relação ao tamanho da tela que, por ser pequena, dificulta a leitura de páginas *Web* completas, além da resolução inadequada entre outros problemas. Esses pontos condizem com o conteúdo dos sumários humanos, em sua maior parte, sendo satisfatório o resultado alcançado.

4.2.5 Conclusões

Nesta subseção, analisamos os resultados obtidos nas Seções 4.2.2, 4.2.3 e 4.2.4 a fim de extrair e discutir as principais características, pontos positivos e negativos de cada técnica utilizada.

Em relação à qualidade dos sumários gerados, comparando com os "sumários dourados", o TextRank teve um desempenho satisfatório apenas no terceiro tópico, por ter informações e comprimento semelhante ao que existe em alguns dos sumários, nos outros casos muitas informações são descartadas, mesmo tendo relevância para o tópico, como ocorreu no sumário da Tabela 5. Por ser um método genérico, dependente do saco de palavras previamente selecionado pelo usuário, os pesos das frases não é distribuída de modo adequado em muitas das vezes, resultando em perda de qualidade na seleção das frases, mesmo que este método apresente sumários consideravelmente mais curtos que os outros métodos testados. Este resultado reforça as conclusões sobre as características do TextRank extraídas na Seção 4.1.6 como reafirma os problemas apontados por trabalhos anteriores como o de Fang *et. al.* [77] e Wang *et. al.*[20].

Os sumários gerados pela técnica LDA, ao contrário dos gerados pelo TextRank, tendem a ser muito longos, como pode ser confirmado ao se comparar o comprimento dos sumários gerados pelo LDA com aqueles gerados pelos outros métodos, com exceção daquele gerado para o tópico da Tabela 9. Conforme explicado neste Capítulo, por El-Kassas *et. al.* (2021), [9] e Wang e Ma (2013) [78], soluções que geram sumários usando a abordagem baseada em tópicos são afetados pelos tópicos escolhidos e a escolha de frases é afetada por consequência. Tendo em mente o fato que apenas as frases mais bem pontuadas são escolhidas, e que a pontuação está relacionada com a quantidade de palavras referente ao tópico que a frase possui, essa característica pode justificar o motivo deste método selecionar frases mais longas que contenham mais palavras contidas no tópico selecionado. Com isso, os sumários são mais longos, mas por vezes com pouco foco no tópico especificado.

A abordagem proposta substitui o conteúdo genérico do saco de palavras, usado pelo TextRank, por termos selecionados pelo LDA. Esta modificação visa aprimorar a seleção de frases ao fornecer ao algoritmo de classificação quais termos devem ser considerados relevantes para o sumário ao mesmo tempo que evita frases longas. Essas características podem ser observadas nos resultados obtidos nas Seções 4.2.2, 4.2.3 e 4.2.4, onde os sumários gerados para os tópicos retiveram informações relevantes, que estão contidas nos sumários humanos, ao tópico ao mesmo tempo que seu comprimento se mantêm entre o TextRank e o LDA. Porém, o problema da seleção dos termos permanece, pois se os termos não forem corretamente selecionados, a seleção de frases é impactada.

Por fim, este experimento demonstra e reforça as características positivas e negativas descritas na Seção 4.1 em relação à tecnicas TextRank, LDA e a abordagem proposta, o comprimento do sumário gerado por cada uma delas e a qualidade das informações contidas em tais sumários.

4.2.6 Desempenho Pela Métrica ROUGE

A métrica ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) ou Substudo Orientado para Recall para Avaliação de Gisting, em tradução livre, é conhecida pelo seu acronimo ROUGE e é uma mética que, de acordo com Lin Chin-Yew, 2004, tem como objetivo determinar a qualidade de um resumo gerado automaticamente comparando-o com outros resumos criados por humano [79].

Neste trabalho são aplicadas as métricas ROUGE-1, ROUGE-2 e ROUGE-L. O conceito básico do ROUGE-1 e ROUGE-2 é o mesmo, sendo a diferença entre eles apenas no conjunto n-gramas que é usado para a avaliação. No ROUGE-1, as palavras são avaliadas individualmente se aparecem na ordem escrita nos "sumários dourados". No ROUGE-2, os bigramas (pares de palavras) sao avaliados e para outras variações ROUGE-n, os n-gramas (sequência de n palavras) são avaliados. O ROUGE-L avalia com base na sequência mais longa de palavras (não necessariamente consecutivas, mas ainda em ordem) que é compartilhada entre o sumário gerado e os sumários "sumários dourados".

Uma vez que um sumário humano é necessário para a comparação, o primeiro experimento, realizado com a base de dados do AirBnb, não pode ser avaliado por esta métrica. Uma possível solução para este problema é a coleta de sumários gerados por meio de voluntários para seleção de, ao menos, um sumário para ser considerado como ideal ou "sumários dourados" como referidos pela base Opinosis.

Ao aplicar as métricas ROUGE para os resultados obtidos para a base de dados Opinosis, os resultados foram agrupados pelo algorítmo utilizado e representados pelas Tabelas 17, 18 e 19.

	ROUGE-1	ROUGE-2	ROUGE-L
Notebook	0.13636363636363638	0.020134228187919462	0.12337662337662336
Kindle	0.07947019867549669	0	0.0662251655629139
Hotel	0.15254237288135591	0.01834862385321101	0.15254237288135591
Tela	0.10207939508506618	0.03461538461538462	0.07894736842105263

Tabela 17 – Avaliação pela métrica ROUGE - TextRank

	ROUGE-1	ROUGE-2	ROUGE-L
Notebook	0.11842105263157895	0.03461538461538462	0.08317580340264649
Kindle	0.07407407407407408	0.008385744234800837	0.05761316872427983
Hotel	0.08661417322834647	0.008016032064128256	0.05511811023622047
Tela	0.10193321616871705	0.003571428571428572	0.06678383128295255

Tabela 18 – Avaliação pela métrica ROUGE - LDA

	ROUGE-1	ROUGE-2	ROUGE-L
Notebook	0.1752577319587629	0.03174603174603175	0.11855670103092784
Kindle	0.11548556430446194	0.01612903225806452	0.06824146981627296
Hotel	0.18430034129692832	0.035211267605633804	0.1501706484641638
Tela	0.15805471124620063	0.0125	0.1094224924012158

Tabela 19 – Avaliação pela métrica ROUGE - Proposta

Avaliando os resultados obtidos, é possivel notar que a abordagem proposta obteve os melhores resultados nos 4 casos para a métrica ROUGE-1. Para a métrica ROUGE-2, a abordagem proposta se sobresai em 2 dos 4 casos, nos outros casos é mediano entre as outras duas abordagens. De modo similar, o resultado obtido com a métrica ROUGE-L, a abordagem proposta se sobresai em 2 dos 4 casos, sendo mediano no restante dos casos.

De modo geral, a abordagem proposta obteve os melhores resultados em 8 das 12 avaliações realizadas, sendo que na métrica ROUGE-1 foi onde ele mais se destacou. Um ponto importante a ser resaltado é que a abordagem proposta conseguiu 91,90% de melhora em relação a segunda melhor abordagem (métrica ROUGE-2 para a linha Hotel), se for considerado apenas onde as três abordagens pontuaram. Se o ROUGE-2 for considerado para o Kindle, a abordagem proposta obteve 92,34% pontos a mais que o LDA.

5 Trabalhos relacionados

Neste capítulo analizaremos os trabalhos de estado-da-arte para sumarizadores de texto extrativos e como o presente trabalho contribui para esta área de pesquisa.

Sharma e Sharma (2021) [80] e Yadav *et. al.* (2022) [81] publicaram pesquisas sobre a sumarização automática de textos onde os métodos extrativos, abstrativos e híbridos foram explorados, suas vantagens e desvantagens explicadas. Ambos os trabalhos apresentam exemplos de aplicações que utilizam estes métodos.

Especificamente, para este trabalho, foi realizada uma revisão das técnicas que se baseiam no uso de tópicos e no uso de grafos, como reportado por Sharma e Sharma (2021) [80].

Neste trabalho, os autores iniciam as análises com o método proposto por Tayal *et. al.* em 2017 [82], na qual o sumarizador é capaz de sumarizar apenas um único documento por vez e as frases que compõem o documento são avaliadas e processadas a fim de se resolver referências anafóricas e as clusterizar de modo que elas possam ser selecionadas com base nas etiquetas previamente treinadas. Para documentos com boa formatação e poucos erros de escrita, gramática ou frases mal-estruturadas este método apresenta bons resultados. No entanto, documentos que não seguem estas condições se provam um desafio para esta proposta.

O trabalho de Steinberger e Jezek (2004) [83] propõe um aprimoramento do LSA que usa os termos mais relevantes do vetor criado como parâmetro para geração do sumário. Nele, os aprimoramentos feitos visam resolver duas desvantagens apontadas por Yihong Gong e Xin Liu (2001) [21], são elas:

- A necessidade de se igualar o número de dimensões com o número de frases no sumário;
- Frases que possuam um valor de indexação alto e que, apesar de serem considerados altos, são escolhidos por não serem os mais altos, mesmo que sejam mais adequados para o sumário.

Seguindo com as técnicas baseadas em grafos, o método proposto por Ferreira *et. al.* (2014)[13] baseia-se em estatística e tratamento linguístico para sistemas sumarizadores multi-documentos, alcançado por meio da clusterização de frases dos múltiplos arquivos. Uma vez feita a clusterização, o algoríto TextRank é usado para classificá-los, para que essa nova abordagem possa selecionar as frases evitando a redundância e aumentando a diversidade. Contudo, Yadav *et. al.*(2022) [81] aponta que como esta proposta trata a

coleção de documentos como um único documento, a diferença nas datas dos diferentes documentos podem levar à ambiguidade.

Na pesquisa realizada por Yadav *et. al.* (2022) [81], os autores apontam ainda o LexRank (Erkan and Radev, 2004) [84] e TextRank (Mihalcea and Tarau, 2004) [38] como sumarizadores populares. Esta afirmação pode ser confirmada pelos próximos trabalhos que ainda os usam em seus métodos ou bases de comparação.

Em 2019, Akhtar *et. al.*, 2019 [85] propõe os métodos TTM (*Two-tiered topic model*) e ETTM (*Enriched Two-tiered topic model*) com base na aplicação de grafos usando TextRank para criar um sumarizador de texto automático. No trabalho citado, os documentos são separados em frases e classificados usando o algoritmo TextRank, o resultado é então usado como parâmetro para o ETTM no processo de geração que resulta no sumário. A desvantagem desta abordagem encontra-se no fato de que ela depende da similariedade das frases para que seja feita a classificação, o que possívelmente a torna uma escolha não muito boa para textos curtos nos quais as frases não se repetem com frequência.

Para este tipo de situação -textos curtos- Wu *et. al.*, 2015[86] apresentam uma combinação do TextRank e LDA para gerar um summarizador para microblogs. Isso é alcançado por meio da conversão de cada frase em um saco de bigramas, então os candidatos são selecionados usando TextRank e TF-IDF (*Term Frequency Inverse Document Frequency*)(ou Frequencia Termo Frequencia Inversa do Documento em tradução livre), estes conjuntos são então filtrados pelo LDA resultando em conjuntos de bigramas-chaves. Estes conjuntos são usados como parâmetro para classificar as frases em dois algoritmos de classificação diferentes, os resultados obtidos são unificados e extraídos.

Uma técnica diferente foi proposta por Belwal *et. al.*, 2021 [87] aplicando ainda o TextRank e o LDA. Nesta proposta o algoritmo cria um sumário utilizando as palavras melhores classificadas dos tópicos gerados pelo LDA em seguida, computando o peso dos vértices no componente TextRank usando dois parâmetros: O primeiro é a similaridade entre os nós que formam as arestas do grafo, o segundo atributo é o peso dado a um componente que representa o quanto uma determinada aresta é semelhante aos tópicos do documento geral para o qual foi incorporada a modelagem de tópicos.

No trabalho de Ali e Malallah (2019) [88], é proposto um sumarizador multi-idiomas utilizando a modelagem de tópicos com o a técnica LDA e uma adaptação do TextRank para que se gerem os sumários. Para testar o sumarizador proposto, foi utilizado uma base de dados que consiste em um conjunto de documentos escritos em 7 idiomas, no entanto apenas 2 são utilizados. Além disso, cada documento usado está escrito completamente em inglês ou árabe e eles são processados separadamente, ou seja, não há ocasião onde um documento em árabe e um em inglês sejam tratados em conjunto. Por esta razão, mesmo que os autores considerem o sistema proposto por eles como múltilingual, de

acordo com a definição que foi apresentada na Seção 2.1.4, o sistema deve ser classificado como mono-idiomático, uma vez que o documento deve ser escrito em um único idioma.

Em comparação com estes métodos, a presente proposta apresenta algumas limitações, são elas:

- O fato de não utilizar alguns termos menos pontuados pode excluí-los da seleção que compõe o conjunto usado pelo TextRank, consequentemente, afetando resultado final;
- O algoritmo foi aplicado em uma base de comentários. É necessário testes com textos mais longos para avaliar o quão eficaz este sistema será nesta condição;
- A dependência de comunicação com a internet, dada por dois fatores: a identificação do idioma das frases e a tradução das mesmas. O que remete aos problemas apontados no Capítulo 3.

Por outro lado, apresenta vários aspectos positivos, listados a seguir:

1. Pode usar documentos escritos em múltiplos idiomas para gerar um sumário em inglês, por meio do uso de um serviço de terceiros para avaliar e traduzir as avaliações que não estejam em inglês para o inglês, tornando possível atuar com qualquer idioma que tenha suporte nesta ferramenta;
2. Os resultados obtidos apontam que os termos selecionados pelo LDA contribuem para que o TextRank selecione frases mais relevantes;
3. Pode atuar de modo a agrupar os documentos para gerar um sumário sobre o conteúdo dos documentos ou tratá-los como documentos individuais e gerar um sumário para cada documento. Isso permite atuar em duas situações diferentes: O primeiro quando todos os documentos são sobre o mesmo assunto e o segundo quando cada documento se refere a um assunto diferente.
4. O algoritmo é flexível, permitindo ao usuário ditar a quantidade de frases a serem selecionadas e quantas palavras cada tópico terá. Isso evita que termos repetidos sejam selecionados para o saco de palavras usado pelo TextRank.

6 Conclusão

A área de pesquisa relacionada aos sumarizadores automático de texto é uma área com grande relevância atualmente, devido ao crescente volume de informações disponibilizadas nas plataformas Web. Uma das abordagens mais populares para a sumarização automática de textos é o TextRank, algoritmo derivado do PageRank.

Porém sua implementação original possui a limitação de não ser capaz de atribuir valores apropriados para cada palavra, o que impacta na qualidade das frases escolhidas para compor o sumário. Uma seleção de palavras-chave mais apropriada pode ter um impacto positivo na qualidade geral do sumário, por meio da escolha de frases mais adequadas.

Este trabalho propôs um sumarizador de textos automático multilingual focado na sumarização de avaliações de usuários por meio da combinação dos algoritmos TextRank e LDA, usando a funcionalidade de seleção de termos do LDA para escolher as palavras-chave usadas para a classificação e seleção de frases realizadas pelo TextRank na composição dos sumários.

A fim de validar a abordagem proposta foram realizados dois experimentos. Ao final do primeiro experimento foi constatado que a qualidade dos sumários gerados pela abordagem proposta é maior que aqueles gerados tanto pelo TextRank, quanto aqueles gerados pelo LDA. Além disso, neste experimento foi possível aplicar a funcionalidade de tradução das frases que funcionou de maneira esperada e satisfatória. O segundo experimento realizado possibilitou a comparação dos sumários gerados com os sumários escritos por humanos, os resultados obtidos neste experimento reforçam os pontos positivos e negativos apresentados no primeiro experimento, validando a proposta deste trabalho.

A comparação do presente trabalho com trabalhos relacionados mostrou que, apesar de poder ser aprimorado, o sistema proposto possui boas qualidades que contribuem para esta área de pesquisa como, por exemplo, a capacidade de gerar sumários com base em documentos escritos em múltiplos idiomas, funcionando mesmo em casos nos quais o documento possua frases escritas em mais de um idioma.

As limitações deste trabalho são: a dependência da comunicação com a Internet para a identificação e tradução das frases e a limitação do tipo de fonte usada nos experimentos.

Portanto, evidencia-se que este trabalho cumpriu seus objetivos, tendo como contribuição o desenvolvimento de um STA flexível, capaz de gerar sumários com base em uma ou mais fontes, onde cada fonte pode possuir frases escritas em mais de um idioma.

Além disso, esta abordagem permite que seja gerado um sumário para todas as fontes fornecidas ou um sumário para cada fonte fornecida.

6.1 Trabalhos futuros

Algumas sugestões de trabalhos futuros são apresentadas nesta Seção. Essas sugestões visam contribuir com o desenvolvimento de novos trabalhos e incentivar novas pesquisas na área de STA:

- Implementar uma função de tradução que não necessite de conexão com a Internet e verificar o impacto desta implementação no sistema;
- Implementar a possibilidade de se usar a data da avaliação como componente do peso da avaliação de modo que comentários mais recentes tenham mais peso que comentários mais antigos;
- Aplicar a abordagem proposta em outros tipos de documentos (artigos, blogs, notícias, etc.).

Referências

1 RAMADHAN, M. R.; ENDAH, S. N.; MANTAU, A. B. J. Implementation of textrank algorithm in product review summarization. In: IEEE. *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*. [S.l.], 2020. p. 1–5. 16

2 KRUMM, J.; DAVIES, N.; NARAYANASWAMI, C. User-generated content. *IEEE Pervasive Computing*, IEEE, v. 7, n. 4, p. 10–11, 2008. 16

3 YAN, Q. et al. E-wom from e-commerce websites and social media: Which will consumers adopt? *Electronic Commerce Research and Applications*, v. 17, p. 62–73, 2016. ISSN 1567-4223. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1567422316300175>. 16

4 OXFORD Advanced Learner's Dictionary. 2023. <https://www.oxfordlearnersdictionaries.com/us/definition/english/summary_1?q=summary>. Accessed: 2023-01-15. 16

5 LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, v. 2, n. 2, p. 159–165, 1958. 16, 19

6 BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, Elsevier, v. 30, n. 1-7, p. 107–117, 1998. 17, 22, 23

7 ZAWARE, S. et al. Text summarization using tf-idf and textrank algorithm. In: IEEE. *2021 5th International conference on trends in electronics and informatics (ICOEI)*. [S.l.], 2021. p. 1399–1407. 17

8 JELODAR, H. et al. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, v. 78, n. 11, p. 15169–15211, 2019. ISSN 1573-7721. Disponível em: <https://doi.org/10.1007/s11042-018-6894-4>. 17, 26

9 EL-KASSAS, W. S. et al. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, v. 165, p. 113679, 2021. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417420305030>. 19, 45, 53

10 CHATTERJEE, N.; MITTAL, A.; GOYAL, S. Single document extractive text summarization using genetic algorithms. In: IEEE. *2012 Third International Conference on Emerging Applications of Information Technology*. [S.l.], 2012. p. 19–23. 19

11 AL-TAANI, A. T. Automatic text summarization approaches. In: *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*. [S.l.: s.n.], 2017. p. 93–94. 19

12 GUPTA, V. K.; SIDDIQUI, T. J. Multi-document summarization using sentence clustering. In: *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*. [S.l.: s.n.], 2012. p. 1–5. 19

13 FERREIRA, R. et al. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, v. 41, n. 13, p. 5780–5787, 2014. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417414001523>. 19, 56

14 MCDONALD, R. A study of global inference algorithms in multi-document summarization. In: AMATI, G.; CARPINETO, C.; ROMANO, G. (Ed.). *Advances in Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 557–564. ISBN 978-3-540-71496-5. 19

15 RAUTRAY, R.; BALABANTARAY, R. C. Cat swarm optimization based evolutionary framework for multi document summarization. *Physica A: Statistical Mechanics and its Applications*, v. 477, p. 174–186, 2017. ISSN 0378-4371. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378437117302121>. 19

16 BHATIA, N.; JAISWAL, A. Automatic text summarization and it's methods - a review. In: *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*. [S.l.: s.n.], 2016. p. 65–72. 19

17 TANDEL, A. et al. Multi-document text summarization - a survey. In: *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. [S.l.: s.n.], 2016. p. 331–334. 20

18 MORATANCH, N.; CHITRAKALA, S. A survey on extractive text summarization. In: *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*. [S.l.: s.n.], 2017. p. 1–6. 20

19 GUI, M. et al. Attention optimization for abstractive document summarization. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. Disponível em: <https://doi.org/10.186532Fv12Fd19-1117>. 20

20 WANG, S. et al. Integrating extractive and abstractive models for long text summarization. In: *2017 IEEE International Congress on Big Data (BigData Congress)*. [S.l.: s.n.], 2017. p. 305–312. 20, 45, 53

21 GONG, Y.; LIU, X. Generic text summarization using relevance measure and latent semantic analysis. In: . New York, NY, USA: Association for Computing Machinery, 2001. (SIGIR '01), p. 19–25. ISBN 1581133316. Disponível em: <https://doi.org/10.1145/383952.383955>. 20, 56

22 GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, v. 47, n. 1, p. 1–66, 2017. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-016-9475-9>. 20

23 WANG, J. et al. A Survey on Cross-Lingual Summarization. *Transactions of the Association for Computational Linguistics*, v. 10, p. 1304–1323, 2022. ISSN 2307-387X. Disponível em: <https://doi.org/10.1162/tacl_a_00520>. 20

24 CHARITHA, S.; CHITTARAGI, N. B.; KOOLAGUDI, S. G. Extractive document summarization using a supervised learning approach. In: *2018 IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*. [S.l.: s.n.], 2018. p. 1–6. 21

25 MOHD, M.; JAN, R.; SHAH, M. Text document summarization using word embedding. *Expert Systems with Applications*, v. 143, p. 112958, 2020. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417419306761>. 21

26 BHAT, I. K.; MOHD, M.; HASHMY, R. Sumitup: A hybrid single-document text summarizer. In: PANT, M. et al. (Ed.). *Soft Computing: Theories and Applications*. Singapore: Springer Singapore, 2018. p. 619–634. ISBN 978-981-10-5687-1. 21

27 HOVY, E.; LIN, C.-Y. *Automated text summarization and the SUMMARIST system*. [S.l.], 1998. 21

28 KIM, S. J.; LEE, S. H. An improved computation of the pagerank algorithm. In: SPRINGER. *Advances in Information Retrieval: 24th BCS-IRSG European Colloquium on IR Research Glasgow, UK, March 25–27, 2002 Proceedings*. [S.l.], 2002. p. 73–85. 22

29 CHUNG, F. A brief survey of pagerank algorithms. *IEEE Trans. Netw. Sci. Eng.*, v. 1, n. 1, p. 38–42, 2014. 22

30 ISHII, H.; SUZUKI, A. Distributed randomized algorithms for pagerank computation: Recent advances. In: _____. *Uncertainty in Complex Networked Systems: In Honor of Roberto Tempo*. Cham: Springer International Publishing, 2018. p. 419–447. ISBN 978-3-030-04630-9. Disponível em: <https://doi.org/10.1007/978-3-030-04630-9_12>. 22

31 PARK, S. et al. A survey on personalized pagerank computation algorithms. *IEEE Access*, v. 7, p. 163049–163062, 2019. 22

32 XING, W.; GHORBANI, A. Weighted pagerank algorithm. In: *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.* [S.l.: s.n.], 2004. p. 305–314. 22

33 LI, J.; WILLETT, P. Articlerank: a pagerank-based alternative to numbers of citations for analysing citation networks. *Aslib Proceedings*, Emerald Group Publishing Limited, v. 61, n. 6, p. 605–618, 2009. ISSN 0001-253X. Disponível em: <https://doi.org/10.1108/00012530911005544>. 22

34 CHEANG, B. et al. Or/ms journals evaluation based on a refined pagerank method: an updated and more comprehensive review. *Scientometrics*, Springer, v. 100, p. 339–361, 2014. 22

35 PATEL, P.; PATEL, K. A review of pagerank and hits algorithms. *Int J Adv Res Eng Sci Technol*, p. 2394–2444, 2015. 22

36 SEN, T.; CHAUDHARY, D. K. Contrastive study of simple pagerank, hits and weighted pagerank algorithms: Review. In: *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*. [S.l.: s.n.], 2017. p. 721–727. 22

37 CHOWDHARY, A.; KUMAR, A. et al. Study of web page ranking algorithms: a review. *Acta Inform. Malaysia*, v. 3, n. 2, p. 01–04, 2019. 23

38 MIHALCEA, R.; TARAU, P. Textrank: Bringing order into text. In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2004. p. 404–411. 24, 25, 57

39 LI, W.; ZHAO, J. Textrank algorithm by exploiting wikipedia for short text keywords extraction. In: IEEE. *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*. [S.l.], 2016. p. 683–686. 24

40 ZHANG, Z.; PETRAK, J.; MAYNARD, D. Adapted textrank for term extraction: A generic method of improving automatic term extraction algorithms. *Procedia Computer Science*, Elsevier, v. 137, p. 102–108, 2018. 24

41 PETASIS, G.; KARKALETSIS, V. Identifying argument components through textrank. In: *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. [S.l.: s.n.], 2016. p. 94–102. 24

42 FAKHREZI, M. F.; BIJAKSANA, M. A.; HUDA, A. F. Implementation of automatic text summarization with textrank method in the development of al-qur'an vocabulary encyclopedia. *Procedia Computer Science*, Elsevier, v. 179, p. 391–398, 2021. 25

43 CHOWDHARY, K. R. Natural language processing. In: _____. *Fundamentals of Artificial Intelligence*. New Delhi: Springer India, 2020. p. 603–649. ISBN 978-81-322-3972-7. Disponível em: <https://doi.org/10.1007/978-81-322-3972-7_19>. 26

44 SUN, S.; LUO, C.; CHEN, J. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, v. 36, p. 10–25, 2017. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253516301117>. 26

45 CHOPRA, A.; PRASHAR, A.; SAIN, C. Natural language processing. *International journal of technology enhancements and emerging engineering research*, Citeseer, v. 1, n. 4, p. 131–134, 2013. 26

46 WU, C. et al. Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, v. 134, p. 104059, 2022. ISSN 0926-5805. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0926580521005100>. 26

47 LAURIOLA, I.; LAVELLI, A.; AIOLLI, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, Elsevier, v. 470, p. 443–456, 2022. 26

48 NATURAL Language Toolkit. 2023. <https://www.nltk.org/>. Accessed: 2023-07-07. 27

49 GENSIM: Topic modelling for humans. 2023. <https://radimrehurek.com/gensim/intro.html>. Accessed: 2023-07-07. 27

50 SPACY. 2023. <https://spacy.io/>. Accessed: 2023-07-07. 27

51 DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, v. 41, n. 6, p. 391–407, 1990. Disponível em: <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199009%2941%3A6%3C391%3A%3AAID-ASI1%3E3.0.CO%3B2-9>. 27, 28

52 BLEI, D.; NG, A.; JORDAN, M. Latent dirichlet allocation. In: DIETTERICH, T.; BECKER, S.; GHAHRAMANI, Z. (Ed.). *Advances in Neural Information Processing Systems*. MIT Press, 2001. v. 14. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf>. 27, 28

53 CHURCHILL, R.; SINGH, L. The evolution of topic modeling. *ACM Computing Surveys*, ACM New York, NY, v. 54, n. 10s, p. 1–35, 2022. 27

54 MUSTAKIM, M. et al. Latent dirichlet allocation for medical records topic modeling: Systematic literature review. In: *2021 Sixth International Conference on Informatics and Computing (ICIC)*. [S.l.: s.n.], 2021. p. 1–7. 28

55 ROSEN-ZVI, M. et al. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*, 2012. 28

56 BLEI, D. M.; LAFFERTY, J. D. Dynamic topic models. In: *Proceedings of the 23rd international conference on Machine learning*. [S.l.: s.n.], 2006. p. 113–120. 28

57 RAMAGE, D. et al. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the 2009 conference on empirical methods in natural language processing*. [S.l.: s.n.], 2009. p. 248–256. 28

58 EVANGELOPOULOS, N.; ZHANG, X.; PRYBUTOK, V. R. Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems*, v. 21, n. 1, p. 70–86, 2012. ISSN 1476-9344. Disponível em: <https://doi.org/10.1057/ejis.2010.61>. 28

59 ZHANG, Y. et al. Delesmell: code smell detection based on deep learning and latent semantic analysis. *Knowledge-Based Systems*, Elsevier, v. 255, p. 109737, 2022. 28

60 SASTRE, F. et al. Method to solve redundant inverse problems based on a latent semantic analysis approach. application to an aerojet engine. *Aerospace Science and Technology*, Elsevier, v. 102, p. 105854, 2020. 28

61 HOBLOS, J. Experimenting with latent semantic analysis and latent dirichlet allocation on automated essay grading. In: IEEE. *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*. [S.l.], 2020. p. 1–7. 28

62 DARWISH, S. M.; MOHAMED, S. K. Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In: SPRINGER. *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4*. [S.l.], 2020. p. 566–575. 28

63 LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, v. 401, n. 6755, p. 788–791, Oct 1999. ISSN 1476-4687. Disponível em: <https://doi.org/10.1038/44565>. 28

64 GAN, J. et al. Non-negative matrix factorization: A survey. *The Computer Journal*, Oxford University Press, v. 64, n. 7, p. 1080–1092, 2021. 28

65 BERRY, M. W. et al. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, Elsevier, v. 52, n. 1, p. 155–173, 2007. 28

66 ZHANG, Z.-Y. Nonnegative matrix factorization: models, algorithms and applications. *Data Mining: Foundations and Intelligent Paradigms: Volume 2: Statistical, Bayesian, Time Series and other Theoretical Aspects*, Springer, p. 99–134, 2012. 29

67 KHERWA, P.; BANSAL, P. Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, v. 7, n. 24, 2019. 29

68 HOFMANN, T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 1999. (SIGIR '99), p. 50–57. ISBN 1581130961. Disponível em: <https://doi.org/10.1145/312624.312649>. 29

69 WANG, M.; HU, F. The application of nltk library for python natural language processing in corpus research. *Theory and Practice in Language Studies*, v. 11, n. 9, p. 1041–1049, 2021. 31

70 ALSHBOUL, Y.; ODAT, W. Text mining to discover design features for cybersecurity tools: The case of password management systems. In: *Proceedings of the 2021 5th International Conference on Software and e-Business*. [S.l.: s.n.], 2021. p. 142–148. 33

71 SUN, W.; TANG, S.; LIU, F. *Examining perceived and projected destination image: A social media content analysis. Sustainability (Switzerland), 13 (6)*. 2021. 33

72 PENG, Q.; SHEN, L. A novel hotel recommendation model. In: ATLANTIS PRESS. *2016 6th International Conference on Machinery, Materials, Environment, Biotechnology and Computer*. [S.l.], 2016. p. 1432–1435. 33

73 MUKRAS, R.; WIRATUNGA, N.; LOTHIAN, R. Selecting bi-tags for sentiment analysis of text. In: SPRINGER. *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. [S.l.], 2007. p. 181–194. 33

74 HASAN, E. et al. An innovative framework for supporting multi-criteria ratings and reviews over big textual data. 2023. 33

75 SHEN, M. et al. Simple yet effective synthetic dataset construction for unsupervised opinion summarization. *arXiv preprint arXiv:2303.11660*, 2023. 33

76 BASTANI, K.; NAMAVARI, H.; SHAFFER, J. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, Elsevier, v. 127, p. 256–271, 2019. 35

77 FANG, C. et al. Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, v. 72, p. 189–195, 2017. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417416306959>. 45, 53

78 WANG, Y.; MA, J. A comprehensive method for text summarization based on latent semantic analysis. In: SPRINGER. *Natural Language Processing and Chinese Computing: Second CCF Conference, NLPCC 2013, Chongqing, China, November 15-19, 2013, Proceedings 2*. [S.l.], 2013. p. 394–401. 45, 53

79 LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: *Text summarization branches out*. [S.l.: s.n.], 2004. p. 74–81. 54

80 SHARMA, G.; SHARMA, D. Automatic text summarization methods: A comprehensive review. *SN Computer Science*, v. 4, n. 1, p. 33, 2022. ISSN 2661-8907. Disponível em: <https://doi.org/10.1007/s42979-022-01446-w>. 56

81 YADAV, D. et al. Feature based automatic text summarization methods: A comprehensive state-of-the-art survey. *IEEE Access*, v. 10, p. 133981–134003, 2022. 56, 57

82 TAYAL, M. A.; RAGHUWANSHI, M. M.; MALIK, L. G. Atssc: Development of an approach based on soft computing for text summarization. *Computer Speech & Language*, v. 41, p. 214–235, 2017. ISSN 0885-2308. Disponível em: <https://www.sciencedirect.com/science/article/pii/S088523081630208X>. 56

83 STEINBERGER, J.; JEZEK, K. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, v. 4, n. 93-100, p. 8, 2004. 56

84 ERKAN, G.; RADEV, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, v. 22, p. 457–479, 2004. 57

85 AKHTAR, N.; BEG, M. S.; JAVED, H. Textrank enhanced topic model for query focussed text summarization. In: IEEE. *2019 Twelfth International Conference on Contemporary Computing (IC3)*. [S.l.], 2019. p. 1–6. 57

86 WU, Y. et al. Tr-lda: a cascaded key-bigram extractor for microblog summarization. *International Journal of Machine Learning and Computing*, IACSIT Press, v. 5, n. 3, p. 172–178, 2015. 57

87 BELWAL, R. C.; RAI, S.; GUPTA, A. A new graph-based extractive text summarization using keywords or topic modeling. *Journal of Ambient Intelligence and Humanized Computing*, Springer, v. 12, n. 10, p. 8975–8990, 2021. 57

88 ALI, Z. H.; MALALLAH, A. P. D. S. Multilingual text summarization based on lda and modified pagerank. *Iraqi Journal of Information Technology. V*, v. 9, n. 3, p. 2018, 2019. 57