UNIVERSIDADE FEDERAL DE JUIZ DE FORA
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

ANDRÉ ROBSON DE ALCÂNTARA SILVA

PESQUISA EXPLORATÓRIA SOBRE REALIDADE AUMENTADA NO BRASIL E EXTERIOR COM O EMPREGO DE TEXT MINING.

JUIZ DE FORA
2021

ANDRE ROBSON DE ALCÂNTARA SILVA

PESQUISA EXPLORATÓRIA SOBRE REALIDADE AUMENTADA NO BRASIL E EXTERIOR COM O EMPREGO DE TEXT MINING.

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Engenheiro de Produção.

Orientador: DSc, Mariana Paes da Fonseca

JUIZ DE FORA
2021

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

de Alcântara Silva, André Robson.
Pesquisa exploratória sobre Realidade Aumentada no Brasil e exterior com o emprego de Text Mining. / André Robson de Alcântara Silva. -- 2021.
58 f. : il.

Orientadora: DSc, Mariana Paes da Fonseca
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Faculdade de Engenharia, 2021.

1. Realidade Aumentada. 2. Pesquisa Exploratória. 3. Text Mining. I. Paes da Fonseca, DSc, Mariana , orient. II. Título.

ANDRÉ ROBSON DE ALCÂNTARA SILVA

Pesquisa exploratória sobre Realidade Aumentada no Brasil e exterior com o emprego de Text Mining.

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia da Universidade Federal de Juiz de Fora, como requisito parcial para a obtenção do título de Engenheiro de Produção.

Aprovada em 8 de setembro de 2021.

BANCA EXAMINADORA

Dsc, Mariana Paes da Fonseca (Orientadora)

UFJF

Dsc, Eduardo Breviglieri Pereira de Castro

UFJF

Dsc, Antônio Ângelo Missiaggia Picorone

UFJF

AGRADECIMENTOS

Primeiramente, à minha avó Hermínia e meu avô Luiz Carlos, que sempre me apoiaram e me proporcionaram todas as oportunidades do mundo.

À minha mãe, Josiana e meu pai José Antônio, cujas lições de vida me incentivaram a nunca desistir apesar de toda adversidade.

Aos meus irmãos Bárbara, Iuri, Nicolly e Nycollas por todo carinho, aos meus amigos de infância, Israel e Ítalo, cuja amizade me acompanhou durante toda essa jornada.

Aos amigos que a universidade me proporcionou, tanto dentro quanto fora dela, que viveram essa experiência ao meu lado e ficarão eternamente na minha vida.

À minha namorada Laila, que acreditou em mim até quando eu não acreditei.

À minha orientadora Mariana, por toda paciência, dedicação e suporte por todo o caminho.

RESUMO

A atual capacidade de difusão de tecnologias é enorme, assim como sua adaptabilidade para a indústria e para o dia-a-dia. Porém, é necessário entender qual o rumo essas tecnologias estão seguindo, principalmente considerando que a adoção das mesmas não se dá de maneira igualitária no mundo. Dentre as tecnologias em expansão, a Realidade Aumentada (RA) é tomada como objeto de estudo, em virtude da sua presença em aplicações cotidianas tanto quanto em aplicações industriais, considerada uma tecnologia de grande importância na Indústria 4.0 (apesar de sua existência datar décadas antes da Quarta Revolução Industrial).

Com base nesse contexto, o presente trabalho buscou-se compreender a expansão do tema Realidade Aumentada, utilizando dados da produção científica para realizar a comparação entre Brasil e Exterior, através pesquisa exploratória dos dados, com o objetivo de levantar semelhanças e diferenças das tendências tecnológicas observadas.

Em razão do alto volume de dados provenientes da produção científica nacional e internacional, fez-se necessário a utilização de ferramentas computacionais capazes de ler, tratar e analisar dados que sem sua utilização seriam humanamente inviáveis. Para isso foi utilizado Text Mining que é um processo automático que usa processamento de linguagem natural para extrair informações de dados textuais. A partir disso, foram utilizados algoritmos de Text Mining e *Machine Learning* para limpeza e exploração dos dados, sendo possível obter informações acerca do tema que apontam para uma adoção tardia da Realidade Aumentada no Brasil, assim como um foco maior para utilização de RA em aplicações de smartphones, principalmente relacionadas a jogos, enquanto internacionalmente se tem maior destaque para aplicações nas áreas da indústria e medicina porém, também com grande ênfase em aplicações para smartphones.

Acredita-se que a pesquisa foi capaz de atingir seu objetivo principal, realizando a comparação entre Brasil e Exterior a fim de compreender características da Realidade Aumentada com base em um conjunto de dados extenso de publicações científicas, apresentando os resultados e concluindo a partir dos mesmos. Também foi possível visualizar propostas de trabalho futuras de grande interesse.

Palavras chave: Realidade Aumentada, tecnologias, Text Mining.

ABSTRACT

The current capacity for technology diffusion is enormous, as well as its adaptability for industry and for the day-to-day. However, it is necessary to understand which direction these technologies are following, especially considering that their adoption does not occur equally in the world. Among the expanding technologies, Augmented Reality (AR) is taken as an object of study, due to its presence in everyday applications as well as in industrial applications, also for being considered a technology of great importance in Industry 4.0 (despite its existence dating back decades before the Fourth Industrial Revolution).

Based on this context, this work sought to understand the expansion of the Augmented Reality subject, using data from scientific production to perform the comparison between Brazil and abroad, through exploratory research of the scientific production data, in order to find similarities and differences in the observed technological trends.

Due to the high volume of data from national and international scientific production, it was necessary to use computational tools capable of reading, processing and analyzing data that without their use would be humanly unfeasible. For this, Text Mining was used, which is an automatic process that uses natural language processing to extract information from textual data. From this, Text Mining and Machine Learning algorithms were used for data cleaning and exploration, making it possible to obtain information on the topic that point to a late adoption of Augmented Reality in Brazil, as well as a greater focus on the use of AR in smartphone applications, mainly related to games, while internationally there is a greater focus on applications in the areas of industry and medicine, but also with a strong focus on smartphone applications.

It is believed that the research was able to achieve its main objective, making a comparison between Brazil and abroad in order to understand the characteristics of Augmented Reality based on an extensive dataset of scientific publications, presenting the findings and drawing conclusions about it. It was also possible to visualize interesting proposals for future work.

Keywords: Augmented reality, technologies, Text Mining.

LISTA DE FIGURAS

Figura 1 – Produção científica nacional 8
Figura 2 – Microsoft Hololens 2 14
Figura 3 – Continuum de virtualidade 15
Figura 4 – Workflow de projetos de ciência de dado 21
Figura 5 – Bibliotecas utilizadas 30
Figura 6 – Recorte da base de dados para ilustração 31
Figura 7 – Limpeza dos dados 32
Figura 8 – Separação das Bases 33
Figura 9 – Top categorias Brasil e Exterior 34
Figura 10 – Top áreas de pesquisas Brasil e Exterior 35
Figura 11 – Publicações por ano – Brasil e Exterior 36
Figura 12 – Palavras-chave mais comuns 37
Figura 13 – Média de citações por documento 38
Figura 14 – Limpeza do texto 39
Figura 15 – Remoção de Stop Words e criação da DTM 39
Figura 16 – Recorte da Document-Term Matrix 40
Figura 17 – Remoção adicional de palavras 41
Figura 18 – Unigrama - Brasil e Exterior 42
Figura 19 – Bigrama - Brasil e Exterior 43
Figura 20 – Trigrama - Brasil e Exterior 43

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

WoS - *Web of Science*

AED - Análise Exploratória de Dados

IA - Inteligência Artificial

IoT - *Internet of Things*

TI - Tecnologia da Informação

RA - Realidade Aumentada

HMD - *Head Mounted Display*

MARTA- *Mobile Augmented Reality Technical Assistance*

RV - Realidade Virtual

RAI - Realidade Aumentada Industrial

NLP - *Natural Language Processing*

NLTK - *Natural Language Toolkit*

SUMÁRIO

1. INTRODUÇÃO 6

1.1 JUSTIFICATIVA 7

1.2 ESCOPO DO TRABALHO 9

1.3 ELABORAÇÃO DOS OBJETIVOS 9

1.4 DEFINIÇÃO DA METODOLOGIA 10

1.5 ESTRUTURA DO TRABALHO 11

2. REVISÃO TEÓRICA 12

2.1 REALIDADE AUMENTADA 13

2.2 PESQUISA EXPLORATÓRIA 17

2.3 CIÊNCIA DE DADOS 18

2.4 DATA E TEXT MINING 23

2.4.1 FERRAMENTAL COMPUTACIONAL 27

3. RESULTADOS 30

4. CONCLUSÃO 45

REFERÊNCIAS 48

ANEXO A - TERMO DE COMPROMISSO 54

1. INTRODUÇÃO

Com o desenvolvimento da sociedade, a produção científica mundial vem apresentando números cada vez maiores, atingindo em 2009 a marca de 50 milhões de publicações desde 1965, e a expectativa de crescimento anual é de 2,5 milhões de publicações (JINHA, 2010). Alguns dos fatores responsáveis por esse crescimento são o número crescente de novos pesquisadores, periódicos e maiores investimentos em ciência e tecnologia.

Historicamente, a produção científica brasileira concentra-se nas universidades e instituições de ensino superior, inclusive em instituições particulares, embora estas em geral tenham foco maior voltado para a formação profissional. Essa produção científica apresenta taxas crescentes ao longo do tempo. Considerando publicações no repositório *Web of Science* (*WoS*) em 2019, o Brasil figurava como o 13º maior produtor científico do mundo, com um crescimento de 30% entre 2013 e 2018 (CLARIVATE ANALYTICS, 2019).

Temas relacionados à evolução da indústria apresentam cada vez mais relevância, tendo em vista seus resultados expressivos, principalmente no viés da produção, com a adoção de tecnologias e sistemas de produção no estado da arte e da técnica. Dentre as tecnologias em ascensão, é de interesse deste trabalho a realidade mista, mais especificamente, a realidade aumentada. A realidade aumentada consiste na utilização de interfaces que possibilitem a combinação de elementos reais com elementos virtuais em tempo real (AZUMA, 1997). Tecnologias de realidade aumentada vem ganhando espaço na indústria, com crescente utilização nas áreas de reparo, manutenção e controle de produção, sendo visto como uma tendência atrelada a Indústria 4.0, como também em aplicações do dia-a-dia, como jogos de smartphones, etc.

Para se estabelecer maior familiaridade com temas em ascensão, a pesquisa exploratória se mostra como um método significativo, permitindo, através da análise bibliográfica do objeto de estudo, que se compreenda características da produção científica relacionada a Realidade Aumentada, como áreas de pesquisas de interesse, categorias, publicações por período de tempo.

Com a evolução da indústria, ocorre a evolução do profissional da indústria. Há a necessidade de se desenvolver novos métodos e ferramentas que, aplicados à produção, trazem benefícios para o modo de se produzir. Cabe ao engenheiro de produção atuar na linha de frente dessas mudanças, sendo capaz de compreender, aplicar e melhorar os sistemas produtivos. É de competência do engenheiro de produção a capacidade de analisar dados e informações

referentes a diferentes processos produtivos. E, com a quantidade de dados gerados cada vez maior, o domínio de novas tecnologias se torna imprescindível.

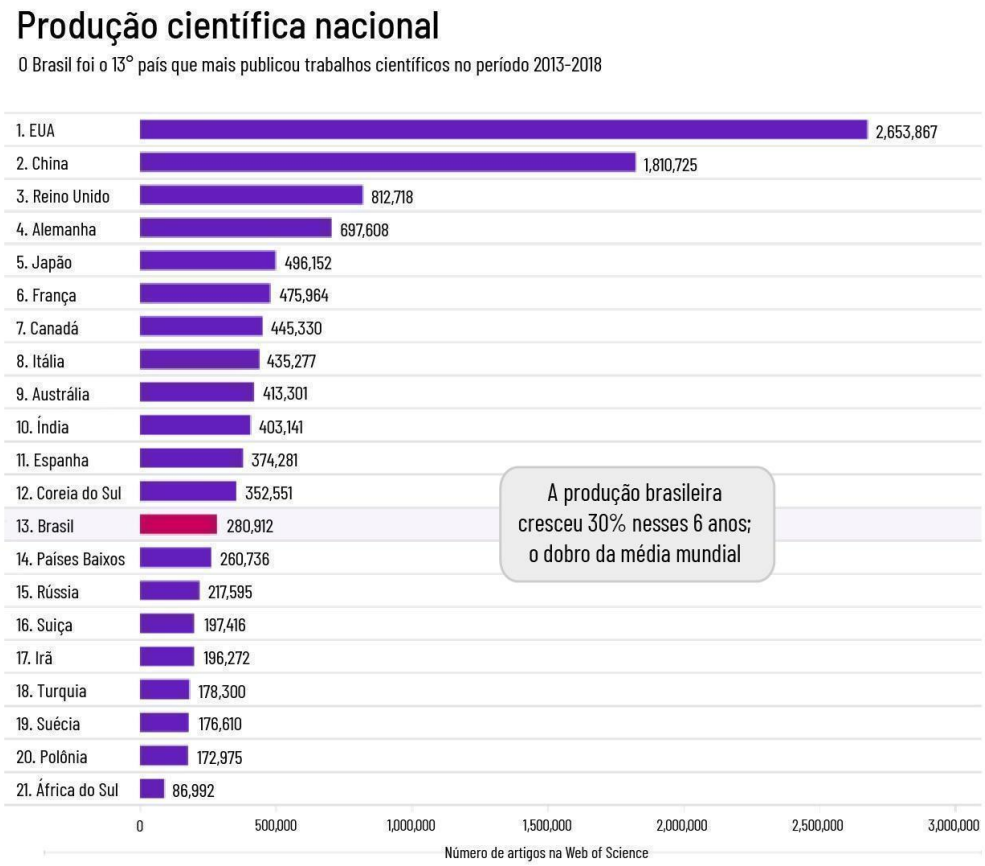
Para lidar com esse fluxo enorme de dados, se faz necessário o uso de ferramentas computacionais, como a Ciência de Dados que permite a análise de bancos de dados extensos, facilitando ao Engenheiro de Produção gerar informações a partir dos dados. Com a utilização de linguagens de programação, e bibliotecas estatísticas para exploração de dados e *Machine Learning*, o engenheiro é capaz de buscar relações interessantes em dados que, sem o auxílio de ferramentas computacionais, seriam inviáveis.

O presente trabalho buscou utilizar técnicas de ciência de dados, em especial Data e Text Mining para analisar um conjunto de dados, composto de publicações científicas retiradas de um repositório internacional, com temas relacionados à Realidade Aumentada. Foi objetivo do autor comparar, através da análise exploratória dos dados, publicações nacionais e internacionais, a fim de compreender características relevantes sobre as mesmas, bem como suas distinções.

1.1 JUSTIFICATIVA

A pesquisa no Brasil apresentou resultados positivos nos últimos anos. De acordo com *Clarivate Analytics* (2019) o país foi o 13º produtor de ciência do mundo em 2019, em número de trabalhos publicados, como apontado pela figura 1. E a produção científica nacional cresceu 30% entre 2013 e 2018. Apesar de notável que o país apresente esse crescimento dado o período de crise econômica, é esperado que os impactos serão visíveis após um intervalo de cinco anos entre o financiamento de pesquisas, a publicação de resultados e o início da contabilização de citações (ESCOBAR, 2019).

Figura 1 - Produção científica nacional



Fonte: Adaptado de ESCOBAR (2019)

O Global Innovation Index GII é um ranking anual publicado pela Cornell University em parceria com diversas instituições, que classifica os países de acordo com capacidade e sucesso em inovação. Apesar de ganhar posições nos últimos anos, o Brasil figurava na 66ª colocação em 2019 (DUTTA et al, 2019).

As instituições de ensino superior e as faculdades de Engenharia no país têm um papel importante no crescimento da produção científica, principalmente em campos relacionados à inovação e tecnologia.

Para o desenvolvimento de conhecimento científico se faz necessário compreender temas que vem crescendo como interesse de pesquisa. A pesquisa exploratória permite familiarizar-se com temas que não se possui tanto conhecimento ou domínio sobre, à maneira de a partir das características apontadas possam-se ampliar o conhecimento e desenvolver trabalhos futuros mais complexos e específicos.

As maiores motivações para a realização do presente trabalho consistem no interesse do autor pela produção científica nacional, área de análise e ciência de dados, e do interesse por

pela Realidade Aumentada, tecnologia que apresenta grande expectativa de adoção no ambiente industrial e cotidiano.

O conhecimento sobre a produção científica possibilita compreender aspectos relevantes sobre o conhecimento desenvolvido no país e compará-lo com a produção estrangeira, tornando mais evidente as diferenças, semelhanças e possíveis particularidades do interesse na tecnologia.

1.2 ESCOPO DO TRABALHO

O presente trabalho foi realizado por meio de análise exploratória e da aplicação de técnicas de *Machine Learning* em artigos da *WoS*, com o emprego de ferramentas estatísticas e computacionais de *Text Mining*, buscando responder às questões:

Quais características da produção científica brasileira sobre Realidade aumentada? E as do exterior? Quais semelhanças e diferenças podem ser apontadas?

Para a pesquisa e extração das publicações foi utilizado um recorte temporal de 30 anos, devido ao aumento da relevância do tema "Realidade Aumentada", que teve grande repercussão na década de 90, com publicações entre as mais citadas nesse campo.

Devido às limitações de tempo e carência de fontes de dados disponíveis, o presente trabalho limitou-se a responder os questionamentos supracitados. Ainda é válido ressaltar que limitações da WoS permitem apenas que sejam baixados dados da produção científica sem o texto das publicações na íntegra, o que permitiria uma análise mais profunda sobre o tema, porém, também demandaria maior capacidade de processamento computacional e técnico.

1.3 OBJETIVOS

O presente trabalho teve como objetivo principal realizar uma pesquisa exploratória em um conjunto de dados científicos sobre Realidade Aumentada, de maneira a apresentar informações sobre suas características, com o emprego de ferramentas de *Data* e *Text Mining*, buscando também comparar aspectos da produção científica brasileira e internacional em um recorte temporal entre 1990 e 2020.

1.4 DEFINIÇÃO DA METODOLOGIA

Para atingir os objetivos da pesquisa, fez-se necessário definir os métodos e ferramentas que serão utilizados em cada etapa, sendo essas a definição da fonte de dados, tratamentos e processamento dos dados para a obtenção de resultados.

Como tratou-se de uma pesquisa que analisa contribuições acadêmicas tanto internacionais quanto nacionais, optou-se por buscar artigos publicados em inglês e, dentre os bancos de dados, destacou-se a base *Web of Science* (*WoS*). A partir da seleção de repositórios foi possível montar uma base de dados com quase 14000 artigos.

Para o download do *dataset* foi utilizada a "Pesquisa Avançada" da plataforma *WoS*, e os campos de pesquisa foram preenchidos da seguinte maneira: resumo = *(augmented reality)*, *all languages, all document types*. Ou seja, publicações que contenham "realidade aumentada" em seu resumo. Não foi especificado o idioma em inglês, pois a quase totalidade dos artigos da base se encontra neste idioma. O período compreendido foi de 1990 a 2020, sendo o *dataset* extraído em julho de 2020.

Para o tratamento e análise dos dados destacou-se a linguagem de programação Python devido à familiaridade prévia com a linguagem e a disponibilidade de ferramentas para análise de *datasets* provenientes especificamente do *WoS*.

Com posse do *dataset*, utilizou-se a biblioteca de Python *Metaknowledge*, para transformar a base de dados em uma *RecordCollection*, formato tabular de registros, que contém diversos campos herdados do *WoS*. Transformou-se então essa *RecordCollection* em um *DataFrame*, objeto da biblioteca de Python *Pandas*, pacote que permite uma variedade de customizações dos dados e possui alto grau de integração com diversas outras bibliotecas de análise de dados. Foram definidos os campos principais a serem utilizados e excluídos os que inicialmente não apresentavam relevância, e os campos mantidos foram: *Pub Type* (tipo de publicação), *Author* (autor), *Author Full* (nome completo do autor), *Title*, *Publication Name* (título da publicação), *Language* (idioma), *Author Keys* (palavras-chave do autor), *WoS Keys* (palavras-chave da *WoS*), *Cit Count* (contagem de citações), *Publisher* (editora), *WoS Categories* (categorias da *WoS*), *Research Areas* (área de pesquisa), *Address* (endereço), *Year* (ano), e *Abstract*. Em seguida, foram retiradas da base registros que continham erros ou valores nulos nos campos desejados.

Após a realização da limpeza dos dados, foi realizada a análise exploratória dos dados (AED), onde foi feito um reconhecimento dos dados, gerando gráficos e séries para

compreender melhor o material que se tinha em mãos. Para a visualização dessas estatísticas foram utilizadas as bibliotecas gráficas de Python *Matplotlib* e *Seaborn*, que oferecem diferentes tipos de gráficos para análise de dados.

Com posse dos dados limpos e maior conhecimento sobre eles, foi possível aplicar algoritmos de *Text Mining*, dessa maneira adquirindo informações sobre o conjunto de textos, como identificação da frequência de palavras e expressões, incidência de palavras-chave, etc. Também realizando a comparação quantitativa de características dos textos brasileiros e estrangeiros. A partir disso, foi possível fazer comparações entre a produção científica nacional e internacional, buscando compreender diferenças, semelhanças, expressões de maior relevância, popularidade do tema dentro do recorte temporal, buscando compreender tendências tecnológicas no conjunto de dados.

1.5 ESTRUTURA DO TRABALHO

O presente trabalho foi dividido em quatro capítulos, onde, no primeiro capítulo é introduzido o tema, apresentando a motivação e objetivos do autor pela sua realização. Além disso, o capítulo introdutório apresenta a relevância do tema Realidade Aumentada com o levantamento de dados da produção científica.

O segundo capítulo é referente a revisão bibliográfica dos temas, onde são apresentados conceitos de realidade aumentada por diversos autores, destacando-se semelhanças e diferenças em suas considerações. Ademais é apresentada a definição de Pesquisa Exploratória, sendo também tratada a conceituação e evolução do tema Ciência de dados, ressaltando suas principais características e as ferramentas que serão utilizadas para a realização do referido trabalho, como *Data* e *Text mining*.

O terceiro capítulo apresenta o desenvolvimento do trabalho, mostrando como se dá a aplicação das ferramentas sobre o banco de dados de maneira que o trabalho possa ser reproduzido e enfatizando a justificativa pela utilização das ferramentas estatísticas e computacionais. Ainda neste capítulo também se expõe os resultados, que podem ser compreendidos como a conclusão da análise exploratória da base de dados e das informações geradas através dos algoritmos de *Machine Learning* aplicados.

Em sequência, o quarto e último capítulo compete à exposição das conclusões produzidas pelo autor, apresentando também oportunidades futuras de desenvolvimento mais a fundo do tema.

2. REVISÃO TEÓRICA

A indústria vive em constante evolução e, certamente, as revoluções industriais se apresentam como marcos históricos que refletem mudanças disruptivas na maneira de se produzir. Com o desenvolvimento da indústria, a engenharia de produção e o papel do engenheiro de produção também se transformaram.

Um diferencial apresentado pela Quarta Revolução Industrial é a velocidade de difusão das tecnologias emergentes e da inovação, sobretudo, ao considerar que as revoluções anteriores ainda não atingiram todos os lugares do mundo (ROSER *et al*, 2015). A Indústria 4.0 é um conceito abrangente, assim como uma nova tendência na produção e está relacionado à coleta e aplicação inteligente de dados e informações em tempo real, visando o aumento da eficiência e eficácia das operações e a redução de custos (STANKOVIC *et al*, 2017).

A partir das tecnologias que dão base a Indústria 4.0, tendências tecnológicas sob do espectro da Quarta Revolução Industrial surgem e se desenvolvem. Dentre as tendências tecnológicas pode-se destacar a Realidade Aumentada, que apesar de não ter seu nascimento atrelado a Indústria 4.0, sua adoção é incentivada e pode propiciar grandes ganhos em produtividade (RUBMANN, 2015; STANKOVIC, 2017).

Aplicações de realidade aumentada (RA) apresentam camadas gráficas de informações sobrepostas a alguma parte do campo de visão do trabalhador. É relevante apontar que tecnologias RA não são de fato novas, seu advento data do início do século XX, entretanto, atualmente a RA industrial apresenta um enorme potencial, já sendo encontrada em uso em algumas indústrias em fases iniciais. A maioria dos sistemas de RA atualmente implantados são baseados em tablet ou smartphone, embora parte desse setor esteja movendo-se agressivamente na direção dos *head-mounted displays* (HMDs), aparelhos de RA como capacetes que apresentam funções de áudio e vídeo, dessa maneira, o operador não necessita interromper sua tarefa para consultar algum dispositivo.

Um exemplo do uso de tecnologia de RA é o caso do IKEA *Mobile App*, onde a empresa do setor mobiliário disponibiliza uma aplicação que permite ao usuário escolher itens do seu catálogo de móveis e com o *smartphone* visualizar como o produto ficaria em suas casas (RUBMANN, 2015; STANKOVIC, 2017).

Com a adoção dessas novas tecnologias pela indústria, a forma que operadores realizam suas tarefas vai mudar. Operadores serão equipados com ferramentas que agem como a interface homem-máquina para comunicação, colaboração e sistemas de apoio à decisão para

otimizar suas ações (FRAGA-LAMAS *et al*, 2018). Apesar de tecnologias de RA já existirem há mais de 50 anos, sua implementação na indústria ainda não é tão expressiva, entretanto, a expectativa é de que o mercado de tecnologias de realidade aumentada cresça significativamente, criando um mercado de 80 bilhões de dólares em 2025 (PALMARINI, 2018; BELLINI, 2016).

2.1 REALIDADE AUMENTADA

A primeira aparição da Realidade Aumentada data os anos 1950 nos Estados Unidos, quando o cinematógrafo Morton Heilig concebe a ideia de que o cinema é uma atividade com a capacidade de atrair o espectador para a atividade na tela, absorvendo todos os sentidos de maneira eficaz. Heilig desenvolve o protótipo de sua visão, chamado de Sensorama, descrito por ele como “o cinema do futuro” (PACKER e JORDAN, 2002).

Em 1968, em Harvard, Ivan Sutherland (chamado de “pai da computação gráfica”) desenvolveu o primeiro sistema de RA em um HMD (*Head Mounted Display*). Anos depois, Tom Caudell e David Mizell, pesquisadores da Boeing, desenvolvem um sistema para auxiliar os trabalhadores a montar fios e cabos para uma aeronave, com um HMD capaz de mostrar no campo de visão do trabalhador as informações necessárias para o desempenho da tarefa. Ainda em seu estudo, Caudell e Mizell ainda apresentam vantagens da Realidade Aumentada *versus* Realidade Virtual. No mesmo ano, L.B Rosenberg desenvolveu um dos primeiros sistemas de RA funcionais, chamado *Virtual Fixture* e demonstrou seu benefício no desempenho humano, enquanto Steven Feiner, Blair MacIntyre e Dorée Seligmann apresentaram o primeiro artigo importante em um protótipo de sistema de RA chamado KARMA (CARMIGNIANI, 2011).

Em 1997, Ronald Azuma fornece uma definição amplamente reconhecida em RA, identificando-a como a combinação do ambiente real e virtual, ambos registrados em 3D e interativos em tempo real. O primeiro jogo para dispositivos móveis em RA foi lançado em 2000, por Bruce Thomas, chamado ARQuake. Nos anos seguintes, o número de aplicações de RA cresce, especialmente em sistemas móveis (CARMIGNIANI, 2011).

Em 2013, grandes montadoras disponibilizam aplicações para auxiliar o reparo de automóveis, com destaque para o MARTA (*Mobile Augmented Reality Technical Assistance*), aplicação da Volkswagen que disponibiliza assistência passo a passo para reparo e permite ao usuário ver como processo de reparo ficará depois de concluído. No ano seguinte, o Google Glass é revelado e disponibilizado para os consumidores. Apesar de não ser tão bem sucedido

quanto o esperado, o *Google Glass* mostrou grande potencial que a RA vestível pode atingir. Mais recentemente, a *Microsoft* apresentou a próxima geração de RA vestível, com o *HoloLens* (figura 2), dispositivo *head-mounted display* intuitivo e fácil de usar, visto como uma evolução quando comparado ao *Google Glass*, já disponível no mercado e utilizado em algumas empresas para serviços de reparo, projeto e design da linha de produção. O próprio aparelho é um computador portátil, vestível, capaz de operar sem estar conectado a uma máquina externa, tornando-o adequado para uma variedade de aplicações (ISBERTO, 2018; GARON et al, 2017).

Figura 2: Microsoft Hololens 2

Fonte: Microsoft (2019)

A Realidade Aumentada é uma variação de Ambientes Virtuais, ou Realidade Virtual (RV) como é comumente conhecida. Por meio de tecnologias de RV o usuário é completamente imerso em um ambiente sintético, sem poder ver o mundo real ao seu redor, enquanto a RA permite que ele veja o mundo real com a superposição de objetos virtuais ou compostos com o mundo real. A Realidade Aumentada complementa a realidade ao invés de substituí-la por completo (AZUMA, 1997).

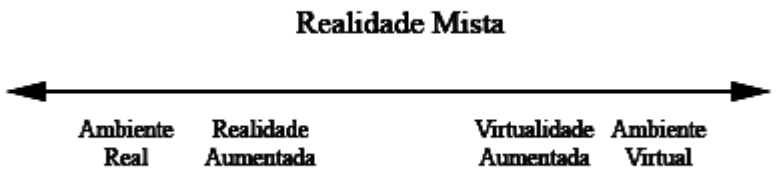
Em um novo estudo, Azuma (2001) define Realidade Aumentada da seguinte forma:

> Um sistema RA complementa o mundo real com objetos virtuais que parecem coexistir no mesmo espaço em que o mundo real, tendo posse de três características: combina objetos reais e virtuais num ambiente real; funciona interativamente, e em tempo real; e, alinha objetos reais e virtuais entre si.

É importante observar que a definição de RA não é restrita a tecnologias de exibição específicas, como HMD, ou ao menos limitada ao senso de visão. RA pode potencialmente se aplicar a todos os sentidos, incluindo audição, toque e olfato (AZUMA, 2001).

Com o crescente interesse na área da Realidade Mista, a definição operacional de Realidade Aumentada se aproxima de qualquer caso em que um ambiente real seja "aumentado" por meio de computação gráfica. A Realidade Aumentada se encontra em algum lugar ao longo do *continuum* de realidade, pois não trata-se da imersão completa num ambiente virtual (RV), porém, apesar de mais próximo não é completamente um ambiente real. O conceito de "*continuum* de virtualidade", apresentado na figura 3, se refere a mistura de classes de objetos apresentados em uma situação específica de exibição, onde ambientes reais estão em uma extremidade, enquanto ambientes virtuais encontram-se no extremo oposto do *continuum* (MILGRAM, KISHINO, 1994).

Figura 3: Continuum de virtualidade



Fonte: adaptado de Milgram e Kishino (1994)

A Realidade Aumentada (RA), e especificamente a Realidade Aumentada Industrial (RAI), é uma tecnologia que fornece ferramentas poderosas para suporte aos operadores na realização de suas tarefas, auxiliando-os nas tarefas de montagem, visualização e interação de dados (como uma Interface Homem-Máquina), localização interna, aplicações de manutenção, controle de qualidade ou gerenciamento de materiais (FRAGA-LAMAS *et al*, 2018).

Uma das aplicações mais comuns da Realidade Aumentada na indústria é a assistência aos operadores em tarefas de manutenção, reparo e controle, principalmente, através de instruções com informação textual, visual ou auditiva, provendo acesso rápido a manuais, plantas e modelos 3D, por exemplo. Esse tipo de aplicação é de grande utilidade, especialmente quando se trata de máquinas que precisam ser reparadas ou monitoradas com o mínimo de pessoas no local (PALMARINI, 2017).

Da mesma forma, o uso industrial da Realidade Aumentada pode auxiliar a tomada de decisão, combinando experiência do operador juntamente com exibição de informações extraídas em tempo real do banco de dados (FRAGA-LAMAS *et al*, 2018).

Uma aplicação de grande importância da RA é a assistência no processo de aprendizagem, podendo reduzir o tempo de treinamento para novos operadores e diminuir os requisitos de habilidades para recém contratados. A RA proporciona um ambiente de estudo

imersivo, facilitando o desenvolvimento destas competências. Além disso, é possível ajustar o treinamento pelo nível de experiência do trabalhador, acelerando esse processo, se concentrando em adquirir as capacidades para a função (FRAGA-LAMAS *et al*, 2018; AKÇAYIR e AKÇAYIR, 2017).

Setores de produção podem obter grandes benefícios com a RA, com o recebimento de informações corretas no tempo preciso, evitando erros e aumentando a produtividade. Isso é algo de grande valor quando as tarefas da produção envolvem atividades perigosas, que podem resultar em o trabalhador se machucar ou equipamentos caros sofrerem danos. Nesse tipo de situação, a Realidade Aumentada pode ser utilizada como uma ferramenta para monitoramento e diagnóstico, colhendo informações disponibilizadas pelos sistemas de controle, gerenciamento e sensores dos equipamentos (PALMARINI, 2017).

A montagem é um setor que pode ter grandes ganhos com uso de RA, reduzindo tempo e custos durante o processo. Utilizando-se do registro espacial de objetos reais e virtuais, interagindo com esses objetos e coordenando e sequenciando o trabalho humano pode-se reduzir movimentação; reduzir a troca de atenção entre o meio de instruções e a tarefa; e, ao relacionar espacialmente informações com objetos e locais no mundo real, a AR provê um forte recurso para a cognição e memória espacial dos operadores (FRAGA-LAMAS *et al*, 2018; TANG *et al*, 2003).

Na logística, a Realidade Aumentada Industrial pode aumentar a eficiência de processos de seleção em um armazém, provendo um sistema de orientação interno, a melhor rota e quais itens selecionar. A importância dessa aplicação é reforçada quando se nota que o custo de seleção de itens pode representar 55% a 65% dos custos de armazenagem (FRAGA-LAMAS *et al*, 2018).

Apesar de apresentar grandes potenciais para a indústria, a RA ainda está em suas fases iniciais de desenvolvimento e apresenta limitações tecnológicas e sociais (RUBMANN, 2015; WANG 2006). As limitações tecnológicas são o maior obstáculo para sistemas de RA, a exemplo da necessidade de sensores de longo alcance e rastreadores precisos de localização, vistos como de grande importância, porque até pequenos erros de rastreio geram erros de alinhamento notáveis entre objetos reais e virtuais (WANG 2006).

Sistemas de RA enfrentam desafios técnicos relacionados a alta resolução, profundidade de cor, luminância, contraste e foco. Interfaces não devem sobrecarregar o usuário com informações, e cuidar para que o usuário não dependa em excesso do sistema, de modo que sinais importantes do ambiente real não sejam percebidos (VAN KREVELEN e POEALMAN,

2010). Entretanto, o desenvolvimento de novos sistemas e avanços tecnológicos continuamente buscam melhorias para esses quesitos. Tecnologias de RA devem se desenvolver para serem cada vez menores, mais leves, portáteis e rápidas o suficiente para exibir gráficos de qualidade (AKÇAYIR e AKÇAYIR, 2017).

Apesar de todos os potenciais benefícios em função da adoção e desenvolvimento da Realidade Aumentada, no Brasil, a presença do tema nas universidades e empresas não é tão comum. O Ranking de Competitividade Digital 2020 mostra o país na 53ª posição de 63 países, estando aquém do potencial no conjunto de variáveis que avaliam o uso e aproveitamento das tecnologias digitais, tanto pelo setor público quanto pelo setor privado, para se tornar um país digital competitivo, que participe ativamente da quarta revolução industrial e promova o desenvolvimento da realidade tecnológica, mudanças são necessárias. (ARRUDA, 2018).

A partir disso, a pesquisa exploratória visa proporcionar maiores informações sobre um assunto investigado, ampliando a visão geral sobre o mesmo, permitindo então formular um problema mais preciso de pesquisa ou criar novas hipóteses (LEÃO, 2017). Dessa maneira, a pesquisa exploratória incentiva o desenvolvimento de ciência mais aprofundada sobre o objeto de estudo.

2.2 PESQUISA EXPLORATÓRIA

O objetivo da pesquisa exploratória é conhecer um assunto ainda pouco conhecido e explorado, com foco em conhecer as características de determinada amostra, baseando-se fortemente nas pesquisas bibliográficas (GIL, 2019).

As pesquisas exploratórias servem como ponto de partida e normalmente antecedem pesquisas explicativas, realizando descrições da situação e procurando levantar relações existentes entre seus elementos. Apesar da margem de erros representar um fator relevante, a contribuição da pesquisa exploratória é significativa, dada a sua aplicação prática (GIL, 2019).

Dentre as finalidades de uma pesquisa exploratória, especialmente quando realizada através da pesquisa bibliográfica, destacam-se a auxílio na delimitação de um tema ou objetivos de trabalho, a obtenção de maiores informações sobre determinado assunto ou descobrir possibilidades de pesquisa dentro de determinado tema (ANDRADE, 2017).

Severino (2016) diz que “a pesquisa exploratória busca apenas levantar informações sobre um determinado objeto, delimitando assim um campo de trabalho, mapeando as

condições de manifestação desse objeto. Na verdade, ela é uma preparação para a pesquisa explicativa".

De acordo com Appolinário (2011), a pesquisa ou estudo exploratório busca "aumentar a compreensão de um fenômeno ainda pouco conhecido, ou de um problema de pesquisa ainda não perfeitamente delineado", podendo ser considerada como um modo de pesquisa prospectiva. Esse tipo de pesquisa se assemelha a uma pesquisa bibliográfica ou documental quando seu objeto de estudo está centrado em documentos, ou também quando se é realizado um levantamento teórico acerca do tema pelo pesquisador (DEL-MASSO; COTTA; SANTOS, 2014).

Pesquisas exploratórias raramente são um fim em si mesmas. Geralmente, elas identificam tendências, determinam o campo de pesquisa, o ambiente, o histórico e a situação e as relações potenciais entre as variáveis. Comparadas com questões descritivas, relevantes ou explicativas, essas questões são mais flexíveis, de escopo mais amplo e diversificado, entretanto envolvem riscos maiores de erro, exigindo avaliação e atenção por parte do pesquisador para sua realização (SAMPIERI; COLLADO; LUCIO, 2013).

Essa modalidade de pesquisa é de grande importância pois permite a familiarização com fenômenos relativamente desconhecidos, e a obtenção de informações sobre possibilidades de pesquisas mais profundas e específicas, identificando possíveis problemas, conceitos ou variáveis para pesquisas futuras.

Considerando que o presente trabalho apresenta uma pesquisa exploratória em um banco de dados extenso, a análise bibliográfica demanda um método que seja capaz de gerar informações a partir desses dados. Para isso, se faz necessário o auxílio de ferramentas computacionais, que possam processar um conjunto de dados de texto, que sem sua utilização levaria um tempo extremamente longo. O método selecionado para a pesquisa exploratória é a ciência de dados, que utiliza algoritmos para analisar conjuntos de dados com a finalidade de extrair informações de relevância para a pesquisa.

2.3 CIÊNCIA DE DADOS

Dados são conjuntos de símbolos que fornecem uma imagem parcial da realidade, descrevendo propriedades de objetos e eventos. Dados podem ou não ser coletados com um propósito específico, porém, geralmente são coletados e disponibilizados em um formato

simples. Para extrair informações desses dados eles devem ser resumidos, organizados e analisados (IN e LEE, 2017).

Como expresso por Ackoff (1999) “informação consiste em dados processados, e o processamento é direcionado ao aumento de sua utilidade”. Assim como os dados, as informações também retratam propriedades de objetos e eventos, entretanto, em um formato mais compacto e útil.

No campo computacional, dados e informações são definidos em concordância com o que foi expresso anteriormente, como defendido por Chen *et al* (2008):

> Dados são representações computadorizadas de modelos e atributos de entidades reais ou simuladas. Informações são dados que representam os resultados de um processo, como análise estatística, para atribuir significados para os dados ou as transcrições de alguns significados atribuídos pelos seres humanos.

A principal diferença entre dados e informação é funcional, não estrutural. Informações consistem na consolidação de dados de forma a fundamentar o conhecimento, capaz de responder questões relacionadas a como, quem, o quê, quando e onde (ACKOFF, 1999).

A partir da necessidade de coleta e análise desses dados para a extração de informações, desenvolve-se a estatística, que consiste em um conjunto de técnicas para organizar, descrever, analisar e interpretar dados. Com o advento do computador, durante a Segunda Guerra Mundial, cálculos mais complexos se tornam possíveis, permitindo o desenvolvimento de modelos computacionais e matemáticos e viabilizando ao computador aprender com os dados, que mais tarde se tornaram a fundação para o *Machine Learning* e para a Inteligência Artificial como empregada atualmente (KELLEHER e TIERNEY, 2018).

O rápido crescimento da internet e a automatização por parte das empresas aumentou o volume de informações geradas e compartilhadas, o que faz com que praticamente todo sistema automatizado gere alguma forma de dados para análise ou diagnóstico. Isso é possível em grande parte pelo desenvolvimento da *Internet of Things*, que permite que equipamentos e sensores constantemente geram e transmitem tais dados a serem armazenados em grandes bancos de dados (BRAMER, 2016; AGGARWAL, 2015).

Essa expansão na variedade, geração e armazenamento de dados consome grande quantidade de recursos custosos, tanto humanos quanto técnicos. Com isso, cria-se a necessidade de melhores formas para gerenciá-los e armazená-los (BLUMBERG e ATRE, 2003). Tal necessidade culmina em tecnologias de *Big Data*, frequentemente definida em termos dos três V’s: Volume extremo de dados; Variedade de tipos de dados; e Velocidade de processamento dos dados (KELLEHER e TIERNEY, 2018).

Ainda em relação ao *Big Data*, Kelleher e Tierney (2018) expressam que:

> No entanto, embora essa flexibilidade representacional permita capturar e armazenar dados em vários formatos, esses dados ainda precisam ser extraídos em um formato estruturado antes que qualquer análise possa ser realizada sobre eles. A existência de *big data* também levou ao desenvolvimento de novas estruturas de processamento de dados.

Para atuar frente a essa vasta quantidade de dados que a ciência de dados se torna crucial. O termo "ciência" implica no ganho de conhecimento através de estudo sistemático, de modo a construí-lo e organizá-lo na forma de explicações e previsões testáveis. De modo geral, a ciência de dados envolve princípios, processos e técnicas para entender fenômenos através da análise automatizada de dados, ou seja, o estudo da extração generalizável do conhecimento por meio dos dados (PROVOST e FAWCETT; DAHR, 2013).

Kelleher e Tierney (2018) definem da seguinte maneira:

> A ciência de dados abrange um conjunto de princípios, definições de problemas, algoritmos e processos para extrair padrões úteis e não óbvios de grandes conjuntos de dados. Está intimamente relacionado aos campos de *Data Mining* e *Machine Learning*, mas de escopo mais amplo.

Tal definição se mostra de acordo com outros autores, por exemplo Bramer (2015), que a descreve como "[...] a extração de informações não-triviais implícitas, previamente desconhecidas e potencialmente úteis de dados, isto é, um processo no qual o *Data Minin*g é apenas uma parte, embora seja central".

Em seu trabalho "50 anos de Ciência de Dados", Donoho (2015) classifica as atividades da ciência de dados em seis áreas: Exploração e preparação de dados; Representação e transformação de dados; Computação com dados; Modelagem de dados; Visualização e apresentação de dados; e Ciência sobre Ciência de dados.

A exploração e preparação dos dados é vista como atividade crucial e que demanda o maior esforço por parte do cientista de dados. É necessário que as propriedades básicas dos dados sejam testadas, para expor falhas que comprometam a análise. Além disso, muitos *datasets* (conjuntos de dados) contém anomalias, tornando imprescindível o pré-processamento para identificar e lidar com tais problemas (MADIGAN *et al*, 2014; DONOHO, 2015).

Como as fontes de dados podem ser variadas, o cientista de dados deve ser capaz de reestruturar os dados originalmente fornecidos em uma forma nova e mais adequada, principalmente por meio de representações matemáticas úteis. Um exemplo disso é a transformação de dados acústicos em uma Transformada de Fourier (DONOHO, 2015).

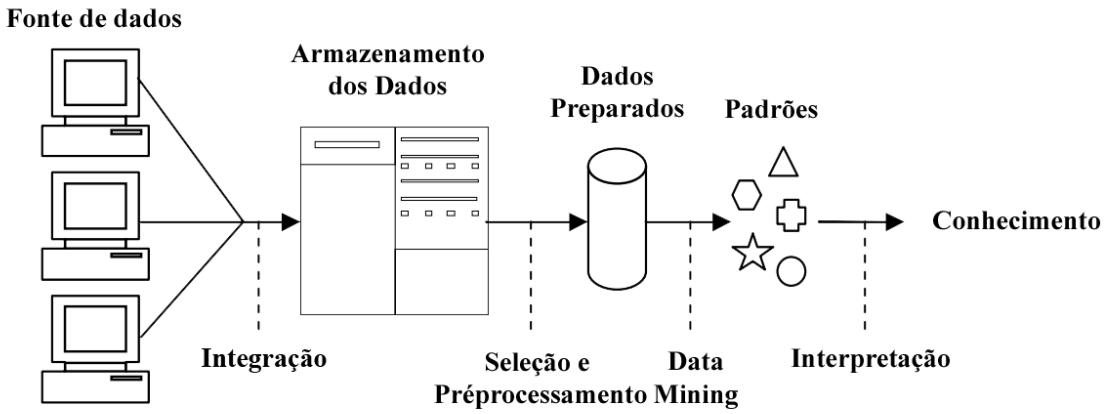
A computação com dados consiste na capacidade de utilizar uma amplitude de ferramentas computacionais, como diferentes linguagens e diferentes tipos de bancos de dados para análise, especialmente quando há necessidade de gerenciar sistemas complexos com linguagens para transformação e manipulação específicas (LEBANON e EL-GEISH, 2018).

A modelagem de dados apresenta tanto conceitos da estatística tradicional, como modelagem generativa, em que se propõe um modelo estocástico, cujo estado é indeterminado com origem em eventos aleatórios e deriva métodos para inferir propriedades, além de modelos preditivos, em que se desenvolve métodos para prever resultados dado o universo de dados, campo de estudo relacionado ao *Machine Learning* (DONOHO, 2015).

Métodos de apresentação dos dados devem levar em consideração o formato dos dados, o método de análise a ser utilizado e a informação que se busca enfatizar. Até quando os dados são os mesmos, a apresentação pode diferir quando se busca destacar algo diferente. Apresentar os dados de maneira inapropriada pode fazer com que os mesmos sejam interpretados de maneira incorreta (IN, LEE, 2017).

O *workflow* de projetos de ciência de dados, apresentado na figura 4, consiste em uma sequência de passos para processar dados de entrada sequencialmente a fim de resolver determinado problema (OJEDA *et al*, 2014).

Figura 4: *Workflow* de projetos de ciência de dados



Fonte: Adaptado de Bramer (2015)

Dados podem ser gerados automaticamente por equipamentos e serviços dotados de *IoT* (*Internet of Things*) ou de alguma outra ferramenta de coleta de dados (BRAMER, 2015). Esses dados são então armazenados, geralmente na nuvem, por meio de serviços de armazenamento como *Oracle*, *Hadoop* e *AWS*.

Os dados então devem ser selecionados e pré-processados. A partir de uma análise de dados exploratória, pode-se conhecer os dados que serão utilizados, verificar se há inconsistências ou problemas que possam afetar o processamento dos dados e a confiabilidade dos resultados. Geralmente, os dados não estão na forma necessária para a análise desejada, podendo apresentar anomalias, como parâmetros ausentes, valores duplicados, recursos irrelevantes etc. Para lidar com esses problemas os dados devem ser limpos a partir da identificação de erros, preenchimento de falhas nos dados, remoção de entradas duplicadas ou corrompidas. Essa etapa do *pipeline* normalmente é a que demanda mais tempo, e é de grande importância para o sucesso da análise (OJEDA *et al*, 2014).

Com posse dos dados preparados, utiliza-se um algoritmo de mineração de dados, que permite a exploração das relações estatísticas entre as variáveis nos dados, gerando saídas na forma de regras ou algum outro tipo de "padrão". Isto possibilita o uso de métodos de *Machine Learning* para categorizar, classificar ou clusterizar os dados e criar modelos preditivos (OJEDA *et al*, 2014; BRAMER, 2015).

A etapa final do *pipeline* da ciência de dados consiste na interpretação das informações extraídas dos dados. O objetivo desta etapa é identificar *insights* relevantes ao problema inicial e, em seguida, relacioná-lo com as descobertas dos dados (OJEDA *et al*, 2014). A participação de especialistas da área do problema em que a ciência de dados propõe a solução é de grande importância, porque esses especialistas podem auxiliar na visualização das descobertas, além de ajudar na comunicação dos resultados a um público não técnico (LIMA, 2009).

Em geral, os termos ciência de dados, *Machine Learning* e *Data Mining* são usados intercambiavelmente, tendo em vista que todas essas áreas têm como foco apoiar a tomada de decisão a partir da análise de dados. Enquanto o *Machine Learning* trata do *design* de algoritmos para a extração de padrões dos dados e o *Data Mining* geralmente lida com a análise de dados estruturados, a ciência de dados engloba conceitos provindos dessas áreas e também enfrenta outros desafios, tais como captura, limpeza e transformação de dados não estruturados da *web* e mídias sociais; armazenamento e processamento de grandes conjuntos de dados com *big data*, e questões relacionadas à ética e regulamentação dos dados (KELLEHER e TIERNEY, 2018).

Apesar de ser uma parte, o *Data Mining* pode ser considerado um dos conceitos mais próximos à ciência de dados, sendo visto como o processo concreto de extração de conhecimento a partir de dados, através de tecnologias que incorporam os princípios da ciência de dados (PROVOST e FAWCETT, 2013). Essa proximidade pode ser percebida por exemplo na conceituação de Hand (2007), que diz:

> O *Data Mining* moderna combina estatística com ideias, ferramentas e métodos de ciência da computação, aprendizado de máquina, tecnologia de banco de dados e outras tecnologias clássicas de análise de dados. Estatísticas clássicas e *Data Mining* diferem em vários aspectos. A estatística é uma disciplina desenvolvida originalmente ao longo do século XX, com base em conjuntos de dados relativamente pequenos. Os computadores só entraram em operação no final daquele século, permitindo a manipulação de conjuntos de dados maiores.

A distinção entre os termos supracitados ainda é fluida na literatura. Deste modo, nos próximos capítulos o conceito de *Data Mining* é melhor delineado, bem como as ferramentas estatísticas e computacionais que serão empregadas neste estudo.

2.4 *DATA* E *TEXT MINING*

Data Mining é o processo automático de vasculhar vastos conjuntos de dados (*datasets*), a fim de descobrir informações úteis. Devido ao tamanho dos *datasets*, técnicas de *Data Mining* se fazem necessárias para identificar relacionamentos e padrões nos dados investigados que de outra maneira continuariam desconhecidos (TAN *et al*, 2014).

Aggarwal (2015) divide o *Data Mining* em duas etapas: Pré-processamento de dados; e Processo analítico. Durante a fase de preparação dos dados, podem ser executadas tarefas de limpeza, redução e transformação de dados (HAN *et al*, 2011). Durante o processo analítico, técnicas de *Data Mining* são empregadas para projetar métodos analíticos eficazes nos dados processados (AGGARWAL, 2015).

Aplicações de *Data Mining* têm uma natureza muito personalizável, o que dificulta a utilização de técnicas gerais. Entretanto, o *Data Mining* apresenta quatro categorias comuns de aplicações: Classificação; Associação; Clusterização; e Detecção de *outliers* (AGGARWAL, 2015).

Aplicações de classificação estão entre as mais comuns no campo do *Data Mining*, onde pode-se classificar atributos de maneira preditiva. É válido ressaltar que alguns algoritmos de *Data Mining* envolvem métodos de *Machine Learning*, exemplificando o quão interligados são esses conceitos como parte da ciência de dados. Utiliza-se dados de treinamento para que o modelo aprenda e reconheça atributos selecionados nos dados, para então determinar a classificação desses atributos em dados não conhecidos. O objetivo desse tipo de modelo é minimizar o erro entre o valor previsto e o valor verdadeiro da variável analisada (AGGARWAL, 2015; TAN *et al*, 2014).

A análise de associação é utilizada para descobrir padrões que descrevem atributos fortemente associados aos dados. Como resultado desse tipo de análise, obtém-se normalmente regras de implicação ou subconjunto de atributos. Um uso comum da análise de associação ocorre em cestas de compras, onde, conhecendo as compras feitas por clientes de uma loja em um período de tempo suficiente, pode-se reconhecer relacionamentos que ajudarão a loja a comercializar produtos mais efetivamente (BRAMER, 2015).

Algoritmos de clusterização analisam os dados em busca de grupos de itens com características semelhantes, de maneira que observações de um mesmo *cluster* sejam mais próximas entre si do que as observações pertencentes a outro grupo (BRAMER, 2015; TAN *et al*, 2014). É válido ressaltar que o processo de clusterização pode ser utilizado no pré-processamento de dados para outras aplicações de *Data Mining* (AGGARWAL, 2015). A segmentação de clientes por determinado atributo, como idade, gasto médio, tipo de produto adquirido é um exemplo de clusterização.

O objetivo de um algoritmo de detecção de *outliers* é encontrar anomalias reais e evitar classificar falsamente uma normalidade como anomalia. Sistemas de detecção de *outliers* são muito utilizados em sistemas bancários para identificar possíveis fraudes em cartões de crédito (TAN *et al*, 2014).

Com frequência, *datasets* registram entradas com valores errôneos ou ausentes, devido a erros de medida, julgamento subjetivo da entrada registrada, mal uso ou funcionamento de equipamentos que registram os dados (BRAMER, 2015). De acordo com Han *et al* (2011) “As rotinas de limpeza de dados trabalham para "limpar" os dados preenchendo valores ausentes, suavizando dados ruidosos, identificando ou/e removendo outliers e resolvendo inconsistências”.

O campo do *Data mining* está em constante evolução, imensos avanços em tecnologias de *hardware* e *software* permitiram que dados de diversos formatos e tamanhos cada vez maiores fossem utilizados (AGGARWAL, 2012). A vasta quantidade e disponibilidade de dados em forma de texto impulsionou o desenvolvimento de algoritmos e métodos para sua análise e processamento (ALLAHYARI, 2017).

Text Mining, também conhecido como *Text Data Mining*, refere-se ao processo de extração de padrões interessantes e não-triviais de um conjunto de dados textuais, podendo ser considerada como uma extensão do *Data Mining* (TAN, 2000). *Text Mining* é um campo multidisciplinar, envolvendo áreas relacionadas à recuperação de informação, extração de informação e também *natural language processing* (NLP - processamento de linguagem

natural) em conjunto com algoritmos e métodos provenientes da ciência de dados, que envolvem *Data Mining*, *Machine Learnin*g e estatística (TAN, 2000; HOTHO, 2005).

O NLP é um subcampo da ciência da computação, inteligência artificial e linguística, e tem como objetivo o entendimento da linguagem natural por computadores (ALLAHYARI, 2017). Técnicas de NLP compreendem desde a simples manipulação de *strings* ao processamento automático de linguagem, e é utilizado extensivamente para *Text Mining* (HOTHO, 2005).

Com o crescimento exponencial de dados textuais online cria-se a necessidade por técnicas de mineração eficazes, pois aplicar *Text Mining* em um conjunto grande de documentos é uma tarefa complexa (AGGARWAL, 2012). Para lidar com documentos de texto, se faz necessário ter uma estrutura de dados que facilite sua análise. Apesar de diversos métodos tentarem explorar a estrutura sintática e semântica dos textos, em geral, aplicações de *Text Mining* representam os documentos na forma de *bag-of-words* (descrito com base no conjunto de palavras neles contidas) (HOTHO, 2002).

Para realizar a análise dos dados é necessário que o conjunto de dados seja bem conhecido pelo analista, para que possa escolher métodos e ferramentas ideais para o processamento posterior, com base na natureza dos dados (AGGARWAL, 2012). Para realizar o processamento dos dados para análise, os dados de entrada devem ser pré-processados, estando em um formato que facilite as tarefas analíticas (HOTHO, 2002). Para selecionar a melhor maneira de realizar o pré-processamento é necessário que os dados sejam familiares ao analista, isso é feito por meio da análise exploratória dos dados.

A análise exploratória dos dados (AED) não apresenta uma definição comum ou procedimentos definidos, sendo um processo extremamente contexto-específico, dependendo também dos detalhes da análise, assim, diferentes aplicações de AED podem ser empregadas (JEBB, 2016).

Jebb (2017) afirma que "AED é caracterizada pela extrema flexibilidade necessária para investigar a gama de fenômenos estatísticos e substantivos que emergem durante a pesquisa empírica" e descreve que as diversas perspectivas sobre AED podem ser compreendidas como "uma atitude analítica abrangente caracterizada como 'trabalho de detetive projetado para revelar a estrutura ou padrões nos dados (*apud* Haig, 2005; Tukey, 1980)'".

A visualização dos dados é de grande importância na análise exploratória dos dados, ferramentas e técnicas como tabelas, gráficos de dispersão, gráficos residuais, entre outras são

utilizadas para tentar compreender a estrutura (ou falta) dos dados (COX, 2017). De acordo com Kanevski & Maignan (2004):

> A maioria das técnicas de AED é de natureza gráfica, com poucas técnicas quantitativas. A razão para a forte dependência de gráficos é que, por sua natureza, o principal papel da AED é explorar, e os gráficos oferecem aos analistas um poder sem paralelo para fazê-lo.

O pré-processamento ocorre em conjunto com a análise exploratória, onde pode-se reconhecer possíveis problemas na qualidade dos dados e resolvê-los para que os resultados sejam confiáveis. Esse pré-processamento de dados tem impacto significante na performance de algoritmos de Data Mining em geral, uma vez que conjuntos de dados com ruídos (dados inconsistentes, *outliers*, fora do padrão) podem diminuir o desempenho dos algoritmos supracitados, fazendo se necessário o tratamento dos dados não-conformes (LIBRALON, 2007).

Além de ruídos, conjuntos de dados muitas vezes contêm atributos redundantes ou que não estão relacionados com a questão da análise. A remoção desses atributos reduz a dimensão dos dados, o que pode proporcionar ao algoritmo maior agilidade para o processamento (KOTSIANTIS *et al*, 2016).

Em relação ao *Text Mining*, a literatura apresenta concordância em relação às tarefas comuns de pré-processamento, sendo elas: remoção de *stop words*, *filtering*, *tokenization*, *lemmatization* e *stemming* (AGGARWAL, 2012).

Stop words são palavras que trazem pouco significado ao texto, tornando mais pesado e difícil de ser analisado (AGGARWAL, 2012). Dentre as palavras mais comuns em um texto em geral estão artigos, preposições e pronomes. Estas palavras são tratadas como *stop words* e removidas do documento durante o pré-processamento (VIJAYARANI, 2015).

Métodos de *filtering* removem palavras dos documentos, sendo a remoção de *stop words* uma das tarefas mais comuns. *Filtering* consiste em retirar dos documentos palavras que carregam pouco ou nenhuma informação, assim como remoção de palavras que aparecem raramente, apresentando pouca relevância estatística (HOTHO, 2002).

Tokenization consiste no processo de dividir uma sequência de textos em pedaços, podendo ser frases, palavras ou outro elemento significativo, chamados de *tokens* (ALLAHYARI, 2017). O objetivo da *tokenization* é a exploração dos elementos isoladamente, gerando uma lista de tokens utilizada para processamento posterior (KANNAN *et al*, 2014).

Métodos de *lemmatization* agrupam palavras em sua forma deflexionada, para determinar seu lema, por exemplo, as palavras "tive", “tenho” e “tem” são formas do mesmo lema “ter” (ALLAHYARI, 2017). Para realizar o processo de *lemmatization* deve-se primeiramente especificar a *part of speech* (processo de associar a cada palavra do texto uma classificação, como nome, verbo, adjetivo, etc., com base em sua definição e contexto). Entretanto, esse processo consome bastante tempo e está propenso a erros e, na prática, aplicam-se, em geral, métodos de *stemming* (HOTHO, 2002).

Métodos de *stemming* buscam obter a palavra raiz de diferentes formas da mesma palavra, por exemplo livro, livreiro, livreco apresentam a mesma raiz “livr” (AGGARWAL, 2012). Algoritmos de *stemming* são dependentes do idioma para que foram projetados, e o *stemmer* de Porter é o mais comumente usado na língua inglesa (ALLAHYARI, 2017). Após os dados serem pré-processados, podem ser alimentados pelo algoritmo de processamento com menor risco de ruídos ou inconsistências.

A fase de processamento em *Text Mining* consiste na aplicação de algoritmos sobre um conjunto de dados pré-processados, gerando informações que podem então ser analisadas pelo cientista de dados. Dentre as linguagens de programação mais utilizadas para análise de dados destacam-se *Python* e R. Para a execução desse trabalho a linguagem escolhida foi *Python*, devido a maturidade das bibliotecas disponíveis para a análise de dados, assim como a vasta disponibilidade de documentação sobre as ferramentas utilizadas para processos de *Text Mining*.

A seguir será apresentado o ferramental computacional para a aplicação de *Text Mining* que será desenvolvida.

2.4.1 FERRAMENTAL COMPUTACIONAL

Python é uma linguagem de programação interpretada, desenvolvida por Guido Van Rossum em 1991, e vem adquirindo destaque nas últimas duas décadas como uma ferramenta para análise e visualização de grandes conjuntos de dados (MCKINNEY, 2010). Apesar de não ter sido projetada especificamente para ciência de dados, a linguagem obteve grande popularidade devido à disponibilidade de bibliotecas poderosas para *Data Mining* e *Machine Learning* (MCKINNEY, 2012).

Dentre as bibliotecas utilizadas para o processamento de dados destacam-se *NumPy*, *SciPy*, *Pandas* e *Scikit-Learn*, ferramentas bem estabelecidas para análise e ricas em

documentação (MCKINNEY, 2012). Entre as bibliotecas específicas para *Text Mining*, a NLTK apresenta destaque, sendo utilizada para lidar com dados de linguagem humana, aplicados ao processamento estatístico de linguagem natural (LOPER e BIRD, 2002). Além da NLTK, a biblioteca *Metaknowledge* é utilizada para suporte para o processamento. Para visualização dos dados serão utilizadas as bibliotecas *Matplotlib* e *Seaborn*, devido a integração com as outras bibliotecas para análise e por apresentarem diversas funcionalidades para criação de gráficos.

De acordo com McKinney (2012) "*NumPy*, abreviação para *Numerical Python* é o pacote fundamental para computação científica em *Python*". A biblioteca *NumPy* traz suporte para *arrays* e matrizes multidimensionais, e possui uma vasta coleção de funções para o tratamento dessas estruturas, como funções matemáticas, lógicas, manipulação dimensional, operações estatísticas básicas, simulações aleatórias, entre outras (OLIPHANT, 2010).

Pandas é uma biblioteca que fornece ferramentas e estruturas de dados de alta performance, aumentando a velocidade e facilidade do processamento de dados, desenvolvendo um poderoso ambiente para análise de dados (MCKINNEY, 2012). Dentre as funcionalidades oferecidas pela biblioteca, destaca-se o *DataFrame*, principal objeto oferecido pela biblioteca, uma estrutura de dados bidimensional, tabular, onde os eixos são rotulados como linhas e colunas, orientada a coluna (MCKINNEY, 2012).

Scikit-Learn é uma biblioteca gratuita para *Python* que fornece diversos algoritmos e ferramentas eficientes para *Machine Learning*, como por exemplo *Support Vector Machine*, *Random Forest*, *k-neighbours* entre outros, além de oferecer suporte para outras bibliotecas para tratamento de dados, como *NumPy* e *SciPy* (VAROQUAUX *et al*, 2020). A lógica por trás dos algoritmos e modelos será apresentada na descrição do ferramental estatístico para o processamento de dados.

Pedregosa *et al* (2011), os criadores da biblioteca, a descrevem em seu artigo, da seguinte maneira:

> O *Scikit-learn* é um módulo *Python* que integra uma ampla gama de algoritmos de aprendizado de máquina de ponta para problemas supervisionados e não supervisionados de média escala. Este pacote se concentra em levar o aprendizado de máquina a não especialistas, usando uma linguagem de alto nível para uso geral.

A popularidade da biblioteca *Scikit-Learn* se dá principalmente devido à sua acessibilidade para não-experts e ao seu design simples e eficiente, além de estar munida de ferramentas para *Machine Learning* e modelagem estatísticas que incluem classificação, regressão, *clustering* e redução de dimensão (PEDREGOSA *et al*, 2011; BUITINCK *et al*, 2013).

Para o processamento de texto em Python foi empregada a biblioteca de NLTK, sigla para *Natural Language Toolkit*. NLTK, originalmente desenvolvida para o ensino, recebeu grande adoção por parte da indústria para pesquisa e desenvolvimento, apresentando-se como ferramenta útil e abrangente para processamento de linguagem natural e análise de texto (PERKINS, 2010). Os desenvolvedores da biblioteca, Bird e Loper (2002), a definem como:

> Conjunto de módulos de programas de código aberto, tutoriais e problemas, que oferecem material didático pronto para uso em linguística computacional. NLTK apresenta ferramentas para o processamento estatístico e simbólico da linguagem natural.

A biblioteca NLTK oferece funções importantes para o pré-processamento de dados de texto, com algoritmos que permitem o tratamento de *stop words*, *tokenization*, *lemmatization* e *stemming*, além de diversos métodos mais avançados para a manipulação dos dados (BIRD, LOPER e KLEIN, 2009).

Ainda para o processamento dos dados, a biblioteca *Metaknowledge* apresenta funcionalidades importantes para auxiliar no tratamento dos textos científicos. Os desenvolvedores da biblioteca McLevey e McIlroy-Young (2016) a descrevem da seguinte maneira:

> Metaknowledge é um pacote Python para pesquisa computacional em ciência da informação, análise de redes e ciência da ciência. Ele é otimizado para ser dimensionado com eficiência para analisar conjuntos de dados muito grandes e foi projetado para integrar-se bem a fluxos de trabalho de pesquisa reproduzíveis e abertos.

Entre as funcionalidades da biblioteca, destacam-se a facilidade de manipular dados de determinados repositórios científicos como *a Web of Science*, *Scopus* e *Pubmed*, além de ferramentas para análise quantitativa, computacional e de *network* (rede) de autores (MCLEVEY e MCILROY-YOUNG, 2016).

Para a visualização dos dados, as bibliotecas *Matplotlib* e *Seaborn* foram utilizadas. De acordo com Barret *et al* (2005) "*Matplotlib* é um pacote de plotagem e imagem 2D portátil destinado principalmente à visualização de dados científicos, de engenharia e financeiros". A biblioteca permite a criação de gráficos de maneira simples, customizável e apresenta compatibilidade com as bibliotecas apresentadas anteriormente.

Seaborn é uma biblioteca para a criação de gráficos estatísticos em *Python*, construída com base no M*atplotlib* e integrada à estruturas de dados da biblioteca *Pandas*, oferecendo uma interface de alto nível, com temas padrão atrativos e customizáveis (WASKOM, 2013).

3. RESULTADOS

O presente trabalho foi realizado utilizando a plataforma *Jupyter Notebook*, software de código aberto, padrões abertos e serviços para o desenvolvimento de computação interativa em dezenas de linguagens de programação. A linguagem de programação utilizada foi o *Python*.

Os dados foram baixados através da plataforma *WoS*, em arquivos no formato de texto *txt*, e em sua totalidade apresentavam 13820 entradas, na forma de artigos científicos com a temática "Realidade Aumentada".

Inicialmente, foram importadas as bibliotecas de *Python* que foram utilizadas para o trabalho, as mesmas estão apresentadas na figura 5.

Figura 5: Bibliotecas utilizadas

```
In [1]: # Bibliotecas para manipulação e análise
        import metaknowledge as mk
        import networkx as nx
        import pandas as pd
        import numpy as np
        from collections import Counter
        import re

        import nltk
        nltk.download('words')

        # Bibliotecas para visualização
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
        import metaknowledge.contour.plotting as mkv
        from wordcloud import WordCloud

        import warnings
        warnings.filterwarnings("ignore")

        [nltk_data] Downloading package words to
        [nltk_data]     C:\Users\Usuario\AppData\Roaming\nltk_data...
        [nltk_data]   Package words is already up-to-date!

Out[1]: True
```

Fonte: O autor

A partir da importação, foi utilizada a função *RecordCollection*() da biblioteca *Metaknowledge* para transformar os dados (que foram baixados como arquivos de texto .txt) em dados do tipo *RecordCollection*. Essa transformação foi empregada para que a base de dados possa ser mais facilmente manipulada, apresentando os dados em uma forma de tabela, com

colunas representando cada atributo e posteriormente transformada em um *DataFrame*, formato mais adequado para a utilização dos algoritmos de *Machine Learning*.

Os atributos contidos na base de dados são: *Pub Type* (tipo de publicação), *Author* (autor), *Author Full* (nome completo do autor), *Title* (título), *Publication Name* (título da publicação), *Language* (idioma), *Author Keys* (palavras-chave do autor), *WoS Keys* (palavras-chave da *WoS*), *Cit Count* (contagem de citações), *Publisher* (editora), *WoS Categories* (categorias da *WoS*), *Research Areas* (área de pesquisa), *Address* (endereço), *Year* (ano), e *Abstract*. A figura 6 apresenta um recorte da base de dados após a transformação em *DataFrame.*

Figura 6: Recorte da base de dados para ilustração

	Author Full	Title	Publication Name	Language	Author Keys	Cit Count	WoS Categories	Research Areas	Address	Year	Abstract
0	[Chang, Hsin-Yi, Yu, Yuan-Tse, Wu, Hsin-Kai, H...	The Impact of a Mobile Augmented Reality Game ...	2016 IEEE 16TH INTERNATIONAL CONFERENCE ON ADV...	English	[mobile game, augmented reality, socioscientif...	2	[Computer Science, Interdisciplinary Applicati...	[Computer Science, Education & Educational Res...	[Natl Taiwan Univ Sci & Technol, Grad Inst Dig...	2016	in this paper we introduce our newly developed...
1	[Kitajima, Muneo, Shimizu, Shono, Nakahira, Ka...	Creating Memorable Experiences in Virtual Real...	2017 3RD IEEE INTERNATIONAL CONFERENCE ON CYBE...	English	[memorable experience, virtual reality, eye mo...	0	[Computer Science, Cybernetics]	[Computer Science]	[Nagaoka Univ Technol, Informat & Management S...	2017	reality is the true situation which consists o...
3	[Chiariotti, Federico, Kucera, Stepan, Zanella...	Analysis and Design of a Latency Control Proto...	IEEE-ACM TRANSACTIONS ON NETWORKING	English	[transport protocols, multipath channels, qual...	3	[Computer Science, Hardware & Architecture, Co...	[Computer Science, Engineering, Telecommunicat...	[Univ Padua, Dept Informat Engn, I-35131 Padua...	2019	as the capacity and reliability of mobile netw...
4	[Lam, Chee Kiang, Sundaraj, Kenneth, Sulaiman,...	Virtual phacoemulsification surgical simulatio...	BMC OPHTHALMOLOGY	English	[cataract surgery, performance assessment, sim...	2	[Ophthalmology]	[Ophthalmology]	[Univ Malaysia Perlis, Sch Mechatron Engn, Kam...	2016	background computer based surgical training is...
5	[Clemente, Francesco, Dosen, Strahinja, Lonini...	Humans Can Integrate Augmented Reality Feedbac...	IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS	English	[augmented reality (ar), motor learning, senso...	12	[Computer Science, Artificial Intelligence, Co...	[Computer Science]	[Scuola Super Sant Anna, BioRobot Inst, I-5602...	2017	tactile feedback is pivotal for grasping and m...

Fonte: O autor

Com posse dos dados, foram realizados procedimentos de limpeza de dados, para evitar que inconsistências fossem propagadas para os resultados. No processo de limpeza foram removidas inicialmente entradas que apresentavam valores nulos para o atributo *Year*, *Author Full* e *Address* (figura 7), pois se fossem mantidos valores inválidos, o algoritmo estaria propenso a erros. Em sequência, foram preenchidas todas as entradas inválidas de *WoS Categories* e *Research Areas* com "*none*" (nenhum). Essas entradas não foram removidas porque indicam que essas informações não foram preenchidas na plataforma *WoS* e sua ausência não geraria impactos significativos na análise.

Figura 7: Limpeza dos dados

```
In [23]: # Remover linhas com entradas problemáticas
         # Year como ''
         base_full.drop(base_full[base_full['Year'] == ''].index, inplace=True)

         # Author como NaN
         base_full.drop(base_full[base_full['Author Full'].isnull()].index, inplace=True)

         # Address como NaN
         base_full.drop(base_full[base_full['Address'].isnull()].index, inplace=True)

         # Preencher WoS Categories como 'None'
         base_full['WoS Categories'] = base_full['WoS Categories'].fillna('None')

         # Preencher Research Areas como 'None'
         base_full['Research Areas'] = base_full['Research Areas'].fillna('None')
```

Fonte: O autor

Ainda durante a limpeza dos dados foram identificados artigos que apresentavam o *Abstract* em inglês e alemão simultaneamente, entretanto, por ser um número muito pequeno de entradas não se fez necessário a remoção das mesmas. Após a limpeza, foram removidas 406 entradas, de maneira que a base de dados ficou com 13444 entradas.

Na última etapa da limpeza, a base de dados foi dividida entre entradas do Brasil e entradas internacionais. Para realizar essa separação foi necessário identificar quais entradas continham "*Brazil*" no atributo "*Address"*. Nessa etapa foi constatou-se que todos os artigos com *"Language = Portuguese"* estavam contidos entre os artigos com o "*Address = Brazil"*, entretanto, o inverso não era verdadeiro. Após a divisão da base de dados, foi percebido que artigos brasileiros representavam cerca de 2% do total de entradas (figura 8), valor significativo considerando que a plataforma possui artigos de diversos países do mundo, sendo majoritariamente artigos norte-americanos.

Figura 8: Separação das Bases

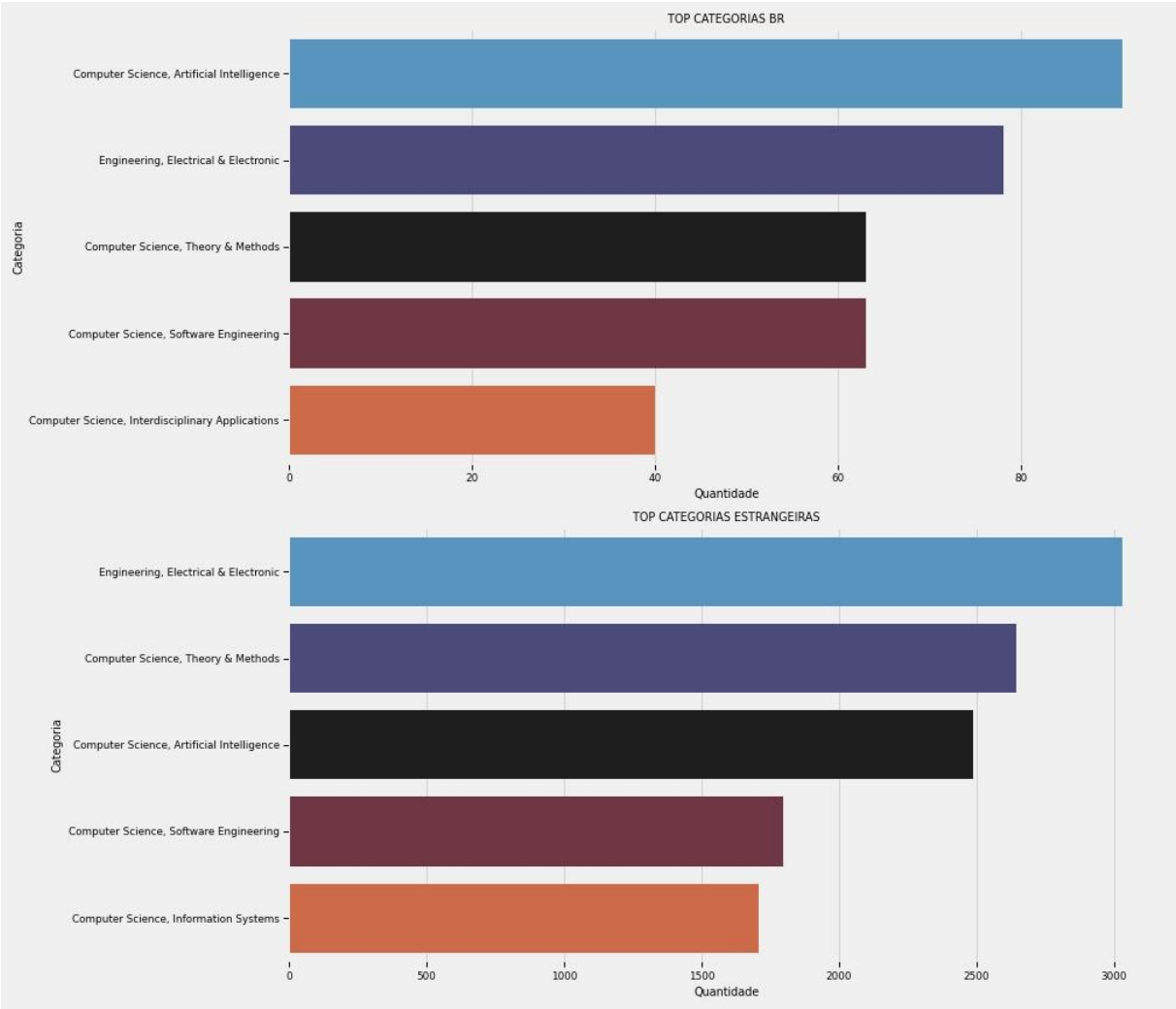
```
In [31]: # Separando as bases
         base_br = base_full.loc[base_full['BR'] == 1]
         base_ext = base_full.loc[base_full['BR'] == 0]

         print('Base original: ', len(base_full))
         print('Base BR: ', len(base_br))
         print('Base Exterior: ', len(base_ext))

         print('Proporção de entradas BR: ', len(base_br)/len(base_full), '%')

         Base original:  13444
         Base BR:  288
         Base Exterior:  13156
         Proporção de entradas BR:  0.021422195775066943 %
```

Fonte: O autor

Após a limpeza dos dados, se iniciou a etapa da análise exploratória dos dados, onde são verificadas características presentes nos dados, como também comparações entre as informações levantadas sobre os dados do Brasil e do Exterior.

A partir da EAD foram comparadas as áreas e categorias das publicações relacionadas ao tema de “Realidade Aumentada”, e utilizadas as bibliotecas *Seaborn* e *Matplotlib* para gerar gráficos exibindo tanto categorias quanto áreas de pesquisas mais comuns dentre os dados analisados. Os resultados obtidos foram apresentados nas figuras 9 e 10.

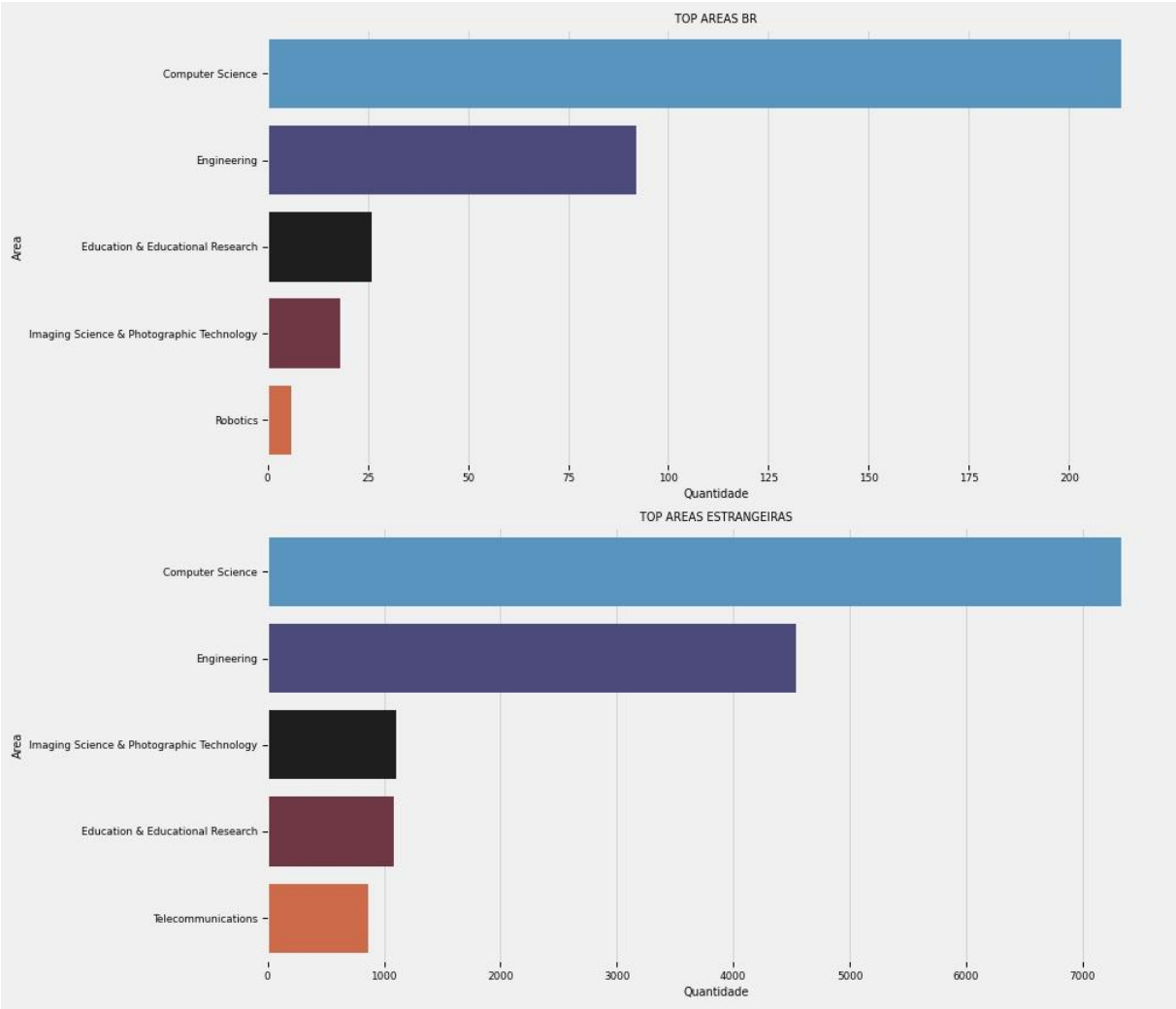
Figura 9: Top categorias Brasil e Exterior



Fonte: O autor

Foi possível perceber a partir do gráfico apresentado na figura 9 a grande semelhança entre as categorias mais comuns das publicações, onde, tanto no Brasil quanto no Exterior, ciências da computação e engenharia figuram como as categorias mais populares. Esse resultado não foge ao esperado, e apresenta indícios sobre a proximidade entre as publicações.

Figura 10: Top áreas de pesquisas Brasil e Exterior

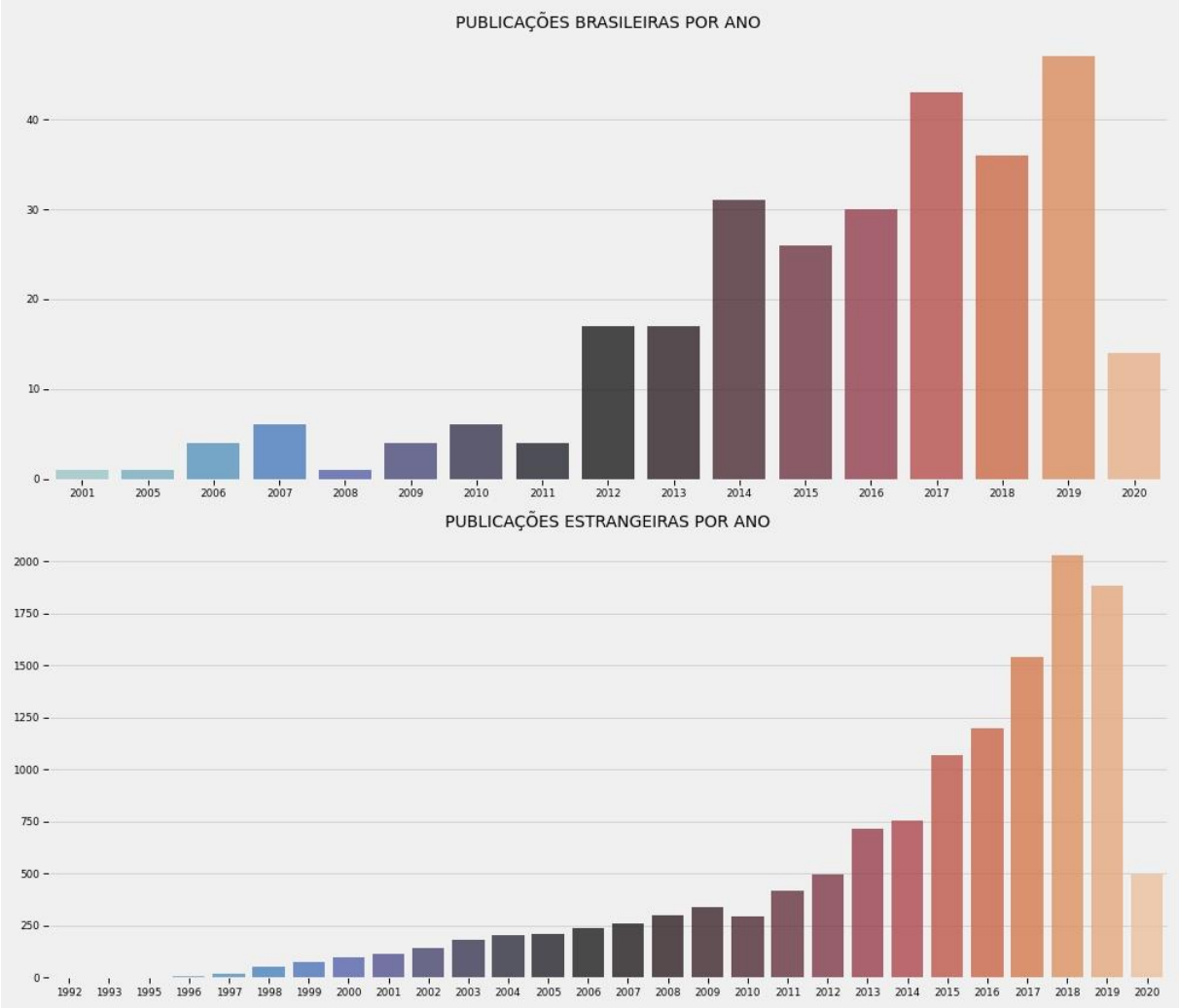


Fonte: O autor

Em relação às áreas também foi possível verificar grande semelhança, sendo também ciência da computação e engenharia as áreas mais populares. A diferença se encontra na quinta área mais popular, onde em publicações brasileiras figura a robótica enquanto nas publicações estrangeiras a telecomunicação apresentou maior destaque. Essa diferença pode ser explicada devido ao crescimento da adoção de robôs no parque industrial do Brasil, que atualmente lidera o parque de robôs industriais na América do Sul, entretanto, o parque robótico brasileiro está muito longe de países desenvolvidos. Carmen Nery aponta em publicação da revista Valor Econômico que “O Brasil tem de 12 a 13 robôs a cada 10 mil trabalhadores, ante 1,3 mil dos EUA, 938 da China, 1,2 mil do Japão e 2,7 mil da Coreia”. Ao passo que países desenvolvidos estão na linha de frente de tecnologias revolucionárias na telecomunicação como o 5G.

Também foi realizada uma análise sobre as publicações por ano, com objetivo de compreender a tendência de publicação através do tempo. A partir dessa análise foi possível notar uma diferença entre o comportamento das publicações, como apresentado na figura 11.

Figura 11: Publicações por ano - Brasil e Exterior



Fonte: O autor

A figura 11 representa a quantidade de artigos internacionais publicados com a temática “Realidade Aumentada” na plataforma *WoS*, onde é notável o comportamento crescente das publicações através do tempo. A queda brusca nas publicações em 2020 foi devido a geração do banco de dados, em março do mesmo ano. O recorte temporal adotado pelo trabalho foi de 30 anos, partindo de 1990, onde o número de publicações ainda era inexpressivo, porém diferente de zero.

A justificativa desse recorte se dá porque apesar da popularidade das pesquisas sobre Realidade Aumentada datar mais expressivamente do início do século XX, desde os anos 90 o tema já era abordado no meio tecnológico. Isso deve-se em grande parte a publicações de P.

Milgram em 1994 e R. Azuma em 1997 e 2001, que figuram como os dois autores mais citados de todo o conjunto de dados.

Já o gráfico para publicações brasileiras por ano (figura 11) difere-se do internacional. A primeira característica observada foi o início da linha temporal em 2001, e uma lacuna entre 2001 e 2005, ocasionada pela ausência de publicações durante esse período. Além disso, diferentemente do que ocorre no exterior, no período entre 2001 e 2011 a tendência do gráfico não era crescente, oscilando entre valores, com 2007 apresentando o maior número de publicações.

A partir de 2011 o gráfico já apresenta uma tendência crescente, apesar de algumas oscilações observadas com a queda na quantidade de publicações em alguns anos em relação ao ano anterior, seguidos prontamente de um aumento no ano seguinte.

Foram comparadas também as palavras chaves para os conjuntos de publicações do Brasil e do Exterior. Essa análise apresentou grandes semelhanças entre as palavras chaves mais comuns em ambos conjuntos de dados, como “educação”, “rastreamento” e “dispositivos móveis”, entretanto, algumas diferenças foram encontradas, entre elas, destacaram-se “jogos”, “usabilidade” e “simulação” presentes na lista de palavras chaves brasileiras mais recorrentes. É válido ressaltar que para ambos conjuntos de dados, entre os resultados mais comuns estavam entradas vazias, entretanto, as mesmas foram removidas, pois apenas indicavam que não haviam sido registradas na plataforma *WoS*. As palavras chaves mais comuns são apresentadas na figura 12.

Figura 12: Palavras-chave mais comuns

```
In [33]: # As areas de pesquisa estão em listas dentro da coluna
base_br['Author Keys'] = base_br['Author Keys'].fillna('None')

keys = []
for ind in base_br.index:
    keys.append(base_br['Author Keys'][ind])

# Flatlist a coluna e achar as entradas mais comuns
x = [item for sublist in keys for item in sublist] # Flatlisting

keys_br = Counter(x)

top_br_keys = pd.DataFrame(keys_br.most_common(14),columns=['Keyword','Quantidade'])

# Remover topwords que estão como "none ou NENHUM"
rem = ['N','E','N','H','U','M', 'n', 'o', 'n', 'e']
for i in rem:
    top_br_keys.drop(top_br_keys[top_br_keys['Keyword']==i].index, inplace=True)
```

```
In [34]: # As areas de pesquisa estão em listas dentro da coluna
base_ext['Author Keys'] = base_ext['Author Keys'].fillna('None')

keys = []
for ind in base_ext.index:
    keys.append(base_ext['Author Keys'][ind])

# Flatlist a coluna e achar as entradas mais comuns
x = [item for sublist in keys for item in sublist] # Flatlisting

keys_ext = Counter(x)

top_ext_keys = pd.DataFrame(keys_ext.most_common(14),columns=['Keyword','Quantidade'])

# Remover topwords que estão como "none ou NENHUM"
rem = ['N','E','N','H','U','M', 'n', 'o', 'n', 'e']
for i in rem:
    top_ext_keys.drop(top_ext_keys[top_ext_keys['Keyword']==i].index, inplace=True)
```

```
TOP KEYWORDS BRASIL
                Keyword  Quantidade
0   augmented reality         161
5     virtual reality          23
6       mixed reality          11
7           education           9
8       serious games           8
9           usability           8
10    computer vision           6
11     mobile devices           6
12         simulation           5
13           tracking           5
----------------------------------------
TOP KEYWORDS EXTERIOR
                       Keyword  Quantidade
0          augmented reality        5602
5            virtual reality        1031
6              mixed reality         342
7                  education         206
8     augmented reality (ar)         194
9   mobile augmented reality         175
10             visualization         175
11           computer vision         166
12                        ar         160
13                  tracking         133
```

Fonte: O autor

Em relação às citações presentes nos textos, foi calculado a média de citações para as publicações brasileiras e internacionais, dividindo o total de citações pela quantidade de documentos presentes em cada grupo. Como resultado, foi observado que a média de citações em publicações brasileiras era de aproximadamente 2,73 enquanto em publicações internacionais era de 7,73. É interessante notar que o número de citações em publicações do exterior é quase três vezes maior, entretanto, parte desse resultado pode ser explicado pela quantidade expressivamente maior de documentos estrangeiros em relação aos brasileiros. A figura 13 abaixo apresenta o cálculo utilizado.

Figura 13: Média de citações por documento

```
print('Média de citações em publicações brasileiras: {:.3}'.format(base_br['Cit Count'].sum()/len(base_br)))
Média de citações em publicações brasileiras: 2.73

print('Média de citações em publicações internacionais: {:.3}'.format(base_ext['Cit Count'].sum()/len(base_ext)))
Média de citações em publicações internacionais: 7.73
```

Fonte: O autor

É de grande importância considerar que apesar dos dados poderem ser manipulados de forma adequada, sua natureza impôs problemas quanto a utilização de algoritmos de *Machine Learning* com exceção de métodos de *Natural Language Processing*. Essa característica foi percebida ao realizar a análise exploratória dos dados, e deu-se principalmente devido à porcentagem baixa de publicações brasileiras em relação ao total, fez com que a utilização de algoritmos tradicionais de *Machine Learning* como de classificação, clusterização e regressão não pudessem ser utilizados de maneira satisfatória, pois não havia dados suficientes para os algoritmos fazerem a distinção entre eles.

Para análise das palavras e expressões mais frequentes no texto dos dados com aplicação da biblioteca *Scikit Learn*, fez-se necessário um passo extra para a adequação dos dados. Inicialmente, utilizando a biblioteca *Regular Expressions* (re). Em seguida todos os textos contidos no *Abstract* foram transformados em uma única *string*, ou seja, todos os textos dos dados brasileiros foram unidos em uma única coluna, formando um único texto. O processo foi repetido também para os dados do exterior e ambos bancos de dados foram unidos, como apresentado na figura 14.

Figura 14: Limpeza do texto

```
In [45]: # Remover pontuação, acentuação, simbolos e palavras com numeros
         def clean_text_round1(text):
             '''Make text lowercase, remove text in square brackets, remove punctuation and remove words containing numbers.'''
             text = text.lower()
             text = re.sub('\[.*?\]', '', text)
             text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
             text = re.sub('\w*\d\w*', '', text)
             text = re.sub('[‘’“”…]', '', text)
             text = re.sub('\n', '', text)
             return text

         round1 = lambda x: clean_text_round1(x)

         data_clean = pd.DataFrame(base['Abstract'].apply(round1))
         data_clean
```

Out[45]:

	Abstract
Brasil	currently due to ease of acquisition of mobile...
Exterior	the quality of atmospheric corrections provide...

Fonte: O autor

Além disso, foram removidas dos textos todas as *stop words*, palavras como “*the*”, “*that*”, “*this*”, “*of*”, entre outras, que não agregam significado para o algoritmo, porém aparecem com muita frequência, utilizando a biblioteca *NLTK*, que provê uma lista pré-determinada de *stop words* na língua inglesa (figura 15).

Figura 15: Remoção de *Stop Words* e criação da DTM

```
In [67]: # Remover stopwords e criar Document Term Matrix (DTM)

         from sklearn.feature_extraction.text import CountVectorizer

         cv = CountVectorizer(stop_words='english')
         data_cv = cv.fit_transform(data_clean['Abstract'])
         data_dtm = pd.DataFrame(data_cv.toarray(), columns=cv.get_feature_names())
         data_dtm.index = data_clean.index
         data_dtm
```

Fonte: O autor

Utilizando a biblioteca *NLTK* para fornecer as *stop words*, foi utilizado o método de *Machine Learning* disponível na biblioteca *Scikit Learn* chamado *CountVectorizer.fit_transform()* que permite ao algoritmo aprender o vocabulário presente no texto (treinar o algoritmo) e converter esses dados em uma *Document-Term Matrix* (DTM), que é uma matriz matemática que descreve a frequência dos termos que ocorrem em uma coleção de documentos.

A *Document-Term Matrix* gerada consiste em uma matriz onde as linhas representam a origem dos textos (Brasil e Exterior) e as colunas representam cada palavra contida no texto. A DTM foi transposta para facilitar a sua visualização (figura 16).

Figura 16: Recorte da Document-Term Matrix

Out[97]:

	Brasil	Exterior
aberrant	0	4
aberration	0	31
aberrationfree	0	1
aberrations	0	18
abhorrence	0	1
abi	0	4
abides	0	2
abilities	5	162
abilitiy	0	1
ability	14	734
abilitytosee	0	2
ablating	0	1
ablation	0	48
ablations	0	2
able	15	925

Fonte: O autor

Com posse da DTM, foram verificadas as palavras mais frequentes para ambos os conjuntos de textos, e foi percebido que algumas palavras e expressões que apareciam com frequência não agregam significado para a análise, como por exemplo “este trabalho apresenta”, nomes de editoras, ou expressões óbvias como “realidade aumentada” que consta em todos os dados. Logo, foi realizada uma nova limpeza nos dados e removidas essas palavras (figura 17) é então gerada uma nova DTM.

Figura 17: Remoção adicional de palavras

```
In [50]: add_stop_words = ['reality', 'augmented', 'ar', 'virtual', 'use', 'user',
                           'used', 'users', 'paper','based', 'present', 'using',
                          'rights', 'reserved', 'elsevier', 'd', 'c']

         from sklearn.feature_extraction import text
         from sklearn.feature_extraction.text import CountVectorizer

         stop_words = text.ENGLISH_STOP_WORDS.union(add_stop_words)

         cv = CountVectorizer(stop_words=stop_words)
         data_cv = cv.fit_transform(data_clean.Abstract)
         data_stop = pd.DataFrame(data_cv.toarray(), columns=cv.get_feature_names())
         data_stop.index = data_clean.index
```

Fonte: O autor

Desse modo, foram retiradas essas palavras dos textos e aplicado o algoritmo para gerar as *n-grams* (n palavras ou expressões que aparecem juntas no conjunto de dados de texto). Os *n-grams* analisados foram de graus 1, 2 e 3 (unigramas, bigramas e trigramas respectivamente). Para melhor compreensão dos *n-grams*, considere o texto “a temática do trabalho é realidade aumentada”, os unigramas desse texto seriam “a”, “temática”, “do”, “trabalho”, “é”, “realidade”, “aumentada”, já os bigramas seriam “a temática”, “temática do”, “do trabalho”, “trabalho é”, “é realidade”, e “realidade aumentada”, de tal maneira que os trigramas seriam “a temática do”, “temática do trabalho”. “do trabalho é”, “trabalho é realidade”, “é realidade aumentada”.

Para a geração dos unigramas, bigramas e trigramas foi utilizada a biblioteca *Regular Expressions*, pois a mesma apresentou o melhor tempo de processamento quando comparado com o método de *Machine Learning* disponível na biblioteca *Scikit Learn.*

Figura 18: Unigrama - Brasil e Exterior

```
In [100]: # UNIGRAM N = 1
for ind in data_clean.index:
    unigram_counts = Counter(ngrams(data_clean.Abstract[ind].split(), 1))
    display(ind, unigram_counts.most_common(10))
```

```
'Brasil'

[(('applications',), 178),
 (('application',), 141),
 (('results',), 141),
 (('information',), 124),
 (('real',), 122),
 (('environment',), 114),
 (('work',), 109),
 (('learning',), 109),
 (('technology',), 106),
 (('mobile',), 105)]

'Exterior'

[(('information',), 5965),
 (('learning',), 5644),
 (('applications',), 5433),
 (('technology',), 5329),
 (('results',), 5323),
 (('mobile',), 5287),
 (('real',), 5276),
 (('data',), 4785),
 (('study',), 4726),
 (('method',), 4444)]
```

Fonte: O autor

Figura 19: Bigrama - Brasil e Exterior

```
In [101]: # BIGRAM N = 2
          for ind in data_clean.index:
              unigram_counts = Counter(ngrams(data_clean.Abstract[ind].split(), 2))
              display(ind, unigram_counts.most_common(10))

'Brasil'

[(('mobile', 'devices'), 34),
 (('real', 'world'), 30),
 (('real', 'time'), 19),
 (('work', 'presents'), 16),
 (('learning', 'process'), 16),
 (('computer', 'vision'), 15),
 (('case', 'study'), 14),
 (('teaching', 'learning'), 13),
 (('mobile', 'device'), 13),
 (('mobile', 'application'), 12)]

'Exterior'

[(('real', 'world'), 1274),
 (('mobile', 'devices'), 866),
 (('real', 'time'), 713),
 (('experimental', 'results'), 619),
 (('proposed', 'method'), 545),
 (('computer', 'vision'), 413),
 (('pose', 'estimation'), 390),
 (('cultural', 'heritage'), 363),
 (('real', 'objects'), 342),
 (('mobile', 'application'), 341)]
```

Fonte: O autor

Figura 20: Trigrama - Brasil e Exterior

```
In [104]: #TRIGRAM N= 3
          for ind in data_clean.index:
              unigram_counts = Counter(ngrams(data_clean.Abstract[ind].split(), 3))
              display(ind, unigram_counts.most_common(10))

'Brasil'

[(('information', 'communication', 'technologies'), 8),
 (('image', 'processing', 'techniques'), 6),
 (('locationbased', 'mobile', 'games'), 6),
 (('eye', 'center', 'location'), 5),
 (('solid', 'geometry', 'concepts'), 5),
 (('head', 'mounted', 'displays'), 4),
 (('computer', 'vision', 'techniques'), 4),
 (('teaching', 'learning', 'process'), 4),
 (('pervasive', 'mobile', 'games'), 4),
 (('focus', 'plus', 'context'), 4)]

'Exterior'

[(('head', 'mounted', 'display'), 155),
 (('authors', 'published', 'bv'), 138),
 (('internet', 'things', 'iot'), 102),
 (('headmounted', 'display', 'hmd'), 98),
 (('optical', 'society', 'america'), 95),
 (('minimally', 'invasive', 'surgery'), 87),
 (('real', 'world', 'objects'), 84),
 (('simultaneous', 'localization', 'mapping'), 82),
 (('experimental', 'results', 'demonstrate'), 80),
 (('objects', 'real', 'world'), 78)]
```

Fonte: O autor

Pode-se perceber a grande semelhança entre os tops *n-grams* em publicações do Brasil e Exterior tanto para n =1 quanto para n = 2, com vários elementos presentes em ambas as

origens. Para n = 1 as listas se assemelham fortemente, já para n = 2 existem algumas diferenças entre os *n-grams* mais comuns como "*cultural heritage*" em publicações estrangeiras, e "*case study*" e "*learning process*" para o Brasil.

As maiores diferenças se encontram em *n-grams* com n = 3, onde pode-se perceber duas das expressões mais populares relacionadas a jogos para publicações brasileiras ("*location-based mobile games* " e "*pervasive mobile games*"). Ao mesmo tempo, em publicações estrangeiras pode-se notar "*internet things iot*", "*minimally invasive surgery*", que podem indicar a tendência de publicações voltadas à indústria e à área médica respectivamente.

4. CONCLUSÃO

A partir do desenvolvimento da presente pesquisa foi possível concluir que a produção científica brasileira sobre o tema “Realidade Aumentada” apresenta grande semelhança à produção internacional, principalmente em relação às áreas pesquisadas. Tal conclusão não foge do esperado, dado que o tema é fortemente atrelado à tecnologia em desenvolvimento, é esperado grande destaque para as áreas de engenharia e ciência da computação.

Entretanto, é interessante perceber a diferença entre a produção científica por ano, enquanto no exterior o tema apresenta uma taxa de crescimento bem definida e ao longo de todo o recorte temporal, para o Brasil, não existem registros de publicações no *WoS* antes dos anos 2000. O que pode ser refletido em um atraso na adoção e desenvolvimento de tecnologias de Realidade Aumentada no país.

Ainda, foi possível encontrar grande variação no número de publicações brasileiras a partir de 2001. Durante uma década o país não apresentou uma taxa estável de crescimento das publicações, ou seja, por um período de 20 anos, o interesse na área se apresentou como inexistente ou baixíssimo. Todavia, o ano de 2012 apresenta um salto no crescimento das publicações se comparado à década anterior. Desse ponto em frente, o número de publicações foi crescente, se assemelhando a produção internacional.

Essa explosão na produção pode ser atribuída ao desenvolvimento de *smartphones* cada vez mais potentes, e capazes de suportar tecnologias de RA. É possível visualizar essa relação quando se analisa as *n-grams* de grau 1 e 2, onde encontram-se em destaque “*mobile*”, “*applications*”, “*mobile application*”, “*mobile device*”, “*mobile devices*”, expressões que se refere a dispositivos móveis como *smartphones*.

Além disso, nos *n-grams* brasileiros o termo “*game*” (jogo) se apresentou com frequência, indicando o interesse de publicações nacionais envolvendo jogos e Realidade Aumentada, principalmente a partir da segunda metade da década de 2010, onde houve a popularização de jogos com componentes de RA, com grande destaque ao título *Pokemon Go*, lançado em 2016, atingindo a marca de mais de um bilhão de *downloads* em 2019, e com uma arrecadação estimada em torno de 6 bilhões de dólares desde seu lançamento.

Outro fator que se sobressai é a presença de “*case study*” (estudo de caso) entre os *n-grams* para publicações brasileiras. O atraso na popularização do tema, visto no gráfico temporal, pode indicar que o Brasil teve grande foco no estudo de casos anteriores, quando a tecnologia ainda não era popularizada, ao passo que publicações do exterior apresentavam

expressões relacionadas às áreas industriais e médicas, como "*iot*" e "*surgery*" respectivamente, indicando novas aplicações potenciais.

Ainda é válido ressaltar a presença de *head mounted display* tanto para publicações do Brasil quanto do Exterior, o que pode indicar que *HMDs* ainda são dispositivos muito utilizados em tecnologias RA, entretanto, pode-se inferir que dispositivos móveis (*mobile*) se mostram como a tendência atual, devido sua presença em *keywords* e *n grams*.

Em relação aos objetivos propostos na formulação do presente trabalho, foi possível comparar a produção científica brasileira e internacional sobre a temática e levantar informações relevantes sobre a produção científica sobre Realidade Aumentada no Brasil, percebendo a grande semelhança em relação ao exterior sobre vocabulário, áreas de pesquisa e categorias da produção, ainda apontando diferenças quanto ao interesse em áreas específicas de pesquisa. Também foi possível compreender que apesar da diferença de tempo na adoção do tema, a produção científica do país aponta crescimento, acompanhando a tendência internacional. Já analisando as citações, nota-se que a média de citações em artigos internacionais é quase três vezes maior que a média nacional, fator que também pode indicar a maior evolução da temática internacionalmente.

As maiores diferenças podem ser apontadas nas análises dos *n grams*, onde a produção científica brasileira aponta para o foco no desenvolvimento de jogos, estudos de caso e educação, enquanto a produção científica internacional, apesar de também apresentar relação com a educação, também aponta forte relação com a área médica, onde tecnologias RA vem se popularizando e também para indústria moderna.

A partir da análise, surgem propostas de trabalho futuras que podem ser desenvolvidas a partir dos resultados obtidos. Com posse das informações reconhecidas, é possível realizar a divisão dos dados em períodos de tempo, e realizar a comparação de modo a compreender as tendências tecnológicas de cada período de tempo no exterior e definir o estado da produção científica brasileira nesta escala. Além disso, se considerar uma comparação entre dois países apenas, pode-se obter o interesse da produção em cada país de maneira mais precisa.

Também seria interessante a análise de produção científica, com técnicas de *Text Mining* em um conjunto de artigos cuja comparação com a análise humana seja possível, de maneira a estabelecer a ferramenta como um método de revisão bibliográfica.

Ainda, em posse de um banco de dados mais extenso (cuja proporção de publicações nacionais fosse maior), seria possível a utilização de algoritmos de *Machine Learning* como de Clusterização, para agrupar conjuntos de dados que apresentam características semelhantes, de

modo a reconhecer a origem de cada texto, também sendo praticável para o reconhecimento do tema, categoria ou outro atributo de interesse.

REFERÊNCIAS

ACKOFF, R. L. **Ackoff's Best**. New York: John Wiley & Sons, p. 170 – 172, 1999.

AGGARWAL, Charu C.; ZHAI, ChengXiang (Ed.). **Mining text data**. Springer Science & Business Media, 2012.

AGGARWAL, C. C.; AGGARWAL, C. C. **Data Classification**. [S.l: s.n.], 2015.

AKÇAYIR, M.; AKÇAYIR, G. **Advantages and challenges associated with augmented reality for education**: A systematic review of the literature, Educational Research Review, v. 20, p. 1–11, 2017.

ALLAHYARI, M.; POURIYEH, S.; ASSEFI, M.; SAFAEI, S.; TRIPPE, E. D.; GUTIERREZ, J. B.; KOCHUT, K. **A Brief Survey of Text Mining**: Classification, Clustering and Extraction Techniques, 2017.

ANDRADE, Maria Margarida de. Introdução à Metodologia do Trabalho Científico. 10. ed. São Paulo: Atlas, 2017.

APPOLINÁRIO, F. **Dicionário de Metodologia Científica**. 2. ed. São Paulo: Atlas, 2011. 295p

ARRUDA, C. **O Brasil está longe da 4ª Revolução Industrial**. Época Negócios, [S. l.], 18 jun. 2018. Disponível em: https://epocanegocios.globo.com/colunas/noticia/2018/06/o-brasil-esta-longe-da-4-revolucao-industrial.html. Acesso em: 10 set. 2021.

AZUMA, R.; BAILLOT, Y.; BEHRINGER, R.; FEINER, S.; JULIER, S.; MACINTYRE, B. **Recent Advances in Augmented Reality Same as Report (SAR)**, n. December, 2001.

AZUMA, R. T. **A survey of augmented reality, Presence: Teleoperators and Virtual Environments**, v. 6, n. 4, p. 355–385, 1997.

BARRETT, P.; HUNTER, J.; MILLER, J. T.; HSU, J.-C.; GREENFIELD, P. **matplotlib -- A Portable Python Plotting Package**, ASP Conference Series, v. 347, n. June, p. 91, 2005.

BELLINI, H. **The Real Deal with Virtual and Augmented Reality**. [*S. l.*], 2016. Disponível em: https://www.goldmansachs.com/insights/pages/virtual-and-augmented-reality.html. Acesso em: 16 jun. 2020.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit.** O'Reilly Media, Inc., 2009.

BORLIDO, D. **Indústria 4.0 – Aplicação a Sistemas de Manutenção**, Faculdade de Engenharia Universidade do Porto, p. 65, 2017.

BUITINCK, L.; LOUPPE, G.; BLONDEL, M.; PEDREGOSA, F.; MUELLER, A.; GRISEL, O.; NICULAE, V.; PRETTENHOFER, P.; GRAMFORT, A.; GROBLER, J.; LAYTON, R.; VANDERPLAS, J.; JOLY, A.; HOLT, B.; VAROQUAUX, G. **API design for machine learning software**: experiences from the scikit-learn project, p. 1–15, 2013.

CARMIGNIANI, J.; FURHT, B.; ANISETTI, M.; CERAVOLO, P.; DAMIANI, E.; IVKOVIC, M. **Augmented reality technologies, systems and applications, Multimedia Tools and Applications**, v. 51, n. 1, p. 341–377, 2011.

CHEN, M.; EBERT, D.; HAGEN, H.; LARAMEE, R.; VAN, R.; CWI, L.; KWAN-LIU MA, A.; RIBARSKY, W.; SCHEUERMANN, G.; SILVER, D. **Data Information Knowledge Visualization**, 2008.

CLARIVATE ANALYTICS. **Research in Brazil**: Funding excellence: Analysis prepared on behalf of CAPES by the Web of Science Group. Clarivate Report, [*S. l.*], 2019.

COX, Victoria. **Exploratory data analysis**. In: Translating Statistics to Make Decisions. Apress, Berkeley, CA, 2017. p. 47-74.

DEL-MASSO, M. C. S.; COTTA; M. A. C.; SANTOS, M. A. P. **Ética em Pesquisa Científica: conceitos e finalidades**. São Paulo, UNESP: 2014.

DONOHO, D. **50 Years of Data Collection**, p. 1–41, 2015.

DUTTA, S; LANVIN, B; WUNSCH-VINCENT, S. **Global Innovation Index**: Creating Healthy Lives-the Future of Medical Innovation. 2019. Cornell University, 2019.

ESCOBAR, Herton. **15 universidades públicas produzem 60% da ciência brasileira**. Jornal da USP, 2019. Disponível em: https://jornal.usp.br/universidade/politicas-cientificas/15-universidades-publicas-produzem-60-da-ciencia-brasileira. Acesso em: 14 jan. 2021.

FRAGA-LAMAS, P.; FERNÁNDEZ-CARAMÉS, T. M.; BLANCO-NOVOA, Ó.; VILAR-MONTESINOS, M. A. **A Review on Industrial Augmented Reality Systems for the Industry 4.0 Shipyard**, IEEE Access, v. 6, p. 13358–13375, 2018.

GARON, M.; BOULET, P. O.; DOIRONZ, J. P.; BEAULIEU, L.; LALONDE, J. F. **Real-Time High Resolution 3D Data on the HoloLens**, Adjunct Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality, ISMAR-Adjunct 2016, p. 189–191, 2017.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 6. ed. São Paulo: Atlas, 2019.

GRUNWALD, A. **Technology Assessment**: Concepts and Methods. Philosophy of Technology and Engineering Sciences. p. 1103–1146, 2009.

HAND, D. J. **Principles of data mining**, v. 30, n. 7, p. 621–622, 2007.

IN, J.; LEE, S. **Statistical data presentation**, Korean Journal of Anesthesiology, v. 70, n. 3, p. 267–276, 2017.

JEBB, A. T.; PARRIGON, S.; WOO, S. E. **Exploratory data analysis as a foundation of inductive research**, Human Resource Management Review, v. 27, n. 2, p. 265–276, 2017.

JINHA, Arif E. **Article 50 million**: an estimate of the number of scholarly articles in existence. Learned Publishing, v. 23, n. 3, p. 258-263, 2010.

JORDAN, M. I.; MITCHELL, T. M. **Machine learning: Trends, perspectives, and prospects**. Science, v. 349, n. 6245, 2015.

JUODYTE, M. **Overview: Data Mining Pipeline**, 2017.

KANNAN, S.; GURUSAMY, V. **Preprocessing techniques for text mining**. International Journal of Computer Science & Communication Networks, v. 5, n. 1, p. 7-16, 2014.

KANEVSKI, M.; MAIGNAN, M. **Analysis and modelling of spatial environmental data**. CRC Press, 2004.

KELLEHER, J. D.; TIERNEY, B. [D.Sci] **Data Science**. [S.l.], The MIT Press, 2018.

KOTSIANTIS, S. B.; KANELLOPOULOS, D. **Data preprocessing for supervised learning**, International Journal of …, v. 1, n. 2, p. 1–7, 2006.

LEÃO, L. M. **Metodologia do Estudo e Pesquisa: facilitando a vida dos estudantes, professores e pesquisadores**. Petrópolis, RJ: Vozes, 2017.

LEBANON, G.; EL-GEISH, M. **Computing with Data**: An Introduction to the Data Industry. Springer, 2018.

LIMA, E.; MUES, C.; BAESENS, B. **Domain knowledge integration in data mining using decision tables**: case studies in churn prediction, Journal of the Operational Research Society, v. 60, n. 8, p. 1096–1106, 2009.

LOPER, E.; BIRD, S. NLTK: **The Natural Language Toolkit**, 2002.

MCKINNEY, W. **Data Structures for Statistical Computing in Python**, Proceedings of the 9th Python in Science Conference, v. 1697900, n. Scipy, p. 51–56, 2010.

MCKINNEY, W. **pandas: powerful Python data analysis toolkit**, 2012.

MCLEVEY, J.; MCILROY-YOUNG, R. **Introducing metaknowledge**: Software for computational research in information science, network analysis, and science of science. Journal of informetrics, v. 11, n. 1, p. 176-197, 2017.

MILGRAM, P.; KISHIMO, F. **A taxonomy of mixed reality**, IEICE Transactions on Information and Systems, v. 77, n. 12, p. 1321–1329, 1994.

OJEDA, T.; TATTAR, P.; MURPHY, S. P.; BENGFORT, B.; DASGUPTA, A. **Practical Data science Cookbook** - Using R. [S.l: s.n.], 2014.

OLIPHANT, T. E. **Guide to NumPy**, Methods, v. 1, p. 378, 2010.

PACKER, R.; JORDAN, K. **Multimedia**: From Wagner to Virtual Reality. New York and London: W. W. Norton & Company, 2002.

PALMARINI, R.; ERKOYUNCU, J. A.; ROY, R.; TORABMOSTAEDI, H. **A systematic review of augmented reality applications in maintenance**, Robotics and Computer-Integrated Manufacturing, v. 49, n. February, p. 215–228, 2018.

PEDREGOSA, F.; VAROQUAX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A., COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. **Scikit-learn**: Machine Learning in Python. Journal of Machine Learning Research, 12, 2011.

PERKINS, J. **Python text processing with NLTK 2.0 cookbook**. Packt Publishing Ltd, 2010.

PROVOST, F.; FAWCETT, T. **Data Science and its Relationship to Big Data and Data-Driven Decision Making**, Big Data, v. 1, n. 1, p. 51–59, 2013.

ROSER, M.; RITCHIE, H.; ORTIZ-OSPINA, E. **Internet**. [*S. l.*], 2015. Disponível em: https://ourworldindata.org/internet. Acesso em: 18 jun. 2020.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P. B. **Metodologia de pesquisa**. 5 ed. Porto Alegre: Penso, 2013.

SEVERINO, A. J. **Metodologia do Trabalho Científico**. 24. ed. São Paulo: Cortez, 2016.

STANKOVIC, M.; GUPTA, R.; FIGUEROA, J. **Industry 4.0** - Opportunities behind the challenge, United Nations Industrial Development Organization, n. December 2017, p. 8–9, 2017.

TANG, A.; OWEN, C.; BIOCCA, F.; MOU, W. **Comparative effectiveness of augmented reality in object assembly**, Conference on Human Factors in Computing Systems - Proceedings, n. 5, p. 73–80, 2003.

VAN KREVELEN, D. W. F.; POELMAN, R. **A Survey of Augmented Reality Technologies**, Applications and Limitations, International Journal of Virtual Reality, v. 9, n. 2, p. 1–20, 2010.

VAROQUAUX, G.; GOUILLART, E.; VAHTRAS, O.; DE BUYL CHRISTOPHER BURNS, P.; CHAUVE, A.; CIMRMAN, R.; COMBELLES RALF GOMMERS, C.; ESPAZE, A.; JĘDRZEJEWSKI-SZMEK VALENTIN HAENEL, Z.; HARTMANN, M.; INGOLD, G.L.;

PEDREGOSA DIDRIK PINTE, F.; ROUGIER, N. P.; VAN DEN BOSSCHE, J.; VIRTANEN, P.; OTHERS, many. **Python Matplotlib SciKits Numpy SciPy IPython**, 2020.

VIJAYARANI, S.; ILAMATHI, Ms J.; NITHYA, Ms. **Preprocessing techniques for text mining-an overview**. International Journal of Computer Science & Communication Networks, v. 5, n. 1, p. 7-16, 2015.

WANG, X.; DUNSTON, P. S. **Design, strategies, and issues towards an Augmented Reality-based construction training platform**, Electronic Journal of Information Technology in Construction, v. 12, n. June 2006, p. 363–380, 2007.

WASKOM, M. **Seaborn**: Statistical data visualization. [*S. l.*], 2013. Disponível em: https://seaborn.pydata.org. Acesso em: 25 jun. 2020.

ANEXO A – Termo de Autenticidade

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA

Termo de Declaração de Autenticidade de Autoria

Declaro, sob as penas da lei e para os devidos fins, junto à Universidade Federal de Juiz de Fora, que meu Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Produção é original, de minha única e exclusiva autoria. E não se trata de cópia integral ou parcial de textos e trabalhos de autoria de outrem, seja em formato de papel, eletrônico, digital, áudio-visual ou qualquer outro meio.
Declaro ainda ter total conhecimento e compreensão do que é considerado plágio, não apenas a cópia integral do trabalho, mas também de parte dele, inclusive de artigos e/ou parágrafos, sem citação do autor ou de sua fonte.
Declaro, por fim, ter total conhecimento e compreensão das punições decorrentes da prática de plágio, através das sanções civis previstas na lei do direito autoral[1] e criminais previstas no Código Penal[2], além das cominações administrativas e acadêmicas que poderão resultar em reprovação no Trabalho de Conclusão de Curso.

Juiz de Fora, 13 de setembro de 2021.

André Robson de Alcântara Silva 201349058
NOME LEGÍVEL DO ALUNO (A) Matrícula

André R. de A. Silva 156 138 257 42
ASSINATURA CPF

[1] LEI N° 9.610, DE 19 DE FEVEREIRO DE 1998. Altera, atualiza e consolida a legislação sobre direitos autorais e dá outras providências.
[2] Art. 184. Violar direitos de autor e os que lhe são conexos: Pena - detenção, de 3 (três) meses a 1 (um) ano, ou multa.