Universidade Federal de Pernambuco
Centro de Informática
Programa de Pós-Graduação em Ciência da Computação

Ihago Henrique Lucena e Silva

Sumarização Automática de Textos de Notícias Baseada na Classe do Documento

Recife

2020

Ihago Henrique Lucena e Silva

Sumarização Automática de Textos de Notícias Baseada na Classe do Documento

Este trabalho foi apresentado à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

Área de Concentração: Inteligência computacional

Orientador: Prof. Dr. Rafael Dueire Lins

Recife

2020

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S586s Silva, Ihago Henrique Lucena e
Sumarização automática de textos de notícias baseada na classe do documento / Ihago Henrique Lucena e Silva. – 2020.
122 f.: il., fig., tab.

Orientador: Rafael Dueire Lins.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2020.
Inclui referências e apêndices

1. Inteligência computacional. 2. Processamento de linguagem natural. I. Lins, Rafael Dueire (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2021 - 17

Ihago Henrique Lucena e Silva

Sumarização Automática de Textos de Notícias Baseada na Classe do Documento

Este trabalho foi apresentado à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

Trabalho aprovado em: 14 de fevereiro de 2020.

Prof. Dr. Rafael Dueire Lins
Centro de Informática - UFPE
Orientador

Prof. Dr. Frederico Luiz Gonçalves de Freitas
Centro de Informática - UFPE
Examinador Titular Interno

Prof. Dr. Rafael Ferreira Leite de Mello
Departamento de Computação - UFRPE
Examinador Titular Externo

Recife
2020

Aos meus pais Luciene Ferreira de Lucena e Silva e Normando José Silva, por sempre estarem comigo em todos os momentos.

AGRADECIMENTOS

Primeiramente, agradeço a **Deus** que me deu saúde, forças, calma e esperança para chegar onde eu cheguei.

Agradeço aos meus pais **Luciene Ferreira de Lucena e Silva** e **Normando José Silva** - aos quais eu também dedico este trabalho - que sempre me aconselharam, incentivaram-me a estudar e a me transformar no profissional que eu sou hoje; por todo cuidado, amor e carinho; por serem os pilares da minha vida. Agradeço, também, ao meu dileto irmão **Augusto César de Lucena e Silva** que, mesmo estando longe, sempre foi um grande suporte para mim e toda a nossa família.

Agradeço, principalmente, ao meu orientador, **Rafael Dueire Lins**, uma das pessoas mais grandiosas e pacientes que já me deparei na vida, sempre sábio e presente. A **Rafael Ferreira Leite de Mello**, **Jamilson Batista Antunes**, **Hilário Tomaz Alves de Oliveira**, **Luciano de Souza Cabral** e **Gabriel de França Pereira e Silva** que, juntamente com meu orientador, auxiliaram-me da melhor maneira possível no desenvolvimento deste trabalho. A **Frederico Luiz Gonçalves de Freitas** pelo auxílio durante a pós-graduação e pela disponibilidade em participar da banca examinadora deste trabalho.

Agradeço a **Pedro Henrique Carneiro dos Santos** por todo apoio emocional e psicológico durante a maior parte dessa árdua, longa e monumental jornada.

Agradeço enormemente aos professores do Centro de Informática da Universidade Federal de Pernambuco por toda sabedoria que me transmitiram ao longo da minha graduação e da minha pós-graduação, em especial, aos professores **Adriano Augusto de Moraes**, **Eduardo Antônio Guimarães Tavares** e **George Darmiton da Cunha Cavalcanti**.

A **Eric Rodrigues Borba** e aos meus amigos e também colegas de trabalho **Gabriel Ramos Falconieri Freitas**, **Ana Carolina Vidal González Amoreira**, **João Vitor Almeida Soares** (**in memoriam**) e **Daniel Marques Oliveira** por sempre me incentivarem a concluir o meu mestrado acadêmico.

À **FACEPE** pelo incentivo financeiro durante parte do meu mestrado para a realização desta dissertação.

Por fim, agradeço aos meus amigos da graduação e pós-graduação e a todos que contribuíram direta ou indiretamente neste trabalho e também a quem somente esteve, de alguma forma, presente nessa etapa decisiva da minha vida.

RESUMO

O crescimento exponencial de documentos textuais na *web* nos últimos anos tem forçado os pesquisadores a descobrir formas de economizar tempo e recursos para encontrar informações relevantes. Muitas soluções na área de Processamento de Linguagem Natural vêm sendo cada vez mais empregadas, principalmente para lidar com esse grande volume de informações não estruturadas. Algumas dessas soluções são a classificação automática de documentos e a sumarização automática de textos. Enquanto a sumarização automática tenta produzir um resumo do texto original, ou seja, um recorte com as informações mais úteis do texto em um determinado cenário, a classificação automática visa categorizar um texto, atribuindo-lhe rótulos (identificadores de classes pré-definidos). Logo, se empregadas conjuntamente, essas soluções distintas podem trazer ganhos significativos do contexto de uma para o contexto da outra. Esta dissertação estuda o quanto a categoria de classificação de um documento oferece um bom critério para escolha das técnicas de sumarização mais adequadas, visto que é muito complexo criar um método genérico o suficiente para resumir diferentes tipos de textos. Também foi realizado um mapeamento das combinações de técnicas que produzissem os melhores resumos para cada uma das classes de documentos empregadas. Por fim, é analisada a eficácia da construção de modelos de classificação de documentos a partir dos próprios resumos dos textos originais gerados pelas técnicas de sumarização.

Palavras-chaves: Processamento de Linguagem Natural. Sumarização Automática de Textos. Sumarização Extrativa. Classificação de Documentos.

ABSTRACT

The exponential growth of the number documents on the web in recent years has forced researchers to find automatic ways to sieve information from the massive amount of data available. Many solutions in the area of Natural Language Processing have been increasingly employed, especially to deal with this large amount of text documents. Automatic document classification and text summarization are possibly the most important of them. While automatic summarization attempts to produce a summary of the original text, automatic classification aims to categorize a text into predefined classes. This M.Sc. dissertation analyzes if the classification of a news document is a good criterion for choosing the most appropriate summarization techniques, as it is very complex to create a generic method to summarize all kinds of texts. Besides that, a mapping of the combinations of techniques that produced the best summaries for each class of documents was also performed. Finally, the effectiveness of the construction of document classification models from the summaries of the original texts generated by the summarization techniques is analyzed.

Keywords: Natural Language Processing. Automatic Text Summarization. Extractive Summarization. Document Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BBC	British Broadcasting Corporation
BOW	*Bag-of-Words*
CNN	Cable News Network
HTML	*Hypertext Markup Language*
IA	Inteligência Artificial
IDC	International Data Corporation
KNN	*K-Nearest Neighbors*
ML	*Machine Learning*
NaN	*Not A Number*
PLN	Processamento de Linguagem Natural
RI	Recuperação de Informação
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*
SVM	*Support Vector Machine*
SUMMAC	*TIPSTER Text Summarization Evaluation*
TF-IDF	*Term Frequency - Inverse Document Frequency*
WEKA	*Waikato Environment for Knowledge Analysis*
XML	*Extensible Markup Language*

LISTA DE FÓRMULAS

SUMÁRIO

1 INTRODUÇÃO

Atualmente, graças à omnipresença de dispositivos computacionais, tem se produzido uma grande quantidade de dados, em sua grande maioria textos não estruturados, tais como *e-mails*, textos de *blogs*, publicações de redes sociais, dentre outros. A IDC[1] estima que em 2025 aproximadamente 80% dos dados serão não estruturados (EMANI; CULLOT; NICOLLE, 2015; KORDE; MAHENDER, 2012). É impossível analisar tais acervos digitais e extrair informação sem ferramentas computacionais adequadas. Há várias soluções que vem sendo utilizadas para organizar essa quantidade massiva de informação como a mineração de textos, a classificação automática de documentos e a recuperação de informações (KHAN et al., 2010). Dentre as muitas outras soluções existentes para esse problema, uma tem sido apontada como extremamente promissora: a sumarização automática de textos (FERREIRA et al., 2014).

Quando um leitor consulta o resumo de um artigo científico antes mesmo de ler o conteúdo integral do artigo, ele está se beneficiando da sumarização, ainda que nesse caso ela seja manual, até mesmo porque o próprio leitor pode ler todo o artigo e ao final não achar o documento relevante (MEENA; GOPALANI, 2014). Outro exemplo é quando alguém resume algum evento para outra pessoa objetivando destacar os pontos principais sob a ótica dele, seja um resumo de um filme, de uma partida de esporte, de uma reunião importante ou de um acontecimento qualquer do seu dia a dia. Alunos, de nível de médio ou superior, frequentemente produzem sumários estruturados de materiais de estudo, seja com a finalidade das próprias atividades curriculares ou como um instrumento mais eficiente de estudo (MARTINS et al., 2001; SANTOS; CORDEIRO, 2012).

A sumarização automática de textos, como o próprio nome já deixa claro, é um processo computacional que visa gerar uma versão mais resumida de um ou mais documentos textuais. Esse processo prima pela manutenção das informações essenciais dos documentos originais em um determinado contexto. Há diversas formas e métodos de obter os pontos mais importantes do documento (GOLDSTEIN et al., 1999). Em muitos contextos a sumarização não é simples nem para humanos, uma vez que aspectos subjetivos podem influenciar na valoração da importância de trechos do texto. A sumarização automática é ainda mais complexa, dado que até hoje, não se tem sistemas de sumarização que façam resumos igualmente semelhantes aos realizados por humanos (MEENA; GOPALANI, 2014; GOLDSTEIN et al., 1999). É importante ainda situar que o processo de sumarização automática de textos é mais uma das inúmeras aplicações da área de Processamento de Linguagem Natural, também conhecida como Linguística Computacional, e, portanto,

[1] https://www.idc.com/

se utiliza de conceitos e técnicas de áreas como Inteligência Artificial, Estatística e Probabilidade e Linguística para tentar emular o processo de sumarização humana da forma mais eficiente possível (NENKOVA; MCKEOWN, 2012).

Existem diversas formas de sumarizar automaticamente um texto. Os modelos e sistemas de sumarização fazem uso de técnicas de sumarização que podem privilegiar determinado tipo de informações em um texto a depender da própria implementação da técnica empregada. Por exemplo, há técnicas que privilegiam sentenças que contenham nomes próprios; enquanto outras, as que contenham informações numéricas. De forma superficial, pode-se acreditar que técnicas de sumarização que priorizam sentenças com a presença de números podem ser mais úteis em textos de conteúdo financeiro que as que priorizam sentenças com nomes próprios. Entretanto, pode-se pensar que as técnicas que privilegiam nomes próprios poderiam ser empregadas em textos de conteúdo financeiro para extrair outros tipos de sumário com outros conjuntos de conhecimentos. Ainda no contexto de tipos de sumarização, há as sumarizações extrativas e abstrativas (MARTINS et al., 2001; WANG et al., 2010). Tais conceitos de sumarização, assim como exemplos práticos da diferença desses tipos de sumarização, serão detalhados nos próximos capítulos.

A sumarização automática possibilita a organização das informações da *web* utilizando soluções de PLN, como a classificação de textos (KHAN et al., 2010). É possível conjecturar benefícios da utilização de soluções híbridas que utilizem tanto a sumarização automática quanto a classificação automática de documentos. Por exemplo, a rotulagem dos documentos de texto de sistemas de classificação automático poderia trazer conhecimento para melhorar a escolha do método de sumarização a ser empregada nos documentos, assim como um documento sumarizado poderia, talvez, ser utilizado no lugar do documento original no momento da classificação para diminuir trechos que pudessem causar confusão entre as classes no momento da classificação. Há várias outras hipóteses que poderiam ser investigadas no uso de uma solução de um domínio no outro, mas antes de expandir esse cenário, faz-se necessário explicar brevemente o que é a classificação automática de textos.

A classificação ou categorização é um processo que visa organizar e/ou diferenciar conceitos e objetos utilizando a similaridade ou a dissimilaridade entre os elementos da classe. É um processo que deve prezar pelo princípio da simplicidade, isto é, as pessoas tendem a preferir categorizar em classes cujas percepções sobre elas e/ou a divergência entre elas seja mais simples de entender. Obviamente, classes sem interseções e com determinadas peculiaridades são menos redundantes e podem ser assimiladas de forma mais simples pelas pessoas (POTHOS; CHATER, 2002). As pessoas constantemente utilizam informações de classificação, as vezes até de forma inconsciente, em várias de suas atividades rotineiras. Por exemplo, quando um leitor acessa um site de notícias e busca pela categoria de esportes, ele espera encontrar reportagens e notícias cujo tema central esteja diretamente relacionado a esse assunto. Também é um exemplo quando institutos de

pesquisa realizam censos demográficos estratificando a população de acordo com critérios como renda, idade e sexo para realizarem projeções em cima dos indivíduos dessas classes. Ou quando pesquisadores e cientistas descobrem novas espécies de animais e catalogam de acordo com o seu reino, ordem, família e gênero (hierarquias de classificação) para entender melhor as relações de similaridade entre espécies.

Retomando o cenário da classificação automática de textos, já é possível imaginar que esse processo essencialmente rotula textos em um conjunto de classes pré-definidas. As classes podem seguir uma hierarquia, apresentar intersecções ou serem completamente distintas. Assim como na sumarização automática de textos, também há diversas formas de classificar um texto, seja pelo seu conteúdo, gênero textual ou por qualquer outro critério. Há diversas abordagens para construir sistemas de classificação automática, sejam utilizando sistemas com um conjunto de regras de análise pré-estabelecidas ou utilizando sistemas que aprendem automaticamente e extraem conhecimento e padrões de documentos de *corpus* (KORDE; MAHENDER, 2012).

Logo, vislumbrando todo esse contexto de classificação e sumarização automática de textos combinado com a grande quantidade de informação disponível nos acervos digitais, fica evidente que essa solução pode trazer benefícios para analisar e selecionar de forma eficiente informações dos documentos armazenados *online*. Cada vez mais se fazem necessárias soluções e sistemas adaptáveis a contextos. Por exemplo, está cada vez maior a necessidade de sistemas sumarizadores multi-documentos, multi-idiomas, sumarizadores que consigam processar textos com *tags* de XML e HTML, explorando essas informações associadas nesses documentos, assim como sumarizadores focados em perguntas e sentimentos (KOULALI; EL-HAJ; MEZIANE, 2013; HAHN; MANI, 2000). Nesse âmbito, considerando a dificuldade de criar um método genérico para resumir diferentes tipos de textos (FERREIRA et al., 2014), o quanto a categoria de classificação de um documento consiste em um bom critério para a escolha das técnicas de sumarização extrativa a serem empregadas? Além disso, quais as combinações de técnicas produzem resumos de melhor qualidade para cada uma das classes de documentos empregadas? Por último, qual a eficiência de modelos de classificação de documentos construídos a partir dos próprios resumos dos textos originais gerados pelas técnicas de sumarização?

Este trabalho tem como base o sistema desenvolvido em (SILVA, 2017) , tendo sido estendido para realizar as investigações aqui propostas.

1.1 Objetivos

O objetivo geral desta dissertação é verificar o quanto a categoria de classificação de um documento consiste em um bom critério para escolha das técnicas de sumarização extrativa a fim de produzir os melhores resumos de textos de notícias. Nesse caso, mapeando

essas relações entre as combinações de técnicas e as classes dos documentos. Também é um ponto de interesse aqui, comparar o resultado da classificação de textos de notícias utilizando os documentos completos e utilizando os documentos sumarizados. Visa-se:

- Selecionar técnicas de sumarização extrativa a partir das classes de documentos para produzir os melhores resumos de textos de notícias.
- Combinar técnicas de sumarização extrativas selecionadas para cada classe de documento distinto.
- Comparar o desempenho entre a classificação de textos de notícias utilizando os documentos completos e utilizando os documentos sumarizados.

1.2 Estrutura

Esta dissertação está dividida em cinco capítulos, sendo o primeiro esta introdução.

No Capítulo 2, é apresentada toda fundamentação teórica para o correto entendimento desta dissertação, desde uma breve abordagem da área de PLN e aplicações correlatas até conceitos e trabalhos mais específicos de classificação automática de textos e de sumarização automática de textos. Na sumarização, é introduzida inicialmente a definição desse campo; em seguida, serão apresentadas as diferenças entre os tipos de sistemas de sumarização, o processo de sumarização extrativa em textos, e, por fim, as metodologias de avaliação da qualidade de sumários. Na classificação, é introduzida inicialmente a definição desse campo; em seguida, apresenta-se o processo de treinamento de modelos de classificação e, por fim, as metodologias de avaliação de modelos de classificação.

O Capítulo 3 descreve a metodologia utilizada: desde a seleção dos *corpora* textuais e dos métodos de classificação e sumarização extrativa, até a avaliação dos sumários gerados. Por último, é descrita a abordagem de verificação de relações entre as classes dos documentos e os sumários gerados.

Os resultados obtidos com base nas análises das medidas de avaliação de modelos de classificação, de avaliação de sumários e na verificação das relações são apresentados no Capítulo 4.

Finalmente, no Capítulo 5, serão apresentadas as conclusões, contribuições, desafios e indicativos de possíveis trabalhos futuros.

2 REVISÃO DA LITERATURA

2.1 Processamento de Linguagem Natural

Soluções de mineração de dados, classificação de textos, incluindo aprendizado de máquina, recuperação de informação e sumarização automática de textos têm sido cada vez mais utilizadas para descobrir padrões e formas eficientes de gerenciar as informações nos meios eletrônicos (KHAN et al., 2010).

O Processamento de Linguagem Natural é uma área de estudo que visa alcançar a compreensão da linguagem natural através da utilização de máquinas e representar os documentos semanticamente para melhorar suas mais diversas aplicações como a classificação de documentos e sumarização de textos, por exemplo (KHAN et al., 2010). Muitas aplicações de PLN fazem uso de técnicas estatísticas para desenvolver de forma generalizada modelos de fenômenos linguísticos, baseando-se em exemplos reais do próprio *corpus* de texto sem necessariamente requerer conhecimento linguístico significativo (LIDDY, 2001; GOLDSTEIN et al., 1999). Convém aqui abordar brevemente algumas áreas correlatas à sumarização automática de textos e à classificação automática de textos, como a de recuperação de informação e de mineração de textos (RINO; PARDO, 2003).

A Recuperação de Informação é uma solução de PLN que busca documentos que contêm respostas para alguma *query*. Para alcançar este objetivo, medidas e métodos estatísticos são utilizados para desempenhar o processamento automático de dados dos textos e comparar com a *query* de busca. Recuperação de informação no sentido mais amplo lida com toda a gama de processamento de informações, desde a recuperação de dados à recuperação de conhecimento (KHAN et al., 2010). Engenhos de busca são os tipos de sistemas de RI mais conhecidos, pois são uma das soluções mais utilizadas para buscar documentos específicos na *web* a partir de vocábulos-chaves, em outras palavras, *queries* de busca (RADOVANOVIĆ; IVANOVIĆ, 2008).

Há abordagens que combinam RI com sumarização automática de textos, sendo a mais comum a de sumarizar documentos recuperados sobre um determinado tópico. Mas há também linhas de pesquisa que visam comprovar melhorias nos sistemas de RI indexando textos sumarizados ao invés do conteúdo completo. Portanto, RI ajuda a reunir apenas documentos relevantes sobre um determinado assunto, enquanto que a sumarização seleciona, de fato, as informações mais importantes deles (LLORET; PALOMAR, 2012).

Baseando-se nas aplicações da mineração de dados que buscam encontrar padrões em grandes conjuntos de dados, a Mineração de Textos tem como principal função identificar, em textos, informações específicas. Essas informações não são necessariamente

as informações relativas às ideias principais dos textos, como ocorre na sumarização automática de textos (RINO; PARDO, 2003).

A análise de sentimentos é uma aplicação muito comum da mineração de textos. O objetivo específico dela é tentar extrair o sentimento geral revelado em um texto: positivo, negativo ou algum outro sentimento intermediário. Exemplificando, um usuário da *web*, ao escrever uma publicação em um *blog* reclamando do visor do seu *smartphone*, muito provavelmente está expressando uma opinião negativa. Por outro lado, se ele tivesse feito um elogio do tempo de duração da bateria, muito provavelmente estaria expressando uma opinião positiva. Tomando como base o exemplo dado, a empresa fabricante do aparelho poderia empregar essa análise em opiniões sobre os seus produtos ou até sobre a sua própria marca para descobrir a opinião geral dos consumidores em relação a ela, aos seus produtos ou ainda a componentes dos seus produtos, como foi o caso do visor e da bateria no exemplo. Comumente, setores como o de *marketing* e de gestão de relacionamento com o consumidor se beneficiam muito desse tipo de análise. Empresas também utilizam isso para conhecer a reputação de seus concorrentes. (YI et al., 2003)

O processamento de linguagem natural fornece teorias e implementações para uma variedade de problemas. De fato, qualquer aplicação que faça uso de algum processamento de texto provavelmente pertence à área de PLN. Há ainda muitas outras aplicações de PLN não mencionadas aqui como extração de informação, tradução de máquina, sistemas de diálogo, entre outras (LIDDY, 2001).

2.2 Sumarização Automática de Textos

Como já mencionado, a sumarização automática é o processo de selecionar as informações mais importantes de um conjunto de fontes, podendo ser um único texto ou um *corpus*, para produzir uma versão resumida (MARTINS et al., 2001; MANI et al., 2002). Os textos podem ser de quaisquer tipos: notícias, artigos científicos, publicações em *blogs*, críticas de filmes, atas de reunião, dentre outros. Também podem ser escritos em qualquer idioma sem necessidade de uma escrita formal ou casta. Obviamente, os métodos de sumarização, incluindo alguns pré-processamentos, podem variar de idioma para idioma, sendo a grande maioria existente para o inglês (CABRAL et al., 2014).

Embora a pesquisa sobre sumarização automática tenha começado no final dos anos 50 do século XX, a área não teve muitos progressos significativos até a década de 90 do século XX, quando foi desenvolvido o SUMMAC (MARTINS et al., 2001). O SUMMAC possibilitou a geração de vários novos métodos para avaliar resumos, estabelecendo que a sumarização era muito eficaz nas tarefas que usam artigos de notícias. (MANI et al., 2002)

Todavia, o crescente volume de texto eletrônico e as recentes pesquisas tem alimentado bastante esse campo, principalmente com o avanço de outras soluções de PLN

(SAGGION, 2008). Várias novas abordagens, incluindo métodos fundamentados de vetores de características, estruturados de grafos e com utilização de *clusters*, foram elaboradas visando melhorar ou até sanar muitos dos problemas da sumarização, especialmente os relacionados à coesão e coerência das sentenças (SARANYAMOL; SINDHU, 2014; MARTINS et al., 2001).

2.2.1 Tipos de Sumarização

No nível mais básico, os resumos se diferem com relação à sua estrutura resultante produzida por técnicas de sumarização extrativas ou abstrativas. As técnicas de sumarização extrativa geram resumos compostos pelas principais sentenças originais do texto sem necessariamente ter uma formação textual coesa entre essas sentenças, isto é, pode se entender que essas técnicas extraem uma lista de tópicos do texto original com exatamente as mesmas palavras (HAHN; MANI, 2000; MARTINS et al., 2001; GAMBHIR; GUPTA, 2017). As técnicas de sumarização abstrativa, por outro lado, produzem resumos mais sofisticados, eliminando redundâncias e esclarecendo o seu contexto, geralmente trazendo material que enriquece o conteúdo do próprio texto de origem. Há ainda os resumos críticos ou resumos de crítica que apresentam não apenas um resumo do texto original de forma interligada, mas uma análise comparativa e até opinativa do conteúdo original do texto com o contexto de outros trabalhos relacionados a esse conteúdo, na área específica em foco. (HAHN; MANI, 2000; FERREIRA et al., 2013b).

Diferente da sumarização extrativa, na sumarização abstrativa a informação presente nos documentos é analisada, reformulada e interligada de várias formas. Apesar de visar melhorar a coerência das sentenças produzidas (FERREIRA et al., 2013b; FERREIRA et al., 2014), isso acarreta em uma dificuldade a mais que a extrativa, uma vez que reformular e/ou interligar conteúdos de sentenças não é algo tão simples (MEENA; GOPALANI, 2014). É possível visualizar a diferença entre essas estruturas na prática com os textos das Tabelas 1, 2 e 3. Esses três textos, nessa ordem, representam o texto original, o texto sumarizado com técnica extrativa com 25% de taxa de compressão e o texto sumarizado com técnica abstrativa com 15% de taxa de compressão. Esses exemplos, intitulados de "*The Gettysburg Address*", foram todos retirados de (MANI, 2001).

Um resumo também pode ser genérico ou focado no usuário (baseado em uma *query* de busca) (WANG et al., 2010; HAHN; MANI, 2000; FERREIRA et al., 2014). As técnicas de resumo genérico geralmente utilizam medidas de pontuação e análise interna do próprio texto. Elas não recebem informações exógenas com relação ao conteúdo do texto, somente o tamanho do resumo desejado, seja expresso pelo número máximo de sentenças ou por um percentual de compactação do texto original.

Em casos onde se deseja dar ênfase ao usuário, algumas abordagens incluem favorecer termos relacionados ao interesse dele, baseando-se na *query* de busca. Dessa

1. Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. **2.** Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. **3.** We are met on a great battlefield of that war. **4.** We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that this nation might live. **5.** It is altogether fitting and proper that we should do this. **6.** But, in a larger sense, we cannot dedicate...we cannot consecrate...we cannot hallow... this ground. **7.** The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. **8.** The world will little note nor long remember what we say here, but it can never forget what they did here. **9.** It is for us, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. **10.** It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth.

Tabela 1 – Exemplo de Texto original com Sentenças Numeradas: parte do discurso de Abraham Lincoln no cemitério de Gettysburg

1. Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. **2.** Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. **7.** The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract.

Tabela 2 – Texto com Sumarização Extrativa com Taxa de 25% de Compressão com Sentenças Numeradas

This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.

Tabela 3 – Texto com Sumarização Abstrativa com Taxa de 15% de Compressão

maneira, é mais provável obter um resumo adaptado à necessidade específica do usuário do que na abordagem genérica (LLORET; PALOMAR, 2012; HAHN; MANI, 2000). Atualmente, o resumo baseado em consultas tem atraído muita atenção devido à sua aplicabilidade imediata em sistemas sociais comerciais, como atendimento automático ao cliente (FERREIRA et al., 2014). Alguns sistemas desse tipo utilizam as *queries* de busca de forma implícita, recolhendo informações do usuário de diversas formas e as mapeando em *queries* internamente sempre que o usuário requisita alguma informação, por exemplo (LLORET; PALOMAR, 2012).

A sumarização não precisa ser realizada apenas com um documento por vez, há também a possibilidade de se resumir uma vasta coleção de documentos de forma unificada. A sumarização de múltiplos documentos estende métodos da sumarização de documentos únicos para uma coleção de documentos por vez. Cada método envolve a análise de cada documento na coleção e, em seguida, mescla as informações entre documentos, sintetizando-as (HAHN; MANI, 2000). Resumo de vários documentos tem problemas de sobrecarga de texto, principalmente em coleções de textos sobre o mesmo assunto, pois muitos documentos compartilham tópicos e sentenças semelhantes, muito mais até que em um único documento de texto. Por exemplo, a matéria da notícia de um incidente terrorista poderia aparecer em formas distintas em fontes diferentes, entretanto, relatando um mesmo acontecimento. Portanto, nesses casos, métodos que evitem redundância são fundamentais (FERREIRA et al., 2014; HAHN; MANI, 2000).

Com base no idioma, existem três tipos de sumários: multilíngue, monolíngue ou ainda translíngue (*cross-lingual*). Considera-se um sistema de resumo monolíngue quando o idioma do documento de origem e destino é o mesmo. Quando o documento de origem está em vários idiomas como inglês, alemão e francês e o resumo também é gerado nesses idiomas, então se considera esse sistema como multilíngue. Por último, se o documento de origem estiver em um idioma e o resumo gerado em qualquer outro idioma, então considera-se esse sistema como translíngue (GAMBHIR; GUPTA, 2017; CABRAL et al., 2014).

2.2.2 Etapas da Sumarização Extrativa

Em geral, os sumários extrativos são produzidos através de alguns passos sequenciais: *i.* Pré-processamento; *ii.* Pontuação e ordenação; *iii.* Seleção; (MEENA; GOPALANI, 2014). Esses passos para geração de extratos de forma genérica a partir de um documento de texto estão exibidos na Figura 1.

Normalmente na etapa de pré-processamento são utilizados procedimentos como segmentação de sentenças em *tokens*, *case folding*, eliminação de *stopwords* e *stemming* para criar representações intermediárias do documento; em seguida, são atribuídas pontuações às sentenças dependendo da implementação do método de sumarização como frequência

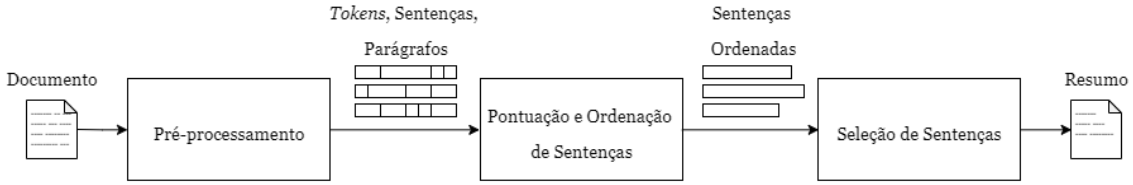

Figura 1 – Esquema Completo de Sumarização Extrativa

de termos do texto, semelhança com o título, presença de nomes próprios, localização da sentença no texto, etc. (MEENA; GOPALANI, 2014). Então, é realizada a ordenação das sentenças com base nas suas pontuações e, finalmente, a elaboração do sumário com base na seleção das sentenças de maior pontuação (NENKOVA; MCKEOWN, 2012; MEENA; GOPALANI, 2014). Geralmente essa seleção escolhe n sentenças de maior pontuação, onde n é um número de sentenças já pré-definido que o sumário deverá apresentar. Mas também há casos que se utiliza um percentual de compactação do texto para determinar o tamanho do sumário a ser produzido, logo, nesse caso, o número de sentenças pode variar bastante a depender do tamanho das próprias sentenças (MANI et al., 2002).

Na segmentação das sentenças, ou melhor, na tokenização de sentenças um documento é tratado como uma sequência e, em seguida, particionado em uma lista de *tokens* (KHAN et al., 2010). Com essa lista, a estrutura do texto torna-se bem mais simples para outros processamentos posteriores como *stemming* e remoção de *stopwords*. *Case folding* é um processamento que converte todos as letras de uma sentença ou *token* para o formato maiúsculo ou minúsculo, sendo mais comum esse último caso (MARTINS et al., 2001). É muito útil a aplicação de *Case folding* para que as técnicas de sumarização possam tratar uma mesma palavra igualmente independente da sua tipografia.

Stopwords são palavras funcionais idiomáticas que ocorrem com frequência nos textos, geralmente artigos, preposições e conjunções. Por exemplo, algumas palavras na língua inglesa são "*a*", "*an*", "*of*", "*the*", etc. (DALAL; ZAVERI, 2011). O objetivo de sua remoção é remover ruídos, uma vez que essas palavras na grande maioria dos casos são vazias no contexto geral, não agregando no processo de sumarização.

Para exemplificar como fica um texto após a remoção de *stopwords*, será mostrado aqui um trecho de um texto antes e depois da aplicação desse processamento. O texto da Tabela 4 mostra um trecho, intitulado de "*Theresa May strikes a Brexit deal*", de uma publicação[1] da BBC[2] antes da remoção de *stopwords*, enquanto que o texto da Tabela 5 mostra esse mesmo trecho após remoção das *stopwords*.

Stemming é a ação de reduzir uma palavra ao seu radical (*stem*). Alguns vocábulos podem chegar até a um terço do seu tamanho original com esse processo. Há *stemmers*

[1] Disponível em https://www.bbc.com/news/uk-46566544
[2] https://www.bbc.com/

1. Our best-read story of the year, overwhelmingly, was the live coverage that ran over five days in November, looking at the fallout after Theresa May struck a Brexit deal with EU members. **2.** Much of her party went into open revolt.

Tabela 4 – Exemplo 1 de Texto da BBC com Sentenças Numeradas

1. Our best-read story year, overwhelmingly, live coverage ran five days November, looking fallout Theresa May struck Brexit deal EU members. **2.** Much party went open revolt.

Tabela 5 – Exemplo 1 de Texto da BBC sem *stopwords* com Sentenças Numeradas

1. Our best-read stori of the year, overwhelmingli, wa the live coverag that ran over five day in Novemb, look at the fallout after Theresa May struck a Brexit deal with EU member. **2.** Much of her parti went into open revolt.

Tabela 6 – Exemplo 1 de Texto da BBC com *stemming* com Sentenças Numeradas

para vários idiomas e até implementações diferentes dentro do mesmo idioma. Na língua inglesa, por exemplo, o *stemmer* de Potter é muito popular (DALAL; ZAVERI, 2011). Para ilustrar o funcionamento dessa técnica utilizando o *stemmer* de Potter, seguem alguns exemplos: as palavras "*plays*", "*playing*"e "*played*"seriam reduzidas para "*play*"; as palavras "*studies*"e "*studying*"seriam reduzidas para "*studi*"; a palavra "*sophisticated*", para "*sophist*"; a palavra "*generalizations*", para "*gener*"(DALAL; ZAVERI, 2011). O objetivo do *stemming* é que palavras com o mesmo radical sejam facilmente relacionadas sem necessidade da análise delas nas suas mais diversas formas flexionadas.

Para demonstrar melhor o funcionamento de um *stemmer*, será mostrado aqui um trecho de um texto antes e depois da aplicação da técnica, utilizando ainda o mesmo exemplo do texto da Tabela 4. Como já visto, o texto da Tabela mostra um trecho de um texto original antes da aplicação de *stemming*, enquanto que o texto da Tabela 6 mostra esse mesmo trecho após a aplicação de *stemming*.

A remoção de *stopwords* e aplicação de *stemming* um após o outro está exemplificada na Tabela 7. Outro exemplo mais completo consiste no texto da Tabela 8 que também mostra um trecho da mesma publicação do texto da Tabela 4, mas dessa vez intitulado de "*Harry and Meghan get married*". Por fim, o texto da Tabela 9 apresenta esse mesmo

1. Our best-read stori year, overwhelmingli, live coverag ran five day Novemb, look fallout Theresa May struck Brexit deal EU member. **2.** Much parti went open revolt.

Tabela 7 – Exemplo 1 de Texto da BBC sem *stopwords* e com *stemming* com Sentenças Numeradas

1. "You look amazing." **2.** Those were the words Prince Harry said to Meghan Markle as they stood together at the altar in Windsor Castle on 19 May. **3.** Their wedding took place at St George's Chapel, and it was attended by 600 guests, including Oprah Winfrey, the Beckhams and Elton John, who performed at the wedding reception. **4.** Meghan wore a dress by British designer Clare Waight Keller and a diamond tiara she borrowed from the Queen, while Harry had gained permission from the Queen to keep his beard (it's customary for anyone in military uniform to be clean-shaven). **5.** Among the most memorable moments of the ceremony: the animated address by American bishop Michael Curry (and the reactions from some of the congregation), the performance by the Kingdom Choir of Ben E King's soul classic Stand by Me, and Prince Charles walking Meghan down the aisle after her father was too unwell to attend. **6.** And then, in October, came some more happy news: the Duke and Duchess of Sussex, titles the couple took after marrying, were expecting a baby, due in Spring 2019.

Tabela 8 – Exemplo 2 de Texto da BBC com Sentenças Numeradas

1. "you look amaz." **2.** Those word princ Harry said Meghan Markle stood togeth altar Windsor castl 19 May. **3.** Their wed took place St George's Chapel, attend 600 guest, includ Oprah Winfrey, Beckhams Elton John, perform wed recept. **4.** Meghan wore dress british design Clare Waight Keller diamond tiara borrow Queen, Harry gain permiss queen keep beard (customari anyon militari uniform clean-shaven). **5.** Among memor moment ceremoni: anim address american bishop Michael Curry (reaction congreg), perform Kingdom Choir Ben E King's soul classic Stand Me, princ Charles walk Meghan aisl father unwel attend. **6.** And, Octob, came happi news: duke duchess Sussex, titl coupl took marri, expect babi, due spring 2019.

Tabela 9 – Exemplo 2 de Texto da BBC sem *stopwords* e com *stemming* com Sentenças Numeradas

trecho após a remoção de *stopwords* e aplicação de *stemming*.

Uma vez que uma representação intermediária é construída após os pré-processamentos, cada sentença recebe uma pontuação que indica sua importância no texto. Os métodos de pontuação são classificados de acordo com três abordagens: baseados em palavras, baseados em sentenças e baseados em grafos. Vários métodos para essas três abordagens foram propostos por diferentes pesquisadores (MEENA; GOPALANI, 2014). Em suma, todos fornecem como uma saída uma pontuação entre 0 e 1 para cada sentença do texto. Inclusive, algumas implementações empregam até uma etapa de normalização de pontuação (FERREIRA et al., 2014). Para algumas abordagens mais complexas que fazem uso de

vários métodos de pontuação, o peso de cada sentença é determinado ponderando as pontuações dos diferentes indicadores (NENKOVA; MCKEOWN, 2012).

Nos métodos baseados em palavras, os primeiros a surgir na literatura, cada palavra recebe uma pontuação e o peso de cada sentença é determinado pela soma de todas as pontuações de suas palavras constituintes. Já os métodos da abordagem baseada em sentenças geralmente analisam as características da própria frase, como a presença de expressões de sinalização. Em inglês, algumas expressões de sinalização bem comuns são, por exemplo, "*in conclusion*", "*in summary*" e "*the most important*". Por fim, nos métodos baseados em grafos, as pontuações das sentenças são determinadas pelas suas relações. Nesse caso, quando uma sentença se refere a outra, é gerada uma ligação com um peso associado entre elas (FERREIRA et al., 2014; FERREIRA et al., 2015).

Após a atribuição de pontuação às sentenças, é realizada uma ordenação decrescente das sentenças com melhores pontuações às de piores pontuações. Nessa etapa, a maioria das abordagens de sumarização procede selecionando o conteúdo sentença por sentença, de acordo com a ordem de pontuação. Alguns processos de verificação de similaridade entre as sentenças escolhidas também costumam ser empregados para evitar a inclusão de frases repetitivas (NENKOVA; MCKEOWN, 2012).

Após todos esses procedimentos, tem-se um resumo com as sentenças mais importantes extraídas do documento original. É esperado que esse resumo seja relativamente menor que o texto original e que contenha as suas ideias principais. Obviamente, as etapas foram explicadas aqui não levando em consideração a sumarização de múltiplos documentos e nem a focada no usuário. Nesses casos, as etapas do processo seriam bem similares, apenas apresentando algumas pequenas adequações.

2.2.3 Métodos de Sumarização Extrativa

Aqui serão descritos alguns dos métodos de pontuação mais utilizados na literatura de cada uma das três abordagens: baseados em palavras, baseados em sentenças e baseados em grafos. Cada método tem o seu funcionamento próprio e pode gerar resumos bem diferentes. Abaixo, são exibidos esses métodos divididos perante a essas três abordagens mencionadas.

2.2.3.1 *Métodos de Pontuação de Palavras*

- *Word Frequency* (Frequência de Palavra): as sentenças são pontuadas como a soma das pontuações dos seus termos constituintes. Quanto mais frequente é um termo no texto, maior será sua pontuação e, consequentemente, o seu impacto na pontuação final das sentenças que o contêm. A suposição deste método é que quanto maior a frequência de uma palavra no texto, maior a probabilidade dela indicar o assunto do

texto. (MEENA; GOPALANI, 2014; FERREIRA et al., 2014; FERREIRA et al., 2015; FERREIRA et al., 2013a).

- TF-IDF (Frequência do Termo – Inverso da Frequência no Documento): a pontuação de cada palavra w é dada pela Fórmula 1. Onde tf é a frequência do termo w na sentença, df é o número de sentenças que a palavra w ocorre em todo o documento, n é o número total de sentenças e $tfidf$ calcula a pontuação atribuída pelo método à palavra w.

$$tfidf(w) = tf \times log(\frac{n}{df}) \tag{1}$$

  Este método assume que palavras mais específicas em uma determinada sentença são relativamente mais importantes. Ele foi proposto para remover o impacto de termos com frequências mais altas que geram resumos incorretos (MEENA; GOPALANI, 2014; FERREIRA et al., 2014; FERREIRA et al., 2013a).

- *Upper Case* (Letras Maiúsculas): atribui pontuações mais altas a palavras que contêm uma ou mais letras maiúsculas. A pontuação de cada sentença s é dada pela Fórmula 3. Onde, d é o documento a ser analisado, cw calcula o número de palavras com letras maiúsculas em uma sentença, tw calcula o número total de palavras em uma sentença, r_{cw} (mostrado na Fórmula 2) calcula a proporção entre o total de palavras com letras maiúsculas em uma sentença e o total de palavras na mesma sentença e uc calcula a pontuação da sentença de acordo com o método em questão (MEENA; GOPALANI, 2014; FERREIRA et al., 2014).

$$r_{cw}(s) = \frac{cw(s)}{tw(s)} \tag{2}$$

$$uc(s) = \frac{r_{cw}(s)}{\max\limits_{s_i \in d}(r_{cw}(s_i))} \tag{3}$$

- *Proper Noun* (Nomes Próprios): é uma especialização do método de *Upper Case*, no qual são levados em consideração apenas palavras que comecem com letras maiúsculas ou outros substantivos como por exemplo, "*she*", "*he*", etc. A ideia deste método é que geralmente as sentenças que contêm um número maior de substantivos são mais importantes (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

- *Word Co-occurrence* (Co-ocorrência de Palavras): mede a probabilidade de dois ou mais termos ocorrerem em uma dada sequência (mesma maneira e posição), ou seja,

este método fornece pontuações mais elevadas para sentenças com mais frequência de co-ocorrência de palavras. Uma maneira simples de implementá-lo é através do uso de *N-grams*, que são uma sequência contínua de N itens de um determinado seguimento de texto. (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a; FARHOODI; YARI, 2010).

- *Lexical Similarity* (Similaridade Léxica): relaciona sentenças através da semelhança entre suas palavras e, também, sentenças que empregam palavras sinônimas ou outras relações semânticas. Este método baseia-se no pressuposto de que sentenças importantes são identificadas por cadeias fortes (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

2.2.3.2 *Métodos de Pontuação de Sentenças*

- *Resemblance to the Title* (Semelhança com o Título): mede a sobreposição entre os termos de uma sentença com as do título do documento. Neste caso, sentenças semelhantes ao título e sentenças que contêm as palavras do título são consideradas importantes. O cálculo deste método para cada sentença s é dado pela Fórmula 4. Onde, tw é número de palavras do título presentes na sentença s, t é o número de palavras do título e rt calcula a pontuação da sentença para este método (FERREIRA et al., 2013a; FERREIRA et al., 2014).

$$rt(s) = \frac{tw}{t} \tag{4}$$

- *Sentence Centrality* (Centralidade da Sentença): verifica a sobreposição de vocabulário entre uma sentença s e todas as outras sentenças no documento. É importante observar que essa abordagem não utiliza nenhum tratamento semântico como similaridade léxica. De forma semelhante ao método *Resemblance to the Title*, o cálculo da centralidade é dado pela Fórmula 5. Onde, dw é número de palavras de outras sentenças do documento presentes na sentença s, d é o número de palavras do documento e sc calcula a pontuação da sentença pelo método (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

$$sc(s) = \frac{dw}{d} \tag{5}$$

- *Sentence Position* (Posição da Sentença): a posição de uma sentença indica sua importância; as mais importantes tendem a estar no início ou no final do texto (em alguns casos). Neste método, uma sentença é pontuada de acordo com sua posição

no parágrafo, sendo considerada a pontuação máxima como 5. Exemplificando, a primeira sentença de um parágrafo tem um valor de pontuação de 5/5, já a segunda sentença tem uma pontuação de 4/5, e assim por diante. Quando a posição da sentença no parágrafo ultrapassa 5, é atribuída a ela a pontuação 0, pois nesse caso, a posição dessas sentenças não é mais significativa (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a; FARHOODI; YARI, 2010).

- *Cue Phrases* (Palavras Sinalizadoras): a pontuação de uma sentença s é dada com base na existência de palavras sinalizadoras nela. Um conjunto de expressões ou palavras-chave como "*in summary*", "*in conclusion*", "*important*", "*in particular*", "*our investigation*", "*the paper describes*", etc. deve ser preparada para a utilização desse método. A ideia é que sentenças com essas palavras possam ser bons indicadores de um conteúdo significativo na sentença. De forma semelhante aos métodos *Resemblance to the Title* e *Sentence Centrality*, o cálculo deste método é dado pela Fórmula 6. Onde, cs é número de palavras de palavras sinalizadoras presentes na sentença s, cd é o número de palavras sinalizadoras presentes no documento e cp calcula a pontuação da sentença para esse método (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

$$cp(s) = \frac{cs}{cd} \tag{6}$$

- *Numerical Data* (Inclusão de Dados Numéricos em Frases): as sentenças que contém dados numéricos (data do evento, transação monetária, porcentagem de danos, etc.) são consideradas mais importantes. Desta forma recebem pesos maiores e provavelmente serão incluídas no resumo final (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

- *Sentence Length* (Comprimento da Sentença): sentenças muito longas ou muito curtas devem ser evitadas, portanto limites podem ser fixados para que essas sentenças sejam penalizadas na atribuição da pontuação (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

2.2.3.3 *Métodos de Pontuação de Grafos*

- *Text Rank* (Classificação do Texto): extrai palavras-chave importantes de um documento para determinar a sua relevância através de um modelo baseado em grafos. Sentenças com uma quantidade maior de palavras-chave obtêm pontuações mais elevadas (FERREIRA et al., 2014; FERREIRA et al., 2013a).

- *Bushy Path of the Node* (Caminho Expresso no Nó): o caminho espesso de um nó (sentença) em um grafo é definido como o número de ligações que o conectam a outros nós no grafo. De uma certa forma é um método semelhante ao *Lexical Similarity*, entretanto enquanto o último considera sentenças com similaridades, esse método aqui considera o número de sobreposição de sentenças (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

- *Aggregate Similarity* (Similaridade Agregada): mede a importância de uma ligação entre sentenças. Isto é, em vez de contar o número de ligações que conectam um nó a outros nós, como ocorre no método *Bushy Path*, a similaridade agregada soma os pesos (semelhanças) das ligações (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a).

2.2.4 Avaliação de Sumários

A avaliação automática de resumos facilitou bastante o desenvolvimento de sistemas de sumarização (OWCZARZAK et al., 2012). Entretanto, há vários aspectos a serem considerados quando se discute a avaliação de sumários, principalmente nos aspectos referentes às divergências entre avaliações manuais, realizadas por humanos, e as automáticas, realizadas por máquinas.

Segundo Goldstein, sistemas de sumarização humanos (manuais) são difíceis de serem projetados e, ainda mais, avaliados, pois os documentos gerados podem apresentar diversas variações como comprimento da sentença e estilo de escrita (GOLDSTEIN et al., 1999). Também há outro fator que precisa ser considerado no momento da avaliação manual de resumos: a subjetividade do julgamento humano, isto é, seres humanos podem avaliar de formas diversas diferentes métricas de qualidade como coerência, concisão, legibilidade e o próprio conteúdo dos textos (LIN, 2004). Deve-se também considerar o fato da não escalabilidade desse tipo de avaliação. Em outras palavras, para avaliar resumos de grandes documentos de texto, seriam exigidas muitas horas de esforço, havendo ainda a possibilidade de erro na determinação da qualidade do sumário.

Dentre as ferramentas de avaliação de resumos automática, o ROUGE (LIN, 2004) é a mais utilizada. O ROUGE dispõe de medidas para avaliar automaticamente a qualidade de resumos gerados, simplesmente comparando com outros (considerados resumos ideias), geralmente criados por humanos. Essas medidas, que inicialmente foram tratadas com muita desconfiança por alguns pesquisadores, mostraram-se como métricas que se correlacionam muito ao julgamento humano.

Explicando brevemente, o ROUGE tem 4 métodos básicos: ROUGE-N, ROUGE-L, ROUGE-W e ROUGE-S. O ROUGE-N é um método de casamento de *n-grams* entre um resumo e um conjunto de resumos de referência. Já o ROUGE-L mede, ao invés de *n-grams*,

a maior subsequência de palavras correspondentes. O ROUGE-W funciona de forma similar ao ROUGE-L, entretanto ele mede a maior subsequência de palavras correspondentes de forma ponderada que favoreçam subsequências de palavras correspondentes consecutivas. Finalmente, o ROUGE-S mede a sobreposição de conjunto de palavras que podem ter lacunas entre elas. Informações mais detalhadas sobre esses métodos são encontrados em (LIN, 2004). Dentre os métodos apresentados, o foco será no ROUGE-N, especificamente no ROUGE-2 (ROUGE-N com N=2), uma vez que os próprios experimentos do artigo (LIN, 2004) demonstram um melhor desempenho dele em relação aos demais.

O cálculo do ROUGE-N é dado pela Fórmula 7. Onde n representa o tamanho de cada um dos *n-grams* ($gram_n$), RS é o conjunto de sumários de referência e $count_{match}(gram_n)$ é o número máximo de *n-grams* de co-ocorrência em um resumo avaliado e o conjunto de resumos de referência (LIN, 2004).

$$ROUGE_N = \frac{\sum_{s \in RS} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{s \in RS} \sum_{gram_n \in S} count(gram_n)} \tag{7}$$

Obviamente, o ROUGE não consegue automatizar completamente o processo, uma vez que para o processo de avaliação automático ser iniciado, é necessário chegar a um resumo linha de base, normalmente um *gold summary*, o qual os sumários a serem comparados deverão se aproximar (OWCZARZAK et al., 2012; LIN, 2004). Interessante pontuar também que, apesar das pesquisas na área de sumarização de textos terem se iniciado desde a década de 50 do século passado, não há até o momento nenhum sistema disponível que possa gerar os *gold summaries*, isto é, essa tarefa ainda é realizada manualmente por humanos (MEENA; GOPALANI, 2014).

Outra medida a ser apresentada é a medida *MATCHES*. Ela consiste em uma avaliação do número de casamentos entre as sentenças de um resumo gerado automaticamente com as sentenças de um resumo de referência. Essa medida já foi utilizada em (SILVA, 2017; FERREIRA et al., 2014; FERREIRA et al., 2013b) e constitui uma medida de qualidade dos resumos gerados. O cálculo dessa medida é dado pela Fórmula 8. Onde ss é o número de sentenças presentes simultaneamente no sumário avaliado e no *gold summary* e sgs é o número total de sentenças presentes no *gold summary* (SILVA, 2017).

$$MATCHES = \frac{ss}{sgs} \tag{8}$$

Independente do tipo de avaliação empregue, avaliar adequadamente a qualidade dos sumários produzidos é uma ótima forma para eleger os melhores sistemas de sumarização para um determinado contexto.

2.3 Classificação Automática de Textos

A classificação automática de textos, como já mencionado brevemente, visa atribuir, de forma automatizada, um ou mais rótulos a documentos textuais escritos em linguagem natural (VILLENA-ROMÁN et al., 2011). Para a realização da classificação automática, podem ser utilizados sistemas especialistas ou algoritmos de aprendizado de máquina. (ROSSI, 2016).

Soluções que fazem uso de sistemas especialistas podem ter alguma vantagem com relação aos que utilizam modelos de aprendizado de máquina em casos que há classes conflitantes, ruidosas e/ou o modelo gerado não conseguiu generalizar bem o suficiente. Entretanto, essas soluções requerem a modelagem de especialistas no domínio, que pode variar bastante (VILLENA-ROMÁN et al., 2011).

As soluções de classificação automática de textos utilizando modelos de aprendizado de máquina podem ser desenvolvidas tanto no paradigma supervisionado, quanto no não-supervisionado ou ainda no semi-supervisionado. Os algoritmos desses três paradigmas se diferenciam pelo tratamento e dependência de dados rotulados. Enquanto que os algoritmos do paradigma supervisionado têm uma completa dependência de dados rotulados, os algoritmos do paradigma semi-supervisionado podem lidar com uma grande quantidade de dados não rotulados, desde que haja um número mínimo de dados rotulados. Por último, os algoritmos do paradigma não supervisionado conseguem trabalhar muito bem na ausência total de rótulos nos dados (KHAN et al., 2010; LIU et al., 2016).

Como o processo de rotular dados geralmente é realizado de forma manual, ele acaba consumindo muito tempo e recursos humanos, o que o torna inviável de ser empregue em alguns contextos (LIU et al., 2016). Além disso, há contextos que nem sequer existe conhecimento o suficiente para organizar e rotular os dados. Logo, a empregabilidade de algoritmos de aprendizado não supervisionado consegue ser muito maior comparada a dos outros dois paradigmas. Normalmente, os algoritmos desse paradigma realizam agrupamento dos textos em *clusters* de acordo com medidas de similaridade ou de dissimilaridade. Alguns desses algoritmos são até empregados em outros processos de análise de dados mais complexos para extrair, num primeiro momento, padrões e conhecimento (SHAFIABADY et al., 2016).

Os algoritmos de aprendizado supervisionado pressupõem a existência de rótulos pré-definidos em todos os dados do conjunto de treinamento o que, dessa forma, possibilita gerar indutivamente modelos com a capacidade de rotular novos dados de forma genérica (KHAN et al., 2010). Contudo, o uso de algoritmos desse paradigma presume a disponibilidade de uma grande quantidade de dados que foram organizados e rotulados corretamente por especialistas humanos para uso na fase de treinamento. Dessa maneira, a menos que os dados em questão existam há algum tempo, o uso de um sistema especialista se torna

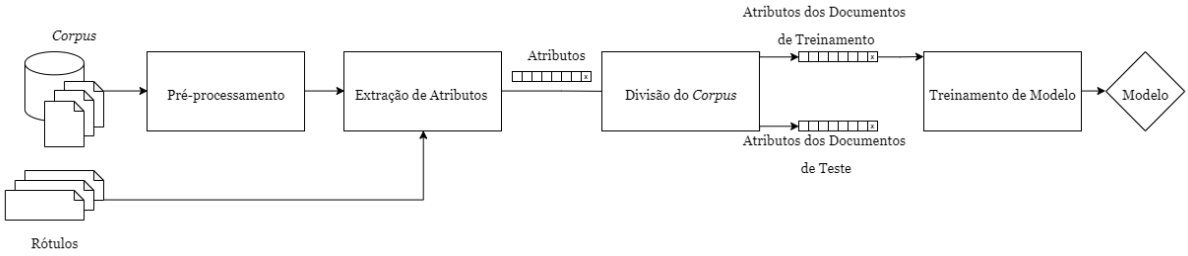

Figura 2 – Esquema de Treinamento de Modelo de ML

inevitável, porque a organização e rotulagem manual de milhares de grupos de dados pode ser uma atividade muito trabalhosa e intelectualmente desafiadora (SHAFIABADY et al., 2016).

O paradigma semi-supervisionado é uma boa opção frente à grande dificuldade de obter dados rotulados, visto que a carência desses dados é um fator impactante na eficiência dos modelos gerados. Isto é, os métodos supervisionados geralmente exigem uma grande quantidade de exemplos de treinamento rotulados para aprender com precisão. Os algoritmos desse paradigma lidam tanto com dados não rotulados, quanto com dados rotulados e geralmente utilizam na etapa de treinamento dos modelos uma quantidade de dados rotulados bem inferior aos algoritmos do paradigma supervisionado. Isso acontece porque geralmente nesses algoritmos são utilizadas técnicas não supervisionadas para agrupar todos os dados, rotulados e não rolutados, em *clusters* e, frequentemente, pontos no mesmo *cluster* compartilham o mesmo rótulo (LIU et al., 2016).

Obviamente não existe uma solução ideal para todo tipo de problema. Como visto, há casos em que sistemas especialistas podem ter um desempenho melhor que algoritmos de aprendizado de máquina, assim como a utilização desses algoritmos, em algum dos paradigmas, pode ser mais viável que a adoção de sistemas especialistas. Existem muitas variáveis que influenciam na escolha da solução como quantidade de dados, existência e/ou quantidade de dados rotulados, interseção entre os rótulos, dentre outras. Às vezes, até os próprios resultados das experimentações podem definir melhor qual solução empregar.

2.3.1 Treinamento de Modelos de Classificação

Basicamente, as principais etapas do treinamento de modelos de ML são: *i.* Pré-processamento; *ii.* Extração de Atributos; *iii.* Divisão do *Corpus*; *iv.* Treinamento de Modelo; *v.* Avaliação de Modelo (DALAL; ZAVERI, 2011). Essa estratégia genérica para treinamento de modelos de ML a partir de *corpus* de documentos é mostrada na Figura 2.

O pré-processamento reduz o tamanho dos documentos de entrada de texto significativamente, uma vez que são executados processamentos como remoção de *stopwords* e *stemming* (DALAL; ZAVERI, 2011; KHAN et al., 2010). Mas há outros pré-processamentos

Código 1 – Exemplo 3 de Texto de Publicação da BBC com *Tags* de HTML

```
<div>
    <div class="control-centre visible">
        <div class="metadata">
            <h2 class="title">3) The death of Stephen Hawking</h2><span class="off-screen
                ">Choose a format. New content will appear below.</span></div>
        <div class="choices">
            <div class="variable">
                <button class="choice  short">
                    <div class="label">short</div><span class="off-screen">text</span></
                        button>
                <button class="choice selected long">
                    <div class="label">long</div><span class="off-screen">text</span></
                        button>
                <button class="choice  video">
                    <div class="label">video</div>
                </button>
            </div>
        </div>
    </div>
    <div class="media-part">
        <div class="image-and-copyright"><img alt="Stephen Hawking" src="https://ichef.
            bbci.co.uk/news/976/cpsprodpb/1248A/production/_86609847_86609846.jpg">
            <div><span class="off-screen">Image copyright</span><span class="copyright">
                AFP</span></div>
        </div>
        <p>"A great scientist and an extraordinary man."</p>
        <p>That was how Stephen Hawking's children Lucy, Robert and Tim remembered him
            after his death aged 76 in March.</p>
        <p>Prof Hawking, one of the most respected and best-known scientists of our time,
             became one of the world's great ambassadors for science.</p>
        <p>Aged 22, he was diagnosed with motor neurone disease, which left him almost
            completely paralysed.</p>
        <p>In his 2013 memoir, Prof Hawking wrote: "At the time, I thought my life was
            over and that I would never realise the potential I felt I had. But now, 50
            years later, I can be quietly satisfied with my life."</p>
        <p>After being told in 1964 he would have only two or three years left to live,
            he went on to write A Brief History of Time - a layman's guide to cosmology,
            which sold more than 10 million copies.</p>
        <p>He was the first person to set out the theory that black holes leak energy and
             fade away to nothing, a phenomenon now known as Hawking radiation.</p>
        <p>"He once said, 'It would not be much of a universe if it wasn't home to the
            people you love'," his children said. "We will miss him forever."</p>
    </div>
</div>
```

que podem ser empregues dependendo do *corpus* utilizado. Um exemplo bem comum é a remoção de *tags* de HTML dos textos extraídos diretamente de páginas da *web* (DALAL; ZAVERI, 2011). Por exemplo, o trecho de texto com *tags* de HTML exibido no Código 1, também extraído da mesma página da BBC que os textos das Tabelas 4 e 8 e intitulado de "*The death of Stephen Hawking*", após um processamento para remoção de *tags*, ficaria igual ao encontrado na Tabela 10.

A extração de atributos identifica as palavras em um documento de texto a serem extraídas, podendo fazer uso de alguma técnica como o TF-IDF, ou simplesmente selecionando todas as palavras do texto pré-processado em uma determinada estrutura (DALAL; ZAVERI, 2011; ALAHMADI; JOORABCHI; MAHDI, 2013). O esquema representacional mais comum na classificação de textos, por exemplo, é o de *Bag-of-Words*, no qual um texto é representado como um vetor de quantidade ou de pesos de palavras. No entanto, essa representação tem uma grande limitação: a quebra de termos de expressões nas suas

palavras constituintes. Por exemplo, a expressão "Sumarização Automática" seria quebrada nas palavras "Sumarização" e "Automática". Esse problema pode trazer uma perda de informações devido à quebra de associação semântica (ALAHMADI; JOORABCHI; MAHDI, 2013).

Para ilustrar a estrutura BOW, faz-se necessário mostrar aqui um trecho de um texto na sua forma original e na estrutura de BOW, de forma semelhante aos exemplos de remoção de *stopwords* e *stemming*. Utilizando ainda os exemplos dos textos das Tabelas 4 e 8 e das suas versões sem *stopwords* e com *stemming* exibidas nas Tabelas 7 e 9, partes das suas respectivas estruturas no formato BOW podem ser visualizadas nas Tabelas 11 e 12.

1. "A great scientist and an extraordinary man."
2. That was how Stephen Hawking's children Lucy, Robert and Tim remembered him after his death aged 76 in March.
3. Prof Hawking, one of the most respected and best-known scientists of our time, became one of the world's great ambassadors for science.
4. Aged 22, he was diagnosed with motor neurone disease, which left him almost completely paralysed.
5. In his 2013 memoir, Prof Hawking wrote: "At the time, I thought my life was over and that I would never realise the potential I felt I had.
6. But now, 50 years later, I can be quietly satisfied with my life."
7. After being told in 1964 he would have only two or three years left to live, he went on to write A Brief History of Time - a layman's guide to cosmology, which sold more than 10 million copies.
8. He was the first person to set out the theory that black holes leak energy and fade away to nothing, a phenomenon now known as Hawking radiation.
9. "He once said, 'It would not be much of a universe if it wasn't home to the people you love','his children said.
10. "We will miss him forever."

Tabela 10 – Exemplo 3 de Texto da BBC sem *tags* de HTML com Sentenças Numeradas

Finalmente, todos os documentos do *corpus* acabam sendo convertidos para uma estrutura única de matriz multidimensional atributo-valor, inclusive o próprio rótulo quando existe acaba virando um atributo nessa estrutura. Essa representação do texto visa reduzir a complexidade para que os próprios algoritmos de aprendizado de máquina possam manusear esses dados mais facilmente. Mas um dos problemas dessa representação acaba sendo a dimensionalidade extremamente alta dos dados; em outras palavras, essas estruturas acabam se tornando grandes matrizes esparsas (KHAN et al., 2010).

Objetivando exemplificar a estrutura única de atributo-valor usada como entrada para os classificadores e utilizando os exemplos das estruturas de BOW das Tabelas 11 e 12, é possível visualizar na Tabela 13 como ficaria essa estrutura. Agora, supondo que esses dois textos fizessem parte de um mesmo *corpus*, cujas únicas classes fossem "*Politics*" e "*Entertainment*", e que os textos das Tabelas 4 e 8 pertencessem às classes "*Politics*" e

Palavra	Frequência
best	1
Brexit	1
coverag	1
day	1
deal	1
...	...
look	1
...	...
may	1
...	...
Theresa	1
went	1
year	1

Tabela 11 – Exemplo 1 de texto de publicação da BBC na estrutura BOW

Palavra	Frequência
address	1
aisl	1
...	...
look	1
...	...
may	1
...	...
Meghan	3
...	...
took	2
uniform	1
...	...
wed	2
...	...
word	1
wore	1

Tabela 12 – Exemplo 2 de texto de publicação da BBC na estrutura BOW

adress	...	*best*	...	*day*	...	*look*	...	*may*	...	Meghan	...	*year*
0	...	1	...	1	...	1	...	1	...	0	...	1
1	...	0	...	0	...	1	...	1	...	3	...	0

Tabela 13 – Exemplos 1 e 2 de textos da BBC na estrutura atributo-valor

Classe	*adress*	...	*best*	...	*day*	...	*look*	...	*may*	...	Meghan	...	*year*
1	0	...	1	...	1	...	1	...	1	...	0	...	1
2	1	...	0	...	0	...	1	...	1	...	3	...	0

Tabela 14 – Exemplos 1 e 2 de textos da BBC na estrutura atributo-valor com as classes

"*Entertainment*", respectivamente, é possível representar essas classes desses documentos na estrutura de atributo-valor. Supondo que na extração de atributos as classes "*Politics*" e "*Entertainment*" tenham sido representadas pelos números 1 e 2, nessa ordem, então a classe poderia ser representada como um atributo cujos valores para os textos das Tabelas 7 e 9 seriam 1 e 2.

A divisão do *corpus* em dados de treinamento e de testes é muito empregada na construção de modelos de ML, havendo também casos de particionamento em que se gera um terceiro conjunto de dados para validação, isto é, dados que se destinam a serem utilizados para ajustar parâmetros de entrada na etapa de treinamento do modelo. Há formas comuns de se realizar a divisão dos dados como, por exemplo, destinar 70% dos dados do conjunto para treinamento e 30% para testes ou valores próximos a essas proporções. Essa divisão geralmente é realizada de forma aleatória e estratificada para que não haja uma predominância de exemplos de classes no treinamento. A ideia é manter a mesma proporção de exemplos de classes do conjunto de dados original (ELKAN, 2012).

Evidentemente, há muitos problemas relacionados ao balanceamento de dados entre as classes que podem causar *overfitting* nos modelos, isto é, quando o modelo perde seu poder de generalização e se ajusta muito ao conjunto de treinamento. Esse sobre-ajuste ao conjunto de treinamento pode ser ruim, porque, em certas circunstâncias, o conjunto de treinamento não representa tão bem os dados do problema real. Nesses casos, o modelo fica muito eficiente para determinar o rótulo de dados bem similares aos de treinamento, mas tem uma performance pobre ao rotular dados mais divergentes. Também existe o caso contrário, no qual o modelo não se adapta sequer aos dados do conjunto de treinamento. Esse caso é chamado de *underfitting* (ELKAN, 2012).

O treinamento do modelo em si depende muito do algoritmo de ML em questão. Várias técnicas de aprendizado de máquina supervisionadas foram propostas na literatura para a classificação automática de documentos de texto como Naïve Bayes, Redes Neurais, SVM, Árvores de decisão, etc. Nenhum algoritmo isolado consegue ser superior a todos os outros para todos tipos de classificação, isso depende muito do escopo do problema.

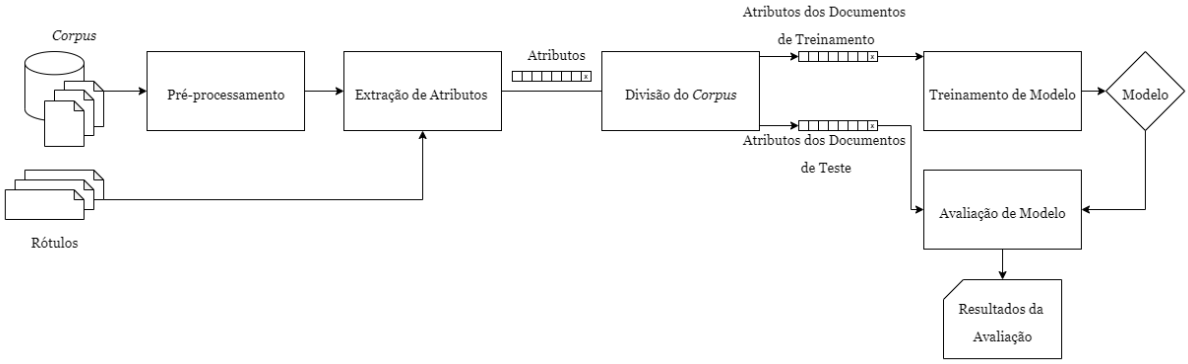

Figura 3 – Esquema de Treinamento e Avaliação de Modelo de ML

Inclusive há a possibilidade também de agrupar esses algoritmos para gerar um sistema de classificação mais robusto, isso sem mencionar a possibilidade de utilizar *ensembles* de classificadores (DALAL; ZAVERI, 2011).

Ensembles são conjuntos de vários classificadores, ou seja, são combinações das saídas de vários classificadores de base para, através de alguma medida de agregação, classificar de forma integrada. Experimentos realizadas em muitos domínios de aprendizado de máquina têm demonstrado a eficácia dessa estratégia. Há ainda *ensembles* homogêneos que são gerados utilizando classificadores base de um mesmo tipo, ou seja, provenientes de um mesmo algoritmo de ML, mas treinando cada um deles com amostras diferentes. Os *ensembles* heterogêneos, em contrapartida, podem fazer uso de classificadores base completamente diferentes, isto é, é possível formar um *ensemble* com SVM, redes neurais e árvores de decisão, por exemplo. (DONG; HAN, 2004).

2.3.2 Avaliação de Modelos de Classificação

Após o treinamento do classificador, são realizados testes com o modelo gerado a partir do conjunto de documentos de testes. Se o desempenho de classificação do modelo treinado é considerada aceitável para o contexto do problema, esse modelo é usado para classificar novas instâncias de texto documentos (DALAL; ZAVERI, 2011). Essa etapa é absolutamente central para medir o desempenho de um classificador em um conjunto de testes independente (ELKAN, 2012). Logo, o treinamento dos modelos pode ser realizado diversas vezes a depender dos resultados da avaliação do próprio modelo gerado com o conjunto de testes. O esquema de treinamento e avaliação do classificador é mostrado na Figura 3.

Alguns dos padrões descobertos pelo classificador podem ser espúrios, ou seja, são válidos nos dados de treinamento devido à aleatoriedade em como esses dados foram particionados, mas não são válidos para toda a população. Um modelo de classificador que se baseia nesses padrões espúrios facilmente poderá apresentar *overfitting*. Para

		Real		
		Positivo	Negativo	Total
Predito	Positivo	a	b	$a+b$
	Negativo	c	d	$c+d$
	Total	$a+c$	$b+d$	n

Tabela 15 – Estrutura de uma matriz de confusão de duas classes

evitar *overfitting* é essencial não realizar treinamentos com o conjunto de testes, ou com o de validação, por mais tentador que isso seja. Pois, somente um conjunto de testes independente pode fornecer uma estimativa adequada do desempenho do modelo. Ou seja, é bom recorrer a técnicas de validação cruzada (*cross-validation*), nas quais os dados são particionados de forma mutuamente exclusivas (ELKAN, 2012).

Aprofundando ainda mais essa discussão, em alguns casos o conjunto de dados é tão pequeno que separar em conjuntos exclusivos de treinamento e testes com uma divisão proporcional aleatória e estratificada das amostras não se mostra uma abordagem satisfatória o suficiente para avaliação do modelo gerado. Então, não apenas nesses casos, mas em quaisquer outros, é possível utilizar a técnica de *k-fold cross-validation* que divide o conjunto de treinamento em k partes ao longo de k iterações, onde a cada iteração será gerado um modelo, no qual utilizará apenas uma dessas partes como conjunto de testes e todas as outras como conjunto de treinamento. O benefício dessa abordagem é testar a capacidade de generalização de predição dos modelos criados, aproveitando todo o conjunto de dados. Contudo, a complexidade temporal dessa técnica é k vezes o tempo de execução do treinamento do algoritmo. Em pesquisas recentes, a escolha mais comum para k é 10 (ELKAN, 2012).

Justamente para avaliar o desempenho do classificador que o conjunto de testes tem que ter os rótulos conhecidos, pois basicamente a avaliação do modelo consiste em comparar as saídas do conjunto de testes com os seus rótulos oficiais e exibir, assim, os resultados de diversas medidas de avaliação do modelo. Em problemas de classificação binária (classes Positivo e Negativo), geralmente todas as medidas são expressas baseadas nos valores da matriz de confusão resultante. Essa matriz 2x2 apresenta 4 valores: quantidade de verdadeiro-positivos, verdadeiro-negativos, falso-positivos e falso-negativos. Usualmente as colunas correspondem à classe predita pelo modelo, enquanto que as linhas correspondem à classe verdadeira dos exemplos. Analisando essa divisão fica fácil entender como os exemplos preditos pelo modelo são dispostos nessa matriz da Tabela 15. Explicando melhor os termos: a é a quantidade de dados cujo rótulo real e predito são Positivo (verdadeiro-positivos); b é a quantidade de dados cujo rótulo real é Negativo, mas o predito é Positivo (falso-positivos); c é a quantidade de dados cujo rótulo real é Positivo, mas o predito é Negativo (falso-negativos); d é a quantidade de dados cujo rótulo real e predito são Negativo (verdadeiro-negativos) (ELKAN, 2012).

Ainda no exemplo de classificação binária, é esperado que: $a + c$ totalize o número de exemplos verdadeiramente da classe Positivo; $b + d$ totalize o número de exemplos verdadeiramente da classe Negativo; $a + b$ totalize o número de exemplos preditos da classe Positivo; $c + d$ totalize o número de exemplos preditos da classe Negativo; $a + b + c + d$ totalize o número total de exemplos do conjunto de treinamento (n). Como também já dito, é possível calcular outras medidas baseadas somente nesses números da matriz de confusão, como é o caso da Precisão (p) na Fórmula 9; Cobertura (r) na Fórmula 10 e Acurácia (a_{cc}) na Fórmula 11. A *F-measure* ($f1$) na Fórmula 12 é uma medida de média harmônica entre o valor da precisão e o valor da cobertura. Já é de se esperar que nenhuma dessas medidas unicamente pode caracterizar o desempenho de um modelo completamente. Então, geralmente quando se faz a avaliação de um classificador, é bom fornecer todas essas medidas ou somente a própria matriz de confusão (ELKAN, 2012).

$$p = \frac{a}{a + b} \tag{9}$$

$$r = \frac{a}{a + c} \tag{10}$$

$$a_{cc} = \frac{a + d}{n} \tag{11}$$

$$f1 = \frac{2 \times p \times r}{p + r} \tag{12}$$

Tanto a matriz de confusão, quanto essas medidas de classificação, podem ser estendidos para problemas de classificação de múltiplas classes. Nesses casos, cada uma das medidas de a, b, c e d serão calculadas por classe, consequentemente as medidas de p, r, a_{cc} e $f1$ também serão. Para efetuar esses cálculos, é necessário considerar no ato da análise para cada classe que a classe Positiva corresponderá a própria classe da análise e a classe Negativa a todas as outras. Isto é, supondo um problema de classificação com 10 classes distintas, se um dado da classe 1 é classificado por um modelo como pertencente à própria classe 1, esse dado será contabilizado como verdadeiro-positivo, caso ele seja classificado em qualquer outra classe pelo modelo, ele será contabilizado como falso-positivo.

Há também na literatura, uma maneira de medir a consistência entre duas classificações. O coeficiente Kappa de modo geral é utilizado para esse caso, isto é, para medir a concordância entre dois avaliadores. No contexto de classificação, ele pode ser

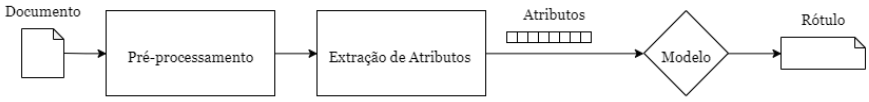

Figura 4 – Esquema de Predição com Modelo de ML

utilizado para medir a concordância entre classificação de um modelo com a classificação real. Um valor do coeficiente Kappa igual a 1 representa uma concordância perfeita entre as classificações, enquanto um valor igual a 0 representa uma total discordância (WANG; XIA, 2019; TEUFEL; SIDDHARTHAN; TIDHAR, 2006; BANERJEE et al., 1999).

Esse coeficiente é calculado através da Fórmula 13. Onde $p_o = \sum_{i=1}^{n} p_{ii}$ é a proporção observada de concordância, $p_c = \sum_{i=1}^{n}(p_{i\cdot} \times p_{\cdot i})$ é a proporção de concordância esperada por acaso, $p_{i\cdot} = \sum_{j=1}^{n} p_{ij}$ denota a proporção de exemplos colocados na i-ésima linha, $p_{\cdot j} = \sum_{i=1}^{n} p_{ij}$ denota a proporção de exemplos colocados na j-ésima coluna e n o número de classes (WANG; XIA, 2019; TEUFEL; SIDDHARTHAN; TIDHAR, 2006; BANERJEE et al., 1999).

$$\kappa = \frac{p_o - p_c}{1 - p_c} \tag{13}$$

Após a definição e escolha do modelo final de classificação, é possível realizar a predição de novos documentos de texto, ou seja, documentos cujos rótulos não são previamente conhecidos. O documento deverá passar pelas mesmas etapas de pré-processamento e extração de atributos para que fique com a mesma estrutura de entrada dos dados de treinamento e testes. Depois, a estrutura de atributos do documento pode ser passada como entrada para o classificador que fará a predição do rótulo daquele documento, atribuindo-o a alguma das possíveis classes existentes (KHAN et al., 2010). O diagrama de predição de novos exemplos é mostrado na Figura 4.

2.4 Documentos de Notícias

Quando se trata da utilização de técnicas de PLN, o tipo do conteúdo dos documentos a serem processados deve ser levado em conta antes da realização de quaisquer experimentos. Isto porque aspectos como o idioma, o tamanho do texto, o gênero textual, a formalidade ou informalidade da linguagem influenciam muito nos pré-processamentos a serem empregados (CONRAD et al., 2009; LLORET; SAGGION; PALOMAR, 2010).

Logo, para o contexto deste trabalho, alguns aspectos favorecem a utilização de *corpus* de textos de notícias em detrimento de *corpus* de outros gêneros textuais. Por exemplo, a sua linguagem formal, objetiva e sem erros gramaticais é um aspecto

positivo. Pois, essa formalidade na disposição do conteúdo e apego às regras da escrita formal do idioma, dispensa esses textos de alguns pré-processamentos empregues em *blogs*, por exemplo (CONRAD et al., 2009; LLORET; SAGGION; PALOMAR, 2010). A disponibilidade de notícias de maneira fácil e acessível por grandes jornais, mídias eletrônicas e serviços de notícias *online* também contribuí para utilização desse tipo de documento (KAUR; BAJAJ, 2016).

Há também alguns pontos que facilitam a utilização desse tipo de documentos num contexto de classificação e sumarização automática simultaneamente. O primeiro ponto é a disponibilidade de resumos, realizados pelos próprios autores das matérias. Resumos considerados "padrão ouro" que podem ser utilizados como base para determinação da qualidade dos resumos gerados automaticamente (FERREIRA et al., 2015). Outro ponto importante é a existência de manchetes como títulos das notícias que possibilita a utilização das técnicas de sumarização extrativa que fazem uso de informações do título (RANA; KHALID; AKBAR, 2014; FERREIRA et al., 2015). Por fim, esses textos muitas vezes já vêm classificados e abordam os mais diversos tópicos possíveis como saúde, economia, esportes, política, etc.

2.5 Trabalhos Relacionados

Agora, faz-se necessário analisar o que já foi feito no escopo de interesse deste trabalho. Todo o levantamento bibliográfico teve como meta principal identificar o que já havia sido desenvolvido no contexto de sumarização automática de textos que servisse como entrada para a classificação e vice-versa, especialmente os que denotassem algum tipo de correlação. Para que esta seção não seja muito exaustiva nas comparações entre os trabalhos relacionados e este trabalho, a Tabela 17 compila todos os trabalhos que serão apresentados aqui. A Tabela 17 apresenta alguns tópicos tidos como fundamentais para diferir este trabalho dos demais. Já a Tabela 16 exibe um resumo dos tópicos mencionados. Cada tópico nessa última tabela apresenta um código, que é utilizado na Tabela 17 para representar, de forma mais compacta, o resultado da análise de cada trabalho relacionado. Os dois últimos tópicos apresentados (T_{MTSC} e $T_{C.COMP}$) são os principais, uma vez que abordam as contribuições finais deste trabalho.

Em (KER; CHEN, 2000) foi proposta uma abordagem de classificação do texto, baseada na sumariação utilizando métodos de frequência e posição de palavras para extrair conhecimento da classificação no campo do título, cujo tempo de computação é muito mais curto. Foi utilizada uma versão do *corpus* de notícias Reuters com aproximadamente 10.000 documentos e 93 categorias. Segundo os próprios autores, os resultados foram promissores, de forma que eles indicam o uso do sistema para classificação de documento *online*.

Em (MIHALCEA; HASSAN, 2005) foi explorado os benefícios da interação entre

classificação automática de textos e sumarização automática utilizando técnicas extrativas baseadas em grafos. Essa combinação é utilizada para reduzir a potencial sobreposição entre documentos pertencentes a diferentes classes objetivando utilizar apenas as seções importantes dos documentos no processo de aprendizagem de máquina para treinamento dos classificadores. Nos experimentos foram utilizados 567 artigos de notícias. Os resultados obtidos demonstraram que a precisão da classificação melhorou em até 19,3%.

A referência (CONRAD et al., 2009) apresenta um sistema de sumarização extrativo multi-documento totalmente automático de sentimentos na *Web* para o domínio jurídico. Baseia-se no processamento de um conjunto de questões legais através de pesquisas em *blogs*. No processo de avaliação dos sumários gerados, as pessoas, profissionais da área, que indicavam o resumo ideal, selecionando os mais informativos. Foi calculada uma estatística Kappa específica para determinar a concordância entre avaliadores, encontrando um valor próximo a 75%. Convencionalmente isso indica uma concordância adequada entre avaliadores. Entretanto, os próprios autores não tiraram conclusões mais ousadas em cima das experimentações, uma vez que, segundo os próprios, foi o primeiro experimento publicado de avaliação da sumarização de sentimentos na blogosfera jurídica.

Lloret e seus colegas em (LLORET; SAGGION; PALOMAR, 2010) investigaram o efeito da sumarização no problema de inferência de classificação em documentos opinativos *online* relacionados a empresas onde os autores buscaram verificar melhorias na classificação de análise de sentimentos através de textos sumarizados. Nos experimentos conduzidos, foi utilizada uma grande variedade de métodos de sumarização, inclusive focados em usuários. Os resultados que os autores obtiveram demostraram que alguns tipos de resumos podem ser tão eficazes ou até melhores que os próprios documentos completos para essa dificuldade. Contudo, um grande problema nesse caso foi a utilização de um *corpus* de dados muito pequeno, principalmente para uma atividade tão complexa.

Dalal e Zaveri (DALAL; ZAVERI, 2013) desenvolveram uma abordagem semi-supervisionada que utiliza mineração de textos, classificação automática e sumarização automática para realizar uma análise de sentimentos (em duas polaridades) em opiniões de blogueiros. Com base nos textos de opiniões e de algumas características extraídas através da mineração de textos, é realizada uma sumarização agrupando informações por produto e, finalmente, classificando o texto de acordo com sua polaridade. Foram utilizados aproximadamente 1.400 *reviews* de 7 produtos (cerca de 200 por produto) para avaliação da abordagem. Resultados empíricos indicaram que a abordagem de várias etapas pode identificar com sucesso o sentimento dos comentários opinativos de usuários com precisão de 86%.

Em (KOULALI; EL-HAJ; MEZIANE, 2013) técnicas de sumarização foram empregadas em textos escritos em árabe para melhorar a classificação em tópicos (Cultura, Economia, Esportes, Internacional, Local e Religião). Nos experimentos realizados, foi

utilizado um *corpus* contendo aproximadamente 20.000 artigos do jornal *online* Al-Wattan. Foram utilizadas as palavras dos documentos no modelo BOW tanto para a sumarização, quanto para a classificação. Para a sumarização extrativa dos textos, foi utilizada a técnica TF-IDF (MEENA; GOPALANI, 2014; FERREIRA et al., 2013a) para atribuir pontuações às sentenças de cada texto; já para a classificação foi utilizada a medida de similaridade de cossenos para determinar o tópico de um novo documento dos *corpus* de testes. Foram comprovados que os resultados obtidos utilizando os sumários são comparáveis à abordagem tradicional de utilizar os textos completos para o *corpus* de teste empregado.

Em (MEENA; GOPALANI, 2014) é realizada uma análise de várias técnicas de sumarização extrativa, criando diversas combinações entre essas técnicas. O ROUGE é utilizado para fazer a avaliação dos resumos extrativos gerados utilizando como base os resumos abstrativos do *corpus* empregado nas experimentações. Para alguns documentos foram obtidos até 65% nos valores da *F-measure*, entretanto os próprios autores consideram esses resultados proeminentes, pois estão comparando o resumo extrativo com o abstrativo.

Em (FERREIRA et al., 2014) foi mostrado que quando a escolha das técnicas de sumarização automática a serem empregadas levam em consideração o contexto do documento, especialmente o domínio linguístico, os resultados obtidos são mais animadores (SAGGION, 2008). De forma similar, os autores aqui propõem um sistema que leva o contexto do documento em consideração antes de sumarizá-lo. O sistema proposto foi comparado com três conjuntos de dados distintos (artigos, notícias e postagens de *blogs*). Além disso, foram encontradas pelos autores as melhores combinações de métodos de sumarização para os textos desses três tipos de documentos. Com base nos experimentos realizados, ficou comprovada que a melhor combinação para textos curtos bem formados de notícias é a de de frequência de palavras junto ao TF-IDF, posição da sentença e semelhança com o título; já para artigos científicos, a melhor combinação é composta pelos métodos palavras sinalizadoras (*Cue phrases*), posição da sentença, TF-IDF e semelhança com o título; para documentos de *blogs* de textos curtos e não estruturados a combinação que alcançou melhores resultados foi a que utilizava frequência das palavras e TF-IDF.

Em (FERREIRA et al., 2015) foi estudada uma maneira eficiente de classificar documentos de forma automatizada através dos resumos provenientes de várias técnicas de sumarização automática de textos. Foi utilizada parte do *corpus* CNN (LINS; MELLO; SIMSKE, 2019) com 1.000 artigos, com textos de 41 sentenças em média, de notícias dividido em 10 categorias diferentes. Os resultados mostram que as melhores técnicas de sumarização para classificação de documentos são métodos de pontuação de palavras. Também mostrou, em alguns pré-testes de desempenho liminar, que a classificação do texto resumido é quase duas vezes mais rápida que os textos dos documentos originais.

Em (JEONG; KO; SEO, 2016) é proposta uma estrutura integrada com técnicas que fazem uso da sumarização e classificação automática de textos de forma complementar.

Dentre as técnicas mais importantes, uma faz uso da categoria do documento para realizar resumos e outra utiliza o resumo para classificar o documento. Na primeira técnica, são utilizadas palavras de domínio de cada classe, de forma ponderada aos algoritmos de sumarização extrativa baseados em palavras e em sentenças, para priorizar sentenças com a presença desses termos. Na segunda técnica, as palavras das sentenças presentes nos sumários dos textos ganham um peso maior como parâmetros de entrada para o classificador de textos. Devido à dificuldade de encontrar *corpora* devidamente rotulados e com a presença de *gold summaries*, os autores selecionaram dois *corpus*, um apenas com a presença de *gold summaries* e o outro apenas com os rótulos dos textos, e anotaram o que estava sem rótulos e sumarizaram o que estava sem *gold summaries*. Os resultados obtidos para essas duas técnicas indicaram um desempenho levemente melhor que suas respectivas *baselines*.

(JO, 2017) propôs uma versão específica do algoritmo KNN que utiliza como parâmetros de comparação os atributos do texto sumarizado ao invés do texto integral. Essa abordagem visa minimizar classificações errôneas derivadas de atributos ruidosos ou que não sejam tão importantes para determinação da classe do texto. Contudo, não há resultados concretos para demonstrar a eficiência do algoritmo.

Em (SILVA, 2017) foi construído um sistema para descobrir se a classificação de um documento consiste em um bom parâmetro para determinação da técnica de sumarização extrativa mais eficiente e, em tal caso, determinar também as técnicas mais adequadas para cada classe. Para evidenciar tal asserção, foram gerados sumários "de uma parte"dos documentos do *corpus* de notícias do CNN por técnicas baseadas em palavras, sentenças e grafos. A qualidade desses sumários gerados foi avaliada utilizando o *framework* ROUGE e a medida $MATCHES$. Contudo, os resultados obtidos demonstraram uma improbabilidade da classe do documento consistir em um bom parâmetro para escolha da técnica de sumarização mais adequada.

(MA et al., 2018) propôs um modelo hierárquico de ponta a ponta para a classificação de sentimentos a partir de um conjunto de textos sumarizados. Nesse modelo o rótulo de classificação é tratado como a "sumarização"adicional dos sumários de textos processados. São utilizados métodos de sumarização focados no domínio a ser analisado, então sentenças como palavras como "ruim"e "bom", por exemplo, tendem a terem suas pontuações priorizadas. Os experimentos realizados com um *corpus* de *reviews* do Amazon indicaram um desempenho melhor que fortes sistemas tidos como *baselines* nesse domínio.

Foram relatados, quase de forma estafante, os trabalhos relacionados mais importantes que integrem soluções de sumarização automática de textos com soluções de classificação automática de textos. Alguns desses tem alguns estudos e análises bem similares a este trabalho, até em outros domínios, como por exemplo em *copus* com textos de *blogs*. Vale ressaltar novamente que os trabalhos aqui apresentados poderão ser visualizados

de forma resumida em tópicos através da Tabela 17.

Código do Tópico	Descrição do Tópico
T_{SUM}	Realiza sumarização automática de textos?
T_{CLA}	Realiza classificação automática de textos?
$T_{S.MET}$	Quais os tipos de métodos de sumarização utilizados?
T_{COR}	Quais tipos de *corpora* utilizados?
$T_{COR.ROT}$	Os *corpora* estão devidamente classificados?
$T_{COR.GS}$	Os *corpora* apresentam *gold sumaries*?
T_{MTSC}	Exibem uma relação de classes por métodos de sumarização ou vice-versa?
$T_{C.COMP}$	Comparam a classificação de textos integrais com os textos sumarizados?

Tabela 16 – Tabela de Tópicos Comparativos para Trabalhos Relacionados

Referência	T_{SUM}	T_{CLA}	$T_{S.MET}$	T_{COR}	$T_{COR.ROT}$	$T_{COR.GS}$	T_{MTSC}	$T_{C.COMP}$
KER; CHEN, 2000	Sim	Sim	Palavras e Sentenças	Notícias	Sim	Não	Não	Sim
MIHALCEA et al., 2005	Sim	Sim	Grafos	Notícias	Sim	Não	Não	Sim
CONRAD et al., 2009	Sim	Sim	Palavras	*Blogs*	Sim	Não	Não	Sim
LLORET et al., 2010	Sim	Sim	Focado no Usuário	*Blogs*	Sim	Não	Não	Sim
DALAL; ZAVERI, 2013	Sim	Sim	Palavras e Sentenças	*Blogs*	Não	Não	Não	Não
KOULALI et al., 2013	Sim	Sim	Palavras	Notícias	Sim	Não	Não	Sim
MEENA et al., 2014	Sim	Não	Palavras, Grafos e Sentenças	Notícias	Não	Sim	Não	Não
FERREIRA et al., 2014	Sim	Sim	Palavras, Grafos e Sentenças	Notícias, Artigos e *blogs*	Não	Sim	Não	Não
FERREIRA et al., 2015	Sim	Sim	Palavras, Grafos e Sentenças	Notícias	Sim	Sim	Não	Sim
JEONG; KO; SEO, 2016	Sim	Sim	Palavras e Sentenças	Notícias	Sim	Sim	Não	Sim
JO, 2017	Sim	Sim	Não Especificado	Não Especificado	Não Especificado	Não Especificado	Não	Não
SILVA, 2017	Sim	Não	Palavras, Grafos e Sentenças	Notícias	Sim	Sim	Sim	Não
MA et al., 2018	Sim	Sim	Focado no Usuário	*Blogs*	Sim	Não	Não	Sim
Este trabalho	Sim	Sim	Palavras, Grafos e Sentenças	Notícias	Sim	Sim	Sim	Sim

Tabela 17 – Tabela Comparativa de Trabalhos Relacionados

Considerando as breves explanações dos trabalhos relacionados aqui, bem como as relações resumidas exibidas na Tabela 17, alguns trabalhos são bem parecidos, de fato, com o trabalho aqui proposto e nos outros, é perceptível as diferenças de objetivos. Então, será descrito aqui, prontamente algumas importantes diferenças deste trabalho com os outros mais semelhantes.

- Em (KOULALI; EL-HAJ; MEZIANE, 2013) são comparadas a eficácia da classificação de textos integrais com textos sumarizados para um *corpus* de documentos de notícias. Todavia, esse *corpus* é escrito em árabe e, como já mencionado, os pré-processamentos aplicados em textos variam de idioma para idioma.
- Apesar de em (MEENA; GOPALANI, 2014) combinações de métodos de sumarização extrativos serem realizadas, o foco não é determinar as melhores combinações de técnicas para cada classe e sim as melhores combinações para os *corpus* analisados.

- Em (FERREIRA et al., 2015) é utilizado o mesmo *corpus* deste trabalho, contudo, aqui são utilizadas classificações mais significativas para esse *corpus*, uma vez que no artigo em questão, havia classes muito genéricas que rotulavam da mesma forma assuntos extremamente diversificados. Nesse artigo também não foi realizada uma análise das classes dos documentos como parâmetros para escolha das melhores técnicas de sumarização.

- Em (JEONG; KO; SEO, 2016), cujo escopo do trabalho é bem similar, é utilizado um sistema cujas técnicas de sumarização apresentam variações e a geração de modelos utiliza parâmetros modificados a partir de cálculos desenvolvidos e/ou remodelados pelos próprios autores.

- Como já dito anteriormente, este trabalho amplia o sistema desenvolvido em (SILVA, 2017). Então, de certa forma, trata-se de uma extensão do trabalho anterior. Entretanto, diferente dos experimentos realizados no trabalho anterior, o trabalho aqui analisa outros cenários para continuação da investigação da utilização de técnicas de sumarização extrativa, e de combinações entre elas, para estabelecer uma relação com cada classe de documento de notícias. Esses outros cenários são: *i.* utilização do *corpus* CNN com outros dois níveis de importância e *ii.* combinação das técnicas de sumarização, 2 a 2, para gerar resumos utilizando o *corpus* CNN com o primeiro nível. Esses detalhes serão explanados melhor no capítulo 3, de Material e Métodos.

3 MATERIAL E MÉTODOS

3.1 Metodologia Geral

Este capítulo descreve a metodologia utilizada neste trabalho. Aqui, nesta seção, será apresentada rapidamente a metodologia de forma resumida objetivando servir como guia das etapas e processos empregados neste trabalho. Obviamente, alguns percalços foram encontrados na execução deste trabalho e eles serão explanados na seção 5.2 do capítulo 5. As seções seguintes deste capítulo aprofundam as etapas fundamentais que aqui, nesta seção, são apenas superficialmente explanadas.

Basicamente, neste trabalho, foram analisadas duas principais conjecturas: *i.* utilização de técnicas de sumarização extrativa, bem como de combinações entre elas, para estabelecer uma relação entre as melhores técnicas e cada classe de documento de notícias; *ii.* comparação do desempenho da classificação de textos integrais com os sumarizados e determinação das melhores técnicas de sumarização extrativa para cada classe de documentos de notícias. Elas serviram como norte para guiar a definição dos experimentos realizados e das análises dos resultados obtidos.

Esta pesquisa iniciou-se com uma revisão bibliográfica aprofundada, com o intuito de comprovar a inexistência de algum trabalho congênere e verificar o que já havia sido realizado em trabalhos semelhantes, além de fundamentar o referencial teórico para a escrita deste trabalho. Os trabalhos correlatos, inclusive, já foram descritos anteriormente na seção 2.5 do capítulo anterior.

Em seguida, foi realizada uma cuidadosa análise em busca de *corpora* de textos de notícias em inglês cuja grande maioria dos documentos, preferencialmente, já estivesse sido corretamente classificada e que apresentasse sumários referenciais, isto é, *gold standards*. Pois dessa forma, seria possível rotular adequadamente todas as classes de documentos a serem trabalhadas e as possíveis relações entre cada classe e as melhores técnicas de sumarização. Entretanto, apesar da busca exaustiva por *corpora* de textos de notícias que pudessem ser utilizados tanto para sumarização automática de textos, quanto para a classificação automática de textos, isto é, por *corpus* com *gold standards* e com rótulos nos textos simultaneamente, somente um foi encontrado. Esse único *corpus* será melhor detalhado na seção 3.2.

Posteriormente, fez-se uma análise e seleção, dentre as mais diversas técnicas de sumarização extrativa de textos, das mais adequadas baseada na literatura. Essas técnicas foram apresentadas na seção 2.2.3 do capítulo anterior. Então, os textos do *corpus* selecionado foram sumarizados através dessas técnicas e de combinações entre elas.

Foram utilizadas durante o processo de avaliação automática as medidas quantitativas do *framework* ROUGE. Especificamente, medidas como precisão, cobertura e *F-measure* do ROUGE-2 (LLORET; PALOMAR, 2012; LIN, 2004), além também da medida *MATCHES*, para comparar a qualidade dos resumos dos textos produzidos automaticamente. Os resultados dessas medidas serviram para evidenciar ou rejeitar uma relação entre as melhores técnicas de sumarização e cada classe de documento de textos.

De posse dos resumos dos textos obtidos através das técnicas de sumarização selecionadas, foram construídos modelos de ML para classificar tanto o texto integral quanto o sumarizado. Para todas as etapas de pré-processamento, treinamento e avaliação dos modelos gerados foi utilizada a biblioteca WEKA[1]. Também, foi realizada a adoção de medidas quantitativas, como precisão, cobertura e *F-measure*, dentre outras para avaliar a qualidade dos modelos gerados.

Ainda foram realizadas verificações no único *corpus* encontrado em busca de possíveis desbalanceamentos entre classes. Esses desbalanceamentos foram verificados baseados nas matrizes de confusão e em outras medidas de avaliação de classificação obtidas de forma preliminar e os documentos foram reclassificados conforme necessário. A metodologia para reclassificação desse *corpus* será melhor detalhada também na seção 3.2.

Concentrando-se nos objetivos deste trabalho, os resultados finais obtidos de ambas as conjecturas foram analisados com o propósito de verificar o quanto categoria de classificação de um documento consiste em um bom critério para escolher quais técnicas de sumarização extrativa empregar e, também, se a eficácia da classificação de documentos a partir dos resumos dos textos originais se equiparava à classificação dos próprios textos integrais.

3.2 *Corpus* de Texto

Como já mencionado, o processo de rotular dados é muito laborioso e sensível a erros tanto quanto o processo de gerar *gold standards*, mais ainda quando ambos os processos (classificação e sumarização automática) tem que ser realizado em um mesmo *corpus*. Logo, apesar da busca exaustiva por *corpora* de textos de notícias que pudessem ser utilizados tanto para sumarização automática de textos, quanto para a classificação automática de textos, isto é, por *corpus* com *gold standards* e com rótulos nos textos simultaneamente, somente um foi encontrado. O único *corpus* textual que atendeu às especificações do escopo deste trabalho foi o da CNN desenvolvido por (LINS et al., 2019; LINS et al., 2012) e utilizado em (FERREIRA et al., 2014; FERREIRA et al., 2015; SILVA, 2017), inclusive na recente competição internacional de sumarização extrativa que ocorreu no ACM Symposium on Document Engineering 2019 (LINS; MELLO; SIMSKE, 2019).

[1] https://www.cs.waikato.ac.nz/ml/weka/

O *corpus* CNN em sua versão atual consiste de 3.000 textos em inglês padrão gramaticalmente correto, extraídos do site da CNN, divididos em 18 classes: "*Arts*", "*Automotive*", "*Business*", "*Computer and Internet*", "*Economy and Finance*", "*Education*", "*Employment and Work*", "*Entertainment*", "*Food and Beverage*", "*Games*", "*Government and Politics*", "*Health*", "*Law*", "*Science*", "*Society and Culture*", "*Sports*", "*Travel*" e "*Weather*". Além dos textos completos, há também os resumos abstrativos, chamados de *highlights*, possivelmente elaborados pelos próprios autores e os *gold standards*, resumos com geralmente 3 a 5 sentenças, gerados com uma rigorosa metodologia de desenvolvimento descrita em (LINS et al., 2019). Por fim, o *corpus* CNN ainda apresenta uma classificação em níveis de importância. Essa classificação em níveis, se distancia da classificação multi-rótulos e classificação hierárquica (LEVATIĆ; KOCEV; DŽEROSKI, 2015; BOUTELL et al., 2004). Posto que não há uma estrutura de organização, no sentido de generalização ou especialização, entre as classes e pelo fato do primeiro nível de importância ser muito mais relevante que o segundo e o segundo, ser mais relevante que o terceiro. Um exemplo de texto de cada classe desse corpus na sua versão integral e seus respectivos *gold standard* e *highlight* pode ser visualizado neste trabalho para real compreensão de documentos dessas classes (Apêndices A - R).

A Tabela 18 exibe a relação da quantidade de documentos por classe nos três níveis de importância do *corpus* CNN. O total de documentos presentes nos totais dos níveis de importância 2 e 3 superam 3.000 devido ao fato de um documento poder estar rotulado em mais de uma classe ao mesmo tempo. Exemplificando, o texto do Apêndice E em seu primeiro nível de importância foi marcado como "*Economy and Finance*"; no seu segundo nível de importância, como "*Government and Politics*", não apresentando classificação no terceiro nível.

Conforme a análise do gráfico mostrado na Figura 5, é possível notar um grande desbalanceamento entre as classes. Há classes com 419 documentos (14% do total) e 833 documentos (28% do total), enquanto há outras com bem menos que 25 documentos (1% do total).

Após algumas análises preliminares dos primeiros experimentos treinando classificadores de texto utilizando as 18 classes originais do *corpus*, foram observadas, através das matrizes de confusão interferências entre algumas classes do *corpus*. Rapidamente, foram realizadas avaliações da classificação dos documentos dentro do *corpus* e novos mapeamentos de classe foram criados objetivando melhorar os modelos de classificação de textos. Foram criadas 3 novas classificações para o *corpus* do CNN utilizando o primeiro nível de importância.

Essas classificações foram criadas visando diminuir a sobreposição entre as classes de documentos. Sobreposições observadas nas matrizes de confusões das análises preliminares pelo autor e professor orientador deste trabalho. Sem mencionar também o fato da

Classe	Documentos no Nível 1	Documentos no Nível 2	Documentos no Nível 3
Arts	275	392	430
Automotive	19	33	35
Business	143	200	215
Computer and Internet	57	79	88
Economy and Finance	33	48	51
Education	101	146	163
Employment and Work	69	113	124
Entertainment	419	598	651
Food and Beverage	22	34	42
Games	27	54	64
Government and Politics	447	544	581
Health	119	160	172
Law	205	299	334
Science	128	167	182
Society and Culture	833	1018	1073
Sports	61	88	96
Travel	26	35	43
Weather	16	31	35

Tabela 18 – Relação entre a Quantidade de Documentos por Classe no *Corpus* do CNN

Classe	Nº de Documentos no Nível 1	% de Documentos no Nível 1
Arts	275	9%
Automotive	19	1%
Business	143	5%
Computer and Internet	57	2%
Economy and Finance	33	1%
Education	101	3%
Employment and Work	69	2%
Entertainment	419	14%
Food and Beverage	22	1%
Games	27	1%
Government and Politics	447	15%
Health	119	4%
Law	205	7%
Science	128	4%
Society and Culture	833	28%
Sports	61	2%
Travel	26	1%
Weather	16	1%

Tabela 19 – Relação entre a Quantidade de Documentos por Classe da Classificação CNN-18 no *Corpus* do CNN

distribuição de documentos por classe que foi estabilizada apenas na última classificação gerada, a que contêm apenas 5 classes.

O mapeamento entre as classes originais e as novas classes de cada classificação foi realizado de forma automatizada, isto é, através de mapeamento direto ($A \rightarrow B$) e pode ser observado na Tabela 20. É muito importante pontuar que a reclassificação dos textos

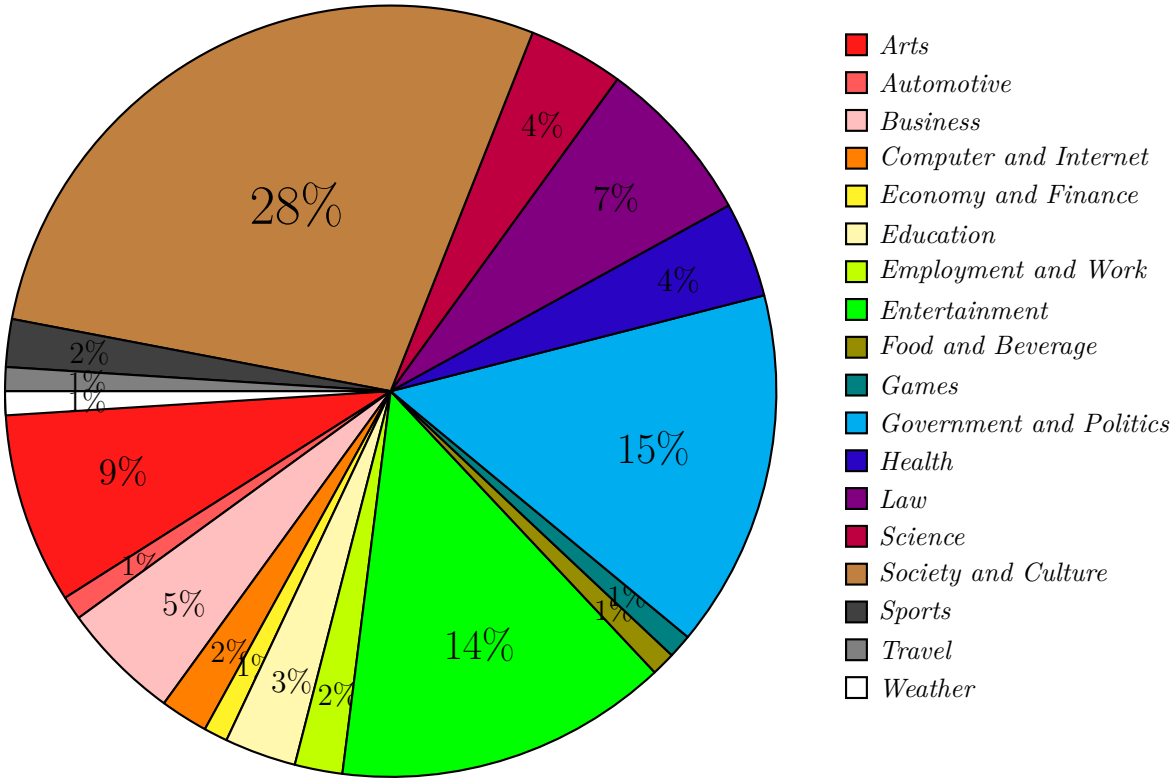

Figura 5 – Proporção das Classes do *Corpus* CNN no Nível 1

não afeta os objetivos a serem alcançados neste trabalho, uma vez que será comparada da mesma forma a performance dos modelos dos textos integrais com a dos textos resumidos nessas novas classificações.

Para simplificar e diferenciar essas 4 classificações, a original e as 3 novas criadas, elas serão indicadas de acordo com o seu número de classes: **CNN-18**, **CNN-11**, **CNN-8** e **CNN-5**. A classificação original **CNN-18** tem 18 classes; enquanto que a nova classificação **CNN-11** tem 11 classes; a classificação **CNN-8**, 8 classes e, por fim, a classificação **CNN-5**, 5 classes.

As classes de todas as classificações, **CNN-18**, **CNN-11**, **CNN-8** e **CNN-5**, bem como a quantidade e porcentagem de documentos por classe no nível de importância 1 estão exibidas nas Tabelas 19, 21, 22 e 23, respectivamente. As proporções entre os documentos de cada classe dessas mesmas classificações podem ser visualizadas nos gráficos das Figuras 5, 6, 7 e 8.

3.3 Métodos de Sumarização Extrativa

Para gerar os resumos a serem avaliados neste trabalho, foram selecionados os métodos de pontuação, já apresentados anteriormente na subseção 2.2.3, das três abordagens: baseados em palavras, baseados em sentenças e baseados em grafos. Para exemplificar as

Classificação CNN-18	Classificação CNN-11	Classificação CNN-8	Classificação CNN-5
Arts	*Arts*	*Arts and Entertainment*	*Arts, Entertainment and Society*
Automotive	*Technology*	*Science*	*Nature and Science*
Business	*Business*	*Business*	*Business*
Computer and Internet	*Technology*	*Science*	*Nature and Science*
Economy and Finance	*Business*	*Business*	*Business*
Education	*Education*	*Education*	*Nature and Science*
Employment and Work	*Business*	*Business*	*Business*
Entertainment	*Recreation*	*Arts and Entertainment*	*Health and Lifestyle*
Food and Beverage	*Lifestyle*	*Lifestyle*	*Health and Lifestyle*
Games	*Recreation*	*Arts and Entertainment*	*Health and Lifestyle*
Government and Politics	*Government and Politics*	*Government and Politics*	*Government and Politics*
Health	*Health*	*Health*	*Health and Lifestyle*
Law	*Government and Politics*	*Government and Politics*	*Government and Politics*
Science	*Science*	*Science*	*Nature and Science*
Society and Culture	*Society and Culture*	*Government and Politics*	*Arts, Entertainment and Society*
Sports	*Science*	*Science*	*Nature and Science*
Travel	*Lifestyle*	*Lifestyle*	*Health and Lifestyle*
Weather	*Environment and Nature*	*Environment and Nature*	*Nature and Science*

Tabela 20 – Mapeamento entre as Classes das 4 Classificações do *Corpus* do CNN

Classe	Nº de Documentos no Nível 1	% de Documentos no Nível 1
Arts	275	9%
Business	245	8%
Education	101	3%
Environment and Nature	16	1%
Government and Politics	652	22%
Health	119	4%
Lifestyle	48	2%
Recreation	507	17%
Science	128	4%
Society and Culture	833	28%
Technology	76	3%

Tabela 21 – Relação entre a Quantidade de Documentos por Classe da Classificação CNN-11 no *Corpus* do CNN

diferenças entre os resumos gerados por métodos diferentes no *corpus* CNN, o texto da Tabela 24 foi sumarizado tanto pelo método "*Cue Phrases*", quanto pelo método "*Word Frequency*". O sumário produzido pelo método "*Cue Phrases*" encontra-se presente na Tabela 25, enquanto que o sumário produzido pelo método "*Word Frequency*" encontra-se presente na Tabela 26.

Classe	Nº de Documentos no Nível 1	% de Documentos no Nível 1
Arts and Entertainment	782	26%
Business	245	8%
Education	101	3%
Environment and Nature	16	1%
Government and Politics	1485	50%
Health	119	4%
Lifestyle	48	2%
Science	204	7%

Tabela 22 – Relação entre a Quantidade de Documentos por Classe da Classificação CNN-8 no *Corpus* do CNN

Classe	Nº de Documentos no Nível 1	% de Documentos no Nível 1
Arts, Entertainment and Society	428	14%
Business	432	14%
Government and Politics	684	23%
Health and Lifestyle	352	12%
Nature and Science	1104	37%

Tabela 23 – Relação entre a Quantidade de Documentos por Classe da Classificação CNN-5 no *Corpus* do CNN

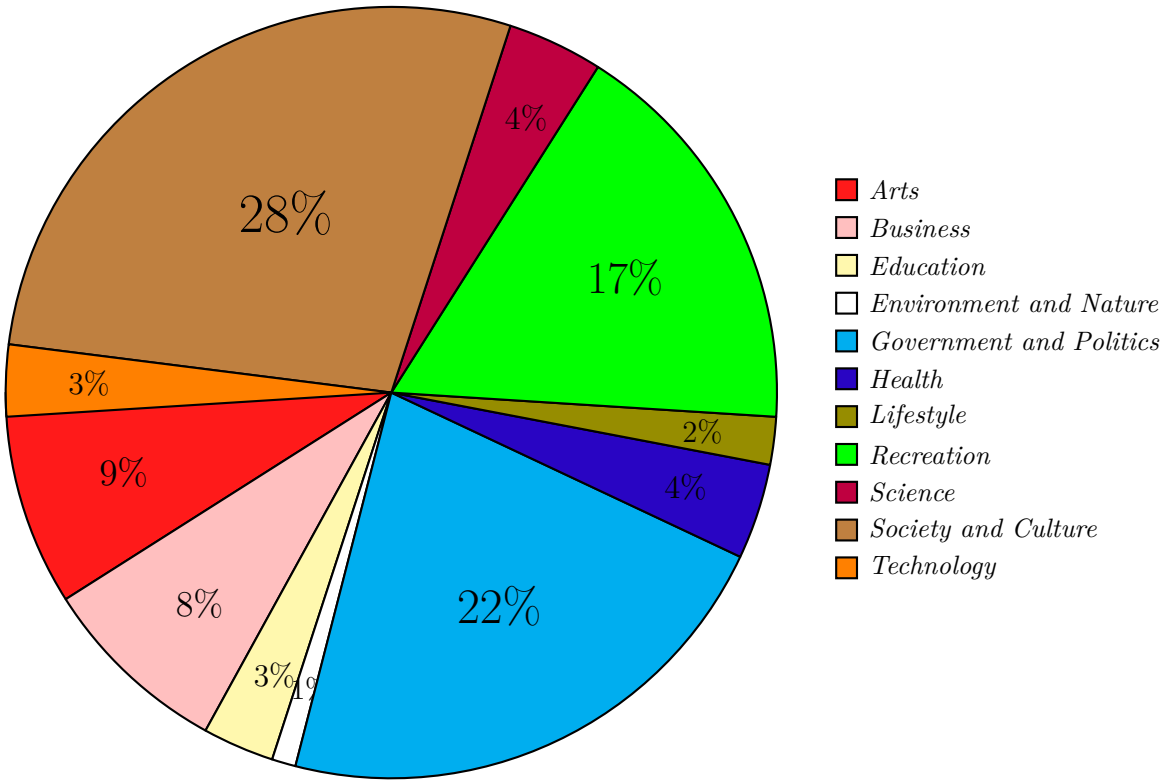

Figura 6 – Proporção das Classes da Classificação CNN-11 do *Corpus* CNN no Nível 1

1. A 17-year-old boy appeared in a California juvenile court on Tuesday in connection with a small fire that occurred during several wildfires last week in San Diego County, authorities said.
2. No further details were available on the nature of the juvenile court appearance for the teen, said spokeswoman Tanya Sierra of the San Diego County District Attorney's Office.
3. The juvenile court judge closed Tuesday's proceedings to the media, and authorities weren't releasing details because of the individual's juvenile status, Sierra said.
4. The 17-year-old boy was arrested last week with a 19-year-old man, but the older teenager was released and there will be no charges filed against him, Sierra said Tuesday.
5. Both individuals are from Escondido in San Diego County, and the small fire that prompted their arrest Thursday was extinguished, authorities said.
6. "We can only file charges when we believe we can prove them beyond a reasonable doubt, and in this case, after a thorough review of the case, the district attorney's office declined to file criminal charges"against the 19-year-old man, Sierra said.
7. Authorities have already charged Alberto Serrato with arson in connection with another small fire last week.
8. He is accused of adding brush onto an existing small fire in the San Luis Rey riverbed area in Oceanside, authorities said.
9. Serrato, 57, pleaded not guilty on Friday, Sierra said.
10. Serrato is the only person charged with arson in last week's wildfires, which at one point numbered about three dozen.
11. In all, nearly 26,000 acres, or about 40 square miles, were burned the past week by nine major fires in San Diego County, Cal Fire spokesman Daniel Berlant said Tuesday.
12. On Tuesday, all but three wildfires were completely contained.
13. The Cocos Fire in San Marco burned 1,995 acres and was 93% contained Tuesday, Cal Fire said.
14. The Pulgas Fire on the Marine Corps base Camp Pendleton burned 14,416 acres and was 99% contained Tuesday afternoon, Cal Fire said.
15. The San Mateo, or Combat Fire, also on Camp Pendleton, burned 1,457 acres and was 99% contained Tuesday afternoon, Cal Fire said.
16. The two fires and others on Camp Pendleton last week burned 21,900 acres, almost 18% of the base, the Marines said.
17. San Diego County authorities have formed a task force to investigate more than a dozen wildfires from last week, said Jan Caldwell, spokeswoman for the San Diego County Sheriff's Department.
18. "The purpose of the task force is to be a clearinghouse for investigators to look for similarities in these cases,"Caldwell said.
19. "They will do trend analysis and look for patterns."
20. A second task force with the district attorney's office and state insurance authorities was also warning homeowners on Tuesday to be wary of criminals who prey on victims of natural disasters, the district attorney's office said.

Tabela 24 – Exemplo de Texto Integral do *Corpus* CNN intitulado "*Boy, 17, appears in juvenile court in California wildfires case*"

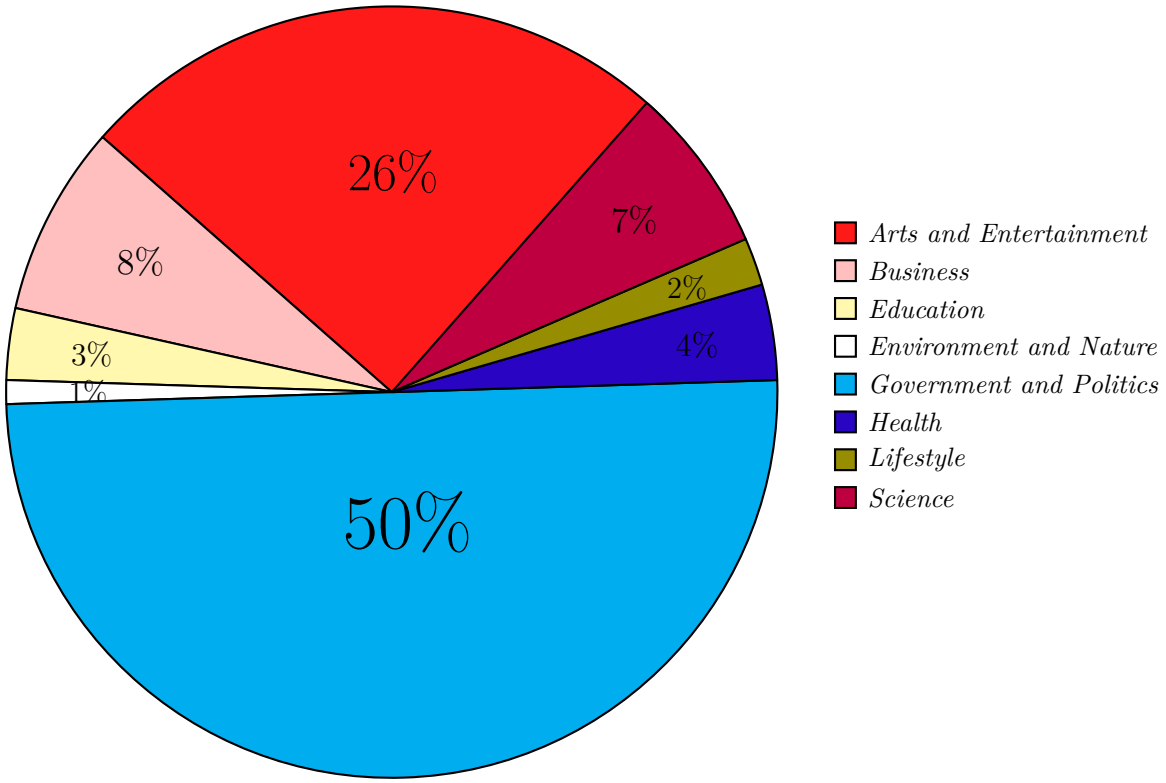

Figura 7 – Proporção das Classes da Classificação CNN-8 do *Corpus* CNN no Nível 1

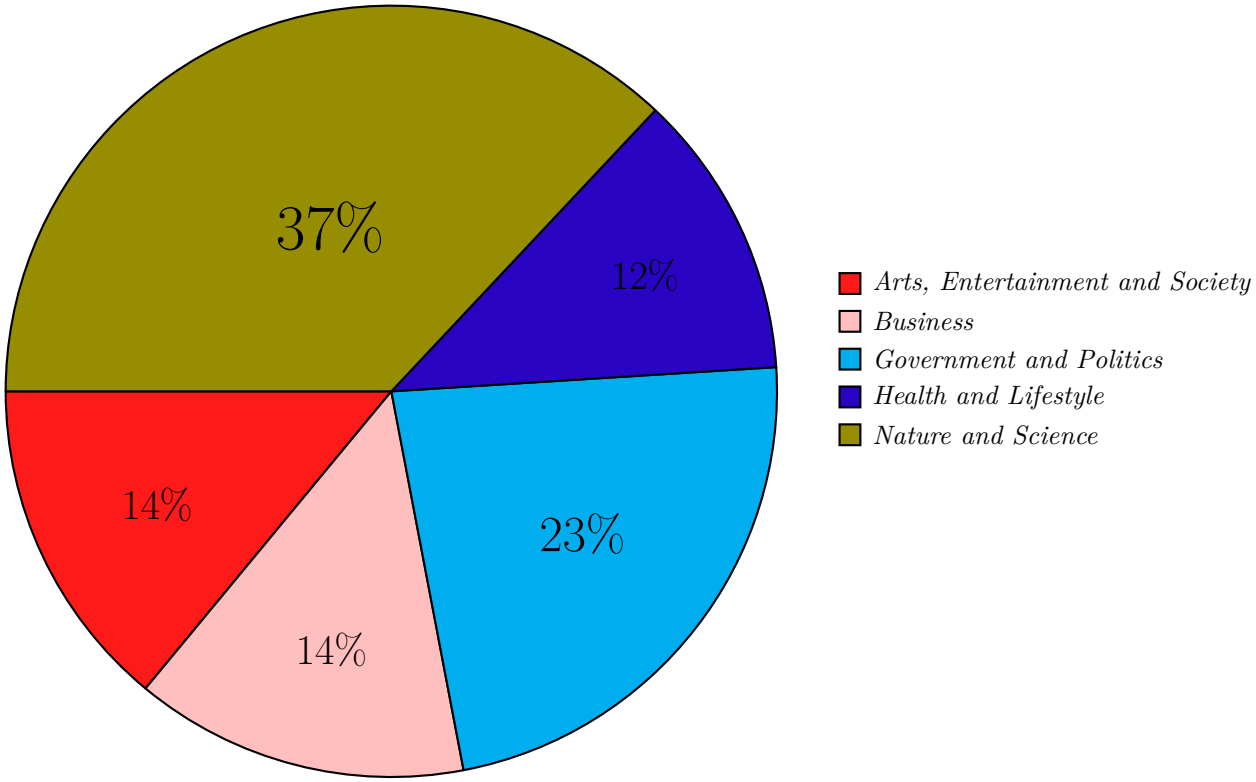

Figura 8 – Proporção das Classes da Classificação CNN-5 do *Corpus* CNN no Nível 1

3. The juvenile court judge closed Tuesday's proceedings to the media, and authorities weren't releasing details because of the individual's juvenile status, Sierra said. **13.** The Cocos Fire in San Marco burned 1,995 acres and was 93% contained Tuesday, Cal Fire said. **14.** The Pulgas Fire on the Marine Corps base Camp Pendleton burned 14,416 acres and was 99% contained Tuesday afternoon, Cal Fire said. **15.** The San Mateo, or Combat Fire, also on Camp Pendleton, burned 1,457 acres and was 99% contained Tuesday afternoon, Cal Fire said. **16.** The two fires and others on Camp Pendleton last week burned 21,900 acres, almost 18% of the base, the Marines said.

Tabela 25 – Exemplo de Texto Sumarizado pelo Método "*Cue Phrases*" do *Corpus* CNN intitulado "*Boy, 17, appears in juvenile court in California wildfires case*"

1. A 17-year-old boy appeared in a California juvenile court on Tuesday in connection with a small fire that occurred during several wildfires last week in San Diego County, authorities said. **11.** In all, nearly 26,000 acres, or about 40 square miles, were burned the past week by nine major fires in San Diego County, Cal Fire spokesman Daniel Berlant said Tuesday. **14.** The Pulgas Fire on the Marine Corps base Camp Pendleton burned 14,416 acres and was 99% contained Tuesday afternoon, Cal Fire said. **15.** The San Mateo, or Combat Fire, also on Camp Pendleton, burned 1,457 acres and was 99% contained Tuesday afternoon, Cal Fire said. **17.** San Diego County authorities have formed a task force to investigate more than a dozen wildfires from last week, said Jan Caldwell, spokeswoman for the San Diego County Sheriff's Department.

Tabela 26 – Exemplo de Texto Sumarizado pelo Método "*Word Frequency*" do *Corpus* CNN intitulado "*Boy, 17, appears in juvenile court in California wildfires case*"

3.4 Avaliação da Qualidade dos Sumários

Os sumários gerados pelas técnicas de sumarização apresentadas na seção anterior foram avaliados através das medidas (precisão, cobertura e *F-measure*) do método ROUGE-2 do *framework* ROUGE. Foram avaliadas as melhores técnicas de sumarização por classes no *corpus* CNN nos seus 3 níveis de importância utilizando a classificação original. Também foram combinadas as técnicas de sumarização 2 a 2 para cada sumário selecionando as 5 sentenças de maior pontuação dessas combinações e avaliando através da medida *MATCHES*, o número de sentenças que casam com o sumário de referência (*gold standard*).

Como existem *gold standards* para todos os textos do *corpus* do CNN, a medida *MATCHES* também foi aplicada para todo o conjunto de textos. Dessa forma foi possível fazer uma avaliação do desempenho geral de cada classe do *corpus*.

Assim como em (SILVA, 2017), neste trabalho também foram processados um

conjunto de documentos por classe. Logo, os resultados serão dispostos de forma macro. Isto é, exibindo valores da média das medidas de avaliação dos textos sumarizados. Os métodos de sumarização mais eficientes para cada classe de documentos do *corpus* CNN, nos 3 níveis de importância, foram obtidos considerando os maiores valores médios da *F-measure* do ROUGE-2 e, para as combinações de métodos 2 a 2, os maiores valores médios da medida *MATCHES*.

3.5 Modelos de Classificação Automática

Os modelos de classificação foram treinados a partir dos textos do *corpus* CNN em suas quatro classificações. Para o treinamento dos modelos de classificação, foram utilizados os textos completos e os resumos criados a partir dos métodos de sumarização extrativa já mencionados anteriormente. Para o pré-processamento dos documentos, foram empregadas os seguintes tratamentos: 1. tokenização dos termos dos documentos; 2. *case folding* para converter todas as palavras em minúsculas; 3. eliminação de *stopwords* e 4. redução dos termos ao seu radical (*stemming*).

Apesar de alguns autores afirmarem que o classificador Naive Bayes multi classes tem um desempenho pobre em classificação automática de textos (KAUR; BAJAJ, 2016; KHAN et al., 2010; KIM et al., 2006), nas análises iniciais, comparando com outros classificadores com SVMs, árvores de decisão e redes neurais, foi o que se mostrou mais escalável para execução dos experimentos. Um classificador Naive Bayes é um algoritmo bem conhecido e prático para gerar modelos classificadores probabilísticos e tem sido empregado com sucesso em muitos cenários como mineração de textos e até categorização de *spams* em *e-mails*. Esse classificador assume que todos os atributos dos documentos são independentes uns dos outros, dado o contexto da classe, ou seja, uma suposição de independência (KHAN et al., 2010; KIM et al., 2006).

Foi justamente essa simplicidade desse classificador que possibilitou testes em larga escala em um espaço de tempo relativamente curto fazendo uso de todos os documentos de texto do *corpus* (KHAN et al., 2010). Os outros algoritmos utilizados em análises iniciais, como SVM, árvores de decisão e redes neurais foram descontinuados, principalmente, devido ao *tradeoff* entre o tempo de treinamento demasiadamente elevado e a qualidade dos resultados obtidos.

3.6 Avaliação dos Modelos de Classificação

Para efetuar a avaliação e comparação dos modelos de classificação, foi utilizada a abordagem de treinamento e avaliação 10-*fold cross-validation* estratificado (k-*fold cross-validation* com k=10) que, como já explicado anteriormente, consiste em dividir o

corpus em 10 partes (10 "*folds*"), onde 9 serão utilizadas para treinamento e 1 parte para testes ao longo de 10 iterações, alternando a cada iteração a parte destinada à testes. A forma estratificada foi utilizada para reduzir a variação de exemplos utilizados entre as classes, isso é, para que em uma determinada divisão aleatória, não fossem utilizadas majoritariamente apenas documentos das classes mais populosas. Então, com esse tipo de variante na abordagem, é possível garantir que cada "*fold*" tenha aproximadamente a proporção correta de exemplos de cada uma das classes (ELKAN, 2012).

As medidas de avaliação utilizadas para mensurar os mais diversos aspectos dos modelos gerados foram: precisão (Fórmula 9), cobertura (Fórmula 10), acurácia (Fórmula 11), *F-measure* (Fórmula 12) e coeficiente Kappa (Fórmula 13). Todas essas medidas foram calculadas através das matrizes de confusão geradas após cada iteração de avaliação dos modelos na abordagem 10-*fold cross-validation* estratificado. É importante informar também que as medidas de precisão, cobertura, acurácia foram analisadas em sua forma macro, isto é, como a média de cada respectiva medida de todas as categorias (TEUFEL; SIDDHARTHAN; TIDHAR, 2006).

Por fim, vale ainda ressaltar que os modelos gerados pelos textos completos não foram utilizados para avaliar os resumos gerados, nem os modelos gerados pelos sumários foram utilizados para avaliar os textos completos. Cada modelo gerado avaliava apenas os documentos extraídos da mesma forma, isto é, os modelos gerados pelos resumos do método TF-IDF, só avaliavam outros resumos gerados pelo mesmo método.

3.7 Análise dos Resultados

Os resultados finais obtidos através da metodologia descrita neste capítulo foram analisados com o propósito de verificar o quanto a categoria de classificação de um documento consiste em um bom critério para escolher quais técnicas de sumarização extrativa empregar. Como já dito, anteriormente, essas possíveis relações entre os melhores métodos de sumarização por classes de documentos serão evidenciadas através dos resultados dos experimentos das duas conjunturas já mencionadas anteriormente.

Há também outra asserção a ser verificada através de experimentos: a comparação da performance da classificação de textos integrais com os sumarizados e, consequentemente, determinação das melhores técnicas de sumarização extrativa para cada classe de documentos de notícias. Nesse cenário que os modelos de classificação automática de textos são avaliados para confrontar a performance dos modelos treinados com textos integrais e textos sumarizados.

4 RESULTADOS E DISCUSSÃO

4.1 Trabalhos Anteriores

Como já mencionado anteriormente, este trabalho tem como base o sistema desenvolvido em (SILVA, 2017) que aqui foi estendido para realizar as investigações propostas neste trabalho. Esse trabalho anterior se propôs a descobrir se a classificação de um documento de notícias era um parâmetro adequado para determinar a melhor técnica de sumarização a ser escolhida para extrair um resumo genérico daquele documento. Os resultados obtidos nele, entretanto, mostraram experimentalmente que apenas a classificação do documento de texto não se mostrou como um bom critério para a escolha do método de sumarização a ser empregado.

Ainda é possível observar, com os resultados obtidos nesse trabalho anterior, uma predominância de técnicas estatísticas (como *Word Frequency* e *TF-IDF*) na geração dos resumos mais bem avaliados. As Tabelas 27 e 28 exibem os métodos cujos sumários gerados, a partir de documentos do *corpus* CNN no nível de importância 1 com a classificação original, demonstraram os melhores resultados perante às medidas de avaliação ROUGE-2 e *MATCHES*, respectivamente.

É importante relembrar que o escopo do trabalho anterior se limitou à utilização do primeiro nível de importância do *corpus* CNN. *Corpus*, este, que apresenta 3 níveis de importância em sua totalidade. Logo, na seção 4.2.1 serão apresentadas as melhores técnicas de sumarização por classes nesse mesmo *corpus* nos seus 3 níveis de importância (reforçando os resultados do primeiro nível novamente) utilizando a classificação original **CNN-18**. Ademais, as 5 melhores combinações das técnicas de sumarização 2 a 2 para cada classe também serão apresentadas.

4.2 Resultados Obtidos

A Tabela 29 exibe um resumo dos métodos de sumarização utilizados nos experimentos. Esses métodos foram apresentados anteriormente. Cada método nessa tabela apresenta um código, que servirá para identificá-lo facilmente na apresentação dos resultados obtidos. De forma análoga, as Tabelas 30, 31, 32 e 33 trazem esses mesmos códigos para as classes das classificações **CNN-18**, **CNN-11**, **CNN-8** e **CNN-5**, respectivamente. Essas classes também já apresentadas anteriormente.

Classe	Método de Sumarização mais Eficiente	$p_{ROUGE-2}$	$r_{ROUGE-2}$	$f1_{ROUGE-2}$
Arts	*TF-IDF*	0,27	0,48	0,34
Automotive	*Sentence Resemblance to the Title*	0,27	0,47	0,33
Business	*Word Frequency*	0,26	0,45	0,32
Computer and Internet	*TF-IDF*	0,24	0,45	0,31
Economy and Finance	*Word Frequency*	0,29	0,54	0,37
Education	*TF-IDF*	0,25	0,46	0,31
Employment and Work	*Word Frequency*	0,31	0,56	0,39
Entertainment	*Word Frequency*	0,26	0,45	0,32
Food and Beverage	*Sentence Resemblance to the Title*	0,39	0,51	0,44
Games	*Proper Noun*	0,37	0,57	0,44
Government and Politics	*Word Frequency*	0,28	0,48	0,35
Health	*Word Frequency*	0,23	0,43	0,29
Law	*TF-IDF*	0,29	0,53	0,37
Science	*Word Frequency*	0,25	0,45	0,32
Society and Culture	*TF-IDF*	0,26	0,46	0,32
Sports	*TF-IDF*	0,37	0,6	0,44
Travel	*TF-IDF*	0,18	0,33	0,23
Weather	*TF-IDF*	0,28	0,49	0,35

Tabela 27 – Melhores Métodos de Sumarização por Classes no *Corpus* do CNN no Nível 1 utilizando o ROUGE-2

Classe	Método de Sumarização mais Eficiente	*MATCHES*
Arts	*TF-IDF*	0,37
Automotive	*Sentence Resemblance to the Title*	0,43
Business	*Word Frequency*	0,36
Computer and Internet	*Upper Case*	0,34
Economy and Finance	*Word Frequency*	0,42
Education	*Word Frequency*	0,37
Employment and Work	*Word Frequency*	0,43
Entertainment	*Word Frequency*	0,36
Food and Beverage	*Sentence Resemblance to the Title*	0,41
Games	*Proper Noun*	0,50
Government and Politics	*Word Frequency*	0,38
Health	*Word Frequency*	0,32
Law	*Word Frequency*	0,42
Science	*Word Frequency*	0,35
Society and Culture	*TF-IDF*	0,34
Sports	*Word Frequency*	0,50
Travel	*TF-IDF*	0,24
Weather	*TF-IDF*	0,37

Tabela 28 – Melhores Métodos de Sumarização por Classes no *Corpus* do CNN no Nível 1 utilizando a Medida *MATCHES*

Abordagem do Método	**Código do Método**	**Método de Sumarização**
Texto Integral	FT	Nenhum
Baseado em Palavras	WF_W	*Word Frequency*
	TF_W	*TF-IDF*
	UC_W	*Upper Case*
	PN_W	*Proper Noun*
	WC_W	*Word Co-occurrence*
	LS_W	*Lexical Similarity*
Baseado em Sentenças	RT_S	*Resemblance to the Title*
	SC_S	*Sentence Centrality*
	SP_S	*Sentence Position*
	CP_S	*Cue Phrases*
	ND_S	*Numerical Data*
	SL_S	*Sentence Length*
Baseado em Grafos	TR_G	*Text Rank*
	BP_G	*Bushy Path*
	AS_G	*Aggregate Similarity*

Tabela 29 – Conjunto de Métodos de Sumarização Extrativa

Código da Classe	Nome da Classe
AR	Arts
AU	Automotive
BS	Business
CO	Computer and Internet
EC	Economy and Finance
ED	Education
EW	Employment and Work
ET	Entertainment
FO	Food and Beverage
GA	Games
GP	Government and Politics
HT	Health
LA	Law
SC	Science
SO	Society and Culture
SP	Sports
TR	Travel
WT	Weather

Tabela 30 – Conjunto de Classes do *Corpus* CNN com a Classificação CNN-18

Código da Classe	Nome da Classe
AR	Arts and Entertainment
BS	Business
ED	Education
EN	Environment and Nature
GP	Government and Politics
HT	Health
LF	Lifestyle
RC	Recreation
SC	Science
SO	Society and Culture
TC	Technology

Tabela 31 – Conjunto de Classes do *Corpus* CNN com a Classificação CNN-11

Código da Classe	Nome da Classe
AR	Arts and Entertainment
BS	Business
ED	Education
EM	Environment and Nature
GP	Government and politics
HT	Health
LF	Lifestyle
SC	Science

Tabela 32 – Conjunto de Classes do *Corpus* CNN com a Classificação CNN-8

Código da Classe	Nome da Classe
AR	Arts, Entertainment and Society
BU	Business
GP	Government and Politics
HL	Health and Lifestyle
NS	Nature and Science

Tabela 33 – Conjunto de Classes do *Corpus* CNN com a Classificação CNN-5

4.2.1 Sumarização de Textos Utilizando Classes de Documentos

Os resultados apresentados aqui nesta subseção são provenientes da análise da primeira conjectura. Esse pressuposto descreve a utilização de técnicas de sumarização extrativa, bem como de combinações entre elas, para estabelecer uma relação entre as melhores técnicas e cada classe de documento de notícias.

Conforme esperado, os resultados exibidos na Tabelas 34 e 35 se igualam exatamente aos resultados já apresentados em (SILVA, 2017). Esses resultados são os valores médios das medidas ROUGE-2 *F-measure* e *MATCHES*, nessa ordem, utilizando o *corpus* do CNN no seu nível de importância 1.

Agora, os resultados exibidos nas Tabelas 36, 37, 38 e 39 são inéditos. Eles seguem a mesma lógica dos resultados das Tabelas 34 e 35, entretanto, fazem o uso de até dois ou três níveis de importância. No caso das Tabelas 36 e 38 são exibidos os resultados da medida ROUGE-2 *F-measure*, enquanto que nas Tabelas 37 e 39 são exibidos os resultados da medida *MATCHES*.

Para efeitos de simplificação, as Tabelas 40, 42 e 44 trazem um compilado dos 3 melhores métodos por classe utilizando a medida ROUGE-2 *F-measure*. Já as Tabelas 41, 43 e 45 trazem um compilado dos 3 melhores métodos por classe utilizando a medida *MATCHES*. São todos os mesmos resultados já exibidos nas Tabelas 34, 35, 36, 37, 38 e 39, contudo, de maneira mais resumida.

Por fim, para não ser mais exaustivo, a Tabela 46 traz o resultado das combinações das técnicas de sumarização. Mais especificamente, essa Tabela 46 exibe os valores médios da medida *MATCHES* das 5 melhores combinações de métodos de sumarização, 2 a 2, para cada uma das classes da classificação original do *corpus* CNN.

Cl.	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,33	0,34	0,27	0,28	0,22	0,13	0,27	0,08	0,26	0,17	0,24	0,26	0,17	0,19	0,18
AU	0,32	0,32	0,30	0,30	0,15	0,23	0,33	0,14	0,24	0,30	0,32	0,26	0,30	0,30	0,29
BS	0,32	0,32	0,25	0,26	0,22	0,13	0,31	0,11	0,27	0,22	0,26	0,26	0,22	0,23	0,22
CO	0,30	0,31	0,3	0,27	0,21	0,13	0,28	0,10	0,26	0,20	0,25	0,27	0,20	0,23	0,20
EC	0,37	0,37	0,29	0,28	0,34	0,21	0,22	0,14	0,28	0,22	0,26	0,32	0,22	0,28	0,26
ED	0,31	0,31	0,26	0,24	0,23	0,16	0,29	0,11	0,26	0,17	0,26	0,25	0,17	0,19	0,19
EW	0,39	0,38	0,32	0,33	0,27	0,15	0,37	0,15	0,30	0,29	0,28	0,32	0,29	0,30	0,28
ET	0,32	0,32	0,27	0,27	0,22	0,14	0,29	0,10	0,26	0,18	0,24	0,27	0,18	0,20	0,19
FO	0,37	0,36	0,30	0,28	0,33	0,16	0,44	0,14	0,29	0,23	0,32	0,32	0,23	0,24	0,28
GA	0,41	0,42	0,42	0,44	0,23	0,20	0,36	0,14	0,35	0,29	0,37	0,31	0,29	0,21	0,17
GP	0,35	0,34	0,31	0,31	0,25	0,16	0,34	0,10	0,30	0,21	0,26	0,29	0,21	0,23	0,22
HT	0,29	0,28	0,24	0,26	0,20	0,13	0,27	0,09	0,26	0,16	0,26	0,25	0,16	0,19	0,19
LA	0,37	0,37	0,33	0,33	0,28	0,18	0,36	0,11	0,32	0,24	0,29	0,32	0,24	0,23	0,23
SC	0,32	0,31	0,25	0,24	0,22	0,13	0,26	0,09	0,23	0,16	0,21	0,25	0,16	0,19	0,18
SO	0,32	0,32	0,26	0,26	0,22	0,13	0,29	0,08	0,26	0,17	0,23	0,26	0,17	0,18	0,17
SP	0,44	0,44	0,35	0,37	0,26	0,12	0,40	0,10	0,37	0,25	0,32	0,36	0,25	0,22	0,19
TR	0,20	0,23	0,16	0,17	0,09	0,09	0,22	0,09	0,11	0,16	0,22	0,22	0,16	0,22	0,18
WT	0,35	0,35	0,24	0,26	0,29	0,20	0,30	0,12	0,32	0,23	0,26	0,32	0,23	0,26	0,26

Tabela 34 – Resultados da Medida ROUGE-2 *F-measure* por Classe das Avaliações dos Sumários com a Classificação CNN-18 no Nível 1

Cl.	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,37	0,37	0,28	0,30	0,00	0,13	0,29	0,09	0,26	0,18	0,25	0,27	0,18	0,20	0,19
AU	0,40	0,39	0,34	0,35	0,14	0,26	0,43	0,16	0,26	0,34	0,38	0,29	0,34	0,33	0,36
BS	0,36	0,35	0,25	0,28	0,20	0,13	0,34	0,11	0,28	0,22	0,29	0,28	0,22	0,24	0,23
CO	0,32	0,34	0,34	0,31	0,20	0,12	0,31	0,11	0,27	0,19	0,27	0,29	0,19	0,23	0,21
EC	0,42	0,42	0,32	0,31	0,35	0,22	0,24	0,15	0,29	0,23	0,28	0,36	0,23	0,30	0,27
ED	0,37	0,35	0,28	0,25	0,21	0,17	0,32	0,11	0,27	0,18	0,28	0,26	0,18	0,19	0,19
EW	0,43	0,42	0,34	0,35	0,28	0,15	0,41	0,18	0,32	0,29	0,30	0,35	0,29	0,31	0,30
ET	0,36	0,35	0,29	0,29	0,21	0,14	0,32	0,11	0,27	0,18	0,25	0,28	0,18	0,21	0,20
FO	0,38	0,36	0,28	0,26	0,27	0,15	0,41	0,15	0,28	0,21	0,30	0,34	0,21	0,24	0,28
GA	0,46	0,48	0,47	0,50	0,24	0,20	0,41	0,12	0,35	0,30	0,39	0,32	0,30	0,23	0,18
GP	0,38	0,38	0,33	0,33	0,24	0,16	0,37	0,11	0,31	0,22	0,27	0,30	0,22	0,24	0,23
HT	0,32	0,30	0,25	0,27	0,18	0,13	0,29	0,10	0,27	0,15	0,27	0,26	0,15	0,19	0,20
LA	0,42	0,42	0,35	0,35	0,28	0,18	0,40	0,12	0,34	0,25	0,32	0,35	0,25	0,24	0,25
SC	0,35	0,34	0,26	0,25	0,20	0,12	0,29	0,08	0,23	0,16	0,21	0,27	0,16	0,19	0,19
SO	0,34	0,34	0,26	0,26	0,20	0,12	0,30	0,08	0,26	0,16	0,23	0,25	0,16	0,17	0,17
SP	0,50	0,50	0,37	0,41	0,24	0,10	0,48	0,09	0,40	0,25	0,34	0,39	0,25	0,23	0,19
TR	0,19	0,24	0,17	0,18	0,07	0,06	0,22	0,07	0,09	0,14	0,23	0,22	0,14	0,23	0,18
WT	0,35	0,37	0,24	0,26	0,28	0,20	0,32	0,16	0,33	0,24	0,28	0,34	0,24	0,27	0,26

Tabela 35 – Resultados da Medida *MATCHES* por Classe das Avaliações dos Sumários com a Classificação CNN-18 no Nível 1

4.2.2 Classificação de Textos Utilizando Sumários de Documentos

Agora, os resultados apresentados aqui nesta subseção são provenientes da análise da segunda conjectura. Esse pressuposto descreve a comparação do desempenho da classificação de textos integrais com os sumarizados e determinação das melhores técnicas de sumarização extrativa para cada classe de documentos de notícias.

As Tabelas 47, 48, 49 e 50 exibem os resultados macro, em forma de médias, das avaliações dos modelos treinados com as classificações **CNN-18**, **CNN-11**, **CNN-8** e **CNN-5**, nessa ordem. Enquanto as Tabelas 51, 53, 55 e 57 exibem uma relação entre os métodos de sumarização e os resultados para cada classe através dos valores da medida *F-measure*. As Tabelas 52, 54, 56 e 58 também exibem uma relação entre os métodos de sumarização e os resultados para cada classe através dos valores da acurácia. Através

Cl.	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,33	0,33	0,26	0,27	0,22	0,32	0,27	0,33	0,23	0,17	0,24	0,26	0,32	0,19	0,18
AU	0,31	0,30	0,29	0,30	0,20	0,31	0,34	0,34	0,25	0,29	0,30	0,27	0,32	0,28	0,27
BS	0,31	0,31	0,25	0,26	0,22	0,31	0,30	0,33	0,24	0,23	0,26	0,26	0,31	0,22	0,22
CO	0,33	0,31	0,29	0,27	0,23	0,31	0,30	0,34	0,21	0,18	0,26	0,28	0,31	0,22	0,19
EC	0,34	0,34	0,27	0,27	0,29	0,32	0,23	0,32	0,22	0,20	0,24	0,29	0,32	0,25	0,23
ED	0,35	0,34	0,27	0,26	0,23	0,32	0,30	0,34	0,24	0,20	0,26	0,26	0,33	0,20	0,19
EW	0,38	0,37	0,31	0,31	0,27	0,35	0,35	0,38	0,28	0,25	0,28	0,32	0,37	0,27	0,27
ET	0,31	0,31	0,26	0,26	0,22	0,30	0,28	0,32	0,22	0,18	0,24	0,25	0,31	0,19	0,19
FO	0,34	0,32	0,28	0,26	0,31	0,32	0,38	0,34	0,24	0,23	0,29	0,27	0,36	0,28	0,32
GA	0,41	0,40	0,37	0,38	0,24	0,38	0,34	0,41	0,25	0,29	0,33	0,32	0,37	0,18	0,19
GP	0,34	0,34	0,30	0,30	0,25	0,33	0,33	0,34	0,27	0,21	0,26	0,29	0,34	0,22	0,22
HT	0,30	0,28	0,25	0,26	0,22	0,27	0,27	0,30	0,23	0,18	0,25	0,25	0,31	0,19	0,19
LA	0,36	0,37	0,33	0,33	0,27	0,35	0,35	0,37	0,30	0,22	0,29	0,31	0,34	0,22	0,22
SC	0,32	0,32	0,26	0,26	0,23	0,31	0,27	0,32	0,21	0,16	0,23	0,25	0,31	0,20	0,20
SO	0,32	0,32	0,26	0,26	0,22	0,30	0,29	0,32	0,22	0,17	0,23	0,26	0,31	0,18	0,18
SP	0,45	0,44	0,35	0,36	0,26	0,40	0,40	0,45	0,30	0,25	0,31	0,35	0,41	0,22	0,19
TR	0,20	0,22	0,17	0,17	0,11	0,19	0,23	0,23	0,12	0,15	0,20	0,22	0,19	0,20	0,16
WT	0,37	0,39	0,33	0,35	0,31	0,37	0,35	0,42	0,31	0,29	0,29	0,38	0,39	0,30	0,30

Tabela 36 – Resultados da Medida ROUGE-2 *F-measure* por Classe das Avaliações dos Sumários com a Classificação CNN-18 no Nível 2

Cl.	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,37	0,37	0,28	0,29	0,22	0,35	0,30	0,37	0,23	0,17	0,25	0,28	0,34	0,21	0,20
AU	0,39	0,36	0,34	0,36	0,20	0,38	0,45	0,41	0,27	0,34	0,35	0,32	0,36	0,32	0,33
BS	0,35	0,35	0,26	0,27	0,20	0,34	0,34	0,37	0,25	0,24	0,29	0,28	0,33	0,23	0,23
CO	0,38	0,37	0,34	0,31	0,24	0,36	0,35	0,40	0,23	0,18	0,29	0,32	0,33	0,24	0,21
EC	0,39	0,39	0,31	0,30	0,31	0,37	0,27	0,36	0,24	0,20	0,26	0,33	0,35	0,28	0,25
ED	0,41	0,39	0,29	0,28	0,21	0,36	0,34	0,39	0,24	0,20	0,28	0,28	0,36	0,21	0,20
EW	0,43	0,43	0,33	0,33	0,28	0,39	0,40	0,44	0,29	0,25	0,31	0,35	0,41	0,28	0,31
ET	0,36	0,35	0,29	0,29	0,21	0,34	0,31	0,36	0,22	0,18	0,26	0,27	0,33	0,21	0,20
FO	0,36	0,34	0,28	0,28	0,28	0,32	0,38	0,36	0,23	0,22	0,30	0,30	0,35	0,30	0,36
GA	0,47	0,47	0,43	0,43	0,25	0,43	0,40	0,47	0,26	0,30	0,34	0,36	0,39	0,19	0,21
GP	0,38	0,37	0,32	0,32	0,24	0,37	0,36	0,38	0,27	0,22	0,27	0,30	0,36	0,23	0,23
HT	0,34	0,32	0,27	0,28	0,22	0,30	0,30	0,35	0,24	0,19	0,28	0,27	0,34	0,21	0,21
LA	0,42	0,42	0,36	0,36	0,27	0,39	0,40	0,42	0,31	0,24	0,32	0,34	0,37	0,23	0,24
SC	0,36	0,36	0,28	0,27	0,22	0,35	0,31	0,37	0,22	0,16	0,23	0,27	0,34	0,22	0,21
SO	0,35	0,35	0,27	0,28	0,20	0,33	0,31	0,35	0,22	0,17	0,24	0,27	0,32	0,18	0,18
SP	0,53	0,51	0,39	0,41	0,25	0,45	0,47	0,53	0,32	0,25	0,34	0,38	0,45	0,24	0,20
TR	0,20	0,23	0,17	0,18	0,09	0,20	0,26	0,25	0,10	0,14	0,21	0,23	0,16	0,22	0,17
WT	0,39	0,43	0,36	0,39	0,31	0,40	0,40	0,45	0,33	0,31	0,33	0,41	0,44	0,32	0,30

Tabela 37 – Resultados da Medida *MATCHES* por Classe das Avaliações dos Sumários com a Classificação CNN-18 no Nível 2

dessas relações é possível identificar em quais classes de cada classificação, os modelos apresentaram os melhores e os piores resultados e, consequentemente, também identificar os métodos de sumarização mais adequados para cada uma das classes.

Os valores NaN presentes em algumas células dessas tabelas são valores não numéricos, isto é, provenientes de resultados de divisão por 0. Pelo cálculo da medida *F-measure*, mostrado na Fórmula 12, essas incidências de NaN ocorreram quando os modelos apresentaram valores de precisão e/ou cobertura iguais à 0. Mais especificamente, em casos que o modelo não conseguiu acertar nenhum exemplo para uma determinada classe.

Para simplificar, as Tabelas 59 e 60 trazem um compilado dos 3 melhores métodos por classe utilizando a medidas *F-measure* e a acurácia, respectivamente. De forma análoga

Cl.	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,33	0,34	0,27	0,28	0,23	0,32	0,28	0,33	0,23	0,17	0,24	0,27	0,33	0,20	0,19
AU	0,31	0,30	0,29	0,30	0,20	0,31	0,35	0,34	0,26	0,30	0,30	0,26	0,33	0,28	0,28
BS	0,32	0,32	0,25	0,26	0,22	0,31	0,30	0,33	0,25	0,23	0,26	0,26	0,31	0,22	0,22
CO	0,33	0,31	0,30	0,27	0,23	0,31	0,30	0,33	0,20	0,18	0,26	0,28	0,31	0,23	0,19
EC	0,34	0,33	0,26	0,27	0,29	0,32	0,24	0,32	0,23	0,21	0,24	0,28	0,32	0,25	0,23
ED	0,36	0,35	0,29	0,27	0,24	0,33	0,31	0,35	0,24	0,20	0,26	0,26	0,33	0,20	0,20
EW	0,38	0,38	0,32	0,31	0,28	0,36	0,35	0,38	0,28	0,25	0,28	0,32	0,38	0,27	0,26
ET	0,32	0,31	0,27	0,27	0,22	0,30	0,28	0,32	0,22	0,18	0,24	0,26	0,31	0,20	0,19
FO	0,34	0,33	0,27	0,27	0,30	0,33	0,35	0,34	0,24	0,22	0,28	0,26	0,34	0,28	0,30
GA	0,41	0,41	0,37	0,38	0,25	0,38	0,36	0,42	0,26	0,27	0,32	0,33	0,37	0,19	0,19
GP	0,34	0,34	0,30	0,30	0,25	0,33	0,33	0,34	0,27	0,22	0,26	0,29	0,34	0,23	0,22
HT	0,30	0,29	0,25	0,26	0,22	0,27	0,27	0,31	0,23	0,18	0,26	0,25	0,31	0,20	0,19
LA	0,36	0,37	0,33	0,33	0,27	0,35	0,35	0,37	0,30	0,23	0,29	0,31	0,34	0,22	0,22
SC	0,32	0,31	0,26	0,26	0,22	0,32	0,27	0,32	0,21	0,16	0,23	0,26	0,31	0,20	0,20
SO	0,32	0,32	0,26	0,26	0,22	0,30	0,29	0,32	0,22	0,17	0,23	0,26	0,32	0,18	0,18
SP	0,45	0,44	0,35	0,36	0,27	0,40	0,41	0,45	0,31	0,24	0,31	0,34	0,41	0,21	0,18
TR	0,21	0,23	0,16	0,16	0,12	0,20	0,22	0,24	0,12	0,13	0,19	0,21	0,18	0,18	0,14
WT	0,39	0,40	0,35	0,37	0,31	0,39	0,36	0,44	0,30	0,30	0,31	0,39	0,42	0,30	0,31

Tabela 38 – Resultados da Medida ROUGE-2 *F-measure* por Classe das Avaliações dos Sumários com a Classificação CNN-18 no Nível 3

Cl.	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,38	0,38	0,29	0,30	0,23	0,36	0,31	0,38	0,23	0,18	0,26	0,29	0,35	0,21	0,21
AU	0,39	0,36	0,34	0,36	0,22	0,37	0,45	0,40	0,30	0,35	0,37	0,31	0,36	0,32	0,33
BS	0,35	0,36	0,26	0,27	0,21	0,34	0,35	0,37	0,26	0,24	0,28	0,28	0,33	0,23	0,23
CO	0,38	0,36	0,34	0,31	0,24	0,36	0,34	0,39	0,21	0,17	0,30	0,32	0,33	0,24	0,21
EC	0,39	0,38	0,30	0,30	0,31	0,37	0,28	0,37	0,25	0,21	0,26	0,32	0,35	0,27	0,25
ED	0,42	0,40	0,30	0,29	0,22	0,37	0,35	0,40	0,25	0,20	0,29	0,28	0,36	0,21	0,20
EW	0,43	0,43	0,33	0,33	0,28	0,39	0,40	0,43	0,29	0,24	0,31	0,35	0,40	0,28	0,29
ET	0,36	0,35	0,29	0,29	0,21	0,34	0,32	0,36	0,23	0,19	0,26	0,28	0,33	0,21	0,21
FO	0,37	0,36	0,29	0,30	0,28	0,35	0,36	0,38	0,25	0,22	0,30	0,29	0,34	0,31	0,35
GA	0,48	0,47	0,43	0,43	0,26	0,45	0,42	0,49	0,27	0,29	0,35	0,38	0,40	0,22	0,22
GP	0,38	0,37	0,32	0,32	0,24	0,37	0,37	0,38	0,27	0,22	0,27	0,31	0,36	0,24	0,23
HT	0,34	0,32	0,27	0,28	0,22	0,30	0,30	0,35	0,24	0,19	0,29	0,27	0,34	0,21	0,20
LA	0,42	0,42	0,36	0,36	0,28	0,39	0,40	0,42	0,31	0,24	0,32	0,35	0,37	0,24	0,24
SC	0,37	0,35	0,28	0,27	0,22	0,36	0,31	0,37	0,22	0,16	0,24	0,28	0,34	0,21	0,22
SO	0,35	0,35	0,27	0,28	0,20	0,33	0,31	0,35	0,22	0,17	0,24	0,27	0,32	0,18	0,18
SP	0,53	0,51	0,39	0,42	0,25	0,45	0,48	0,53	0,34	0,24	0,35	0,38	0,46	0,23	0,19
TR	0,20	0,23	0,16	0,15	0,10	0,19	0,23	0,25	0,12	0,11	0,19	0,22	0,16	0,19	0,15
WT	0,43	0,45	0,40	0,42	0,31	0,42	0,42	0,49	0,33	0,32	0,35	0,43	0,47	0,32	0,32

Tabela 39 – Resultados da Medida *MATCHES* por Classe das Avaliações dos Sumários com a Classificação CNN-18 no Nível 3

à tabelas com os 3 melhores resultados da conjectura anterior, essas duas tabelas contêm de maneira mais resumida os mesmos resultados já exibidos nas Tabelas 51 e 52.

No final, ainda é apresentado na Figura 9 um gráfico de barras agrupando os valores do coeficiente Kappa para cada uma das 4 classificações do *corpus* CNN nos métodos de sumarização automática de textos. Nesse gráfico ainda é possível visualizar quais classificações, no geral, apresentaram uma adequação melhor dos modelos.

4.3 Considerações Finais

Para efeitos práticos, os resultados apresentados na seção anterior são suficientes para avaliar as duas principais conjecturas investigadas neste trabalho. Foi verificada a

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W (0,34)	WF_W (0,33)	PN_W (0,28)
AU	RT_S (0,33)	WF_W (0,32)	TF_W (0,32)
BS	WF_W (0,32)	TF_W (0,32)	RT_S (0,31)
CO	TF_W (0,31)	UC_W (0,30)	WF_W (0,30)
EC	WF_W (0,37)	TF_W (0,37)	WC_W (0,34)
ED	TF_W (0,31)	WF_W (0,31)	RT_S (0,29)
EW	WF_W (0,39)	TF_W (0,38)	RT_S (0,37)
ET	WF_W (0,32)	TF_W (0,32)	RT_S (0,29)
FO	RT_S (0,44)	WF_W (0,37)	TF_W (0,36)
GA	PN_W (0,44)	TF_W (0,42)	UC_W (0,42)
GP	WF_W (0,35)	TF_W (0,34)	RT_S (0,34)
HT	WF_W (0,29)	TF_W (0,28)	RT_S (0,27)
LA	TF_W (0,37)	WF_W (0,37)	RT_S (0,36)
SC	WF_W (0,32)	TF_W (0,31)	RT_S (0,26)
SO	TF_W (0,32)	WF_W (0,32)	RT_S (0,29)
SP	TF_W (0,44)	WF_W (0,44)	RT_S (0,40)
TR	TF_W (0,23)	ND_S (0,22)	BP_G (0,22)
WT	TF_W (0,35)	WF_W (0,35)	SP_S (0,32)

Tabela 40 – 3 Melhores Resultados da Medida ROUGE-2 *F-measure* por Classe na Classificação CNN-18 no Nível 1

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W (0,37)	WF_W (0,37)	PN_W (0,30)
AU	RT_S (0,43)	WF_W (0,40)	TF_W (0,39)
BS	WF_W (0,36)	TF_W (0,35)	RT_S (0,34)
CO	UC_W (0,34)	TF_W (0,34)	WF_W (0,32)
EC	WF_W (0,42)	TF_W (0,42)	SL_S (0,36)
ED	WF_W (0,37)	TF_W (0,35)	RT_S (0,32)
EW	WF_W (0,43)	TF_W (0,42)	RT_S (0,41)
ET	WF_W (0,36)	TF_W (0,35)	RT_S (0,32)
FO	RT_S (0,41)	WF_W (0,38)	TF_W (0,36)
GA	PN_W (0,50)	TF_W (0,48)	UC_W (0,47)
GP	WF_W (0,38)	TF_W (0,38)	RT_S (0,37)
HT	WF_W (0,32)	TF_W (0,30)	RT_S (0,29)
LA	WF_W (0,42)	TF_W (0,42)	RT_S (0,40)
SC	WF_W (0,35)	TF_W (0,34)	RT_S (0,29)
SO	TF_W (0,34)	WF_W (0,34)	RT_S (0,30)
SP	WF_W (0,50)	TF_W (0,50)	RT_S (0,48)
TR	TF_W (0,24)	BP_G (0,23)	ND_S (0,23)
WT	TF_W (0,37)	WF_W (0,35)	SL_S (0,34)

Tabela 41 – 3 Melhores Resultados da Medida *MATCHES* por Classe na Classificação CNN-18 no Nível 1

conjectura relativa à utilização de técnicas de sumarização extrativa, e combinações entre elas, para estabelecer uma relação entre as melhores técnicas e cada uma das classes de documentos de notícias. Por fim, a de comparação da performance da classificação de textos integrais com os sumarizados e consequente determinação das melhores técnicas de sumarização extrativa para cada classe de documentos de notícias.

Através da análise dos resultados das Tabelas 34, 35, 36, 37, 36, 37 e 46, foi possível observar que quando o *corpus* foi utilizado em seu segundo e até terceiro níveis de

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W (0,33)	SC_S (0,33)	WF_W (0,33)
AU	RT_S (0,34)	SC_S (0,34)	TR_G (0,32)
BS	SC_S (0,33)	TF_W (0,31)	WF_W (0,31)
CO	SC_S (0,34)	WF_W (0,33)	TF_W (0,31)
EC	TF_W (0,34)	WF_W (0,34)	SC_S (0,32)
ED	WF_W (0,35)	TF_W (0,34)	SC_S (0,34)
EW	SC_S (0,38)	WF_W (0,38)	TF_W (0,37)
ET	SC_S (0,32)	WF_W (0,31)	TF_W (0,31)
FO	RT_S (0,38)	TR_G (0,36)	SC_S (0,34)
GA	SC_S (0,41)	WF_W (0,41)	TF_W (0,40)
GP	SC_S (0,34)	TR_G (0,34)	WF_W (0,34)
HT	TR_G (0,31)	SC_S (0,30)	WF_W (0,30)
LA	TF_W (0,37)	SC_S (0,37)	WF_W (0,36)
SC	WF_W (0,32)	SC_S (0,32)	TF_W (0,32)
SO	TF_W (0,32)	WF_W (0,32)	SC_S (0,32)
SP	SC_S (0,45)	WF_W (0,45)	TF_W (0,44)
TR	RT_S (0,23)	SC_S (0,23)	TF_W (0,22)
WT	SC_S (0,42)	TR_G (0,39)	TF_W (0,39)

Tabela 42 – 3 Melhores Resultados da Medida ROUGE-2 *F-measure* por Classe na Classificação CNN-18 no Nível 2

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W (0,37)	WF_W (0,37)	SC_S (0,37)
AU	RT_S (0,45)	SC_S (0,41)	WF_W (0,39)
BS	SC_S (0,37)	TF_W (0,35)	WF_W (0,35)
CO	SC_S (0,40)	WF_W (0,38)	TF_W (0,37)
EC	TF_W (0,39)	WF_W (0,39)	LS_W (0,37)
ED	WF_W (0,41)	TF_W (0,39)	SC_S (0,39)
EW	SC_S (0,44)	WF_W (0,43)	TF_W (0,43)
ET	SC_S (0,36)	WF_W (0,36)	TF_W (0,35)
FO	RT_S (0,38)	SC_S (0,36)	WF_W (0,36)
GA	WF_W (0,47)	SC_S (0,47)	TF_W (0,47)
GP	WF_W (0,38)	SC_S (0,38)	TF_W (0,37)
HT	SC_S (0,35)	WF_W (0,34)	TR_G (0,34)
LA	TF_W (0,42)	SC_S (0,42)	WF_W (0,42)
SC	SC_S (0,37)	WF_W (0,36)	TF_W (0,36)
SO	WF_W (0,35)	TF_W (0,35)	SC_S (0,35)
SP	WF_W (0,53)	SC_S (0,53)	TF_W (0,51)
TR	RT_S (0,26)	SC_S (0,25)	TF_W (0,23)
WT	SC_S (0,45)	TR_G (0,44)	TF_W (0,43)

Tabela 43 – 3 Melhores Resultados da Medida *MATCHES* por Classe na Classificação CNN-18 no Nível 2

importância, em ambas as medidas (ROUGE-2 *F-measure* e do *MATCHES*) foi notada uma melhora significativa na performance dos métodos LS_W, SC_S e TR_G. Essa melhora, inclusive, ergueu, no contexto geral, o método SC_S como um dos melhores para diversas classes. Visto que, a partir do segundo nível de importância, tem-se uma predominância desse método junto aos métodos estatísticos WF_W e TF_W que desde (SILVA, 2017) já se mostravam indispensáveis.

Pode-se ainda notar que da passagem de utilização do segundo para o terceiro

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W (0,34)	SC_S (0,33)	WF_W (0,33)
AU	RT_S (0,35)	SC_S (0,34)	TR_G (0,33)
BS	SC_S (0,33)	TF_W (0,32)	WF_W (0,32)
CO	SC_S (0,33)	WF_W (0,33)	LS_W (0,31)
EC	WF_W (0,34)	TF_W (0,33)	TR_G (0,32)
ED	WF_W (0,36)	TF_W (0,35)	SC_S (0,35)
EW	SC_S (0,38)	WF_W (0,38)	TF_W (0,38)
ET	SC_S (0,32)	WF_W (0,32)	TF_W (0,31)
FO	RT_S (0,35)	SC_S (0,34)	WF_W (0,34)
GA	SC_S (0,42)	WF_W (0,41)	TF_W (0,41)
GP	SC_S (0,34)	WF_W (0,34)	TR_G (0,34)
HT	TR_G (0,31)	SC_S (0,31)	WF_W (0,30)
LA	SC_S (0,37)	TF_W (0,37)	WF_W (0,36)
SC	WF_W (0,32)	SC_S (0,32)	LS_W (0,32)
SO	TF_W (0,32)	SC_S (0,32)	WF_W (0,32)
SP	WF_W (0,45)	SC_S (0,45)	TF_W (0,44)
TR	SC_S (0,24)	TF_W (0,23)	RT_S (0,22)
WT	SC_S (0,44)	TR_G (0,42)	TF_W (0,40)

Tabela 44 – 3 Melhores Resultados da Medida ROUGE-2 *F-measure* por Classe na Classificação CNN-18 no Nível 3

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W (0,38)	WF_W (0,38)	SC_S (0,38)
AU	RT_S (0,45)	SC_S (0,40)	WF_W (0,39)
BS	SC_S (0,37)	TF_W (0,36)	WF_W (0,35)
CO	SC_S (0,39)	WF_W (0,38)	TF_W (0,36)
EC	WF_W (0,39)	TF_W (0,38)	LS_W (0,37)
ED	WF_W (0,42)	TF_W (0,40)	SC_S (0,40)
EW	SC_S (0,43)	WF_W (0,43)	TF_W (0,43)
ET	SC_S (0,36)	WF_W (0,36)	TF_W (0,35)
FO	SC_S (0,38)	WF_W (0,37)	TF_W (0,36)
GA	SC_S (0,49)	WF_W (0,48)	TF_W (0,47)
GP	SC_S (0,38)	WF_W (0,38)	TF_W (0,37)
HT	SC_S (0,35)	WF_W (0,34)	TR_G (0,34)
LA	SC_S (0,42)	TF_W (0,42)	WF_W (0,42)
SC	WF_W (0,37)	SC_S (0,37)	LS_W (0,36)
SO	WF_W (0,35)	TF_W (0,35)	SC_S (0,35)
SP	WF_W (0,53)	SC_S (0,53)	TF_W (0,51)
TR	SC_S (0,25)	RT_S (0,23)	TF_W (0,23)
WT	SC_S (0,49)	TR_G (0,47)	TF_W (0,45)

Tabela 45 – 3 Melhores Resultados da Medida *MATCHES* por Classe na Classificação CNN-18 no Nível 3

nível de importância, tanto a média do ROUGE-2 *F-measure*, quanto a do *MATCHES* se mantiveram as mesmas ou sofreram uma leve queda. Mas no geral, os métodos bons em um nível de importância mantinham a boa performance nos outros níveis. Esse comportamento não se manteve no contexto das combinações de métodos, no qual algumas das melhores combinações não apresentavam necessariamente combinações entre os melhores métodos por classe separadamente. Inclusive, observando os resultados específicos da Tabela 46 é possível observar que os métodos WF_W e TF_W estão presentes na maioria das melhores combinações.

Cl.	1º Melhor Comb.	2º Melhor Comb.	3º Melhor Comb.	4º Melhor Comb.	5º Melhor Comb.
AR	TR_G+TF_W (0,37)	CP_S+TF_W (0,37)	TF_W+WF_W (0,37)	RT_S+WF_W (0,37)	TR_G+WF_W (0,37)
AU	PN_W+TF_W (0,48)	PN_W+WF_W (0,47)	ND_S+WF_W (0,46)	TF_W+UC_W (0,46)	UC_W+WF_W (0,46)
BS	TR_G+WF_W (0,36)	CP_S+WF_W (0,36)	TR_G+TF_W (0,35)	CP_S+TF_W (0,35)	RT_S+TF_W (0,35)
CO	SL_S+TF_W (0,38)	TF_W+UC_W (0,38)	UC_W+WF_W (0,37)	PN_W+WF_W (0,36)	PN_W+TF_W (0,36)
EC	TR_G+WF_W (0,42)	SL_S+WF_W (0,42)	CP_S+WF_W (0,42)	TF_W+WF_W (0,42)	TR_G+TF_W (0,42)
ED	RT_S+WF_W (0,37)	TR_G+WF_W (0,37)	CP_S+WF_W (0,37)	TF_W+WF_W (0,36)	TR_G+TF_W (0,35)
EW	SL_S+WF_W (0,43)	TR_G+WF_W (0,43)	RT_S+WF_W (0,43)	CP_S+WF_W (0,43)	PN_W+WF_W (0,43)
ET	RT_S+WF_W (0,36)	TR_G+WF_W (0,36)	CP_S+WF_W (0,36)	TF_W+WF_W (0,35)	RT_S+TF_W (0,35)
FO	CP_S+RT_S (0,41)	RT_S+TR_G (0,41)	SL_S+TF_W (0,4)	SL_S+WF_W (0,4)	RT_S+TF_W (0,4)
GA	PN_W+UC_W (0,51)	PN_W+TF_W (0,5)	CP_S+PN_W (0,5)	PN_W+TR_G (0,5)	TF_W+UC_W (0,49)
GP	RT_S+WF_W (0,4)	RT_S+TF_W (0,39)	TR_G+WF_W (0,38)	CP_S+WF_W (0,38)	TF_W+WF_W (0,38)
HT	RT_S+WF_W (0,32)	TR_G+WF_W (0,32)	CP_S+WF_W (0,32)	TF_W+WF_W (0,31)	UC_W+WF_W (0,31)
LA	RT_S+TF_W (0,44)	TR_G+WF_W (0,42)	CP_S+WF_W (0,42)	RT_S+WF_W (0,42)	TF_W+WF_W (0,42)
SC	TR_G+WF_W (0,35)	CP_S+WF_W (0,35)	RT_S+WF_W (0,35)	TR_G+TF_W (0,34)	CP_S+TF_W (0,34)
SO	RT_S+WF_W (0,35)	RT_S+TF_W (0,34)	TF_W+WF_W (0,34)	TR_G+TF_W (0,34)	CP_S+TF_W (0,34)
SP	TR_G+WF_W (0,5)	CP_S+WF_W (0,5)	TR_G+TF_W (0,5)	CP_S+TF_W (0,5)	TF_W+WF_W (0,49)
TR	BP_G+ND_S (0,26)	RT_S+TF_W (0,25)	SL_S+TF_W (0,25)	TR_G+TF_W (0,24)	CP_S+TF_W (0,24)
WT	WC_W+RT_S (0,4)	RT_S+TF_W (0,39)	BP_G+TF_W (0,39)	RT_S+SL_S (0,38)	RT_S+WF_W (0,38)

Tabela 46 – Resultados das 5 Melhores Combinações 2 a 2 de Métodos de Sumarização para cada Classe do *Corpus* CNN com a Classificação CNN-18 no Nível 1

Método	Acurácia	Precisão	Cobertura	*F-measure*	Kappa
FT	0,93	0,39	0,39	0,39	0,31
WF_W	0,93	0,35	0,35	0,35	0,22
TF_W	0,93	0,36	0,36	0,36	0,23
UC_W	0,93	0,35	0,35	0,35	0,22
PN_W	0,93	0,35	0,35	0,35	0,21
WC_W	0,93	0,34	0,34	0,34	0,20
LS_W	0,93	0,35	0,35	0,35	0,21
RT_S	0,93	0,34	0,34	0,34	0,21
SC_S	0,93	0,36	0,36	0,36	0,23
SP_S	0,93	0,35	0,35	0,35	0,21
CP_S	0,93	0,34	0,34	0,34	0,19
ND_S	0,93	0,33	0,33	0,33	0,19
SL_S	0,93	0,36	0,36	0,36	0,22
TR_G	0,93	0,37	0,37	0,37	0,23
BP_G	0,93	0,35	0,35	0,35	0,2
AS_G	0,93	0,36	0,36	0,36	0,21

Tabela 47 – Resultados Macro das Avaliações dos Modelos com a Classificação CNN-18 no Nível 1

Levando em consideração todas as informações e os levantamentos recém expostos, foi construída a Tabela 61 com os 3 melhores métodos de sumarização por classe do *corpus* CNN. Vale ressaltar que para a concepção dessa tabela, também foi levado em conta a quase inexistência de variações entre os melhores métodos pelas medidas ROUGE-2 *F-measure* e $MATCHES$.

Após a análise da conjectura anterior, deve-se também verificar comparativamente o desempenho dos modelos de classificação de documentos integrais e sumarizados. Logo, os resultados obtidos e exibidos nas Tabelas 47, 48, 49, 50 são suficientes para perceber visualmente que no geral a classificação de documentos de notícias a partir dos textos integrais e dos textos sumarizados se equipara. Observando mais a fundo, é possível

Método	Acurácia	Precisão	Cobertura	*F-measure*	Kappa
FT	0,86	0,21	0,21	0,21	0,11
WF_W	0,85	0,20	0,20	0,20	0,10
TF_W	0,85	0,20	0,20	0,20	0,10
UC_W	0,85	0,19	0,19	0,19	0,10
PN_W	0,85	0,19	0,19	0,19	0,09
WC_W	0,85	0,18	0,18	0,18	0,09
LS_W	0,85	0,19	0,19	0,19	0,10
RT_S	0,85	0,19	0,19	0,19	0,10
SC_S	0,86	0,21	0,21	0,21	0,10
SP_S	0,85	0,17	0,17	0,17	0,08
CP_S	0,85	0,16	0,16	0,16	0,07
ND_S	0,85	0,18	0,18	0,18	0,08
SL_S	0,85	0,17	0,17	0,17	0,07
TR_G	0,86	0,20	0,20	0,20	0,10
BP_G	0,85	0,17	0,17	0,17	0,08
AS_G	0,85	0,16	0,16	0,16	0,07

Tabela 48 – Resultados Macro das Avaliações dos Modelos com a Classificação CNN-11 no Nível 1

Método	Acurácia	Precisão	Cobertura	*F-measure*	Kappa
FT	0,85	0,40	0,40	0,40	0,19
WF_W	0,86	0,43	0,43	0,43	0,22
TF_W	0,86	0,43	0,43	0,43	0,22
UC_W	0,86	0,44	0,44	0,44	0,22
PN_W	0,86	0,44	0,44	0,44	0,22
WC_W	0,84	0,38	0,38	0,38	0,18
LS_W	0,85	0,42	0,42	0,42	0,21
RT_S	0,85	0,41	0,41	0,41	0,20
SC_S	0,86	0,43	0,43	0,43	0,22
SP_S	0,85	0,38	0,38	0,38	0,18
CP_S	0,84	0,36	0,36	0,36	0,17
ND_S	0,85	0,39	0,39	0,39	0,19
SL_S	0,85	0,41	0,41	0,41	0,21
TR_G	0,85	0,40	0,40	0,40	0,20
BP_G	0,84	0,38	0,38	0,38	0,18
AS_G	0,84	0,37	0,37	0,37	0,18

Tabela 49 – Resultados Macro das Avaliações dos Modelos com a Classificação CNN-8 no Nível 1

Método	Acurácia	Precisão	Cobertura	*F-measure*	Kappa
FT	0,82	0,56	0,56	0,56	0,43
WF_W	0,81	0,52	0,52	0,52	0,36
TF_W	0,81	0,52	0,52	0,52	0,37
UC_W	0,80	0,51	0,51	0,51	0,35
PN_W	0,81	0,51	0,51	0,51	0,36
WC_W	0,80	0,49	0,49	0,49	0,33
LS_W	0,80	0,51	0,51	0,51	0,35
RT_S	0,80	0,50	0,50	0,50	0,33
SC_S	0,81	0,51	0,51	0,51	0,36
SP_S	0,79	0,48	0,48	0,48	0,31
CP_S	0,80	0,49	0,49	0,49	0,31
ND_S	0,80	0,49	0,49	0,49	0,32
SL_S	0,81	0,52	0,52	0,52	0,36
TR_G	0,80	0,51	0,51	0,51	0,35
BP_G	0,80	0,49	0,49	0,49	0,32
AS_G	0,80	0,50	0,50	0,50	0,33

Tabela 50 – Resultados Macro das Avaliações dos Modelos com a Classificação CNN-5 no Nível 1

notar também casos em que os textos sumarizados apresentam resultados melhores que os originais, como no caso dos resultados da classificação **CNN-8** ou casos em que os

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,42	0,31	0,33	0,33	0,30	0,31	0,33	0,24	0,30	0,27	0,26	0,25	0,28	0,30	0,26	0,25
AU	0,09	NaN	0,10	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
BS	0,30	0,24	0,21	0,23	0,22	0,19	0,23	0,21	0,22	0,19	0,16	0,19	0,22	0,21	0,17	0,19
CO	0,42	0,20	0,23	0,15	0,14	0,15	0,17	0,20	0,17	0,03	0,10	0,14	0,21	0,27	0,09	0,13
EC	0,25	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0,06	NaN	0,06	NaN	NaN	0,11	NaN	NaN
ED	0,06	0,06	0,03	0,03	0,01	0,04	0,08	0,05	0,10	0,03	NaN	0,04	0,07	0,04	0,06	0,07
EW	0,12	0,07	0,05	0,02	NaN	NaN	0,02	NaN	0,05	0,03	NaN	0,05	0,05	0,03	NaN	NaN
ET	0,21	0,24	0,27	0,25	0,23	0,26	0,22	0,24	0,26	0,26	0,23	0,25	0,26	0,24	0,24	0,25
FO	0,07	NaN	0,08	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0,09	NaN	NaN
GA	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
GP	0,55	0,43	0,43	0,45	0,45	0,43	0,43	0,45	0,44	0,46	0,43	0,43	0,44	0,46	0,45	0,44
HT	0,50	0,43	0,42	0,43	0,40	0,37	0,45	0,45	0,47	0,35	0,37	0,35	0,42	0,48	0,42	0,41
LA	0,41	0,27	0,30	0,29	0,30	0,26	0,31	0,29	0,31	0,30	0,25	0,30	0,30	0,33	0,24	0,29
SC	0,39	0,33	0,30	0,23	0,26	0,24	0,35	0,30	0,29	0,19	0,24	0,22	0,29	0,33	0,20	0,28
SO	0,37	0,46	0,46	0,45	0,46	0,45	0,44	0,45	0,47	0,46	0,45	0,44	0,48	0,48	0,47	0,48
SP	0,51	0,42	0,42	0,38	0,36	0,33	0,37	0,42	0,42	0,40	0,38	0,34	0,43	0,34	0,35	0,29
TR	0,36	0,07	0,35	0,14	0,25	0,12	0,18	0,07	0,24	0,07	0,13	0,13	0,19	0,20	0,14	0,06
WT	0,30	NaN	NaN	NaN	NaN	NaN	0,10	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Tabela 51 – Resultados da Medida *F-measure* por Classe das Avaliações dos Modelos com a Classificação CNN-18 no Nível 1

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,88	0,85	0,86	0,86	0,87	0,87	0,87	0,86	0,86	0,86	0,87	0,85	0,86	0,87	0,86	0,85
AU	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
BS	0,93	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94
CO	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98
EC	0,98	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
ED	0,95	0,95	0,95	0,95	0,96	0,95	0,95	0,95	0,95	0,96	0,96	0,96	0,96	0,95	0,96	0,96
EW	0,97	0,97	0,97	0,97	0,98	0,98	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,97	0,98	0,98
ET	0,83	0,79	0,80	0,78	0,78	0,78	0,78	0,76	0,79	0,79	0,78	0,76	0,79	0,79	0,78	0,79
FO	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
GA	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
GP	0,83	0,81	0,82	0,81	0,82	0,82	0,81	0,81	0,81	0,82	0,80	0,81	0,82	0,82	0,83	0,83
HT	0,95	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,97	0,96	0,96
LA	0,88	0,90	0,91	0,91	0,91	0,91	0,91	0,91	0,91	0,92	0,91	0,92	0,92	0,91	0,91	0,92
SC	0,95	0,95	0,95	0,95	0,95	0,95	0,96	0,95	0,95	0,95	0,96	0,95	0,95	0,95	0,95	0,95
SO	0,73	0,66	0,65	0,65	0,63	0,62	0,63	0,65	0,67	0,62	0,61	0,64	0,64	0,65	0,62	0,63
SP	0,97	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98
TR	0,98	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
WT	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99

Tabela 52 – Resultados da Acurácia por Classe das Avaliações dos Modelos com a Classificação CNN-18 no Nível 1

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,21	0,24	0,25	0,24	0,22	0,24	0,25	0,22	0,25	0,23	0,20	0,23	0,19	0,22	0,22	0,22
BS	0,21	0,20	0,20	0,20	0,17	0,21	0,21	0,20	0,22	0,14	0,16	0,18	0,16	0,16	0,17	0,16
ED	0,06	0,06	0,04	0,04	0,03	0,09	0,07	0,05	0,05	0,06	0,06	0,04	0,05	0,08	0,03	0,01
EN	0,01	NaN	0,01	NaN	NaN	NaN	0,01	0,01	0,02	NaN	0,02	NaN	NaN	0,02	NaN	NaN
GP	0,37	0,34	0,32	0,34	0,32	0,29	0,33	0,35	0,37	0,27	0,27	0,27	0,31	0,34	0,26	0,25
HT	0,16	0,13	0,08	0,01	0,01	0,10	0,14	0,08	0,12	0,08	0,08	0,15	0,08	0,14	0,05	0,08
LF	0,05	0,05	0,05	0,04	0,07	0,03	0,01	NaN	0,03	0,02	0,03	0,04	0,01	0,04	0,04	0,04
RC	0,25	0,26	0,27	0,27	0,26	0,22	0,25	0,25	0,25	0,25	0,21	0,26	0,24	0,27	0,26	0,23
SC	0,18	0,14	0,14	0,13	0,13	0,13	0,14	0,15	0,14	0,12	0,12	0,11	0,12	0,14	0,13	0,13
SO	0,17	0,16	0,17	0,16	0,16	0,16	0,14	0,15	0,17	0,16	0,16	0,17	0,13	0,18	0,15	0,16
TC	0,04	0,07	0,05	0,06	0,04	0,02	0,08	0,06	NaN	0,05	0,05	0,06	0,05	0,11	0,11	0,06

Tabela 53 – Resultados da Medida *F-measure* por Classe das Avaliações dos Modelos com a Classificação CNN-11 no Nível 1

textos integrais obtêm resultados relativamente melhores, como no caso dos resultados da classificação **CNN-11**. Há também casos um pouco mais discrepantes como é o caso dos resultados das classificações **CNN-18** e **CNN-5**.

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,90	0,90	0,90	0,89	0,89	0,90	0,90	0,89	0,90	0,89	0,89	0,89	0,90	0,89	0,90	0,90
BS	0,82	0,86	0,86	0,86	0,86	0,88	0,86	0,87	0,86	0,87	0,88	0,88	0,86	0,86	0,87	0,87
ED	0,96	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95
EN	0,87	0,94	0,94	0,95	0,95	0,96	0,93	0,94	0,94	0,96	0,96	0,95	0,95	0,94	0,95	0,95
GP	0,72	0,74	0,74	0,75	0,74	0,76	0,74	0,76	0,75	0,76	0,75	0,75	0,75	0,74	0,75	0,75
HT	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,96	0,95	0,95	0,95	0,95	0,95	0,95	0,95
LF	0,96	0,95	0,95	0,96	0,96	0,95	0,94	0,94	0,95	0,94	0,95	0,96	0,95	0,95	0,95	0,95
RC	0,81	0,80	0,80	0,79	0,79	0,80	0,80	0,79	0,79	0,79	0,78	0,79	0,79	0,80	0,80	0,80
SC	0,78	0,68	0,66	0,63	0,62	0,57	0,67	0,67	0,68	0,57	0,56	0,6	0,59	0,68	0,57	0,57
SO	0,68	0,68	0,68	0,69	0,69	0,69	0,68	0,68	0,68	0,69	0,69	0,68	0,68	0,68	0,68	0,69
TC	0,97	0,96	0,97	0,97	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96

Tabela 54 – Resultados da Acurácia por Classe das Avaliações dos Modelos com a Classificação CNN-11 no Nível 1

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,16	0,39	0,38	0,41	0,41	0,35	0,38	0,36	0,39	0,39	0,34	0,38	0,38	0,36	0,36	0,34
BS	0,35	0,30	0,32	0,30	0,29	0,28	0,30	0,30	0,31	0,21	0,26	0,26	0,28	0,28	0,29	0,29
ED	0,16	0,13	0,13	0,1	0,12	0,05	0,11	0,11	0,15	0,12	0,09	0,10	0,11	0,12	0,10	0,12
EM	0,01	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0,02	NaN	NaN	NaN
GP	0,66	0,61	0,62	0,62	0,62	0,57	0,60	0,60	0,62	0,57	0,54	0,57	0,61	0,59	0,56	0,56
HT	0,02	0,06	0,05	0,03	0,03	0,02	0,06	0,03	0,03	NaN	0,03	0,02	0,02	0,03	0,02	NaN
LF	0,06	0,05	0,04	0,03	NaN	NaN	0,05	0,02	NaN	NaN	0,03	NaN	0,04	0,03	0,03	NaN
SC	0,29	0,21	0,21	0,22	0,22	0,18	0,20	0,19	0,22	0,18	0,18	0,20	0,21	0,20	0,19	0,19

Tabela 55 – Resultados da Medida *F-measure* por Classe das Avaliações dos Modelos com a Classificação CNN-8 no Nível 1

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,75	0,77	0,76	0,77	0,77	0,76	0,77	0,76	0,77	0,76	0,75	0,76	0,77	0,76	0,77	0,76
BS	0,88	0,83	0,84	0,85	0,85	0,83	0,83	0,84	0,84	0,83	0,83	0,81	0,83	0,83	0,84	0,83
ED	0,97	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,97	0,97	0,96	0,96	0,96	0,96	0,97
EM	0,71	0,97	0,96	0,98	0,98	0,97	0,96	0,97	0,97	0,98	0,98	0,98	0,96	0,97	0,97	0,97
GP	0,67	0,64	0,65	0,64	0,65	0,63	0,64	0,64	0,65	0,63	0,63	0,64	0,65	0,64	0,63	0,63
HT	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96	0,96
LF	0,98	0,97	0,97	0,97	0,98	0,97	0,97	0,97	0,97	0,97	0,98	0,98	0,97	0,98	0,98	0,97
SC	0,89	0,75	0,76	0,75	0,75	0,67	0,75	0,72	0,75	0,67	0,63	0,69	0,73	0,72	0,65	0,65

Tabela 56 – Resultados da Acurácia por Classe das Avaliações dos Modelos com a Classificação CNN-8 no Nível 1

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,55	0,53	0,53	0,50	0,50	0,51	0,52	0,50	0,53	0,50	0,47	0,45	0,53	0,54	0,49	0,50
BU	0,51	0,45	0,42	0,45	0,42	0,43	0,44	0,40	0,42	0,36	0,41	0,44	0,45	0,42	0,4	0,43
GP	0,58	0,53	0,54	0,50	0,52	0,48	0,52	0,49	0,53	0,47	0,48	0,47	0,52	0,52	0,50	0,50
HL	0,53	0,48	0,51	0,49	0,50	0,45	0,48	0,45	0,47	0,42	0,44	0,46	0,47	0,46	0,42	0,44
NS	0,57	0,54	0,56	0,54	0,55	0,53	0,54	0,55	0,55	0,54	0,54	0,55	0,55	0,55	0,54	0,54

Tabela 57 – Resultados da Medida *F-measure* por Classe das Avaliações dos Modelos com a Classificação CNN-5 no Nível 1

Cl.	FT	WF_W	TF_W	UC_W	PN_W	WC_W	LS_W	RT_S	SC_S	SP_S	CP_S	ND_S	SL_S	TR_G	BP_G	AS_G
AR	0,85	0,86	0,87	0,86	0,86	0,86	0,86	0,86	0,87	0,86	0,86	0,85	0,87	0,87	0,86	0,86
BU	0,85	0,84	0,83	0,84	0,84	0,84	0,84	0,83	0,83	0,82	0,84	0,84	0,84	0,83	0,83	0,84
GP	0,8	0,79	0,79	0,77	0,78	0,76	0,79	0,78	0,79	0,75	0,77	0,76	0,78	0,78	0,77	0,78
HL	0,88	0,88	0,88	0,88	0,88	0,87	0,88	0,88	0,87	0,87	0,88	0,88	0,88	0,87	0,87	0,88
NS	0,73	0,67	0,68	0,65	0,66	0,65	0,65	0,66	0,67	0,65	0,64	0,65	0,66	0,66	0,65	0,64

Tabela 58 – Resultados da Acurácia por Classe das Avaliações dos Modelos com a Classificação CNN-5 no Nível 1

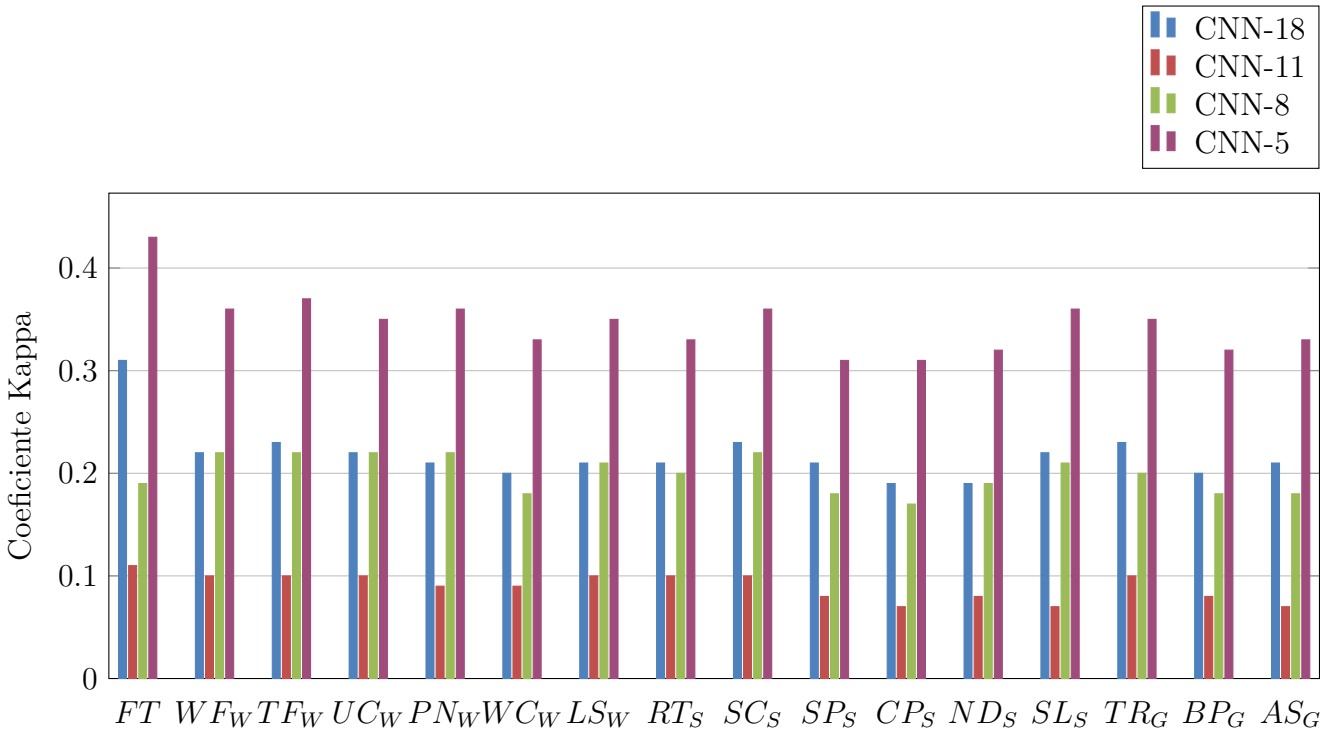

Figura 9 – Comparação dos Valores do Coeficiente Kappa para Todas as Classificações do *Corpus* CNN no Nível 1

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W (0,33)	LS_W (0,33)	UC_W (0,33)
AU	TF_W (0,10)	-	-
BS	WF_W (0,24)	UC_W (0,23)	LS_W (0,23)
CO	TR_G (0,27)	TF_W (0,23)	SL_S (0,21)
EC	TR_G (0,11)	SC_S (0,06)	CP_S (0,06)
ED	SC_S (0,10)	LS_W (0,08)	SL_S (0,07)
EW	WF_W (0,07)	ND_S (0,05)	SC_S (0,05)
ET	TF_W (0,27)	SP_S (0,26)	SC_S (0,26)
FO	TR_G (0,09)	TF_W (0,08)	-
GA	-	-	-
GP	TR_G (0,46)	SP_S (0,46)	UC_W (0,45)
HT	TR_G (0,48)	SC_S (0,47)	RT_S (0,45)
LA	TR_G (0,33)	LS_W (0,31)	SC_S (0,31)
SC	LS_W (0,35)	TR_G (0,33)	WF_W (0,33)
SO	AS_G (0,48)	SL_S (0,48)	TR_G (0,48)
SP	SL_S (0,43)	SC_S (0,42)	WF_W (0,42)
TR	TF_W (0,35)	PN_W (0,25)	SC_S (0,24)
WT	LS_W (0,10)	-	-

Tabela 59 – 3 Melhores Resultados da Medida *F-measure* por Classe dos Modelos na Classificação CNN-18 no Nível 1

Analisando as Tabelas 51, 53, 55 e 57 junto com as anteriores, é percebido que, no geral, não há muita discrepância entre os métodos de sumarização no contexto comparativo entre os modelos de classificação. Outro aspecto importante a ser observado é a quantidade de NaN nas avaliações dos modelos das 4 classificações analisadas. Em algumas dessas classificações, a presença desse termo é relativamente alta. Esse fato pode indicar

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	WC_W (0,87)	TR_G (0,87)	LS_W (0,87)
AU	TF_W (0,99)	LS_W (0,99)	SC_S (0,99)
BS	ND_S (0,94)	UC_W (0,94)	PN_W (0,94)
CO	SL_S (0,98)	TR_G (0,98)	TF_W (0,98)
EC	SC_S (0,99)	CP_S (0,99)	TR_G (0,99)
ED	BP_G (0,96)	SL_S (0,96)	ND_S (0,96)
EW	ND_S (0,98)	PN_W (0,98)	RT_S (0,98)
ET	TF_W (0,80)	SP_S (0,79)	WF_W (0,79)
FO	TR_G (0,99)	WF_W (0,99)	TF_W (0,99)
GA	UC_W (0,99)	PN_W (0,99)	RT_S (0,99)
GP	AS_G (0,83)	BP_G (0,83)	TR_G (0,82)
HT	TR_G (0,97)	BP_G (0,96)	SC_S (0,96)
LA	SL_S (0,92)	SP_S (0,92)	ND_S (0,92)
SC	CP_S (0,96)	LS_W (0,96)	AS_G (0,95)
SO	SC_S (0,67)	WF_W (0,66)	TF_W (0,65)
SP	RT_S (0,98)	SL_S (0,98)	SP_S (0,98)
TR	TF_W (0,99)	PN_W (0,99)	TR_G (0,99)
WT	WF_W (0,99)	UC_W (0,99)	PN_W (0,99)

Tabela 60 – 3 Melhores Resultados da Acurácia por Classe dos Modelos na Classificação CNN-18 no Nível 1

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W	SC_S	WF_W
AU	RT_S	SC_S	TR_G
BS	SC_S	TF_W	WF_W
CO	SC_S	WF_W	LS_W
EC	TF_W	WF_W	SC_S
ED	WF_W	TF_W	SC_S
EW	SC_S	WF_W	TF_W
ET	SC_S	WF_W	TF_W
FO	RT_S	SC_S	WF_W
GA	SC_S	WF_W	TF_W
GP	SC_S	TR_G	WF_W
HT	TR_G	SC_S	WF_W
LA	TF_W	SC_S	WF_W
SC	WF_W	SC_S	TF_W
SO	TF_W	WF_W	SC_S
SP	SC_S	WF_W	TF_W
TR	RT_S	SC_S	TF_W
WT	SC_S	TR_G	TF_W

Tabela 61 – Melhores Métodos de Sumarização por Classe do *Corpus* CNN com a Classificação CNN-18 no Nível 1

um treinamento insuficiente dos modelos para as classes que apresentam pouquíssimos documentos.

As Tabelas 52, 54, 56 e 58 apresentaram valores de acurácias relativamente altos. Isso é comum em problemas com classes desbalanceadas e ocorre porque o classificador acaba tratando da mesma forma diferentes erros. Logo, esse comportamento comumente prejudica a identificação de exemplos pertencentes à classes minoritárias, mas essa estratégia acaba não afetando a acurácia, que apresenta valores elevados.

Dentre as 4 classificações, a que mostrou os melhores resultados foi a **CNN-5**. Comparando os resultados das avaliações dos modelos dessa classificação com as avaliações

dos modelos de outras, é possível observar valores relativamente mais altos de precisão, cobertura, *F-measure* e coeficiente Kappa para a grande maioria dos modelos. Outro fator importante foi a ausência de NaN nos resultados da Tabela 57. Isso se deve ao fato dessa classificação ter uma proporção de documentos bem mais balanceada que as outras, como é possível ver nos gráficos das Figuras 5, 6, 7 e 8. Ainda é possível comparar os valores dos coeficientes Kappa de todos os métodos de sumarização e dos textos completos nas 4 classificações através do gráfico da Figura 9. Nessa comparação é possível ver que no aspecto geral, o coeficiente do texto completo não é muito diferente dos textos sumarizados.

Toda informação obtida até aqui, foi sumarizada na Tabela 62, semelhante à Tabela 61, com os três melhores métodos de sumarização por classe, no contexto de treinamento e avaliação de modelos de classificação, do *corpus* CNN com sua classificação original. É importante pontuar ainda que não foi possível estabelecer para todas as classes os três melhores métodos ou até qualquer método utilizando apenas os valores da medida *F-measure*, já que alguns desses valores foram iguais a NaN. Para esses casos utilizou-se os valores da acurácia, mas ainda sim, alguns métodos apresentaram desempenhos iguais.

Uma última análise a ser realizada é comparar as Tabelas 61 e 62 para verificar em cada uma das conjecturas analisadas, quais os melhores métodos de sumarização por classes. Percebe-se que há algumas diferenças entre os melhores métodos em cada cenário avaliado.

Cl.	1º Melhor Método	2º Melhor Método	3º Melhor Método
AR	TF_W	LS_W	UC_W
AU	TF_W	$AS_G/SL_S/SC_S/LS_W$	$AS_G/SL_S/SC_S/LS_W$
BS	WF_W	UC_W	LS_W
CO	TR_G	TF_W	SL_S
EC	TR_G	SC_S	CP_S
ED	SC_S	LS_W	SL_S
EW	WF_W	SL_S	TF_W
ET	TF_W	SL_S	SP_S
FO	TF_W	TR_G	$CP_S/ND_S/SL_S$
GA	RT_S	UC_W	SC_S
GP	SP_S	TR_G	UC_W
HT	TR_G	SC_S	RT_S
LA	TR_G	LS_W	SC_S
SC	LS_W	TR_G	WF_W
SO	AS_G	TR_G	SC_S
SP	SL_S	SC_S	WF_W
TR	TF_W	PN_W	SC_S
WT	LS_W	WF_W/UC_W	WF_W/UC_W

Tabela 62 – Melhores Métodos de Sumarização por Classe para Treinamento de Classificadores do *Corpus* CNN com a Classificação CNN-18 no Nível 1

5 CONCLUSÕES

5.1 Conclusões

A sumarização ajuda as pessoas no dia a dia em várias atividades rotineiras, possibilitando uma maior dinamicidade em suas vidas ao entregar informações importantes de forma mais suscita. Os sistemas de sumarização automática de textos não funcionam de forma diferente, eles ajudam as pessoas a lidar com a sobrecarga de informação disponível hoje em dia. Mesmo que ainda não seja possível criar resumos de forma automatizada como os seres humanos fazem, esses sistemas trazem diversos benefícios com suas limitações. Dessa forma, evoluções nesses geram impactos muito significativos, atenuando, inclusive diversos outros problemas de áreas correlatas.

Desde antes do início deste trabalho, benefícios da integração das áreas de classificação automática de textos e sumarização automática de textos foram vislumbrados e explorados. Este trabalho buscou verificar o quanto a categoria de classificação de um documento poderia consistir em um bom parâmetro para determinação de quais técnicas de sumarização extrativa empregar. T Também foi um objetivo desta dissertação, mapear as melhores técnicas de sumarização, ou combinações, que produzissem os melhores resumos. Contemplando essa análise, os mapeamentos obtidos para produção de novos resumos evidenciaram, no geral, uma grande eficiência do método "*Sentence Centrality*" e dos métodos estatísticos como "*Word Frequency*" e "*TF-IDF*".

Por fim, a última hipótese analisada foi a de comparar a eficiência da classificação de documentos de modelos de ML a partir dos próprios resumos dos textos originais gerados pelas técnicas de sumarização com o conteúdo dos textos integralmente. Para testar essa hipótese foi utilizado o *corpus* de notícias do CNN, mesmo *corpus* utilizado na avaliação dos resumos gerados. Os problemas de desbalanceamento de classes desse *corpus* no treinamento de modelos foram atenuados através da reclassificação dos documentos. Seguidamente, foi possível comparar a performance dos modelos treinados e assemelhar a eficiência entre os modelos treinados com os textos integrais e com somente os resumos.

Portanto, as principais contribuições desse trabalho foram evidenciar que assim como em trabalhos anteriores, até o momento, apenas a classificação de um texto por si só não configura um bom parâmetro para escolha da técnica de sumarização a ser empregue em documentos de notícias. Obviamente, esse entendimento decorre dos resultados dos experimentos realizados com o *corpus* do CNN em seus diversos níveis de importância. Outra importante contribuição foi equiparar a classificação de modelos de ML treinados com os textos integrais com os textos resumidos pelas mais diversas técnicas de sumarização.

Neste caso, ficou comprovado a eficácia da utilização dos resumos ao invés dos documentos originais para o treinamento dos modelos.

5.2 Desafios

Desafios e problemas são encontrados em qualquer tipo de trabalho e, em muitos contextos de pesquisa, são tão importantes quantos os próprios resultados obtidos e as conclusões. Relatar tais problemas serve também para dar um panorama de possíveis dificuldades que podem ocorrer até no desenvolvimento de trabalhos futuros na mesma linha de pesquisa.

Neste trabalho alguns desafios foram encontrados durante todo o seu desenvolvimento. A maior parte deles foram superados ou mitigados. Entretanto, alguns devem ser destacados como os mais significantes e impactantes.

Talvez o mais marcante desses desafios foi a ausência de outros *corpora* textuais que pudessem ser utilizados tanto para a sumarização automática de textos quanto para a classificação automática. Outro contratempo foi o grande investimento de tempo na geração de modelos de classificação que não apresentaram resultados tão eficientes. Isto se deve ao *tradeoff* entre o tempo de treinamento excessivamente elevado e a qualidade dos resultados obtidos, principalmente para modelos como árvores de decisão, redes neurais e SVMs. Além disso, o tempo desprendido em análises e validações de novos modelos de classificação a partir de pré-processamento de dados foi bem mais elevado do que se estimou.

Por fim, também houve uma certa carência de equipamentos mais potentes, com mais recursos computacionais como poder de processamento e memória, para executar mais rapidamente os experimentos de forma completa. Isto é, ausência de dispositivos capazes de realizar o treinamento de modelos de classificação com todo o volume de dados do *corpus* textual.

5.3 Trabalhos Futuros

Para trabalhos futuros outros *corpora* podem ser incluídos nas experimentações e algoritmos de classificação utilizando *deep learning* também podem ser empregados para melhorar os índices de classificação dos testes no geral. Sistemas de abordagem híbrida de classificação de textos mesclando algoritmos de aprendizado de máquina e sistemas especialistas podem ser utilizados como uma possível solução para melhorar os índices de classificação dos textos. Também poderão ser empregadas metodologias de avaliações manuais e de forma qualitativa com uma parte significativa dos resumos produzidos pelas melhores técnicas selecionadas para cada uma das classes.

REFERÊNCIAS

ALAHMADI, A.; JOORABCHI, A.; MAHDI, A. E. A new text representation scheme combining bag-of-words and bag-of-concepts approaches for automatic text classification. In: IEEE. *2013 7th IEEE GCC Conference and Exhibition (GCC)*. [S.l.], 2013. p. 108–113. Citado 2 vezes nas páginas 37 e 38.

BANERJEE, M. et al. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, Wiley Online Library, v. 27, n. 1, p. 3–23, 1999. Citado na página 44.

BOUTELL, M. R. et al. Learning multi-label scene classification. *Pattern recognition*, Elsevier, v. 37, n. 9, p. 1757–1771, 2004. Citado na página 53.

CABRAL, L. d. S. et al. A platform for language independent summarization. In: ACM. *Proceedings of the 2014 ACM symposium on Document engineering*. [S.l.], 2014. p. 203–206. Citado 2 vezes nas páginas 22 e 25.

CONRAD, J. G. et al. Query-based opinion summarization for legal blog entries. In: ACM. *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. [S.l.], 2009. p. 167–176. Citado 3 vezes nas páginas 44, 45 e 46.

DALAL, M. K.; ZAVERI, M. A. Automatic text classification: a technical review. *International Journal of Computer Applications*, International Journal of Computer Applications, 244 5 th Avenue,# 1526, New . . . , v. 28, n. 2, p. 37–40, 2011. Citado 5 vezes nas páginas 26, 27, 36, 37 e 41.

DALAL, M. K.; ZAVERI, M. A. Semisupervised learning based opinion summarization and classification for online product reviews. *Applied Computational Intelligence and Soft Computing*, Hindawi Publishing Corp., v. 2013, p. 10, 2013. Citado na página 46.

DONG, Y.-S.; HAN, K.-S. A comparison of several ensemble methods for text categorization. In: IEEE. *IEEE International Conference onServices Computing, 2004.(SCC 2004). Proceedings. 2004*. [S.l.], 2004. p. 419–422. Citado na página 41.

ELKAN, C. Evaluating classifiers. *San Diego: University of California*, Citeseer, 2012. Citado 5 vezes nas páginas 40, 41, 42, 43 e 62.

EMANI, C. K.; CULLOT, N.; NICOLLE, C. Understandable big data: a survey. *Computer science review*, Elsevier, v. 17, p. 70–81, 2015. Citado na página 17.

FARHOODI, M.; YARI, A. Applying machine learning algorithms for automatic persian text classification. In: IEEE. *2010 6th International Conference on Advanced Information Management and Service (IMS)*. [S.l.], 2010. p. 318–323. Citado 2 vezes nas páginas 31 e 32.

FERREIRA, R. et al. A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, Elsevier, v. 41, n. 13, p. 5780–5787, 2014. Citado 2 vezes nas páginas 23 e 25.

FERREIRA, R. et al. Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, Elsevier, v. 40, n. 14, p. 5755–5764, 2013. Citado 5 vezes nas páginas 30, 31, 32, 33 e 47.

FERREIRA, R. et al. A four dimension graph model for automatic text summarization. In: IEEE. *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. [S.l.], 2013. v. 1, p. 389–396. Citado 2 vezes nas páginas 23 e 34.

FERREIRA, R. et al. A context based text summarization system. In: IEEE. *2014 11th IAPR International Workshop on Document Analysis Systems*. [S.l.], 2014. p. 66–70. Citado 11 vezes nas páginas 17, 19, 23, 28, 29, 30, 31, 32, 34, 47 e 52.

FERREIRA, R. et al. Automatic document classification using summarization strategies. In: *Proceedings of the 2015 ACM Symposium on Document Engineering*. [S.l.: s.n.], 2015. p. 69–72. Citado 6 vezes nas páginas 29, 30, 45, 47, 50 e 52.

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, Springer, v. 47, n. 1, p. 1–66, 2017. Citado 2 vezes nas páginas 23 e 25.

GOLDSTEIN, J. et al. Summarizing text documents: Sentence selection and evaluation metrics. Carnegie Mellon University, 1999. Citado 3 vezes nas páginas 17, 21 e 33.

HAHN, U.; MANI, I. The challenges of automatic summarization. *Computer*, IEEE, v. 33, n. 11, p. 29–36, 2000. Citado 3 vezes nas páginas 19, 23 e 25.

JEONG, H.; KO, Y.; SEO, J. How to improve text summarization and classification by mutual cooperation on an integrated framework. *Expert Systems with Applications*, Elsevier, v. 60, p. 222–233, 2016. Citado 2 vezes nas páginas 47 e 50.

JO, T. K nearest neighbor for text summarization using feature similarity. In: IEEE. *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*. [S.l.], 2017. p. 1–5. Citado na página 48.

KAUR, G.; BAJAJ, K. News classification and its techniques: a review. *IOSR JOURNAL OF COMPUTER ENGINEERING (IOSR-JCE)*, v. 18, n. 1, p. 22–26, 2016. Citado 2 vezes nas páginas 45 e 61.

KER, S. J.; CHEN, J.-N. A text categorization based on summarization technique. In: *the 38th Annual Meeting of the Association for Computational Linguistics IR&NLP workshop, Hong Kong*. [S.l.: s.n.], 2000. Citado na página 45.

KHAN, A. et al. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, Academy Publisher, PO Box 40 Oulu 90571 Finland, v. 1, n. 1, p. 4–20, 2010. Citado 9 vezes nas páginas 17, 18, 21, 26, 35, 36, 38, 44 e 61.

KIM, S.-B. et al. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, IEEE, v. 18, n. 11, p. 1457–1466, 2006. Citado na página 61.

KORDE, V.; MAHENDER, C. N. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, Academy & Industry Research Collaboration Center (AIRCC), v. 3, n. 2, p. 85, 2012. Citado 2 vezes nas páginas 17 e 19.

KOULALI, R.; EL-HAJ, M.; MEZIANE, A. Arabic topic detection using automatic text summarisation. In: IEEE. *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*. [S.l.], 2013. p. 1–4. Citado 3 vezes nas páginas 19, 46 e 49.

LEVATIĆ, J.; KOCEV, D.; DŽEROSKI, S. The importance of the label hierarchy in hierarchical multi-label classification. *Journal of Intelligent Information Systems*, Springer, v. 45, n. 2, p. 247–271, 2015. Citado na página 53.

LIDDY, E. D. Natural language processing. 2001. Citado 2 vezes nas páginas 21 e 22.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://www.aclweb.org/anthology/W04-1013>. Citado 3 vezes nas páginas 33, 34 e 52.

LINS, R. et al. The cnn-corpus: A large textual corpus for single-document extractive summarization. In: ACM. *In ACM Symposium on Document Engineering 2019 (DocEng '19)*. [S.l.], 2019. v. 1, p. 389–398. Citado 2 vezes nas páginas 52 e 53.

LINS, R. D.; MELLO, R. F.; SIMSKE, S. J. Doceng'19 competition on extractive text summarization. In: ACM. *In ACM Symposium on Document Engineering 2019 (DocEng '19)*. [S.l.], 2019. v. 1, p. 1–2. Citado 2 vezes nas páginas 47 e 52.

LINS, R. D. et al. A multi-tool scheme for summarizing textual documents. In: *Proc. of 11st IADIS International Conference WWW/INTERNET 2012*. [S.l.: s.n.], 2012. p. 1–8. Citado na página 52.

LIU, C.-L. et al. Semi-supervised text classification with universum learning. *IEEE transactions on cybernetics*, IEEE, v. 46, n. 2, p. 462–473, 2016. Citado 2 vezes nas páginas 35 e 36.

LLORET, E.; PALOMAR, M. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, Kluwer Academic Publishers, v. 37, n. 1, p. 1–41, 2012. Citado 3 vezes nas páginas 21, 25 e 52.

LLORET, E.; SAGGION, H.; PALOMAR, M. Experiments on summary-based opinion classification. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. [S.l.], 2010. p. 107–115. Citado 3 vezes nas páginas 44, 45 e 46.

MA, S. et al. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. *arXiv preprint arXiv:1805.01089*, 2018. Citado na página 48.

MANI, I. *Automatic summarization*. [S.l.]: John Benjamins Publishing, 2001. v. 3. Citado na página 23.

MANI, I. et al. Summac: a text summarization evaluation. *Natural Language Engineering*, Cambridge University Press, v. 8, n. 1, p. 43–68, 2002. Citado 2 vezes nas páginas 22 e 26.

MARTINS, C. B. et al. Introdução à sumarização automática. *Relatório Técnico RT-DC*, v. 2, n. 1, p. 35, 2001. Citado 5 vezes nas páginas 17, 18, 22, 23 e 26.

MEENA, Y. K.; GOPALANI, D. Analysis of sentence scoring methods for extractive automatic text summarization. In: *Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies*. New York, NY, USA: Association for Computing Machinery, 2014. (ICTCS '14). ISBN 9781450332163. Disponível em: <https://doi.org/10.1145/2677855.2677908>. Citado 12 vezes nas páginas 17, 23, 25, 26, 28, 30, 31, 32, 33, 34, 47 e 49.

MIHALCEA, R.; HASSAN, S. Using the essence of texts to improve document classification. In: CITESEER. *Proceedings of RANLP*. [S.l.], 2005. Citado na página 45.

NENKOVA, A.; MCKEOWN, K. A survey of text summarization techniques. In: *Mining text data*. [S.l.]: Springer, 2012. p. 43–76. Citado 3 vezes nas páginas 18, 26 e 29.

OWCZARZAK, K. et al. An assessment of the accuracy of automatic evaluation in summarization. In: *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*. Montréal, Canada: Association for Computational Linguistics, 2012. p. 1–9. Disponível em: <https://www.aclweb.org/anthology/W12-2601>. Citado 2 vezes nas páginas 33 e 34.

POTHOS, E. M.; CHATER, N. A simplicity principle in unsupervised human categorization. *Cognitive Science*, Wiley Online Library, v. 26, n. 3, p. 303–343, 2002. Citado na página 18.

RADOVANOVIĆ, M.; IVANOVIĆ, M. Text mining: Approaches and applications. *Novi Sad J. Math*, v. 38, n. 3, p. 227–234, 2008. Citado na página 21.

RANA, M. I.; KHALID, S.; AKBAR, M. U. News classification based on their headlines: A review. In: IEEE. *17th IEEE International Multi Topic Conference 2014*. [S.l.], 2014. p. 211–216. Citado na página 45.

RINO, L. H. M.; PARDO, T. A. S. A sumarização automática de textos: principais características e metodologias. In: *Anais do XXIII Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2003. v. 8, p. 203–245. Citado 2 vezes nas páginas 21 e 22.

ROSSI, R. G. *Classificação automática de textos por meio de aprendizado de máquina baseado em redes*. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado na página 35.

SAGGION, H. Automatic summarization: an overview. *Revue française de linguistique appliquée*, Pub. linguistiques, v. 13, n. 1, p. 63–81, 2008. Citado 2 vezes nas páginas 23 e 47.

SANTOS, Â.; CORDEIRO, J. *Sumarização automática de texto*. Dissertação (Mestrado) — Departamento de Informática, UBI, Portugal, 2012. Citado na página 17.

SARANYAMOL, C.; SINDHU, L. A survey on automatic text summarization. *Int. J. Comput. Sci. Inf. Technol*, v. 5, n. 6, p. 7889–7893, 2014. Citado na página 23.

SHAFIABADY, N. et al. Using unsupervised clustering approach to train the support vector machine for text classification. *Neurocomputing*, Elsevier, v. 211, p. 4–10, 2016. Citado 2 vezes nas páginas 35 e 36.

SILVA, I. H. L. e. *Sistema de Sumarização Automática de Textos Baseado em Classes de Documentos*. 50 f. Monografia (Graduação) — Centro de Informática, Universidade Federal de Pernambuco, Recife, 2017. Citado 9 vezes nas páginas 19, 34, 48, 50, 52, 60, 63, 66 e 71.

TEUFEL, S.; SIDDHARTHAN, A.; TIDHAR, D. Automatic classification of citation function. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 2006 conference on empirical methods in natural language processing*. [S.l.], 2006. p. 103–110. Citado 2 vezes nas páginas 44 e 62.

VILLENA-ROMÁN, J. et al. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In: *Twenty-Fourth International FLAIRS Conference*. [S.l.: s.n.], 2011. Citado na página 35.

WANG, J.; XIA, B. Relationships of cohen's kappa, sensitivity, and specificity for unbiased annotations. In: *Proceedings of the 2019 4th International Conference on Biomedical Signal and Image Processing (ICBIP 2019)*. [S.l.: s.n.], 2019. p. 98–101. Citado na página 44.

WANG, S. et al. A survey on automatic summarization. In: IEEE. *2010 International Forum on Information Technology and Applications*. [S.l.], 2010. v. 1, p. 193–196. Citado 2 vezes nas páginas 18 e 23.

YI, J. et al. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: IEEE. *Third IEEE international conference on data mining*. [S.l.], 2003. p. 427–434. Citado na página 22.

APÊNDICE A – EXEMPLOS DE TEXTO DA CLASSE "*ARTS*" DO *CORPUS* DO CNN

While digging through a box that belonged to his father, Robert Ondrovic uncovered a collection of vintage photos that brought him back to the 1964 World's Fair. Armed with a Yashica camera, Ondrovic's father captured everything from family portraits to the modern architecture, which was considered futuristic at the time, said the New York resident. Ondrovic was just 4 years old when his family visited the fair at Flushing Meadows Corona Park in Queens. Today, the photos are a throwback to a different era: Many of the women are decked out in dresses (though Ondrovic's mother is sporting mint green capris), a man smokes a cigarette as he strolls through the park, and the Zeppelin-like IBM Pavilion heralds the future of technology. "It's nice to look back at these and remember all these memories,"said Ondrovic, who originally posted the photos to CNN iReport. "Then photography was a luxury because film was expensive. My dad took everything and put it on slides." Ondrovic's father introduced both his sons to photography when he bought them Brownie cameras. (Ondrovic's brother, Richard, was allowed to bring his camera to the fair. He's six years older.). This week marked the 50th anniversary of the grand opening of the fair, which ran April 22 through October 18, 1964, and April through October 1965. The fair's theme of "peace through understanding in a shrinking globe and in an expanding universe,"is most easily seen through the 12-story steel globe dubbed the Unisphere. It's one of two structures still standing. A group of local volunteers has vowed to preserve the other structure, the New York State Pavilion. So far, the New York State Pavilion Paint Project has managed to clean and repaint the iconic building. Vintage photos like these evoke a certain sense of nostalgia. Do you have vintage photographs that are special to you? Share your '60s throwback photos and memories with CNN iReport.

Tabela A.1 – Texto intitulado "*From the vault Vintage World's Fair photos*"

While digging through a box that belonged to his father, Robert Ondrovic uncovered a collection of vintage photos that brought him back to the 1964 World's Fair. Today, the photos are a throwback to a different era: Many of the women are decked out in dresses (though Ondrovic's mother is sporting mint green capris), a man smokes a cigarette as he strolls through the park, and the Zeppelin-like IBM Pavilion heralds the future of technology. This week marked the 50th anniversary of the grand opening of the fair, which ran April 22 through October 18, 1964, and April through October 1965. Do you have vintage photographs that are special to you?.

Tabela A.2 – *Gold Standard* do Texto intitulado "*From the vault Vintage World's Fair photos*"

Robert Ondrovic found his father's photos from the 1964 World's Fair. The photos are a throwback to a time of retro clothes and "modern"architecture. This week marks the 50th anniversary of the opening of the fair. Do you have special vintage photos? Share them with CNN iReport.

Tabela A.3 – *Highlight* do Texto "*From the vault Vintage World's Fair photos*"

APÊNDICE B – EXEMPLOS DE TEXTO DA CLASSE "*AUTOMOTIVE*" DO *CORPUS* DO CNN

What happens when you ask a super-car designer to create a super luxurious train?
Magic, judging by the above images released by Japanese rail company JR East.
Yamagata-born Ken Okuyama, well known in automobile design circles, was brought in to style the company's super slick new Cruise Train, due to start chugging down the tracks in spring 2017.
Holding a maximum of 34 passengers, the Cruise Train will have 10 carriages made up of five suites, one deluxe suite, two glass-walled observation cars, a dining car and lounge.
Okuyama has worked as a chief designer for General Motors, a senior designer for Porsche AG and design director for Pininfarina, the company behind the Ferrari Enzo and Maserati Quattroporte.
According to his company, Ken Okuyama Design, he wanted to create a train that would allow passengers "to appreciate the flow of the time and space,"while enjoying Japan's landscapes and culture throughout the journey.
One of the more unique features is the observation carriage at the front of the train, which allows passengers to see onto the tracks ahead.
The other observation car is at the end of the train.
JR Rail says the train will be fitted with furniture that conveys the nobility of traditional Japanese culture.
The Lounge, for instance, is wrapped in graceful curves and features decor inspired by trees.
All suites will have a private bathroom with a shower and toilet, but the top sleeping space is the split level deluxe suite, which sleeps four.
On the bottom are two double beds, on the top a traditional Japanese dining area, with seats on the floor.
The train will be able to run on both electric and non-electric rails.
Can't wait till 2017?
JR East's upcoming Cruise Train won't be the first Japanese train to take the super-luxury route.
JR Kyushu's Seven Stars train, which features Japanese and Western design elements, hit the tracks in the fall of 2013.
This one only travels through the island of Kyushu and has 14 luxury guest rooms, two deluxe suites, three presidential suites, a lounge car, dining car and bar.
Guests can choose either the two- or three-night experience.
The name "Seven Stars"represents the seven prefectures of Kyushu, the seven carriages of the train and the seven main tourist attractions of the island (nature, cuisine, hot springs, history/culture, spiritual sites, local hospitality and sightseeing).
Those who want to take a Seven Stars journey need to apply online.
Prices start from ¥180,000 ($1,765) per person for the two-night trip.
No first come first serve here.
The company says that in the event that applications exceed available places, a lottery will be conducted to select participants.

Tabela B.1 – Texto intitulado "*Ferrari designer creates super luxurious train*"

Yamagata-born Ken Okuyama, well known in automobile design circles, was brought in to style the company's super slick new Cruise Train, due to start chugging down the tracks in spring 2017. Holding a maximum of 34 passengers, the Cruise Train will have 10 carriages made up of five suites, one deluxe suite, two glass-walled observation cars, a dining car and lounge. Okuyama has worked as a chief designer for General Motors, a senior designer for Porsche AG and design director for Pininfarina, the company behind the Ferrari Enzo and Maserati Quattroporte.

Tabela B.2 – *Gold Standard* do Texto intitulado "*Ferrari designer creates super luxurious train*"

Yamagata-born Ken Okuyama designed JR East's new Cruise Train, due to debut in 2017. Okuyama has designed for Porshe and Pininfarina – the company that designs Ferrari and Maserati. Cruise Train to have 10 carriages that hold maximum of 34 passengers.

Tabela B.3 – *Highlight* do Texto "*Ferrari designer creates super luxurious train*"

APÊNDICE C – EXEMPLOS DE TEXTO DA CLASSE "*BUSINESS*" DO *CORPUS* DO CNN

Turkey and Japan have agreed to a $22 billion deal to build a nuclear power plant in Turkey, the semi-official Turkish news agency Anadolu reported. The deal, signed Friday, represents a step toward recovery for Japan's nuclear industry, left reeling by the devastating earthquake and tsunami that triggered a disaster at the Fukushima Daiichi plant in 2011. Turkish Prime Minister Recep Tayyip Erdogan and Japan's Prime Minister Shinzo Abe, speaking at a joint news conference, said the deal would place relations between the two nations on a different level. Japanese and French companies lead the consortium which will deliver the contract signed by the two governments. The third-generation ATMEA1 reactor will be built in Turkey's northern Sinop province, which borders the Black Sea, said GDF SUEZ, part of the consortium. Erdogan said lessons had been learned from the Fukushima disaster, the worst nuclear accident in a generation. "After the Fukushima incident, people said negative things about Japanese technology,"Erdogan said, quoted by Anadolu. But, he said, in response to that criticism he drew a parallel with what happens after an air crash. "We should consider there is a risk of accident, but we need that technology. With an advanced technology we will take better steps." Turkey, like Japan, is in an active earthquake zone. The two prime ministers said they hoped it would take less than a decade to complete the project. The two countries have also agreed to found a Turkish-Japanese Technical University in Turkey, Anadolu reported, with plans to follow up with a Japan-based counterpart in the future. "A step like this between Turkey and Japan is really important,"said Erdogan. The tsunami that hit Fukushima Daiichi after Japan's historic earthquake knocked out power and coolant systems at the plant, resulting in meltdowns in three reactors. The result was the most serious nuclear accident since Chernobyl, as the crippled reactors spewed enormous amounts of radioactive particles into the environment.

Tabela C.1 – Texto intitulado "*Japan signs deal with Turkey to build nuclear plant*"

Turkey and Japan have agreed to a $22 billion deal to build a nuclear power plant in Turkey, the semi-official Turkish news agency Anadolu reported. The third-generation ATMEA1 reactor will be built in Turkey's northern Sinop province, which borders the Black Sea, said GDF SUEZ, part of the consortium. Erdogan said lessons had been learned from the Fukushima disaster, the worst nuclear accident in a generation. Turkey, like Japan, is in an active earthquake zone.

Tabela C.2 – *Gold Standard* do Texto intitulado "*Japan signs deal with Turkey to build nuclear plant*"

Japan and Turkey agree a $22 billion contract for a nuclear reactor. The reactor will be built in Turkey's northern Sinop province, on the Black Sea. Turkey's prime minister says technology has advanced since Japan's Fukushima disaster. Turkey, like Japan, is in an active earthquake zone.

Tabela C.3 – *Highlight* do Texto "*Japan signs deal with Turkey to build nuclear plant*"

APÊNDICE D – EXEMPLOS DE TEXTO DA CLASSE "*COMPUTER AND INTERNET*" DO *CORPUS* DO CNN

WhatsApp, the globally popular texting app that Facebook just acquired for a whopping $19 billion, is adding phone calls to its list of services. At the Mobile World Congress in Barcelona, Spain, WhatsApp CEO Jan Koum said the voice service will be free and begin rolling out to users within the next few months. Currently, WhatsApp offers unlimited text and voice-mail messages between users. Its service is free for the first year, then costs 99 cents annually. "We want to make sure people always have the ability to stay in touch with their friends and loved ones really affordably,"Koum said in a speech at Mobile World Congress. As reported by multiple news outlets, Koum also announced that WhatsApp now has 465 million monthly users and 330 million daily users. The latter is 15 million more than what was made public last week when Facebook announced the purchase. Voice service will come first to Apple devices and Google's Android operating system, with Windows phones and Blackberry to follow. The move puts WhatsApp in competition not only with other messaging apps that offer voice but chat tools such as Skype and even mobile carriers. WhatsApp's unlimited texting has already helped establish it in places where smartphones and fancy data plans are less common. It has 40 million users in India and another 38 million in Brazil, two countries highly coveted by tech companies such as Facebook for their large populations and emerging mobile customer base. WhatsApp hasn't released figures for the United States, where it is less popular. Last week, Facebook shocked the business world when it announced it was buying WhatsApp for up to $19 billion in cash and stock – by far the social network's largest acquisition to date.

Tabela D.1 – Texto intitulado "*Facebook's WhatsApp adding voice calls*"

WhatsApp, the globally popular texting app that Facebook just acquired for a whopping $19 billion, is adding phone calls to its list of services. At the Mobile World Congress in Barcelona, Spain, WhatsApp CEO Jan Koum said the voice service will be free and begin rolling out to users within the next few months. Last week, Facebook shocked the business world when it announced it was buying WhatsApp for up to $19 billion in cash and stock – by far the social network's largest acquisition to date.

Tabela D.2 – *Gold Standard* do Texto intitulado "*Facebook's WhatsApp adding voice calls*"

Facebook bought popular texting service WhatsApp last week. WhatsApp adding voice calls to its messaging service. CEO announces the plan at Mobile World Congress in Spain.

Tabela D.3 – *Highlight* do Texto "*Facebook's WhatsApp adding voice calls*"

APÊNDICE E – EXEMPLOS DE TEXTO DA CLASSE "*ECONOMY AND FINANCE*" DO *CORPUS* DO CNN

Never underestimate the capacity of the Eurozone to shoot itself in both feet. "Breath-taking,Staggering,Bewildering." Just some of the adjectives we could use to describe the latest Eurozone fiasco where the troika – made up of the European Commission, the European Central Bank and the International Monetary Fund – has managed to snatch defeat from the jaws of victory. The decision to "bail-in"depositors in Cyprus – to make ordinary people pay for the mistakes of the banks – is extraordinary. Nothing like this was done in Greece, Portugal or Ireland. Why, oh why, it should be up to the Cypriots to test drive this dangerous and maverick policy remains unknown. To be sure, the presence of large Russian and offshore funds in Cypriot banks is cause for a different format than those used in other peripheral countries. For example, you could not realistically have the money of German housewives bailing out Russian oligarchs. There were more sophisticated and rapier like ways to achieve the same goal, taxing deposits under €100,000 at over 6% is effectively punishing the Cypriot people. Cypriots are already going to feel the ferocity from the effects of a recession caused by lower wages and high unemployment, they do not need the additional confiscatory measures of their deposits being taken. Once again the Eurozone has shown it is better at creating crises than confronting them. In the past few hours, I've spoken to bank CEOs and top economists. Words like "great mistake"and "disastrous"are being used. So far I have not spoken to anyone who thinks this is a good idea. At all levels, for depositors, future foreign investors and Eurozone policy, the decision on tiny Cyprus is going to have huge ramifications.

Tabela E.1 – Texto intitulado "*Cypriot disastrous decision How to turn a drama into a crisis*"

For example, you could not realistically have the money of German housewives bailing out Russian oligarchs. There were more sophisticated and rapier like ways to achieve the same goal, taxing deposits under €100,000 at over 6% is effectively punishing the Cypriot people. Once again the Eurozone has shown it is better at creating crises than confronting them.

Tabela E.2 – *Gold Standard* do Texto intitulado "*Cypriot disastrous decision How to turn a drama into a crisis*"

The money of German housewives should not bail out Russian oligarchs, writes Quest. Quest: Taxing deposits under €100,000 at over 6% is effectively punishing the Cypriot people. Quest: The Eurozone has shown it is better at creating crises than confronting them.

Tabela E.3 – *Highlight* do Texto "*Cypriot disastrous decision How to turn a drama into a crisis*"

APÊNDICE F – EXEMPLOS DE TEXTO DA CLASSE "*EDUCATION*" DO *CORPUS* DO CNN

"I have rights. I have the right of education,"Malala Yousufzai boldly asserted during an interview with CNN last year. Now the 14-year-old girl from Pakistan is slowly recovering after being shot in the head by the Taliban for blogging against them and defending the right of girls to go to school. Her plight has inspired people far beyond her home in the Taliban-heavy Swat Valley. Large crowds are rallying around the world to show support for Yousufzai and her cause. Before the attack, Yousufzai was in the process of starting a charity, the Malala Education Development Organization, to promote female education in northern Pakistan. Other organizations are also working in the region to turn her dream into a reality for all girls in Pakistan. UNICEF condemned the assault, calling Yousufzai a "courageous voice"who speaks for millions of girls "desperate to receive and education." To make a donation to UNICEF's Stand with Malala campaign and support education programs in Pakistan, visit the organization's website. The Citizens Foundation has worked to improve education in Pakistan since 1995 and started 830 schools, according to the organization's website. The group says it encourages girls to enroll in its schools and works to ensure that approximately half of its students are female. Go online to make a donation. Developments in Literacy also operates schools and provides teacher training in Pakistan. The organization says that more than 17,000 students are enrolled in its schools, approximately 68% of them girls, according to its website. To make a donation in honor of Yousufzai, visit the group's website. Be sure to write "Malala"in the notes. "I Am Malala"is an online petition honoring Yousufzai and calling for Pakistan and countries worldwide to ensure all children have access to the education. The initiative was launched by the Office of the U.N. Special Envoy for Global Education. To sign the petition, visit the website. You can also share your story and promote girls' education on CNN iReport. Girls + Education = #BasicMath is spreading the message that educating girls in developing nations can change the world.

Tabela F.1 – Texto intitulado "*Pakistani teen inspires others to fight for education*"

Now the 14-year-old girl from Pakistan is slowly recovering after being shot in the head by the Taliban for blogging against them and defending the right of girls to go to school. Other organizations are also working in the region to turn her dream into a reality for all girls in Pakistan. "I Am Malala"is an online petition honoring Yousufzai and calling for Pakistan and countries worldwide to ensure all children have access to the education.

Tabela F.2 – *Gold Standard* do Texto intitulado "*Pakistani teen inspires others to fight for education*"

Malala Yousufzai is a 14-year-old Pakistani activist fighting for the right of girls to go to school. Yousufzai was shot in the head by the Taliban for blogging against them. Nonproft organizations are working in Pakistan to help girls gain access to education. Sign the "I Am Malala"petition or submit an iReport to show your support for Yousufzai.

Tabela F.3 – *Highlight* do Texto "*Pakistani teen inspires others to fight for education*"

APÊNDICE G – EXEMPLOS DE TEXTO DA CLASSE "*EMPLOYMENT AND WORK*" DO *CORPUS* DO CNN

London UK police arrested four people Tuesday on suspicion of traveling to Syria or supporting the fighting there. Two men, ages 29 and 18, and a 21-year-old woman were arrested in Manchester, northwest England, on suspicion of being involved in the commission, preparation or instigation of acts of terrorism, Greater Manchester Police said. A fourth person, a 29-year-old man, was arrested in Oxford on the same charge, a police statement said. All four are now being questioned by counterterror officers in Manchester. Detective Chief Superintendent Tony Mole, head of the counterterror unit, said there was no imminent threat to anyone in Manchester or Britain. "The operation has been running since autumn 2013, since we first became aware of a number of individuals traveling from the northwest to the battlefields of Syria,"he said. Mole said even those who travel to the region with the intention of providing humanitarian aid are putting themselves in danger. "We know that some have already lost their lives or been detained by the regime and badly treated,"he said. "There are serious concerns that anyone traveling to Syria, whether for humanitarian reasons or because of a desire to support the Syrian opposition, may be targeted by extremist groups who want to recruit them. This could have serious repercussions for the safety of the individual concerned." Syria's civil war will this month have dragged on for three painful years. More than 100,000 people have died and more than 680,000 others have been wounded, the United Nations has said. At least 6.5 million have been internally displaced and nearly 2.5 million people have fled to other countries.

Tabela G.1 – Texto intitulado "*UK police arrest 4 on suspicion of Syria-linked terror offenses*"

London UK police arrested four people Tuesday on suspicion of traveling to Syria or supporting the fighting there. All four are now being questioned by counterterror officers in Manchester. "The operation has been running since autumn 2013, since we first became aware of a number of individuals traveling from the northwest to the battlefields of Syria,"he said..

Tabela G.2 – *Gold Standard* do Texto intitulado "*UK police arrest 4 on suspicion of Syria-linked terror offenses*"

Four people are arrested by UK police on suspicion of Syria-related terror offenses. The three men and a woman are being questioned by counterterror officers in Manchester. Police say a number of people have traveled to the battlefields of Syria.

Tabela G.3 – *Highlight* do Texto "*UK police arrest 4 on suspicion of Syria-linked terror offenses*"

APÊNDICE H – EXEMPLOS DE TEXTO DA CLASSE "*ENTERTAINMENT*" DO *CORPUS* DO CNN

It's not hard to dash out a Facebook status update saying how much you're digging "House of Cards"or the new Lorde song. But Facebook wants to make it even easier. The social network on Wednesday said it's rolling out an audio-recognition feature that lets you automatically tag music, TV shows or movies in status updates. The feature employs your phone's microphone to identify the song or TV show while it's playing and tag it in your post, saving you the trouble of typing it yourself. In this way, the new tool acts sort of like Shazam, the mobile app that can tell you what song is playing on the radio. Facebook said the unnamed feature will be available to U.S. users on Android and iOS devices "in the coming weeks." Of course, the more information Facebook users share about themselves and their tastes, the more Facebook can target ads at its 1.2 billion users. Here's how it works: If you've turned the feature on, you'll see an audio icon jiggling on your phone's screen as you write a status update. That means the feature is listening and trying to find a match; if it does, you can then add the song, TV show or movie to your post. As with any Facebook post, you can choose which of your friends can see it. You can also turn the feature off at any time by clicking an audio icon at the top right of the screen. Facebook said that if you choose to share a song, your friends can listen to a 30-second snippet. For TV shows, Facebook said your News Feed post will highlight the specific season and episode you're watching, "so you can avoid any spoilers and join in conversations with your friends after you've caught up." That's probably for the best – nobody wants overeager commenters revealing unforseen deaths in "Game of Thrones." For movies, the feature will presumably work best for people watching at home, since most movie houses (and moviegoers) frown on glowing screens in the theater.

Tabela H.1 – Texto intitulado "*Facebook adds new way to tag updates about music, TV*"

The social network on Wednesday said it's rolling out an audio-recognition feature that lets you automatically tag music, TV shows or movies in status updates. The feature employs your phone's microphone to identify the song or TV show while it's playing and tag it in your post, saving you the trouble of typing it yourself. Facebook said the unnamed feature will be available to U.S. users on Android and iOS devices "in the coming weeks.".

Tabela H.2 – *Gold Standard* do Texto intitulado "*Facebook adds new way to tag updates about music, TV*"

New Facebook feature helps you tag music, TV shows or movies in updates. Feature employs phone's microphone to identify song or TV show as it's playing. Feature will be available on Android and iOS devices "in the coming weeks".

Tabela H.3 – *Highlight* do Texto "*Facebook adds new way to tag updates about music, TV*"

APÊNDICE I – EXEMPLOS DE TEXTO DA CLASSE "*FOOD AND BEVERAGE*" DO *CORPUS* DO CNN

One of New York City's most exclusive restaurants is in a real pickle after being served a "C"grade by the New York City Department of Health. The restaurant, Per Se, was slammed by health inspectors after racking up 42 violation points during its inspection on February 19, city health department records showed. Per Se, one of only seven in New York City to earn three Michelin stars, previously had an "A"rating before the inspection. Violations listed in the latest health inspection included no hand-washing facility or soap in the food-prep area, hot and cold items held in improper temperatures, and eating or drinking in the food-prep area and tobacco use, all of which qualify as "critical"violations, according to the records. In its last inspection in June 2013, Per Se had only one violation worth 7 points, but previous inspections in 2013 and 2011 also fell into the 40-point range. Inspectors give an "A"for 0 to 13 points, "B"for 14 to 27 points and "C"for 28 or more, according to the health department's website. In a statement, Per Se's chef, Thomas Keller, said: "We look forward to the opportunity to address the allegations with the Department of Health in the upcoming Oath Tribunal. At that time our final grade will be determined. As with all of our restaurants, we continue to maintain the highest standards at Per Se."The restaurant will have a chance to argue the inspection at a hearing but must post a sign that reads "Grade Pending"until then. Keller oversaw the cuisine Sunday night at the Vanity Fair Oscar party, where guests enjoyed chicken pot pie and truffle lasagna, according to a menu that Keller tweeted. Per Se is considered the East Coast interpretation of Keller's French Laundry restaurant in Northern California. For those interested in putting their money where their mouth is, the eatery offers a nine-course tasting menu for $310 per person.

Tabela I.1 – Texto intitulado "*Posh NYC restaurant roasted by health inspectors*"

One of New York City's most exclusive restaurants is in a real pickle after being served a "C"grade by the New York City Department of Health. Violations listed in the latest health inspection included no hand-washing facility or soap in the food-prep area, hot and cold items held in improper temperatures, and eating or drinking in the food-prep area and tobacco use, all of which qualify as "critical"violations, according to the records. The restaurant will have a chance to argue the inspection at a hearing but must post a sign that reads "Grade Pending"until then.

Tabela I.2 – *Gold Standard* do Texto intitulado "*Posh NYC restaurant roasted by health inspectors*"

Per Se, one of New York's most exclusive restaurants, gets a "C"grade. Violations included no hand-washing facility or soap in food prep area. Per Se will have a chance to argue the inspection at a hearing.

Tabela I.3 – *Highlight* do Texto "*Posh NYC restaurant roasted by health inspectors*"

APÊNDICE J – EXEMPLOS DE TEXTO DA CLASSE "*GAMES*" DO *CORPUS* DO CNN

The delayed stadium in the Brazilian city of Curitiba has retained its World Cup status after satisfying FIFA that all was being done to get the Arena da Baixada ready for June's finals. "It is essential that the works are maintained at the required levels and that a collective effort by all the stakeholders involved in Curitiba continues,"he said. Spain-Australia aside, Curitiba is also set to host the following group games: Iran-Nigeria, Honduras-Ecuador and Algeria-Russia.

Tabela J.1 – *Gold Standard* do Texto intitulado "*Curitiba stadium retains World Cup status*"

Curitiba's Arena da Baixada retains World Cup status after satisfying FIFA. FIFA Secretary General says it is 'essential' that progress is maintained. Stadium set to host four World Cup group games..

Tabela J.2 – *Highlight* do Texto "*Curitiba stadium retains World Cup status*"

The delayed stadium in the Brazilian city of Curitiba has retained its World Cup status after satisfying FIFA that all was being done to get the Arena da Baixada ready for June's finals.
Last month, world football's governing body gave local organizers a deadline of 18 February by which to have made significant improvements or risk losing its four World Cup games.
One of these matches includes the final group game for world champions Spain against Australia on June 23.
"The special committee instigated by Brazil's Ministry of Sports following an emergency meeting on January 21, consisting of representatives of Atletico Paranaense, the state of Parana and the city of Curitiba, has managed ... to develop a comprehensive recovery plan which includes the solving of the financial challenges involved,"said FIFA in a statement.
The Arena de Baixada venue, home to Atletico Paranaense in the southern state of Parana, is being expanded for the World Cup with new seats added alongside the pitch and capacity raised to 40,000.
Officials claim the stadium should now be ready by May 15, with work set to intensify yet further and a minimum of 1,500 workers guaranteed to be on-site.
Nonetheless, FIFA Secretary General Jerome Valcke warned that the pace of improvement must not falter. "It is essential that the works are maintained at the required levels and that a collective effort by all the stakeholders involved in Curitiba continues,"he said.
"It is a race against a very tight timeline and will require regular monitoring, but we are counting on the commitment made by the Atletico Paranaense, the city and the state of Curitiba."
Some may question the decision to intensify work when six construction workers have died in the rush to meet FIFA's World Cup deadlines.
But Luis Fernandes, Brazil's Deputy Sports Minister, said he was delighted to see the "three measures plan"working out.
These cover the progress on construction, improved financial guarantees as well as increased commitments by local organizers.
"It is great to see the significant progress made since our last visit.
It's a city which lives and breathes football,"said a man who is also the executive coordinator within the government for the FIFA World Cup.
Curitiba is one of four stadiums that missed FIFA's December deadline for completion.
Aside from simply finishing the stadium, local officials must also carry out a number of security tests at new arenas to ensure that they are both safe and fully operational.
Spain-Australia aside, Curitiba is also set to host the following group games: Iran-Nigeria, Honduras-Ecuador and Algeria-Russia.
Preparations for the World Cup have been controversial in Brazil.
Protesters are outraged at what they consider lavish spending on the World Cup as well as the 2016 Olympic Games.
Brazil has not hosted the World Cup since 1950 – when it lost 2-1 in the deciding match to Uruguay.
The 2014 tournament is due to open on June 12 with Brazil taking on Croatia in Sao Paulo's Arena Corinthians, a stadium which has also had its own renovation issues.

Tabela J.3 – Texto intitulado "*Curitiba stadium retains World Cup status*"

APÊNDICE K – EXEMPLOS DE TEXTO DA CLASSE "*GOVERNMENT AND POLITICS*" DO *CORPUS* DO CNN

A U.S. airstrike near Baghdad on Monday marked a new phase in the fight against ISIS. The airstrike southwest of the city appears to be the closest the U.S. airstrikes have come to the capital of Iraq since the start of the campaign against ISIS, a senior U.S. military official told CNN. And U.S. Central Command said in a statement that it was the first strike as part of "expanded efforts"to help Iraqi forces on the offensive against ISIS. Monday's airstrike destroyed an ISIS fighting position that had been firing at Iraqi forces, Central Command said. It occurred about 35 km (22 miles) southwest of Baghdad, another U.S. official said. The United States began targeted airstrikes against ISIS in Iraq last month to protect American personnel and support humanitarian missions. Last week, U.S. President Barack Obama said new airstrikes would aim to help Iraqi forces on the offensive against the Islamist militants. Obama also said airstrikes would include ISIS targets in Syria. And last week he also asked Congress for authorization to train and equip moderate Syrian rebels. The authority comes under Title 10 of the U.S. code, which deals with military powers, and Congress could vote on granting it this week. Approval also would allow the United States to accept money from other countries for backing the Syrian opposition forces. A senior administration official told reporters Monday that Obama has been making calls to Democratic and Republican members of Congress, asking them to pass the authorization. U.S. Secretary of State John Kerry courted Middle Eastern leaders over the weekend to join a coalition in the fight against the Islamist militant group, which calls itself the Islamic State. More than two dozen nations, the Arab League, the European Union and United Nations met in the French capital Monday, calling ISIS a threat to the international community and agreeing to "ensure that the culprits are brought to justice." The United States has conducted more than 150 airstrikes in Iraq against ISIS, and Kerry has said nearly 40 nations have agreed to contribute to the fight against the militants. But it remains unclear which countries are on that list and the precise roles they'll play.

Tabela K.1 – Texto intitulado "*US airstrike aims at ISIS near Baghdad*"

The airstrike southwest of the city appears to be the closest the U.S. airstrikes have come to the capital of Iraq since the start of the campaign against ISIS, a senior U.S. military official told CNN. And U.S. Central Command said in a statement that it was the first strike as part of "expanded efforts"to help Iraqi forces on the offensive against ISIS. Monday's airstrike destroyed an ISIS fighting position that had been firing at Iraqi forces, Central Command said.

Tabela K.2 – *Gold Standard* do Texto intitulado "*US airstrike aims at ISIS near Baghdad*"

The U.S. military says an airstrike near Baghdad is the first in "expanded efforts". Appears to be closest U.S. airstrikes have come to capital in campaign against ISIS. It destroyed an ISIS position that had been firing at Iraqi forces, Central Command says.

Tabela K.3 – *Highlight* do Texto "*US airstrike aims at ISIS near Baghdad*"

APÊNDICE L – EXEMPLOS DE TEXTO DA CLASSE "*HEALTH*" DO *CORPUS* DO CNN

Hungry? Grab a handful of nuts. Not only are they packed with protein, but it turns out they may be the food for longevity. At least, that's the conclusion of the largest study to date looking at the relationship between eating nuts and longer lives. Nuts are high in unsaturated fats, protein and vitamins, as well as antioxidants that are thought to be linked to a lower risk of heart disease. Researchers from Brigham and Women's Hospital and Harvard Medical School looked at nut consumption and deaths from all causes among 76,464 women participating in the Nurse's Health Study and 42,498 men involved in the Health Professionals Follow-up Study. They asked the participants about their nut consumption, including how many almonds, cashews, hazelnuts, macadamias, pecans, pine nuts, pistachios or walnuts they typically ate. Those who reported regularly consuming nuts were less likely to die from a variety of diseases, most significantly cancer, heart disease and respiratory diseases. People who ate nuts seven or more times a week, in fact, enjoyed a 20% lower death rate after four years than individuals who did not eat nuts. Nut eaters also tended to be leaner, more physically active, and non-smokers. Prior studies found similar connections between nuts and longer life, but the large size of this study gives the association more support. The study was partially funded by the International Tree Nut Council Nutrition Research & Education Foundation, a nonprofit organization representing nine tree nut industries, but the group played no role in the research or results, said Maureen Ternus, executive director. How many nuts does it take to extend lifespan? That's not clear, and the scientists say that the findings don't imply any cause and effect relationship between nuts and later death, but the correlation is worth investigating further. Nuts are part of the balanced diet that public health officials recently outlined in the Dietary Guidelines for Americans – the government group advised that adults eat about five to six ounce of protein (which could include nuts) a day. This story was initially published on TIME.com.

Tabela L.1 – Texto intitulado "*Eat nuts, live longer*"

Researchers from Brigham and Women's Hospital and Harvard Medical School looked at nut consumption and deaths from all causes among 76,464 women participating in the Nurse's Health Study and 42,498 men involved in the Health Professionals Follow-up Study. Prior studies found similar connections between nuts and longer life, but the large size of this study gives the association more support. That's not clear, and the scientists say that the findings don't imply any cause and effect relationship between nuts and later death, but the correlation is worth investigating further.

Tabela L.2 – *Gold Standard* do Texto intitulado "*Eat nuts, live longer*"

A large study finds an association between longevity and nut consumption. The study links nut consumption to fewer disease-related deaths. More research is needed, the study authors say.

Tabela L.3 – *Highlight* do Texto "*Eat nuts, live longer*"

APÊNDICE M – EXEMPLOS DE TEXTO DA CLASSE "*LAW*" DO *CORPUS* DO CNN

Former New England Patriots star Aaron Hernandez used "coded messages"in jailhouse calls to discuss the killing of Odin Lloyd, Massachusetts prosecutors said in court papers. Hernandez, 24, is being held on first-degree murder and weapons charges in the shooting death last year of his friend Lloyd. Hernandez has pleaded not guilty. In court papers filed Thursday, the Bristol County District Attorney's Office asked a state court to order the Sheriff's Office to turn over recordings of jailhouse calls and records of the people who visited Hernandez since the former tight end was arrested in Lloyd's killing in June. Prosecutors said Hernandez used coded messages when discussing with friends allegations that he planned the June 17 killing of Lloyd in a North Attleborough, Massachusetts, industrial park, where Lloyd was found shot to death. Hernandez's lawyers moved to quash the request, calling it a "fishing expedition." "There is absolutely no basis to provide the Commonwealth with access to all of the defendant's recorded phone calls, past, present, and future,"his attorneys, Michael Fee and Jamie Sultan, said in their court filing. Those who have visited Hernandez included his fiancee, Shayanna Jenkins, and cousin, Tanya Cummings Singleton, both of whom face charges in connection with the Lloyd homicide, the court papers said. "Both of these co-defendants have been charged as accessories in the underlying offense of murder on the theory that they provided assistance to the defendant after the commission of that offense,"Bristol Assistant District Attorney Roger L. Michel Jr. wrote in a court papers requesting the tapes and records. Michel wrote that in jailhouse phone conversations Hernandez discussed "matters directly relevant to the circumstances surrounding the murder of Odin Lloyd; viz: the defendant's subjective belief about his criminal liability; his use of coded messages to communicate with persons outside of jail; related prior offenses; inculpatory denials of ownership of a vehicle connected with the investigation; the extent of his control over persons charged as accessories; other matters relating to his codefendants, including their whereabouts and likely criminal liability.". Both detainees and people they call are warned by the jail that their conversations are recorded. Authorities have said that Hernandez and two other men picked Lloyd up from his Boston apartment in a rental car shortly before he was found shot to death June 17. Surveillance cameras then captured the rental car leaving the crime scene and Hernandez carrying a gun as he returned to his home minutes later.

Tabela M.1 – Texto intitulado "*Prosecutors say Aaron Hernandez discussed killing in jailhouse calls*"

Hernandez, 24, is being held on first-degree murder and weapons charges in the shooting death last year of his friend Lloyd. Hernandez has pleaded not guilty. In court papers filed Thursday, the Bristol County District Attorney's Office asked a state court to order the Sheriff's Office to turn over recordings of jailhouse calls and records of the people who visited Hernandez since the former tight end was arrested in Lloyd's killing in June. Prosecutors said Hernandez used coded messages when discussing with friends allegations that he planned the June 17 killing of Lloyd in a North Attleborough, Massachusetts, industrial park, where Lloyd was found shot to death.

Tabela M.2 – *Gold Standard* do Texto intitulado "*Prosecutors say Aaron Hernandez discussed killing in jailhouse calls*"

NEW: Defense attorneys move to block request for Aaron Hernandez's jailhouse calls. The former NFL star is being held on first-degree murder charge in death of Odin Lloyd. Prosecutor says Hernandez used "coded messages"while discussing killing during calls. He has pleaded not guilty to the charge stemming from the June killing.

Tabela M.3 – *Highlight* do Texto "*Prosecutors say Aaron Hernandez discussed killing in jailhouse calls*"

APÊNDICE N – EXEMPLOS DE TEXTO DA CLASSE "*SCIENCE*" DO *CORPUS* DO CNN

From Earth, the sun appears as a constant circle of light, but when viewed in space a brilliant display of motion is revealed. Flares that light up the galaxy and eruptions that can be as large as 30 times the Earth's surface occur regularly. During the peak of the 11-year solar cycle, these events can happen several times a day. The flares and eruptions are collectively known as space weather and although they create dazzling visuals in space, it isn't just a harmless fireworks show for the galaxy. Each burst of energy can have a disrupting effect on systems we rely on every day. With their headquarters next to the Rocky Mountains in the state of Colorado, a team of forecasters aims to minimize that impact. "The Space Weather Prediction Center (SWPC) essentially watches the sun, watches for activity on the sun originating from sun spots,"explains Bob Rutledge, Forecast Office lead. "That's really where the magnetic fields of the sun poke through the surface and kind of hold that part of the surface in place allowing it to cool – that's why it appears dark." Gas rolls up and down the sun's outer layer, similar to the bubbles in boiling water. When the magnetic field around a sun spot breaks, magnetic energy explodes in the solar atmosphere like a pot boiling over. The size and position of sun spots can give forecasters a clue as to when or where a solar flare may bubble up. They produce daily forecasts that are important to the industries most vulnerable. "Space weather can have a variety of impacts across many different customer bases – commercial aviation, precision GPS use, power grid operations – all these are really critical,"says Rutledge. The sun is currently at its "solar maximum-- the point in its cycle where it is at peak activity – but the SWPC says that activity is modest compared to recent cycles. Nonetheless, last week the center reported that the sun had produced a "moderate-level"solar flare, which had "short-lived impacts to high frequency radio communications on the sunlit side of Earth." Solar flares can send blasts of radiation through space that can interfere with satellites and even harm astronauts during spacewalks. "So when an eruption happens – when we have that flash of light, those radio waves – that takes eight minutes to get from the sun to the Earth. So as soon as we start the measurement, it's already affecting the sunlit side of the Earth,"explains Rutledge. Innovations in spacecraft by NASA are showing us some of the best images of the sun we've ever seen – giving us a clearer picture and hopefully a better understanding of space weather. But there is still much mystery to the 4.5 billion-year-old star and the emissions that are blasted through space, so scientists and forecasters will continue to watch every movement.

Tabela N.1 – Texto intitulado "*Space weather Fine, with a chance of solar flares*"

Each burst of energy can have a disrupting effect on systems we rely on every day. "The Space Weather Prediction Center (SWPC) essentially watches the sun, watches for activity on the sun originating from sun spots,"explains Bob Rutledge, Forecast Office lead. The sun is currently at its "solar maximum-- the point in its cycle where it is at peak activity – but the SWPC says that activity is modest compared to recent cycles.

Tabela N.2 – *Gold Standard* do Texto intitulado "*Space weather Fine, with a chance of solar flares*"

Space Weather Prediction Center watches skies for solar activity. Coronal mass ejections can disrupt satellites and power grids. The sun is at its "solar maximum-- but its activity is described as "modest".

Tabela N.3 – *Highlight* do Texto "*Space weather Fine, with a chance of solar flares*"

APÊNDICE O – EXEMPLOS DE TEXTO DA CLASSE "*SOCIETY AND CULTURE*" DO *CORPUS* DO CNN

The parents of an American journalist missing in Syria have a new message for his captors: "Let us be a whole family again." In a statement published Thursday on the McClatchy Newspapers website, Austin Tice's parents say he went to Syria to share the stories of the country's people. "We urge you, whoever you are: Let Austin come home for Christmas,"Marc and Debra Tice wrote. "Let us hug him, laugh and cry with him, love him in person." Austin Tice, who was working as a freelancer for McClatchy and other news outlets, last contacted his family on August 13 while in Syria reporting on the uprising against President Bashar al-Assad's government. He was reportedly preparing to leave Syria for Lebanon when he went missing, according to his family. In Thursday's statement, Marc and Debra Tice describe their son as someone who has "a special affinity for the people of the Middle East." "He is especially attracted to your tradition of hospitality,"the statement said. "He deeply connects with your intense loyalty to family, faith and ideals." The U.S. State Department has said they believe Tice was detained by Syrian officials in August as he was preparing to leave the country. He had smuggled himself into the country to report on the uprising. In November, Marc Tice told reporters that the Syrian government had told his family that it doesn't know where their son is. In October, a shaky video surfaced on YouTube showing a man believed to be Tice surrounded by armed men walking him up a hill. State Department officials have questioned the veracity of the video, which purports to show Tice in the custody of rebels fighting the Syrian government. Earlier this month, Tice's parents told CNN they do not want to speculate about who is holding him; they just want their son back home. Austin is the oldest of the couple's seven children. "He likes to know what's going on in the world,"Debra Tice said earlier this month, and he was frustrated by the lack firsthand reporting from Syria's civil war. She said her son had reassured her that it was worth it to travel to Syria. "I'm someone that can go,"he told her. "I can face that danger because this story is important.'"

Tabela O.1 – Texto intitulado "*Missing American journalist's parents Send our son home from Syria for Christmas*"

"We urge you, whoever you are: Let Austin come home for Christmas,"Marc and Debra Tice wrote. Austin Tice, who was working as a freelancer for McClatchy and other news outlets, last contacted his family on August 13 while in Syria reporting on the uprising against President Bashar al-Assad's government. The U.S. State Department has said they believe Tice was detained by Syrian officials in August as he was preparing to leave the country. In November, Marc Tice told reporters that the Syrian government had told his family that it doesn't know where their son is.

Tabela O.2 – *Gold Standard* do Texto intitulado "*Missing American journalist's parents Send our son home from Syria for Christmas*"

Marc and Debra Tice: "We urge you, whoever you are: Let Austin come home for Christmas". Austin Tice last contacted his family on August 13 while in Syria. The State Department has said it believes he was detained by Syrian officials. The Syrian government has said it doesn't know where the journalist is.

Tabela O.3 – *Highlight* do Texto "*Missing American journalist's parents Send our son home from Syria for Christmas*"

APÊNDICE P – EXEMPLOS DE TEXTO DA CLASSE "*SPORTS*" DO *CORPUS* DO CNN

Barcelona have been handed a boost ahead of next week's European Champions League quarterfinal tie against Paris Saint-Germain after the club confirmed coach Tito Vilanova is to return to Catalonia following cancer treatment. Vilanova has been in New York for two months undergoing treatment, with assistant coach Jordi Roura assuming first team duties on a temporary basis. A statement released by the four-time European champions said Vilanova, who was assistant to Josep Guardiola between 2008 and 2012, is heading home to Spain this week. "Barcelona coach Tito Vilanova will be returning home this week after being in New York for the last two months,"read a statement on the Spanish league leader's official website. However the statement did not specify when Vilanova would return to the dugout for the four-time European champions. "Tito traveled to New York on January 21 for treatment and this week he will be returning home to the Catalan capital. "Jordi Roura has been in charge of the team over the last couple of months and has been in permanent contact with Vilanova to agree on important decisions. "But Vilanova's place on the bench has always been kept for him and this week the team will welcome their boss back. Welcome home Tito." Vilanova, who was assistant to former Barca coach Josep Guardiola between 2008 and 2012, has been undergoing treatment following a recurrence of cancer of his parotid gland, which is located in the mouth. Bounced back. The 44-year-old Vilanova initially underwent surgery to remove a tumor in November 2011. For the first time since Guardiola took over Barca in 2008, recently the Catalan team have faced questions over their performances in the time Vilanova has been away. Barca were beaten over two legs by Real Madrid in a Copa del Rey semifinal, including a crushing 3-1 home defeat at the Nou Camp stadium, while Jose Mourinho's side also beat the Catalans in the Spanish league. Between the two cup ties with Real, Barca were also beaten 2-0 by AC Milan in the first leg of their Champions League round of 16 tie. But Barca bounced back emphatically, steamrollering Milan in the second leg with a 4-0 triumph. The first leg of Barca's last eight tie with PSG takes place in Paris on April 2, before the return fixture on April 10.

Tabela P.1 – Texto intitulado "*Barcelona boosted by Tito Vilanova return*"

Barcelona have been handed a boost ahead of next week's European Champions League quarterfinal tie against Paris Saint-Germain after the club confirmed coach Tito Vilanova is to return to Catalonia following cancer treatment. Vilanova has been in New York for two months undergoing treatment, with assistant coach Jordi Roura assuming first team duties on a temporary basis. Vilanova, who was assistant to former Barca coach Josep Guardiola between 2008 and 2012, has been undergoing treatment following a recurrence of cancer of his parotid gland, which is located in the mouth.

Tabela P.2 – *Gold Standard* do Texto intitulado "*Barcelona boosted by Tito Vilanova return*"

Barcelona coach Tito Vilanova returning to Catalonia. Vilanova has been in New York for two months undergoing cancer treatment. Assistant coach Jordi Roura has been in temporary charge of the team. Vilanova took over from Josep Guardiola in June 2012.

Tabela P.3 – *Highlight* do Texto "*Barcelona boosted by Tito Vilanova return*"

APÊNDICE Q – EXEMPLOS DE TEXTO DA CLASSE "*TRAVEL*" DO *CORPUS* DO CNN

Irreverent? Obscene? Glorious? While some parts of the world are instituting bikini bans and campaigning for travelers to cover up, elsewhere, travelers are encouraging others to take it off on social media for everyone to see. What began as a spontaneous idea and Instagram post has sparked a stream of travelers to send in fun, topless photos – taken from the back – to be featured on the social media channels of the Topless Tour, which began as an Instagram project. Judging from the enthusiastic response and growing number of participants, there's something about posing topless that heightens the feeling of liberation and adventure while traveling. How it started. The Topless Tour was started two years ago by friends Olivia Edginton, 20, Lydia Buckler, 21, and Ingvild Marstein Olsen, 20, students at Trinity Laban Conservatoire of Music & Dance in London. After taking an impromptu topless dip in a cold lake in Olsen's hometown in Norway, the three friends wanted to capture the moment of freedom. "It simply just happened, no planning involved,"Edginton wrote in an e-mail. They continued to travel and post similar photos throughout Europe and New York, all featuring themselves in topless poses against dramatic backgrounds. Then they called for others to send in their topless photos from around the world. From Idaho to Cape Town to Thailand, the mildly salacious images poured in. Among the more amazing submissions: a woman bungee jumping topless in the Swiss Alps. More to come. The Topless Tour currently has nearly 20,000 followers on Instagram. That number is increasing quickly as the trio and a growing community continue to document their travels with bare backs. Although the topless photos come predominantly from women, many men have submitted as well. "We never imagined it would have such global reach and appreciation,"Edginton said. "It was always just something we would hope would catch on, never really thinking it would happen." For the trio leading the Tour, the best part of the experience has been hearing people's stories from all over the globe, and how the project has helped people "feel themselves again, be proud of who they are and love their beautiful bodies,"said Edginton. "One of our followers recently said these lovely words: 'All that pressure and judging went away. And so did my shirt. I felt free, happy, and me.' "

Tabela Q.1 – Texto intitulado "*Topless travel photos latest travel rage*"

The Topless Tour was started two years ago by friends Olivia Edginton, 20, Lydia Buckler, 21, and Ingvild Marstein Olsen, 20, students at Trinity Laban Conservatoire of Music & Dance in London. Then they called for others to send in their topless photos from around the world. The Topless Tour currently has nearly 20,000 followers on Instagram.

Tabela Q.2 – *Gold Standard* do Texto intitulado "*Topless travel photos latest travel rage*"

Three students in UK start topless travel photo trend. Travelers send in topless photos from around the world. The Topless Tour has nearly 20,000 followers on Instagram.

Tabela Q.3 – *Highlight* do Texto "*Topless travel photos latest travel rage*"

APÊNDICE R – EXEMPLOS DE TEXTO DA CLASSE "*WEATHER*" DO *CORPUS* DO CNN

Tackling the effects of climate change could cost governments around the world more than $100 billion a year, a United Nations panel of experts said Monday. A report by the U.N.'s Intergovernmental Panel on Climate Change, says that a temperature rise of 2 degrees Celsius will wipe out up to 2% of the world's income by 2050. But it says the price tag could grow even higher if the world's governments fail to address the looming climate change. "If we get up to 4 degrees temperature rise, which most scientists now expect would happen if we carry on emitting greenhouse gasses as we do, then the cost could be much more severe,"Chris Hope, a climate change researcher at Cambridge University said. The combined cost of crop losses, rising sea levels, higher temperatures and fresh water shortages could mount of to between $70 and $100 billion a year, the report said. But these estimates do not account for catastrophic scenarios, which researchers said tend to have the most devastating effect. Typhoon Haiyan, which swept through Philippines in November, killed 6,000 people and cost more than $10 billion. When severe floods hit parts of the UK earlier this year, the Federation of Small Businesses estimated the overall cost to businesses to be $1.3 billion. And according to the U.S. Department of Agriculture, the 2012 drought – the worst in 25 years – pushed up poultry prices by 5.5% and egg prices by 7%. The report says crop yields will fall by 2% per decade, as the rising temperature affects some of the world's major crops – such as rice, maze or wheat. Hope said that if people continue to emit greenhouse gasses into the atmosphere, the bill will grow for everyone. "It looks as though it's about $125 worth of extra impact for every one more ton of Carbon Dioxide we put in the atmosphere – that comes up to around $0.20 per a liter of gasoline,"he said. "Businesses must expect that, if we are serious about climate change, at some point they are going to be charged that kind of money if they carry on using gas coal, oil, gas, fossil fuels which emit those kind of gasses to the atmosphere,"he added. The report, released in Yokohama, Japan, is the second part of the IPCC's benchmark assessment of climate change, a document released every six years. Nearly 1,000 scientists contributed to it.

Tabela R.1 – Texto intitulado "*Climate change could cost $100 billion a year*"

Tackling the effects of climate change could cost governments around the world more than $100 billion a year, a United Nations panel of experts said Monday. Hope said that if people continue to emit greenhouse gasses into the atmosphere, the bill will grow for everyone. The report, released in Yokohama, Japan, is the second part of the IPCC's benchmark assessment of climate change, a document released every six years.

Tabela R.2 – *Gold Standard* do Texto intitulado "*Climate change could cost $100 billion a year*"

Climate change could cost more than $100 billion a year, the UN panel of experts said. Experts say the bill could be much higher if emissions continue at the current pace. The report is the second part of the IPCC's benchmark assessment of climate change.

Tabela R.3 – *Highlight* do Texto "*Climate change could cost $100 billion a year*"