Rafael José de Alencar Almeida

Suporte à análise interativa de discussões *online* combinando técnicas de mineração de dados

São João del-Rei

2020

Rafael José de Alencar Almeida

Suporte à análise interativa de discussões *online* combinando técnicas de mineração de dados

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São João del-Rei como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Universidade Federal de São João del-Rei - UFSJ

Departamento de Ciência da Computação

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Dárlinton Barbosa Feres Carvalho

São João del-Rei

2020

Ficha catalográfica elaborada pela Divisão de Biblioteca (DIBIB)
e Núcleo de Tecnologia da Informação (NTINF) da UFSJ,
com os dados fornecidos pelo(a) autor(a)

A447s

Almeida, Rafael José de Alencar.
Suporte à análise interativa de discussões online combinando técnicas de mineração de dados / Rafael José de Alencar Almeida ; orientador Dárlinton Barbosa Feres Carvalho. -- São João del-Rei, 2020.
107 p.

Dissertação (Mestrado - Ciência da Computação) -- Universidade Federal de São João del-Rei, 2020.

1. Discussões online. 2. Descoberta de Conhecimento. 3. Mineração de Dados. 4. Teoria Fundamentada em Dados. 5. Design Science Research. I. Carvalho, Dárlinton Barbosa Feres, orient. II. Título.

Rafael José de Alencar Almeida

Suporte à análise interativa de discussões *online* combinando técnicas de mineração de dados

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São João del-Rei como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Trabalho aprovado. São João del-Rei, 12 de fevereiro de 2020:

Prof. Dr. Dárlinton Barbosa Feres Carvalho
UFSJ (Orientador)

Prof. Dr. Clodis Boscarioli
UNIOESTE

Prof. Dr. Edimilson Batista dos Santos
UFSJ

São João del-Rei
2020

Agradecimentos

Agradeço ao meu orientador professor Dr. Dárlinton Barbosa Feres Carvalho, pela disponibilidade, apoio, motivação e compreensão em todas as etapas do mestrado, que foram fundamentais para o desenvolvimento desta dissertação.

Agradeço à minha mãe Rachel Alencar por me estimular a sempre me dedicar aos estudos e dar o melhor de mim em todas atividades, e à minha esposa Alice Cristina pelo companheirismo.

Agradeço aos professores membros da banca, que deram valiosa contribuição na estruturação da versão final desta dissertação.

Agradeço aos meus colegas docentes do Núcleo de Informática do Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais - *Campus* Barbacena, por me ajudarem a flexibilizar meus horários de trabalho durante as disciplinas que cursei no mestrado.

Agradeço aos colegas da Universidade Federal de São João del-Rei, pela convivência, troca de conhecimentos e amizade, e a todos que de alguma forma contribuíram com o desenvolvimento deste trabalho de dissertação.

"Sem dados você é apenas mais uma pessoa com uma opinião."
(William Edwards Deming)

Resumo

O crescente volume de discussões *online*, contextualizadas e espontâneas dos seus participantes, vem possibilitando pesquisas científicas de larga escala. Todavia, para investigar estes grandes conjuntos de mensagens é necessário utilizar um processo analítico iterativo denominado descoberta de conhecimento em bases de dados, que é fundamentado em técnicas computacionais de mineração de dados para identificação de padrões e relacionamentos nos textos. Cada etapa deste processo envolve diversas escolhas técnicas, que impactam significativamente no resultado final da análise. Nesta dissertação, buscou-se contribuir com melhorias no suporte à realização destas pesquisas por meio da criação de uma ferramenta centrada no pesquisador, a qual mapeia as principais escolhas para controles interativos e visualizações gráficas, bem como seu processo de uso embasado na Teoria Fundamentada em Dados, de modo a produzir resultados cientificamente relevantes. Na realização deste trabalho foi empregado o método *Design Science Research*, o qual se baseia na pesquisa, desenvolvimento e avaliação de um artefato e das teorias relacionadas. O artefato desenvolvido (ferramenta e respectivo processo de uso) foi avaliado por meio dos métodos *System Usability Scale* e UTAUT 2, evidenciando uma excelente usabilidade e aceitação pelos usuários. Também foi conduzido um estudo de caso na análise de um fórum sobre neurociência, apresentando a aplicação da ferramenta e contribuindo com a produção de novos conhecimentos científicos na área. A avaliação positiva dos usuários e sua capacidade de responder de forma satisfatória às questões de pesquisa do estudo de caso demonstram a aplicabilidade e a relevância deste trabalho no suporte à análise interativa de discussões *online*.

Palavras-chave: Discussões *online*, Descoberta de Conhecimento, Mineração de Dados, Teoria Fundamentada em Dados, *Design Science Research*

Abstract

The growing volume of online discussions, which are contextualized and spontaneously generated by its participants, is enabling unprecedented large-scale scientific research. However, to investigate these large sets of messages requires the use of an iterative analytical process called knowledge discovery in databases, relying on computational data mining techniques that support the identification of patterns and relationships in the texts. Each step of this process involves several technical choices, that significantly impact the final result of the analysis. In this dissertation, we sought to contribute with improvements in support of these researches through the creation of a tool centered on the researcher, which maps the leading choices onto interactive controls and graphic visualizations, as well as its process of use based on the Grounded Theory to produce scientifically relevant results. In carrying out this work, the Design Science Research method, which is based on research, development, and evaluation of an artifact and related theories, was employed. The assessment of the created artifact (tool and respective usage process) applied the System Usability Scale and UTAUT 2 methods, showing excellent usability and acceptance by users. A case study analyzing a forum on neuroscience illustrates the tool application and provides new knowledge in the area. The positive user evaluation and their ability to respond satisfactorily to the case study research questions show the applicability and relevance of this work in supporting the interactive analysis of online discussions.

Keywords: Online Discussions, Knowledge Discovery, Data Mining, Grounded Theory, Design Science Research

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	*Application Programming Interface*
BoW	*Bag of Words*
CBOW	*Continuous Bag Of Words*
CSV	*Comma Separated Values*
DSR	*Design Science Research*
GT	*Grounded Theory*
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
LDA	*Latent Dirichlet Allocation*
LeIA	Léxico para Inferência Adaptada
MAE	*Mean Absolute Error*
MVC	*Model View Controller*
NMF	*Non-negative Matrix Factorization*
PLS	*Partial Least Square*
POS	*Parts Of Speech*
SUS	*System Usability Scale*
TF	*Term Frequency*
TF–IDF	*Term Frequency – Inverse Document Frequency*
UTAUT	*Unified Theory of Acceptance and Use of Technology*
VADER	*Valence Aware Dictionary and sEntiment Reasoner*

Sumário

1 Introdução

Com o advento da Internet e a popularização dos dispositivos móveis, bem como das mídias sociais, ambientes virtuais de discussão vêm possibilitando uma contínua e intensa troca de conhecimentos e experiências entre usuários sobre os mais variados temas e aspectos de suas vidas (FAN; GORDON, 2014). Representando importantes pontos de encontro para as pessoas no mundo virtual e abrangendo os mais variados assuntos, estes espaços para discussões *online* aproximam usuários com valores e interesses compartilhados, constituindo agregações sociais emergentes, estabelecidas pelo sentimento de pertencimento a um grupo.

O crescente volume de informações disponíveis publicamente nestes ambientes para discussão virtual, como fóruns e sites de redes sociais, possibilita a condução de pesquisas científicas de larga escala, buscando compreender a percepção dos participantes em relação aos assuntos discutidos. Isto se deve pelo conteúdo dessas comunidades ser vasto, contextualizado, detalhado e disponível em seu habitat natural (KOZINETS, 2002), onde discussões públicas podem ser coletadas com facilidade de forma natural e pouco intrusiva – representando manifestações espontâneas dos usuários, em oposição a outras formas de coleta como questionários, enquetes e entrevistas.

Nesse contexto, apresenta-se o campo de pesquisa da Ciência Social Computacional (do inglês, *Computational Social Science*), combinando Ciência da Computação e Ciências Sociais, o qual traz possibilidades inéditas de pesquisa graças à capacidade computacional de se coletar e analisar dados em um volume, profundidade e escalas sem precedentes (LAZER et al., 2009). O método de destaque deste campo é a Netnografia, focada no estudo qualitativo de comunidades *online* (O'DONOHOE, 2010). Inspirada na estrutura metodológica da Etnografia, a Netnografia aproveita-se da possibilidade de observação natural das manifestações de usuários em discussões *online*, trazendo um conjunto de práticas de pesquisa para compreensão da interação social nos contextos da comunicação virtual – de forma inconcebível para métodos tradicionais de pesquisa.

Transpassando diversos campos acadêmicos, a Netnografia demonstra-se extremamente útil para revelar estilos de interação, narrativas pessoais, trocas comuns, regras *online*, práticas, estilos discursivos, formas inovadoras de colaboração e organização e manifestações de criatividade (KOZINETS, 2015). Dessa forma, seu emprego na análise de discussões *online* se estende de pesquisas acadêmicas multidisciplinares à investigação de discussões de consumidores para inovação de produtos por empresas, apresentando grande adoção atual na indústria como método científico de pesquisa (BARTL; KANNAN; STOCKINGER, 2016).

Consistindo em um processo sistemático de pesquisa, coleta e análise de dados, a Netnografia fundamenta-se em seis passos metodológicos (KOZINETS, 2015):

1) Definição do campo de pesquisa, incluindo o contexto a ser pesquisado e as questões de pesquisa a serem investigadas;

2) Identificação e seleção da comunidade, envolvendo a busca da comunidade cuja discussão *online* será analisada, avaliando seu potencial para responder às questões de pesquisa propostas;

3) Observação da comunidade e coleta de dados;

4) Ética da pesquisa, buscando respeitar a identidade dos participantes das discussões coletadas e proteger sua privacidade, garantindo o anonimato e a confidencialidade dos membros da discussão;

5) Análise dos dados, podendo ser manual ou assistida por *software*, tendo como objetivo identificar padrões e relacionamentos;

6) Descobertas e soluções, consistindo na consolidação das descobertas resultantes da análise das discussões e sua capacidade de responder às questões de pesquisa propostas, além da possibilidade de gerar novas soluções ou produtos.

A partir do conjunto de práticas estabelecidas pela Netnografia, tem-se um referencial metodológico para a elaboração e condução de análises de discussões *online*, visando coletar e analisar dados em busca de novos conhecimentos, enquanto preserva a privacidade dos participantes. Entretanto, apesar do grande volume e do fluxo crescente de dados das mídias sociais possibilitarem novas maneiras de se fazer ciência, sua análise manual por especialistas apresenta-se inviável devido ao grande tempo consumido no processo – requerendo inovações em pesquisas na área de computação (SHNEIDERMAN et al., 2011). Nesse sentido, técnicas de mineração de dados vêm sendo adotadas para suporte à análise de discussões *online* em diversas áreas do conhecimento, com destaque para os desafios de seleção do conteúdo a ser analisado, análise dos sentimentos expressos nas discussões e sumarização visual dos resultados (FAN; GORDON, 2014).

A aplicação de técnicas de mineração de dados fundamenta-se no ciclo do processo de Descoberta de Conhecimento em Bases de Dados, do inglês *Knowledge Discovery in Databases* (KDD), um processo iterativo e com o envolvimento do especialista no domínio de problema analisado, que visa extrair conhecimento útil de grandes volumes de dados (FAYYAD et al., 1996). No KDD, o processo de mineração de dados é definido na forma de um conjunto de etapas aplicadas sucessivamente em busca de identificar padrões e relacionamentos ocultos nos dados e produzir valor por meio da descoberta de novos conhecimentos (Figura 1). As cinco etapas do processo de KDD são:

1) Seleção, que envolve coletar e formatar os dados de entrada, que no caso da

análise de discussões *online* são conjuntos de textos;

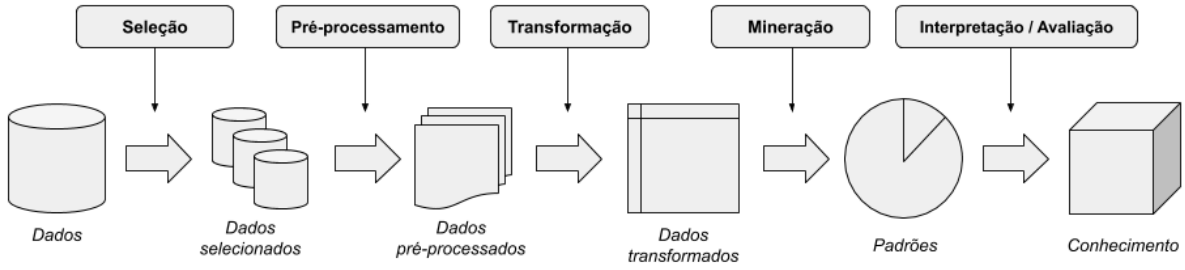
2) Pré-processamento, onde os dados selecionados para análise são pré-processados para garantir sua consistência, validade, formatação e normalização;

3) Transformação, onde os dados pré-processados são transformados em uma representação numérica requerida pelos algoritmos de mineração de dados;

4) Mineração, na qual são aplicados algoritmos para identificação de padrões;

5) Interpretação/Avaliação, onde o resultado do processo é apresentado de forma visual e intuitiva para o pesquisador.

Figura 1 – Etapas do processo de KDD



Fonte: elaborada pelo autor com base em Fayyad et al. (1996)

DEBUSE et al. (2001) observam que cada estágio deste processo apresenta numerosas escolhas do usuário especialista – como técnicas, modelos e parâmetros –, que podem impactar significativamente nos resultados finais da pesquisa. Entretanto, a participação do especialista em todo o processo apresenta-se como um desafio em análises multidisciplinares de discussões *online*, em que o domínio do problema e os pesquisadores muitas vezes não pertencem à área de mineração de dados. Nesse sentido, além de ser um processo não-trivial, iterativo, interativo e de múltiplos passos (GEIST, 2002), a complexidade dos algoritmos e métodos empregados cria uma barreira técnica adicional para a efetiva participação dos pesquisadores em todas as etapas da análise.

De fato, observa-se que trabalhos e ferramentas propostos focam no envolvimento do pesquisador especialista no domínio de problema predominantemente na etapa final do processo, ficando a cargo dos especialistas em mineração de dados definir, para uma única execução, os passos e parâmetros a serem utilizados nas demais etapas do processo de KDD (ZEMMOURI et al., 2012). Além disso, observa-se nos trabalhos relacionados a ausência de um enfoque na especificação de processos metodológicos bem definidos para as análises, limitando seu rigor e reprodutibilidade. Dessa forma, no contexto onde são produzidos cada vez mais dados sociais *online* relevantes para pesquisa multidisciplinar – possibilitando a produção de valor para a ciência e a indústria –, apresenta-se o desafio de se especificar e avaliar com rigor científico um método para inserção efetiva de especialistas de diversos domínios em todas as etapas da descoberta de conhecimento.

No sentido de apresentar soluções para os problemas e questões relatadas, o presente trabalho apresenta uma pesquisa que culminou com a especificação e o desenvolvimento de uma ferramenta de mineração de dados para análise interativa de discussões *online*, bem como seu processo de uso, com foco na participação do usuário pesquisador em todas as etapas da análise. Para garantir o rigor científico do desenvolvimento da ferramenta e seu processo, é empregado o método *Design Science Research* (DSR), no qual o conhecimento e a compreensão de um domínio do problema e sua solução são alcançados graças à construção e aplicação de um artefato projetado. O fluxo de uso integrado ao processo de KDD proposto para aplicação da ferramenta desenvolvida fundamenta-se na análise rigorosa e sistemática dos dados para construção de teorias, proveniente da Pesquisa Fundamentada em Dados (do inglês *Grounded Theory*), combinada com o conjunto de práticas de pesquisa para compreensão da interação social nos contextos da comunicação virtual, da Netnografia.

A efetiva contribuição da ferramenta e do seu processo de uso para pesquisas em discussões *online* centradas no pesquisador é avaliada em relação à sua usabilidade e aceitação, por meio dos métodos *System Usability Scale* (SUS) e *Unified Theory of Acceptance and Use of Technology* 2 (UTAUT 2), aplicados a usuários reais na exploração de uma discussão *online* sobre programação de computadores. Também é realizado um estudo de caso exploratório aplicando a ferramenta e seu processo de uso à análise de uma discussão *online* sobre neurociência, demonstrando a aplicabilidade da pesquisa desenvolvida e contribuindo com a produção de novos conhecimentos científicos no contexto de sua comunidade.

Dessa forma, pode-se destacar como principais contribuições – tecnológica e científicas – deste trabalho de dissertação:

- Especificação de um novo artefato tecnológico (ferramenta interativa) para mineração de dados de discussões *online*, capaz de incluir o pesquisador especialista no domínio do estudo em todas as etapas do processo de KDD, com seu desenvolvimento pautado no método *Design Science Research*;
- Especificação de um processo de análise de discussões *online* baseado na Teoria Fundamentada em Dados e integrado a todas as etapas de KDD, de modo a possibilitar a condução de pesquisas éticas, rigorosas e reprodutíveis;
- Avaliação da usabilidade e aceitação da ferramenta e de seu processo de uso, por meio de verificação com usuários no contexto de uma análise de discussão *online* real;
- Estudo de caso exploratório, aplicando a ferramenta e seu processo de uso à análise de uma discussão *online* sobre neurociência, demonstrando a aplicabilidade da pesquisa de dissertação desenvolvida;

- Produção e comunicação científica dos conhecimentos obtidos a partir da análise da comunidade de discussões *online* sobre neurociência.

A seguir, é apresentada a organização dos capítulos desta dissertação:

O Capítulo 2 apresenta o referencial teórico para o desenvolvimento da pesquisa, envolvendo o método utilizado (*Design Science Research*) e os trabalhos relacionados. O método DSR é apresentado no contexto da pesquisa em Sistemas de Informação, por meio de um panorama geral de sua abordagem científica e pragmática, e do seu processo de aplicação. Este método baseia-se na conjunção de ciclos de relevância, rigor e *design*, e por etapas para garantir contribuições científicas e tecnológicas. A seção de trabalhos relacionados apresenta o estado da arte em relação à análise interativa de discussões *online*, por meio da investigação dos trabalhos relacionados à pesquisa. São analisadas e mapeadas suas principais características e contribuições – como recursos, técnicas e modelos – bem como suas limitações. Por meio delas são identificadas as necessidades de pesquisas relevantes, possibilitando delimitar as contribuições desenvolvidas nesta dissertação.

No Capítulo 3 é apresentada a especificação e o desenvolvimento da ferramenta proposta, chamada Topic Insights, bem como seu processo de uso. São investigadas as principais técnicas de mineração de dados aplicadas à análise de discussões *online* e suas relações com o processo de KDD. Para cada etapa são apresentadas suas principais características e seu relacionamento com a ferramenta de análise e seu processo de uso. Também é apresentado o projeto do desenvolvimento técnico da ferramenta, destacando sua arquitetura, tecnologias e protocolos empregados. Buscando padronização e interoperabilidade, são utilizadas linguagens de programação, bibliotecas e formatos de dados abertos e padronizados, por não demandarem licenças de uso e por possuírem comunidades ativas de desenvolvedores. Integrado à ferramenta, o processo de uso baseia-se na Teoria Fundamentada em Dados, e no rigor de pesquisa da Netnografia.

O Capítulo 4 apresenta o processo de avaliação da ferramenta desenvolvida, composto por um experimento com usuários e um estudo de caso. A avaliação com usuários envolve a análise de uma comunidade *online* real sobre programação, empregando os métodos SUS e UTAUT 2 para avaliar a usabilidade e a aceitação. Já o estudo de caso envolve a análise de uma discussão *online* do mundo real, visando contribuir com novos conhecimentos sobre a mesma e validar a aplicabilidade da pesquisa. Para sua condução foi selecionado um fórum de discussão *online* sobre neurociência – um assunto que faz interseção com diversas áreas do conhecimento, possuindo tópicos de discussão ricos em conteúdo multidisciplinar e com grandes volumes de dados a serem explorados.

A discussão sobre os resultados do trabalho é descrita no Capítulo 5, onde são apresentadas as descobertas, contribuições e limitações da pesquisa, sendo também propostas sugestões de trabalhos futuros para aperfeiçoar e estender a pesquisa.

2 Referencial teórico

2.1 *Design Science Research*

Diante do caráter de produção de novos conhecimentos aliado à inovação tecnológica no presente trabalho, seu desenvolvimento fundamenta-se no método de pesquisa *Design Science Research* (DSR). Este método, indicado para pesquisas científicas em Sistemas de Informação (PEFFERS et al., 2007), envolve a pesquisa, o desenvolvimento e avaliação de um artefato – objeto cuja construção segue métodos científicos –, bem como das teorias que o viabilizam. No presente trabalho, o artefato consiste na ferramenta e no seu processo de uso para apoio à análise interativa de discussões *online*, centradas no pesquisador.

2.1.1 Histórico

O método DSR fundamenta-se na metodologia de pesquisa de Ciência do Projeto, do inglês *Design Science* (DS), a qual se apresenta como o conjunto de conhecimentos para concepção e desenvolvimento (*designing*) de projetos produzidos por pesquisas rigorosas e sistemáticas (DRESCH; LACERDA; JÚNIOR, 2015). O conceito de DS foi introduzido por FULLER (1957), como uma forma sistemática de conceber e conduzir projetos que possuam interface com o mundo real. Posteriormente, GREGORY (1966) retomou o conceito, reforçando a distinção entre a concepção de um projeto (*design*) e a concepção por um método científico (*Design Science*), e a relevância deste para a produção de novos conhecimentos.

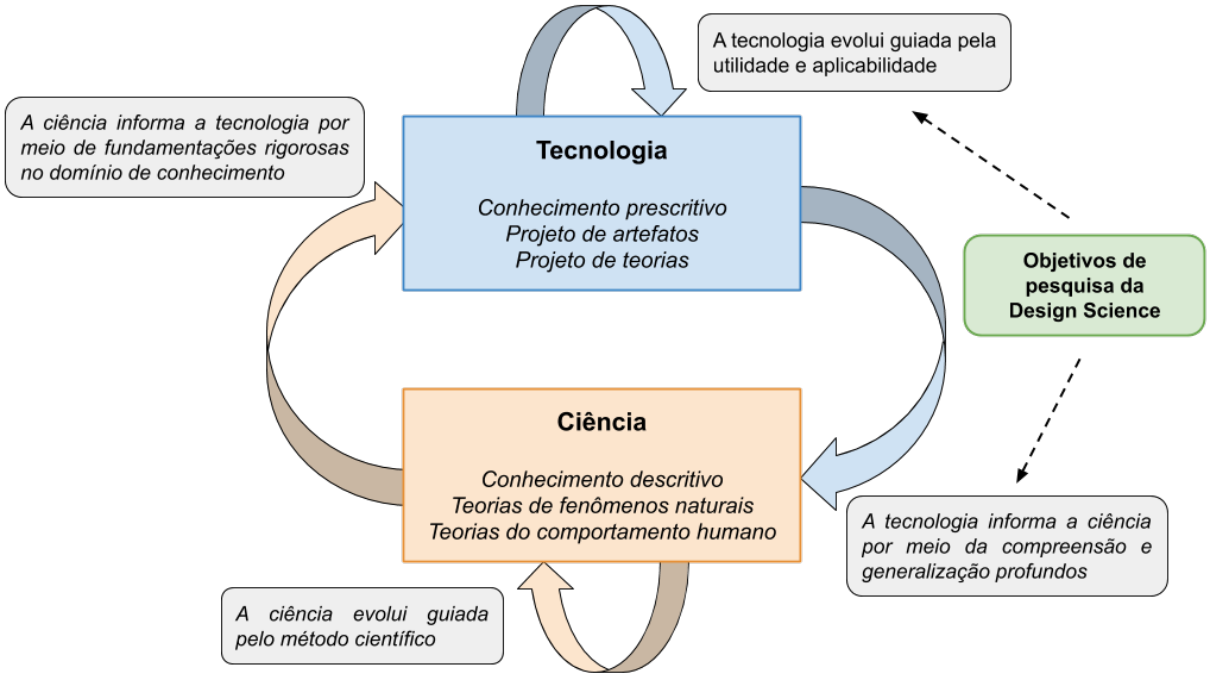
Apesar de já ser estudada há décadas, foi com SIMON (1996) que o termo *Design Science* se popularizou, como uma abordagem científica do processo de criação no contexto do mundo do artificial – em oposição ao mundo natural. Nesta ciência do projeto, a geração de conhecimento está atrelada ao processo de concepção de artefatos artificiais, como *software*, processos e produtos físicos. Na área de Sistemas de Informação, a *Design Science* é introduzida por JR; CHEN; PURDIN (1990), que apresentam e discutem projetos de pesquisa para o desenvolvimento de sistemas baseados nesta metodologia.

A primeira formalização de um método baseado na *Design Science* é proposta por TAKEDA; VEERKAMP; YOSHIKAWA (1990), onde são apresentadas as ideias de um ciclo de *design*, com etapas para conduzir o desenvolvimento de um artefato. HEVNER et al. (2004) desenvolvem o DSR como método para a área de Sistemas de Informação, destacando o rigor e a relevância que devem estar presentes em projetos que a utilizam – sendo seu trabalho a base para diversos outros autores da área. Por rigor, entende-se a fundamentação e as contribuições científicas esperadas do projeto de pesquisa, enquanto a relevância relaciona-se à contribuição prática para o contexto do projeto.

O trabalho de Hevner é estendido por WIERINGA (2014), onde é proposto um arcabouço para a pesquisa com o método DSR. Este é composto por um conjunto de etapas em um ciclo regulador, que itera pela investigação do problema, passando pelo projeto, validação e implementação de uma solução, até a avaliação da mesma. O autor também observa que uma pesquisa baseada no método DSR é composta por problemas práticos e problemas teóricos, mutuamente aninhados (dependentes). A solução do problema prático contribui com a produção de um conhecimento prescritivo, trazendo uma contribuição prática, e produz conhecimento a partir da criação e aplicação dessa solução. Já o problema teórico, quando solucionado, produz um conhecimento descritivo, o qual se propõe a explicar e ampliar o conhecimento no contexto da pesquisa.

Dessa forma, combinando artefatos úteis e rigor científico, uma pesquisa baseada no método DSR traz dois tipos de contribuições principais: projetos de artefatos e teorias do projeto – ambas viabilizadas pela interação entre ciência e tecnologia. Nesse contexto, BASKERVILLE et al. (2018) destacam que os objetivos da ciência são ampliar o conhecimento descritivo do mundo natural e do comportamento humano, por meio da aplicação do método científico – buscando uma melhor compreensão de como o mundo funciona. Já os objetivos da tecnologia são aumentar o conhecimento prescritivo, por meio do desenvolvimento de artefatos projetados para melhorar as capacidades humanas. O método DSR posiciona-se neste ciclo de interação entre ciência e tecnologia, como ilustrado na Figura 2.

Figura 2 – Interação entre ciência e tecnologia no contexto do método DSR

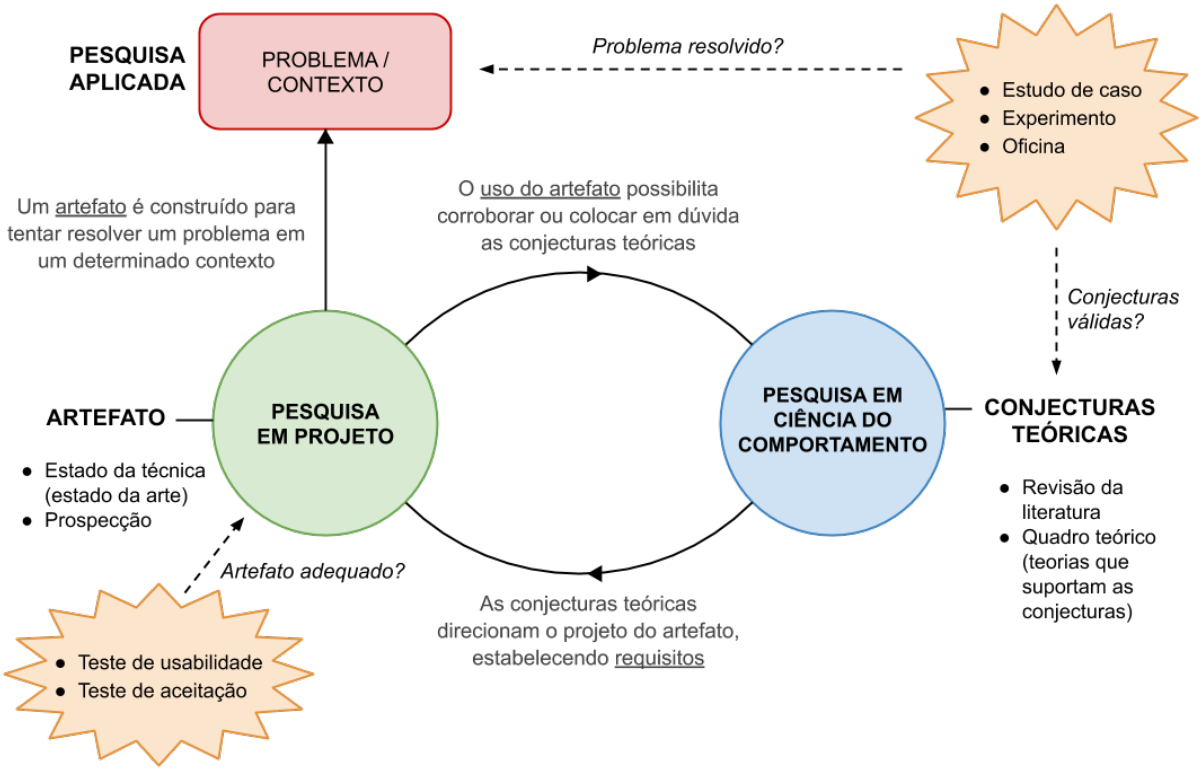


Fonte: elaborada pelo autor com base em Baskerville et al. (2018)

Um esquema de pesquisa conduzida com o método DSR é apresentado por PIMENTEL (2016), ilustrando como a pesquisa em projeto (*design*) e a pesquisa em ciência

iteram na busca de uma solução para o problema de um determinado contexto (Figura 3). A pesquisa em ciência traz as fundamentações e conjecturas teóricas para compreensão do problema e definição de uma solução válida, por meio da revisão de literatura e do uso de teorias e modelos. Dessa forma são estabelecidos os requisitos para o artefato, que direcionarão a pesquisa em projeto. Esta envolve o desenvolvimento do artefato, sua avaliação e sua aplicação em um contexto onde suas contribuições possam ser avaliadas – validando ou descartando as conjecturas teóricas. Dessa forma, como destacam MARCH; STOREY (2008), o uso do método DSR contribui como estratégia de pesquisa capaz de orientar tanto a construção do conhecimento, quanto aprimorar as práticas em sistemas de informação e de várias disciplinas relacionadas ao campo gerencial e tecnológico da Ciência da Computação.

Figura 3 – Esquema do método de pesquisa DSR



Fonte: elaborada pelo autor com base em Pimentel (2016)

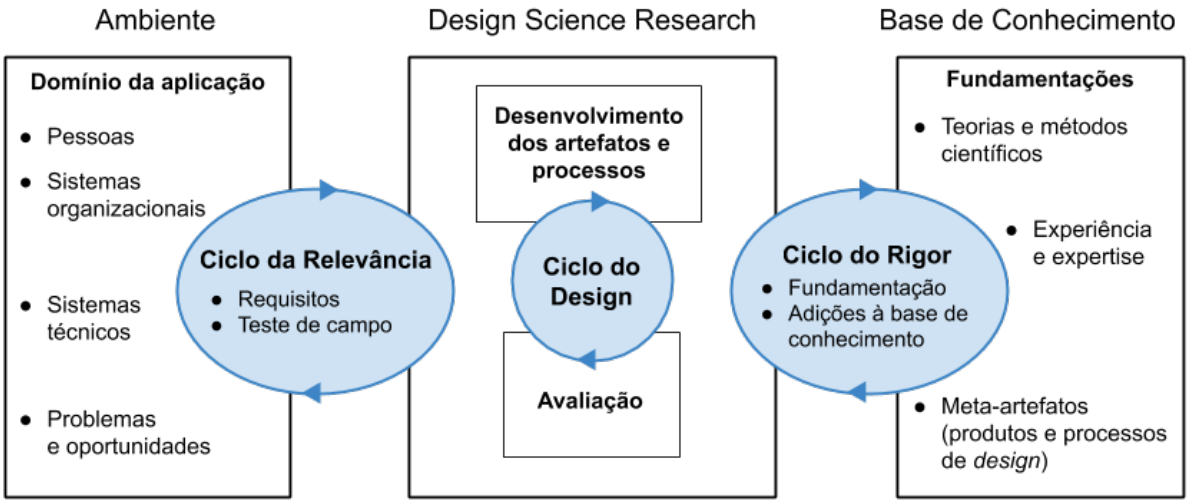
2.1.2 Ciclos do método DSR

Com uma abordagem científica pragmática, o método DSR busca identificar e compreender problemas do mundo real e propor soluções apropriadas e úteis, fazendo avançar o conhecimento teórico da área (HEVNER et al., 2004). Ele envolve a especificação, construção, investigação, validação e avaliação de artefatos, os quais podem ser construtos (entidades e relações), modelos (abstrações e representações), métodos (algoritmos e práticas) e instanciações (implementações de sistemas e protótipos), a fim de resolver novos problemas práticos (MARCH; SMITH, 1995). Dessa forma, o DSR estabelece um processo

metodológico com o rigor científico necessário para a produção de novos conhecimentos acadêmicos, garantindo também a relevância e aplicabilidade prática da pesquisa.

A aplicação do DSR na pesquisa científica pode ser vista como a conjunção de três ciclos reguladores de atividades relacionadas (HEVNER, 2007): Ciclo da Relevância, Ciclo do Rigor e Ciclo do *Design* (Figura 4). Estes ciclos orientam a parte técnica da pesquisa enquanto permitem gerar novos conhecimentos respondendo às questões teóricas envolvidas – por meio do estabelecimento de interfaces entre o ambiente social e suas necessidades de negócios, e a base de conhecimento científico para sua fundamentação. Dessa forma, o método DSR possibilita a criação de conhecimentos descritivos e prescritivos, a partir da inter-relação dos aspectos sociais e tecnológicos que compõem o problema a ser investigado.

Figura 4 – Ciclos do método DSR



Fonte: elaborada pelo autor com base em Hevner et al. (2007)

O Ciclo da Relevância inicia a pesquisa no contexto do problema a ser analisado e seu ambiente, possibilitando avaliar os requisitos para a pesquisa, os critérios de aceitação e os resultados esperados para a investigação. Para isso ele se fundamenta no domínio de aplicação do problema, que é constituído por pessoas, sistemas organizacionais e sistemas técnicos que interagem no contexto analisado. Após o desenvolvimento e avaliação científicos do artefato, o Ciclo da Relevância envolve a aplicação do mesmo em testes de campo, de modo a avaliar sua capacidade de solucionar o problema e determinar se serão necessárias iterações adicionais.

O Ciclo do Rigor envolve a fundamentação teórica para a pesquisa e o estudo da literatura científica da área investigada, de modo a assegurar suas efetivas contribuições e inovação. Neste ciclo o pesquisador deve realizar uma investigação profunda na base de conhecimento do contexto do problema. Esta se caracteriza como o conjunto de trabalhos relacionados de referência na literatura científica da área, bem como fundamentações teóricas e metodológicas para a pesquisa a ser desenvolvida. Os processos desse ciclo têm

como objetivo garantir que os artefatos produzidos sejam contribuições de pesquisa efetivas, e não apenas projetos de rotina baseados em processos já propostos e estabelecidos na área (HEVNER et al., 2004). Após o desenvolvimento e avaliação científicos do artefato, o Ciclo do Rigor adiciona os novos conhecimentos gerados pela pesquisa à base de conhecimento. Estes são constituídos pelos artefatos e processos desenvolvidos, bem como as experiências obtidas na aplicação prática dos mesmos durante sua validação.

Já o Ciclo do *Design* envolve as atividades principais de construção do artefato, sua instanciação e avaliação, constituindo o ciclo central do processo de pesquisa. Ele itera entre a construção do artefato (nesta dissertação uma ferramenta e seu processo de uso) e sua avaliação, analisando se os objetivos para uma solução foram atendidos. O *feedback* da avaliação é analisado de modo a aprimorar e refinar o desenvolvimento do artefato, até que o mesmo constitua uma solução adequada para resolver o problema investigado na pesquisa.

2.1.3 Etapas do método DSR

Durante o ciclo de pesquisa DSR, os artefatos são instanciados para os contextos a serem investigados e avaliados em relação às soluções que se propõem alcançar – trazendo como contribuições científicas os novos conhecimentos para sua definição e os resultados das análises de suas instanciações. Dessa forma, como destaca ÇAĞDAŞ; STUBKJÆR (2011), o DSR representa um processo rigoroso de projetar artefatos para resolver problemas, avaliar o que foi projetado e comunicar cientificamente os resultados obtidos. Este processo pode ser sumarizado em seis etapas fundamentais (PEFFERS et al., 2007): 1) Identificação do problema e motivação; 2) Definição dos objetivos para uma solução; 3) *Design* e desenvolvimento; 4) Avaliação; 5) Demonstração; 6) Comunicação.

O desenvolvimento do presente trabalho fundamenta-se nos ciclos do método DSR, estruturando-se de acordo com suas etapas. A primeira etapa, de identificação do problema e motivação, é contemplada na introdução deste trabalho. Nela é apresentada o contexto e o potencial das análises de discussões *online* em pesquisas multidisciplinares, com base nos princípios da Netnografia e no processo de descoberta de conhecimento em bases de dados. Também são discutidas as limitações das pesquisas atuais em relação ao envolvimento do pesquisador em todas as etapas da descoberta de conhecimento, bem como a ausência de processos de análise bem fundamentados metodologicamente.

A segunda etapa do DSR consiste na definição dos objetivos para uma solução, constituindo a análise de trabalhos relacionados e dos parâmetros configuráveis do processo de KDD. Nela é realizada uma revisão da literatura (base de conhecimento, na terminologia do método DSR), aprofundado os conceitos de mineração de dados e processo de descoberta de conhecimento, identificando propostas de ferramentas para análise interativa de discussões *online*. Também são mapeadas técnicas, métodos, recursos e limitações, possibilitando delimitar as características e os resultados esperados de uma solução adequada para o

problema de se conduzir pesquisas centradas no pesquisador.

Na terceira etapa, de *design* e desenvolvimento, são apresentadas as fundamentações teóricas e técnicas para o desenvolvimento da ferramenta e de seu processo de uso. Ela abrange a implementação da ferramenta, envolvendo as escolhas tecnológicas como linguagens, modelos e protocolos, bem como seu processo de uso. Este baseia-se na Teoria Fundamentada em Dados e na Netnografia, e possui um fluxo de execução integrado à ferramenta desenvolvida.

A quarta etapa, de avaliação, envolve o emprego de técnicas de avaliação com usuários para verificar a usabilidade, aceitação e utilidade da solução desenvolvida. Visando avaliar se os objetivos para a solução do problema foram atendidos, nesta etapa são aplicados os métodos *System Usability Scale* e *Unified Theory of Acceptance and Use of Technology* 2, para analisar, respectivamente, sua usabilidade e aceitação. Estes métodos são quantitativos convergentes, com ênfase igual e coleta simultânea, e os resultados dessa avaliação mista são relacionados por meio da técnica de triangulação – possibilitando uma discussão aprofundada em relação à avaliação geral do artefato. A pesquisa de avaliação com usuários obteve aprovação de Comitê de Ética em Pesquisa Envolvendo Seres Humanos, conforme registrado na Plataforma Brasil.

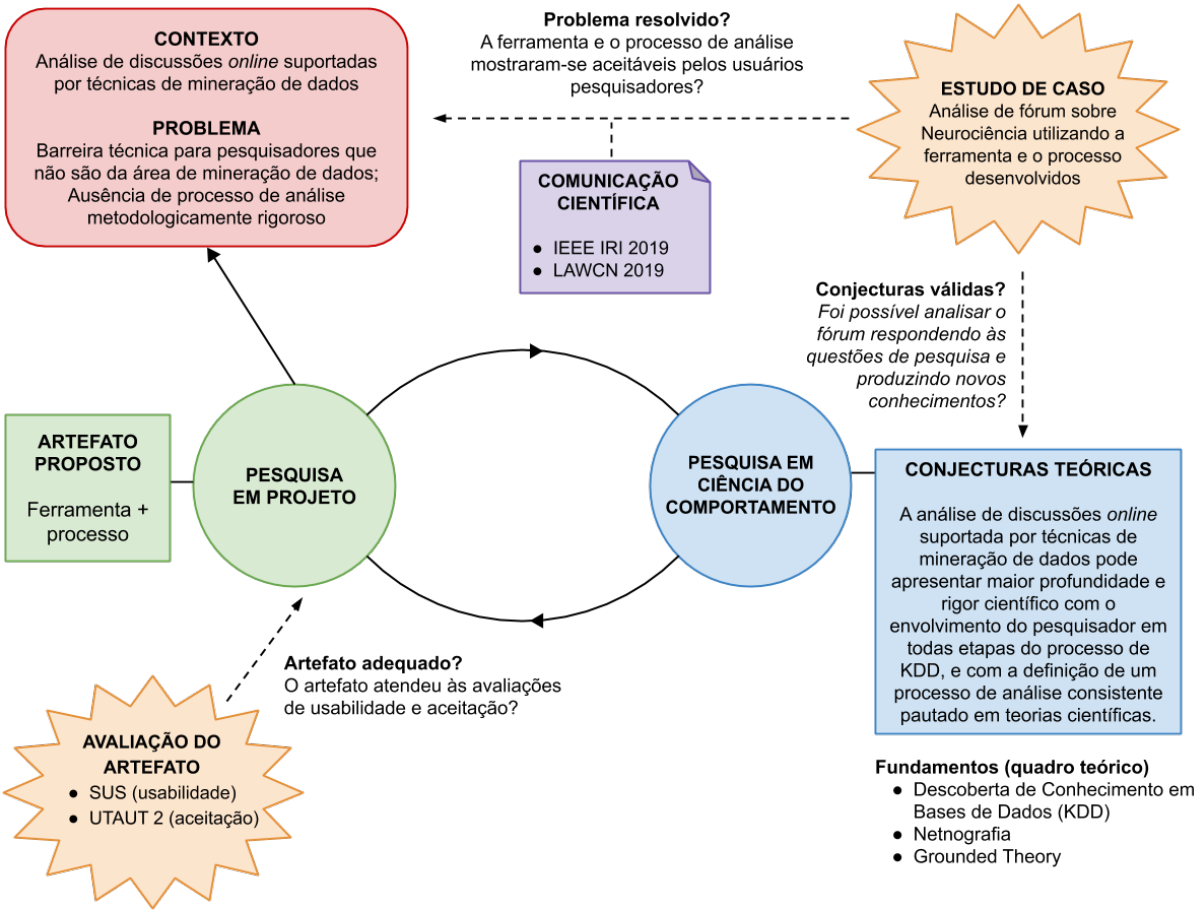
A etapa cinco, de demonstração, é apresentada no estudo de caso, onde a solução desenvolvida é aplicada a um contexto do mundo real, buscando validar as conjecturas teóricas. Sua aplicabilidade é demonstrada por meio de um estudo de caso exploratório, em que a ferramenta e seu processo de uso são utilizados na análise de uma discussão *online*, buscando responder às questões de pesquisa propostas para produção de novos conhecimentos.

Na etapa final, de comunicação, são comunicados cientificamente os resultados da pesquisa desenvolvida com o método DSR – envolvendo os conhecimentos descritivos e prescritivos obtidos. Esta etapa é atendida pelo presente trabalho de dissertação, bem como por meio das publicações de resultados parciais em eventos internacionais, apresentadas na conclusão do trabalho.

Seguindo a estrutura do método de pesquisa DSR proposta por PIMENTEL (2016) e com base nas etapas apresentadas, esta pesquisa pode ser sumarizada por meio do diagrama da Figura 5. O contexto do trabalho envolve a análise de discussões *online* suportada por técnicas de mineração de dados, buscando soluções para o desafio de se envolver pesquisadores que não são da área de mineração de dados em todas as etapas da descoberta de conhecimento, por meio de um processo interativo e metodologicamente rigoroso. Fundamentando-se no processo de KDD, na Netnografia e na Teoria Fundamentada em Dados, são propostas melhorias no contexto do problema, envolvendo a especificação e o desenvolvimento de uma ferramenta interativa e seu respectivo processo de uso. Eles constituem o artefato, e são avaliados em relação a sua usabilidade e aceitação, e também aplicados a um estudo de caso com dados reais – de modo a avaliar a adequação da

pesquisa, bem como a validade das conjecturas propostas.

Figura 5 – Esquema do método DSR no contexto da pesquisa desenvolvida



Fonte: elaborada pelo autor com base em Pimentel (2016)

2.2 Trabalhos relacionados

Diferentes técnicas de mineração de dados podem ser aplicadas em pesquisas envolvendo a análise multidisciplinar de discussões *online*. Em meio à grande variedade de técnicas aplicáveis, FAN; GORDON (2014) apresentam uma visão geral do processo de análise de mídias sociais, descrevendo as principais técnicas de mineração de dados relevantes para o contexto: modelagem de tópicos, análise de sentimentos e análise visual. A modelagem de tópicos é uma técnica de sumarização automática, que consiste na identificação de assuntos semânticos (tópicos) de um conjunto dos textos, possibilitando identificar de forma não-supervisionada os temas que compõem a discussão. A análise de sentimentos consiste na identificação da polaridade de um texto ou conjunto de textos, normalmente classificando-os como positivos, negativos ou neutros. Aplicada ao conjunto de textos de cada tópico, esta técnica permite identificar o sentimento predominante de cada assunto por ele expresso – contribuindo com uma percepção geral da comunidade sobre o tema. Já as técnicas de visualização de dados possibilitam a identificação visual de

padrões e outras informações descobertas na etapa de mineração, de forma mais intuitiva para o pesquisador, sendo composta por diversos elementos visuais como: gráficos indicando o volume de postagens de um tópico ao longo do tempo, nuvem de palavras com os termos semânticos de maior destaque de cada tópico e gráficos indicando a distribuição dos sentimentos de um tópico.

A viabilidade e o potencial da combinação dessas técnicas na análise de discussões *online* em diferentes domínios vem sendo pesquisada pelo autor da presente dissertação em trabalhos anteriores. Em (LUIZ et al., 2018), é proposto um arcabouço para análise de avaliações de aplicativos móveis por usuários em lojas virtuais. Identificou-se no trabalho que a combinação das técnicas de modelagem de tópicos, análise de sentimentos e visualização possibilitou a descoberta automática e intuitiva de funcionalidades, *bugs* e requisitos dos aplicativos analisados, representando uma alternativa de análise semântica superior ao tradicional sistema de avaliação por estrelas. Em uma segunda publicação do presente autor (ALMEIDA et al., 2019), o contexto analisado são discussões *online* sobre linguagens de programação, e é adicionado às técnicas de mineração de dados já utilizadas o componente tempo, possibilitando não só a identificação automática de assuntos relevantes e seus sentimentos associados, como também sua evolução temporal. Como resultado, demonstrou-se a possibilidade da identificação de alterações realizadas em linguagens de programação, bem como a percepção dos usuários em relação às mesmas.

Os trabalhos descritos foram aplicados à análise de discussões *online* na área de Computação (aplicativos e linguagens) utilizando artefatos de código não interativos, de modo que os autores não só conduziram o processo de KDD fazendo as escolhas técnicas em cada etapa, como também foram os especialistas no domínio do problema. Para análises de outros contextos, onde os especialistas no domínio não são da área de Computação, surge a necessidade de desenvolver ferramentas interativas que possibilitem aos mesmos conduzir o processo de forma visual e intuitiva. Esta abordagem de envolver o pesquisador no fluxo de análise (*human-in-the-loop*) torna-se essencial pelo fato do KDD constituir um processo iterativo analítico e sistemático, onde a descoberta de conhecimento impõe um modelo de trabalho em que suas etapas são revisitadas à medida que os dados são explorados em busca de novos conhecimentos (SILVA; PERES; BOSCARIOLI, 2017). Além disso, cada etapa exige escolhas de parâmetros pelo pesquisador, demandando uma ferramenta e um processo de análise interativos para os sucessivos ajustes durante a pesquisa.

A revisão de literatura para identificar os trabalhos discutidos a seguir foi realizada por meio de buscas às bases do Google Acadêmico[1] e do Portal de Periódicos da CAPES[2], por meio de consultas utilizando os termos “human in loop topic modeling” e “interactive topic modeling”. Esta escolha se deu pelo fato de que o recurso de modelagem de tópicos

[1] https://scholar.google.com
[2] http://www.periodicos.capes.gov.br/

empregado na sumarização automática dos assuntos é um componente essencial para a análise de discussões *online*, bem como a interatividade do processo. A partir da leitura dos trabalhos identificados, foram selecionados aqueles que possibilitam a análise de discussões virtuais, e que possuem pelo menos dois recursos interativos configuráveis pelo usuário pesquisador.

A técnica principal de análise de discussões *online*, a modelagem de tópicos, auxilia na compreensão de grandes volumes de textos, mas os tópicos produzidos por sua execução inicial nem sempre fazem sentido para os usuários, sendo necessária uma abordagem centrada no usuário (HU et al., 2014). Isso se deve pelas características de cada *dataset* de entrada, que para uma melhor análise, pode demandar a alteração de etapas e parâmetros. Por exemplo, no caso do nível de granularidade (quantidade) dos tópicos: uma discussão virtual com muitos assuntos, necessita de uma maior segmentação de tópicos produzidos para que possam ser captados adequadamente. Nesse contexto, HOQUE; CARENINI (2016) formalizam e investigam o conceito de modelagem de tópicos com o humano no processo, buscando resolver o problema de tópicos inconsistentes produzidos pelo algoritmo por meio da introdução de supervisão humana. Seu trabalho avalia os impactos na qualidade dos tópicos produzidos quando o algoritmo tem a possibilidade de modificá-los a partir do *feedback* do usuário, o qual realiza a tarefa por meio de uma interface gráfica interativa. Avaliando esta abordagem, os autores observam que a mesma é capaz de aprimorar a habilidade dos usuários da ferramenta em identificar tópicos mais consistentes. Além disso, os usuários avaliados preferiram o uso da ferramenta interativa apresentada em detrimento de outras ferramentas estáticas para geração e visualização de tópicos – indicando a relevância da interatividade no processo de análises de discussões *online*.

Aprimorando seu trabalho anterior, HOQUE; CARENINII (2018) apresentam a ferramenta MultiConVisIT, a qual estende a produção de tópicos com *feedback* do usuário por meio da integração de novas técnicas de mineração e visualização de dados no processo de análise de discussões *online*. Dentre os novos recursos introduzidos e aplicados aos tópicos gerados – e com sua relevância validada por meio de avaliação com usuários –, destacam-se: controle de granularidade, evolução temporal e análise de sentimentos. O controle de granularidade possibilita ao usuário escolher a quantidade de tópicos em que a discussão será segmentada. Menos tópicos forçam textos com pequenas semelhanças fazerem parte de um mesmo grupo, de modo que os tópicos resultantes captarão os assuntos globais da discussão virtual. Já com mais tópicos, torna-se possível captar mais nuances, resultando na identificação de assuntos semânticos mais específicos.

Já o recurso de evolução temporal analisa o volume de postagens de cada tópico ao longo do ciclo de vida da discussão, caracterizando uma análise longitudinal do tópico, a qual possibilita identificar como um determinado assunto variou ao longo do tempo. Isto permite identificar assuntos emergentes, recorrentes, constantes e esparsos – tornando a

análise do tópico multifacetada e trazendo mais riqueza para a pesquisa realizada. Este recurso de linha do tempo também é identificado na ferramenta interativa VISTopic (YANG; YAO; QU, 2017), na qual defendem a utilidade desse recurso como uma visão evolutiva da tendência dos tópicos, capaz de ajudar os pesquisadores a obterem mais *insights* ao comparar múltiplos tópicos.

Técnicas de análise de sentimentos aplicadas às postagens que fazem parte de cada tópico possibilitam a rápida identificação da polaridade dos assuntos, permitindo ao usuário pesquisador compreender o sentimento geral sobre determinado tópico e segmentar os assuntos de acordo com a manifestação dos usuários. O emprego da análise de sentimentos nos tópicos identificados também mostra-se relevante para a identificação de assuntos controversos entre os participantes de uma discussão virtual, possibilitando o entendimento de como surgem divergências em tais conversas e como as diferenças de opiniões iniciam e evoluem para diferentes questões controversas (HOQUE; ABID, 2019).

Outros recursos identificados na literatura para aprimorar o processo de análise interativa de discussões *online* utilizando mineração de dados são: anotação dos tópicos produzidos para etapas futuras de análise e uso de *n-grams* para criação de tópicos mais ricos. Por se tratarem de distribuições probabilísticas de palavras, tópicos são representados por sequências de termos, ordenados por seu peso semântico, e cabe ao usuário inferir o assunto a partir dos n termos mais relevantes do tópico. Por exemplo, para os três termos "vaga", "emprego" e "programação", cabe ao pesquisador que analisa essa saída dos dados rotular esse tópico como "Vaga de emprego em programação". Nesse contexto, SKEPPSTEDT et al. (2018) e ZOU; HOU (2014) apresentam ferramentas interativas para análise de discussões *online* que introduzem o recurso de rotular manualmente os tópicos resultantes do processo de análise, para auxiliar na sua identificação e possibilitar que os mesmos sejam salvos para análises futuras.

Devido à característica dos tópicos consistirem em sequências de termos, termos compostos como "bom dia" e "não gostei" acabam sendo tratados individualmente na modelagem de tópicos, perdendo seu valor semântico. Cientes dessa característica, WANG et al. (2019) destacam que esta representação na forma de conjuntos de unigramas geralmente resulta em descrições ambíguas dos tópicos – levando a uma interpretação pobre dos mesmos pelos pesquisadores. Seu trabalho apresenta um algoritmo para definição automática de quando preservar termos em conjunto para modelagem de tópicos e um protótipo de ferramenta interativa implementando a proposta. Sua avaliação com usuários indica que alguns tópicos baseados em *n-grams* preservam maior valor semântico, mas observam que o custo computacional de seu algoritmo para calcular automaticamente quando utilizar esta técnica representa um desafio para análise de grandes volumes de dados.

A Tabela 1 mapeia os principais recursos das ferramentas propostas nos trabalhos

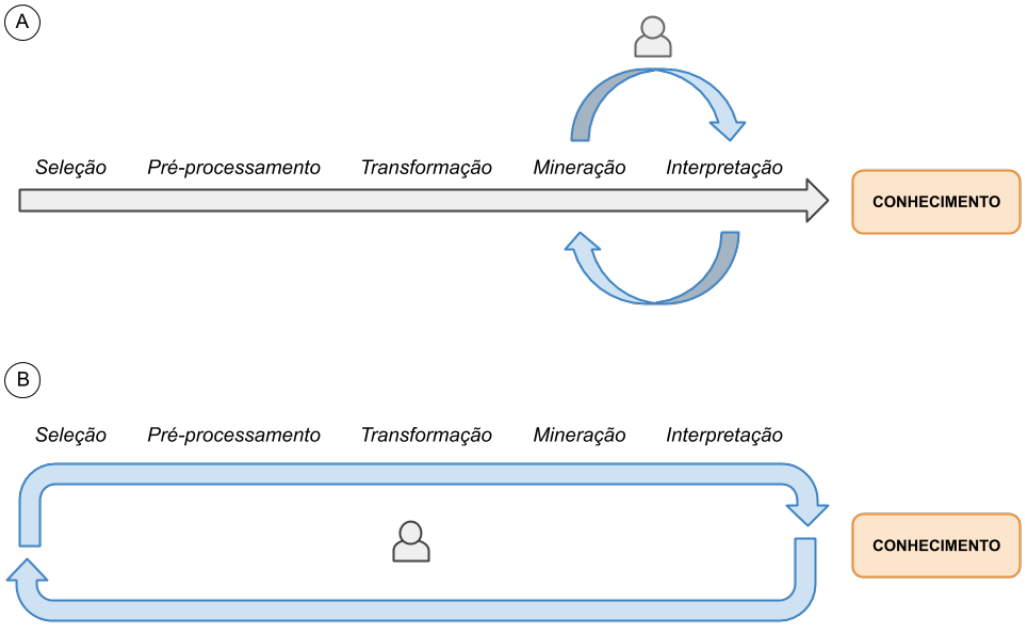
investigados, que são tomados como base para a especificação e o desenvolvimento do artefato apresentado nessa dissertação.

Tabela 1 – Recursos dos principais trabalhos relacionados

Ferramenta / Recurso	LDAAnalyzer (2014)	VisTopic (2017)	Topics2Themes (2018)	MultiConVisIT (2018)	Wang (2019)	Hooke (2019)
Modelagem de Tópicos	X	X	X	X	X	X
Análise de Sentimentos			X	X		X
Anotação do tópico	X		X			
Linha do tempo		X		X		X
Granularidade				X		
N-grams					X	
Visualização	X	X	X	X	X	X

Enquanto cada ferramenta interativa analisada possui recursos característicos para apoiar a análise de discussões *online*, todas dão maior enfoque nos parâmetros e especificidades do recurso de modelagem de tópicos. Desta forma, a interatividade e a inserção do pesquisador especialista no domínio da discussão analisada ficam limitadas às etapas de mineração e avaliação do processo de KDD – não explorando todo potencial de envolvê-lo e se beneficiar da expertise de suas decisões em todas as etapas (Figura 6).

Figura 6 – Interatividade do processo de KDD nos trabalhos avaliados (A) e em uma situação onde o pesquisador participa de todas as etapas (B)



Fonte: autor

Outra limitação observada é que, apesar da maioria das ferramentas identificadas serem avaliadas por usuários, os trabalhos não propõem um processo de uso rigoroso e bem definido para o emprego das mesmas, que possibilite reforçar a consistência e o rigor científico dos resultados das análises realizadas em discussões *online*. Observa-se um enfoque predominante nas ferramentas e algoritmos, não sendo apresentada a metodologia e os critérios para o desenvolvimento das ferramentas, bem como do seu processo de uso.

Destaca-se também que os trabalhos investigados apresentam resultados de avaliação de uso das ferramentas com usuários, mas não apresentam o resultado da aplicação das mesmas em conjuntos de dados do mundo real, de modo a demonstrar sua efetiva aplicabilidade na produção de novos conhecimentos a partir da análise de discussões *online*. Os recursos e as limitações identificados nos trabalhos relacionados são consideradas no presente trabalho, que propõe uma ferramenta interativa e seu processo de uso para apoiar análises de discussões *online* envolvendo o pesquisador e todas as etapas do processo de KDD. No próximo capítulo é apresentada a ferramenta desenvolvida (Topic Insights), sendo discutidas em profundidade as etapas do processo de KDD, mapeando os parâmetros que podem ser ajustados pelo usuário pesquisador da ferramenta em cada etapa. Também são apresentados os aspectos do desenvolvimento técnico da Topic Insights, descrevendo sua arquitetura, bem como as tecnologias e protocolos utilizados na sua implementação.

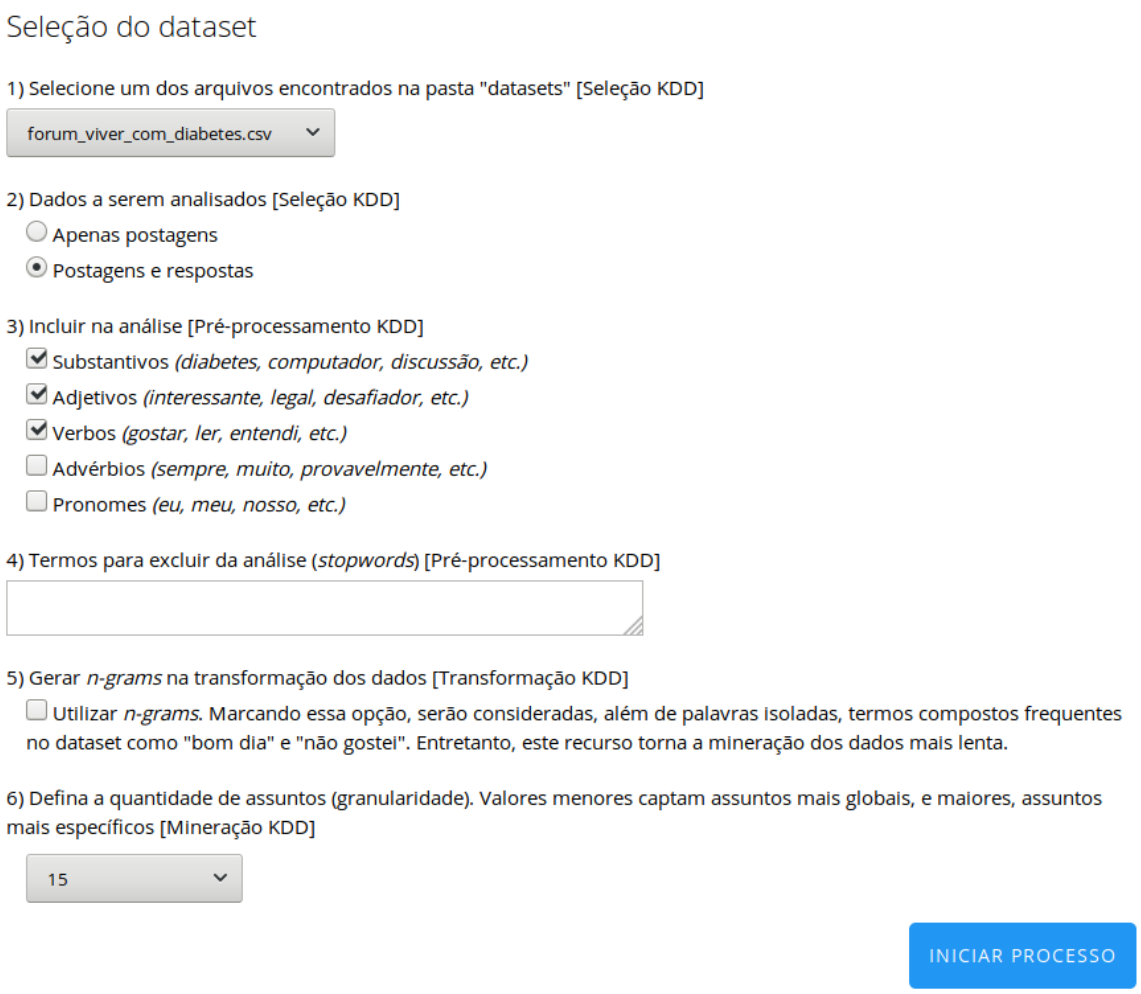
3 Ferramenta Topic Insights

3.1 Interatividade no processo de KDD

O artefato de *software* desenvolvido, chamado Topic Insights, integra os principais recursos identificados nas ferramentas analisadas na revisão de literatura, e busca ser agnóstico em relação à implementação das técnicas discutidas, podendo ser instanciado com diferentes algoritmos. Também são consideradas as principais técnicas de processamento de dados textuais no processo de KDD, comuns às diversas possibilidades de análise no processamento de linguagem natural (VIJAYARANI; ILAMATHI; NITHYA, 2015).

A Figura 7 apresenta a tela do artefato desenvolvido, na qual o usuário pesquisador pode configurar as opções interativas do processo de KDD. Estas opções permitem ajustar parâmetros de cada uma das cinco etapas do processo, que serão discutidas a seguir.

Figura 7 – Tela de seleção de opções das etapas de KDD na ferramenta Topic Insights



Fonte: autor

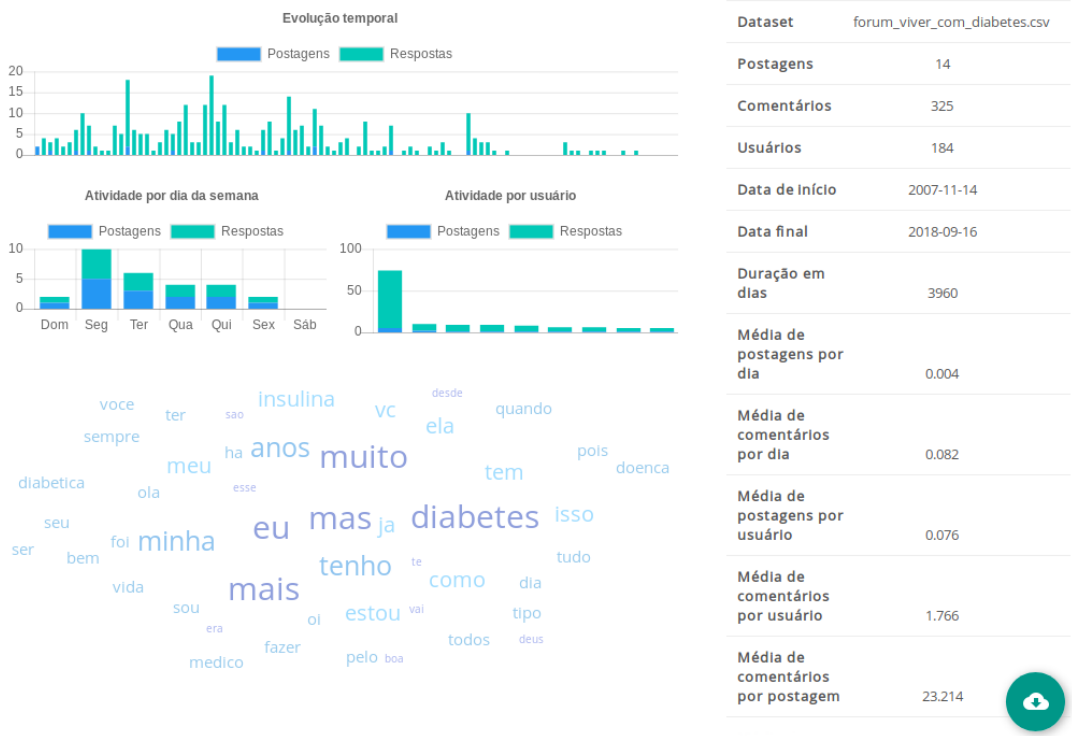
3.1.1 Etapa 1: Seleção

Na primeira etapa do processo de KDD, de seleção, são definidos os dados de entrada para a mineração de dados. Por envolver a análise de discussões *online*, a ferramenta desenvolvida recebe como entrada dados textuais, representando cada postagem do fórum ou rede social analisada. O arquivo de entrada, contendo o conteúdo da discussão, deve estar no formato de interoperabilidade *Comma Separated Values* (CSV), por ser um padrão simples (um registro por linha, com os campos separados por vírgulas), bem definido e fácil de se exportar a partir de planilhas e linguagens de programação.

Além do texto da postagem, o formato de entrada de dados contém outras meta-informações, como: identificador, data da postagem, autor da postagem (identificador anonimizado) e identificador da postagem pai (no caso da postagem ser um comentário a outra postagem). Essas informações possibilitam analisar de forma mais rica as discussões *online*, identificando o volume de postagens e comentários por tópico, a distribuição da participação dos usuários na discussão e a evolução das interações ao longo do tempo.

Os dados selecionados são utilizados pela ferramenta para realizar uma análise descritiva da discussão virtual a ser investigada, produzindo estatísticas e visualizações (Figura 8). O objetivo é possibilitar ao pesquisador uma visão geral da discussão analisada, apresentando, dentre outras informações, o volume de postagens e comentários ao longo do tempo, a atividade dos usuários por dia da semana e os termos de maior frequência.

Figura 8 – Tela de análise descritiva de uma discussão virtual selecionada na ferramenta Topic Insights



Dataset	forum_viver_com_diabetes.csv
Postagens	14
Comentários	325
Usuários	184
Data de início	2007-11-14
Data final	2018-09-16
Duração em dias	3960
Média de postagens por dia	0.004
Média de comentários por dia	0.082
Média de postagens por usuário	0.076
Média de comentários por usuário	1.766
Média de comentários por postagem	23.214

Fonte: autor

Como opção interativa da ferramenta desenvolvida, nesta primeira etapa do processo de KDD o usuário pode selecionar se a análise será efetuada apenas nos textos do tipo postagem ou em todos os textos, incluindo os comentários às postagens. No primeiro caso, serão captados tópicos globais das discussões, uma vez que postagens costumam ser mais sucintas e objetivas, enquanto que na segunda opção serão captadas mais nuances, relativas aos desdobramentos das discussões nos comentários das postagens.

3.1.2 Etapa 2: Pré-processamento

No pré-processamento, segunda etapa do processo de KDD, os dados selecionados na etapa anterior são normalizados para garantir que sigam uma formatação válida e consistente. Nesta etapa ocorre a normalização de letras maiúsculas e minúsculas e remoção de acentos, de modo a garantir que uma palavra grafada de formas diferentes possa ser identificada como a mesma palavra. Também ocorre a remoção de *stopwords*, que são termos frequentes nos textos mas com pouco valor semântico para a análise, como preposições e artigos.

Por depender do contexto da análise, ao invés de se usar uma lista de *stopwords* predefinida, o pesquisador usuário da ferramenta pode definir seus próprios termos, e avaliar os resultados com *stopwords* diferentes a cada iteração do ciclo de KDD. O pesquisador também é envolvido nesta etapa com a opção de selecionar as estruturas gramaticais (*parts-of-speech*) de interesse para a análise. Dessa forma, se estiver em busca de identificar entidades e suas ações, pode manter na análise apenas substantivos e verbos, e se estiver interessado em opiniões, pode selecionar substantivos e adjetivos.

3.1.3 Etapa 3: Transformação

Os dados pré-processados são utilizados na terceira etapa do processo de KDD, de transformação. Nesta etapa, os dados textuais já selecionados, normalizados e pré-processados são transformados para uma representação numérica, requerida pelos algoritmos de mineração de dados da etapa seguinte. Inicialmente os textos passam pelo processo de tokenização, no qual são transformados em listas de termos (*tokens*). Existem diferentes formas de se representar os conjuntos de termos que constituem um texto de forma numérica, mas no geral este processo ocorre mapeando-os para um espaço vetorial. Dessa forma, cada texto de uma postagem é transformado em um vetor de números, onde cada número representa um dos termos.

A técnica mais simples de vetorização é a *Term Frequency* (TF), na qual é produzido um vetor onde cada entrada corresponde à frequência de cada termo do *corpus* (conjunto total de todos os textos), como demonstrado na Figura 9. Para balancear a discrepância de termos pouco frequentes, pode-se normalizar esse vetor de contagem de termos, por

Figura 9 – Representação vetorial Term Frequency (TF)

	termo1	termo2	termo3	termo4	termo5
doc1	0	0	2	1	0
doc2	3	1	5	1	1
doc3	0	0	3	0	0
doc4	1	0	4	1	3

Fonte: autor

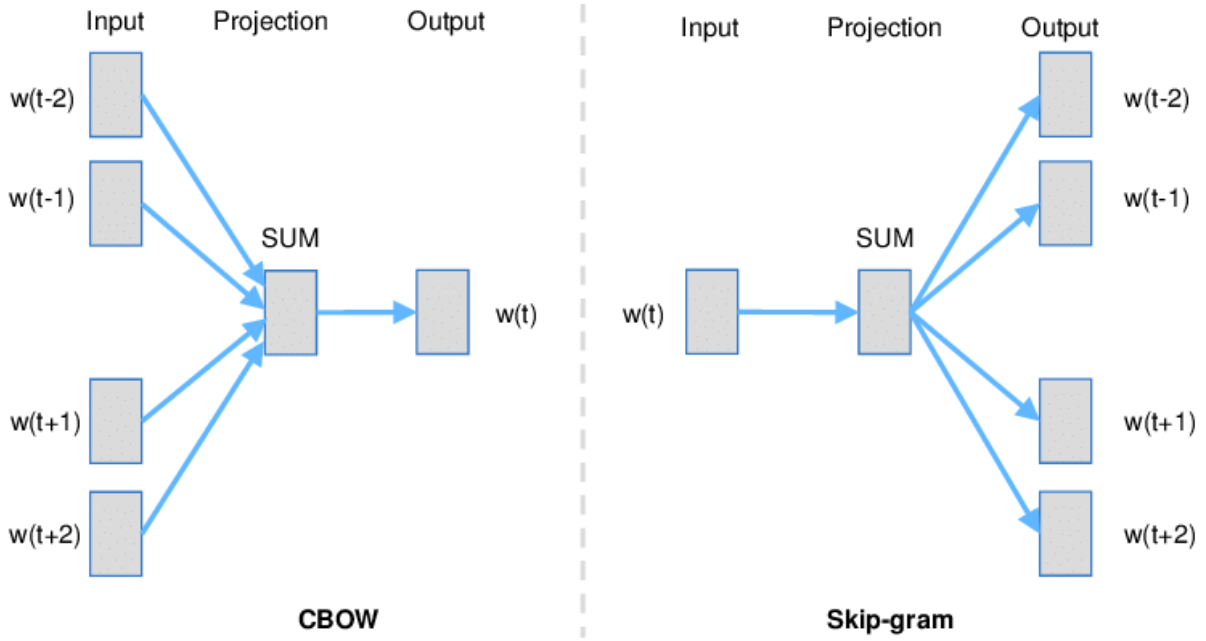
meio da técnica *Term Frequency – Inverse Document Frequency* (TF-IDF). Nela, o valor numérico de um termo aumenta proporcionalmente à medida que aumenta o número de ocorrências do mesmo em um documento, mas esse valor é equilibrado pela frequência do termo no *corpus*. Isso impede que termos raros ou muito frequentes em apenas um documento tenham muita predominância de valor em relação aos outros termos.

As técnicas TF e TF-IDF constituem um modelo do tipo *Bag-of-Words* (BoW), onde a ordem dos termos das frases é perdida, e as suas representações vetoriais produzem matrizes esparsas (devido à existência de uma coluna para cada termo possível). Outra técnica de vetorização de textos alternativa aos *Bag-of-Words* são os *word embeddings* (MIKOLOV et al., 2013). Nesta técnica, são gerados vetores densos treinando uma Rede Neural Artificial (RNA) com cada termo e seu contexto (termos que aparecem em volta do termo analisado), agregando maiores informações semânticas a cada termo, e possibilitando análises mais ricas no espaço vetorial.

Existem dois modelos de arquiteturas principais para construir um *word embedding*: *Continuous Bag Of Words* (CBOW) e *Skip-gram* (Figura 10). No *Continuous Bag Of Words*, para cada termo no *corpus*, a rede neural artificial é treinada para prever este termo, recebendo como entrada o contexto (os termos que envolvem o termo a ser previsto). Já no *Skip-gram*, a RNA é treinada para receber como entrada um termo, e tentar prever os termos do seu contexto. O modelo CBOW é mais rápido para treinar e possui maior acurácia para conjuntos de textos grandes e palavras com grande frequência (LANDTHALER et al., 2017), enquanto o modelo Skip-gram funciona melhor para conjuntos de dados pequenos ou que possuam palavras pouco frequentes (MIKOLOV et al., 2013).

Independente da técnica de vetorização utilizada, por padrão elas trabalham com termos individuais, não captando termos compostos como “Bom dia” e “Não gostei” – impossibilitando que na etapa de modelagem de tópicos um tópico possa ser descrito por meio de um termo composto. Essa característica, que em determinados contextos pode ser útil para definição de termos descritivos mais ricos para a modelagem de tópicos, pode ser aplicada por meio do uso de *n-grams*. Esta técnica consiste no uso de uma janela deslizante de tamanho n durante o processo de tokenização, onde para $n=1$ são geradas entradas

Figura 10 – Arquiteturas de *word embeddings* CBOW e Skip-gram



Fonte: Landthaler et al. (2017)

no vetor para os termos individuais (unigramas), *n=2* são geradas tuplas de termos que co-ocorrem, como os bigramas "Bom dia" e "Muito melhor", e assim sucessivamente.

A ferramenta desenvolvida pode ser instanciada utilizando modelos do tipo *Bag-of-Words* ou *word embeddings*, e no seu componente interativo o usuário pesquisador participa da etapa de transformação do processo de KDD definindo se a análise utilizará termos individuais ou se levará também em consideração *n-grams*. Enquanto neste caso o consumo de memória e o custo computacional aumentam significativamente devido à necessidade de se representar os termos individuais e suas combinações, o resultado da mineração pode ficar mais rico em discussões onde são utilizados muitos termos compostos.

3.1.4 Etapa 4: Mineração de dados

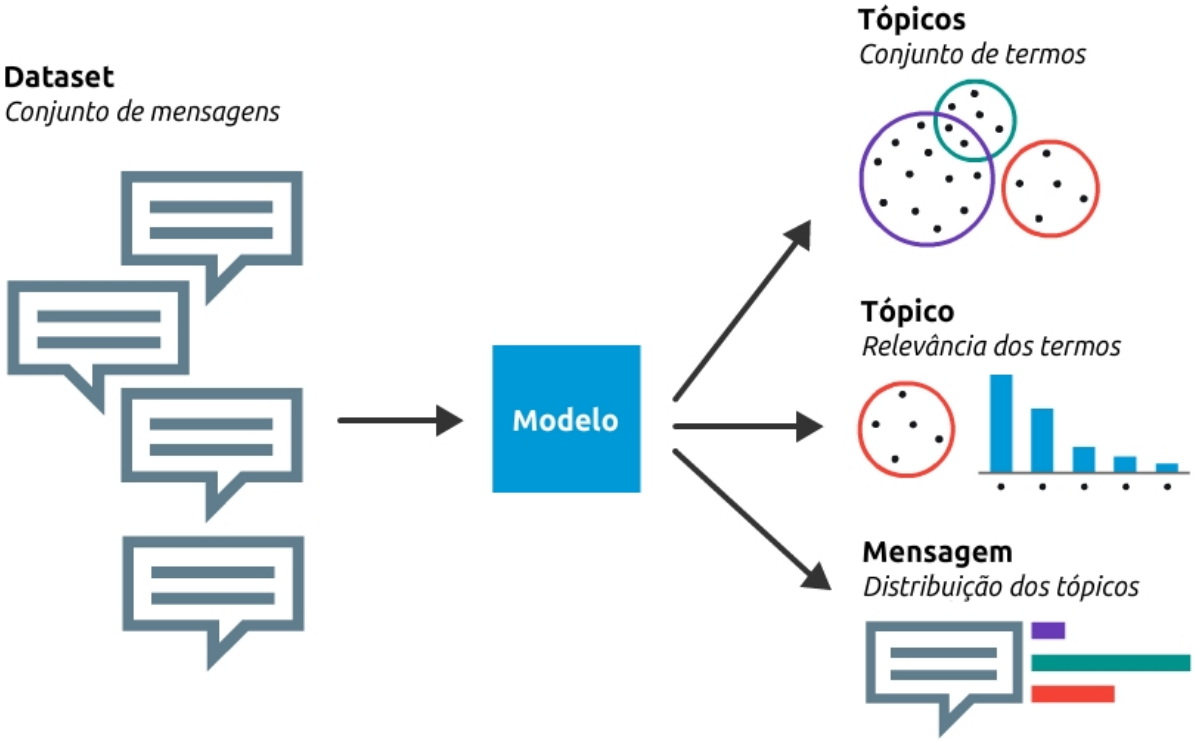
Na quarta etapa do processo de KDD, de mineração de dados, os textos convertidos para uma representação numérica por um dos processos de vetorização discutidos são analisados à procura de padrões e relacionamentos relevantes. As técnicas de mineração de dados aplicadas nesta etapa são a modelagem de tópicos e a análise de sentimentos.

3.1.4.1 Modelagem de tópicos

A modelagem de tópicos consiste no agrupamento automático de um conjunto de textos, por meio da identificação de relações semânticas estabelecidas pela co-ocorrência dos seus termos. A partir da identificação dessas relações, torna-se possível sumarizar os principais assuntos (tópicos) que compõem um texto, apresentando uma solução para

o desafio da inviabilidade da leitura completa de grandes volumes de textos de uma discussão *online* a serem analisados. Na modelagem de tópicos, as postagens coletadas de uma discussão são automaticamente agrupadas por relação semântica (LIU, 2012), onde cada grupo (tópico) pode ser caracterizado por seus termos mais relevantes, e cada texto (postagem) pode ser caracterizada pelos seus tópicos mais relevantes (Figura 11). Dessa forma, um tópico consiste em uma coleção de termos representativos dos textos que têm mais relação com o mesmo. Analisando esses termos, é possível identificar sobre qual assunto o tópico se refere, e rotulá-lo.

Figura 11 – Esquematização da técnica de modelagem de tópicos



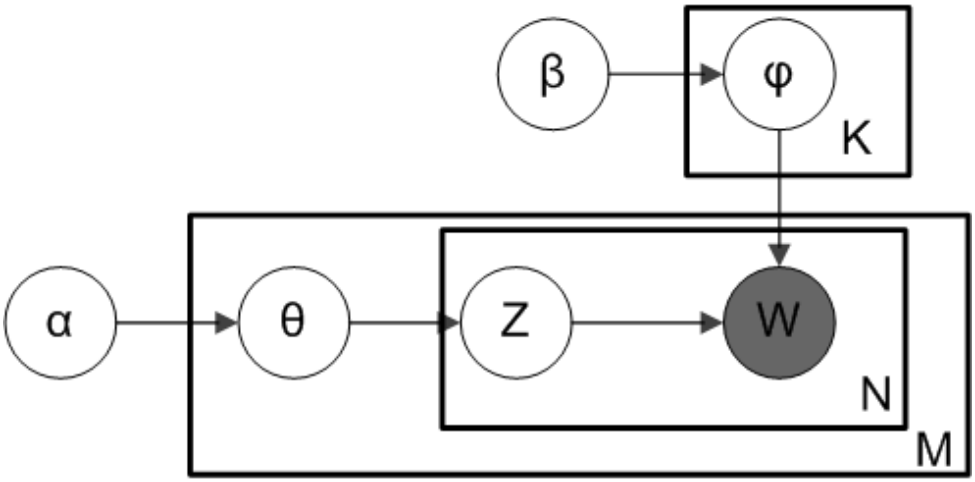
Fonte: elaborada pelo autor com base em Nabli et al. (2018)

Para o contexto das discussões virtuais, que têm por característica textos predominantemente curtos das mensagens, destacam-se na literatura dois algoritmos principais de modelagem de tópicos (CHEN et al., 2019): *Latent Dirichlet Allocation* (LDA) e *Non-negative Matrix Factorization* (NMF). Ambos recebem como entrada um conjunto de textos e a quantidade de tópicos esperada, e retornam como saída os termos mais relacionados a cada tópico e os tópicos mais relacionados a cada texto. Entretanto, seu funcionamento para modelar essas relações difere-se de forma considerável, sendo o LDA baseado em distribuições estatísticas e o NMF em operações com matrizes.

O LDA é um modelo probabilístico, que identifica tópicos com base na frequência dos termos em um conjunto de documentos (BLEI; NG; JORDAN, 2003). O algoritmo pressupõe que cada documento é composto por uma mistura, em diferentes proporções, dos K tópicos a serem identificados. Ele pode ser representado por meio da notação de

placas (*plate notation*), a qual representa variáveis que se repetem em um modelo de grafo que representa a estrutura de dependências condicionais probabilísticas entre variáveis (Figura 12).

Figura 12 – Representação do modelo LDA



Fonte: Blei et al. (2003)

Na execução do algoritmo, inicialmente é definida a quantidade K de tópicos a serem identificados. Feito isso, o LDA itera sobre cada termo W em todos os documentos da coleção M, associando o termo, aleatoriamente, a um dos K tópicos. Ao final da iteração, tem-se uma distribuição de termos por tópicos (φ) e de documentos por tópicos (ϑ), calculada com base na porcentagem de termos de cada tópico presente em cada documento. Por ter sido gerada de forma aleatória, essa representação obtida não é a mais otimizada de modo a refletir a real distribuição dos tópicos.

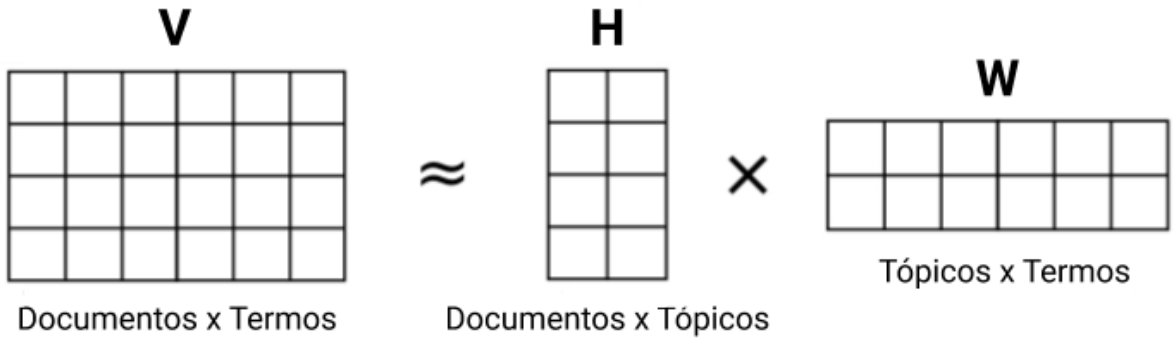
Para otimizar a representação produzida, para cada documento é analisada a porcentagem de termos do mesmo que foi atribuída a cada tópico. E, para cada termo do documento, a porcentagem que o mesmo foi atribuído a cada tópico, levando-se em conta todos os documentos nos quais ele ocorre. Caso o termo tenha mais probabilidade geral de estar em um tópico diferente daquele no qual ele foi associado inicialmente pelo algoritmo, este termo é movido para o tópico mais provável. Este processo é executado iterativamente, até que não haja mais termos a serem reassociados a outros tópicos – momento em que o LDA converge para uma representação otimizada para descoberta semântica de informações (NABLI; DJEMAA; AMOR, 2018).

Na Figura 12, K representa a quantidade de tópicos, M a quantidade de documentos e N a quantidade de termos por documento. O termo W é representado com fundo sombreado para indicar que é a única variável observável no sistema, enquanto as demais são variáveis latentes – ou seja, derivadas probabilisticamente.

Operando de forma diferente para segmentação dos textos em tópicos semânticos, o NMF é um modelo baseado no uso de álgebra linear, que utiliza fatoração de matrizes

(CICHOCKI; PHAN, 2009). Este algoritmo recebe como entrada uma matriz V representando os textos da discussão *online* vetorizados, e realiza a fatoração da mesma em duas matrizes resultantes W e H, onde o produto das mesmas aproxima-se de V (Figura 13).

Figura 13 – Representação do modelo NMF



Fonte: autor

Enquanto a matriz inicial V representa os textos (linhas) e os termos (colunas), a matriz resultante W representa os textos e os tópicos, e a matriz resultante H os tópicos e os termos. Pode-se analisar os valores numéricos da matriz W para identificar o peso de cada texto em relação a cada tópico, associando-o ao tópico de maior valor. E, de forma semelhante, identificar o peso de cada termo em relação a cada tópico, de modo a identificar quais são os termos de maior destaque (maior valor numérico) de cada tópico. Dessa forma, cada tópico identificado pode ser representado pelos termos mais relacionados numericamente ao mesmo na matriz H, cabendo ao usuário do algoritmo definir um texto descritivo para o mesmo.

Deve-se observar que, para os algoritmos LDA e NMF, a quantidade adequada de tópicos em que as postagens serão fatoradas depende significativamente das características do *dataset* e das escolhas realizadas nas etapas de processamento anteriores do processo de KDD. Esta escolha é incluída na ferramenta desenvolvida por meio do recurso interativo de definição da quantidade de tópicos, permitindo que o usuário pesquisador regule a granularidade dos assuntos produzidos pela análise – explorando diferentes configurações de tópicos.

3.1.4.2 Análise de sentimentos

Enquanto a técnica de modelagem de tópicos possibilita segmentar os assuntos de uma discussão *online*, a análise de sentimentos possibilita inferir a polaridade dos textos das postagens, em faixas numéricas ou classes, normalmente identificando sentimentos como positivos, negativos ou neutros. Assim como a modelagem de tópicos, a análise de sentimentos pode ser aplicada por meio de diferentes técnicas, sendo as mais comuns o treinamento supervisionado com aprendizado de máquina e o uso de léxicos (ANANDARAJAN; HILL; NOLAN, 2019).

Na análise de sentimentos baseada em aprendizado de máquina, são treinados modelos para aprenderem a partir de exemplos, recebendo como entrada amostras de texto (sua representação vetorial), e tendo como objetivo de aprendizado (*target*) a classe à qual pertencem (classificação) ou um valor numérico ao qual estão associados (regressão). Uma vez treinados, os modelos são capazes de generalizar o aprendizado, realizando previsões para novos textos de entrada que não faziam parte do conjunto de treinamento.

Existem diversos algoritmos que podem ser treinados para uso na análise de sentimentos, como redes neurais artificiais, árvores de decisão e máquinas de vetor de suporte. Apesar de terem implementações e características e desempenho diferentes, todos são algoritmos supervisionados, com uma etapa inicial de treinamento antes de realizar as previsões. Essa abordagem tem como vantagem permitir treinar o modelo com textos e seus respectivos termos específicos do domínio de problema a ser analisado. Entretanto, essa abordagem traz o desafio da criação manual de uma base inicial de treinamento, associando um grande volume de textos diversos ao sentimento predominante de cada um.

Uma opção de uso de análise de sentimentos baseada em aprendizado de máquina sem necessitar de treinamento para cada conjunto de textos a ser analisado consiste em se utilizar um modelo já treinado com grandes volumes de textos diversos. No contexto da análise de sentimentos aplicada a textos proveniente de mídias sociais como fóruns de discussão, destaca-se o Sentiment140 (GO; BHAYANI; HUANG, 2009), o qual combina a previsão de diferentes classificadores de aprendizado de máquina treinados com um grande volume de postagens do Twitter.

Já a análise de sentimentos baseada em léxicos utiliza um dicionário associando termos e pesos numéricos (léxico), os quais representam a polaridade e intensidade do sentimento de cada termo. Uma implementação de léxico para análise de sentimentos de mídias sociais é o *Valence Aware Dictionary and sEntiment Reasoner* (VADER), produzido a partir da análise de textos por dez avaliadores humanos independentes (HUTTO; GILBERT, 2014). Utilizando-se o VADER, para cada texto de entrada, ele é capaz de retornar individualmente a intensidade dos sentimentos positivo, negativo e neutro na faixa de 0 a 1, bem como um valor composto agregando os três valores, para identificação do sentimento predominante.

As técnicas de análise de sentimentos baseadas em aprendizado de máquina e em léxicos também podem ser integradas em modelos híbridos, que combinam as previsões das duas técnicas, buscando uma maior generalização nos resultados. RIBEIRO et al. (2016) avaliam diversas ferramentas de análise de sentimentos – baseadas em aprendizado de máquina, léxicos e híbridas – constatando que a taxa de acerto de cada técnica varia de acordo com as características do *dataset* a ser analisado. Dessa forma, como observam os autores, não é possível indicar a melhor técnica e ferramenta, de modo que sua escolha dependerá do contexto e das características do projeto em que serão empregadas. Uma vez

que a análise de sentimentos é aplicada aos textos segmentados pela modelagem de tópicos e não possui configurações para sua execução, a ferramenta não possui opção interativa com parâmetros a serem escolhidos pelo usuário pesquisador para esta técnica.

3.1.5 Etapa 5: Interpretação e Avaliação

Finalizando o ciclo do processo de KDD, na etapa de Interpretação e Avaliação, o resultado da mineração de dados é apresentado de forma visual e intuitiva. A sumarização dos padrões identificados possibilita a análise exploratória pelo pesquisador, selecionando os tópicos com potencial para responder às questões de pesquisa – com a possibilidade de uma nova execução do processo de KDD, reajustando-se os parâmetros com base nos resultados identificados. A ferramenta desenvolvida combina gráficos tradicionais como o de barras com visualizações populares e intuitivas comumente utilizadas em mineração de dados, como nuvens de palavras e *sparklines*[1] (ALENCAR; OLIVEIRA; PAULOVICH, 2012).

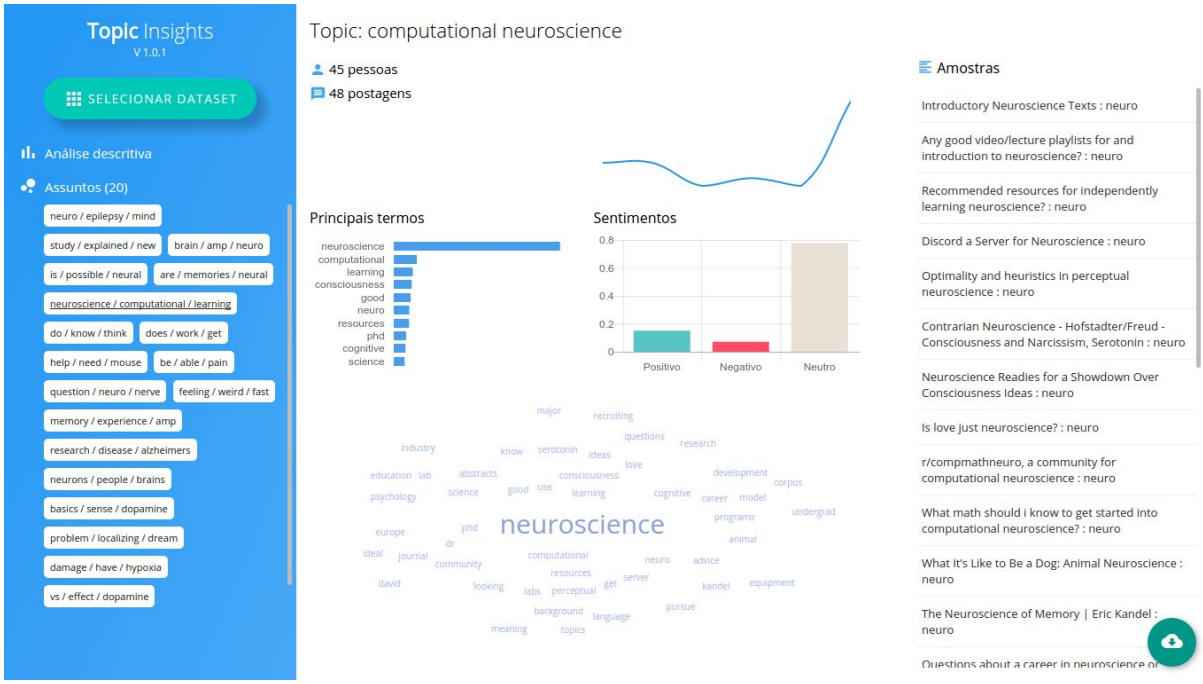
A Figura 14 apresenta o resultado da execução da ferramenta para uma determinada configuração das etapas de KDD por meio do seu processo de uso. À esquerda são apresentados os tópicos identificados, e à direita a visualização de um dos tópicos selecionado. O gráfico de barras horizontal apresenta os termos semânticos descritivos de cada tópico, por ordem de relevância. A partir desses termos é possível ao pesquisador identificar e anotar a descrição ou observações sobre o tópico identificado. Enquanto o gráfico de barras apresenta os poucos termos necessários para se identificar e descrever o tópico, a nuvem de palavras é composta pelos termos de maior frequência dos textos relacionados ao tópico, possibilitando identificar o contexto geral em que ele está inserido.

O gráfico de barras vertical apresenta a proporção dos sentimentos positivo, negativo e neutro (barras verde, vermelha e cinza, respectivamente), possibilitando identificar o sentimento predominante das discussões naquele tópico. O gráfico de linha compacto (*sparkline*) apresenta a variação temporal do volume de postagens do tópico ao longo do ciclo de vida da discussão. Constituindo um gráfico condensado capaz de sintetizar informações e transmitir a essência dos dados (TIRRONEN; WEBER, 2010), ele possibilita a identificação de tendências, como tópicos emergentes, perdendo popularidade ou recorrentes.

Também são exibidas amostras dos textos mais relevantes de cada tópico para inspeção pelo pesquisador, e os resultados produzidos nesta etapa também podem ser exportados no formato CSV. Dessa forma, os conhecimentos produzidos na análise podem ser persistidos para usos futuros em publicações e, por serem exportados em um formato interoperável, podem ser importados por editores de planilhas, linguagens de programação

[1] Gráficos condensados capazes de sintetizar informações e transmitir a essência dos dados.

Figura 14 – Visualização de dados de um tópico identificado pela ferramenta Topic Insights



Fonte: autor

e ferramentas de análise de dados como o Tableau[2] e o Power BI[3].

[2] https://www.tableau.com/pt-br
[3] https://powerbi.microsoft.com/en-us/

3.2 Implementação

Neste capítulo são discutidos os aspectos técnicos da etapa de implementação da ferramenta Topic Insights para análise de discussões *online* centrada no pesquisador. Como linguagem de programação principal para o desenvolvimento da ferramenta, optou-se pela linguagem Python[4], por ter licença de uso livre, código-fonte aberto e uma ampla comunidade de desenvolvedores. Além disso, ela é uma das linguagens de programação mais populares da atualidade, com grande destaque na área de análise de dados e desenvolvimento *web* (CASS, 2019). Buscando padronização e interoperabilidade, foram utilizados protocolos e formatos de dados abertos e padronizados, por não demandarem licenças de uso e possuírem comunidades ativas de desenvolvedores.

3.2.1 Arquitetura

A ferramenta foi desenvolvida utilizando a arquitetura Modelo-Visão-Controle, do inglês *Model-View-Controller* (MVC), um padrão de projeto de desenvolvimento de *software* que consiste em desacoplar um sistema em três componentes: camada de representação e manipulação de dados (*model*), camada de visualização e interação com o usuário (*view*) e camada de controle de requisições (*controller*). Essa divisão em camadas independentes que trocam informações entre si por meio de mensagens facilita o desenvolvimento e a manutenção do sistema, uma vez que isola abstrações de código e desacopla as regras de negócio da interface com o usuário. A arquitetura MVC foi implementada utilizando-se a biblioteca Python Flask[5], um *microframework web* Python, leve e flexível para o desenvolvimento de APIs *web*.

Na camada de modelo, é implementado o modelo de dados do sistema e as regras de negócio, compostos pela representação e manipulação dos dados de entrada das discussões *online* e pelos algoritmos de processamento e mineração de dados. Cada discussão *online* é representada na forma de um *dataset* no formato CSV, que deve ser copiado para o diretório */datasets* do servidor da aplicação. Este *dataset* estará visível e poderá ser analisado remotamente pelos pesquisadores, que não necessitarão de possuir uma cópia do mesmo em seus computadores.

A camada de controle expõe uma API RESTful de acesso à camada de modelo – a qual se caracteriza por utilizar os recursos do protocolo HTTP, como dialetos de mensagem (GET, POST, PUT, etc.) para informar a ação da chamada e códigos de informação (200, 404, 500, etc.) para indicar o *status* do retorno da chamada. A camada de controle intermedia as consultas dos usuários para acesso à camada de modelo, executando métodos remotos capazes de recuperar, modificar e realizar operações sobre os dados da aplicação.

[4] https://www.python.org
[5] https://palletsprojects.com/p/flask/

Os dados processados e/ou recuperados são retornados para a camada de visão, a qual constitui o ponto de interação dos usuários com o sistema. Constituindo uma interface gráfica com o usuário, ela transforma as ações do usuário em chamadas para a API exposta pela camada de controle. Por ser independente da lógica do sistema, a camada de visão pode ser implementada de diversas formas, como: uma aplicação *desktop*, uma página *web* ou um *app* para dispositivo móvel – contanto que façam a requisição à camada de controle seguindo o protocolo HTTP[6].

Seguindo a terminologia do desenvolvimento de aplicações *web*, a camada de visão constitui o *front-end* do sistema, composto pela interface com o usuário. Já as camadas de controle e modelo constituem o *back-end* do sistema, onde reside a lógica e os modelos de dados. Nas próximas seções são apresentadas as escolhas de projeto e tecnologias empregadas em cada parte do sistema, bem como seus protocolos.

3.2.2 *Back-end*

A modelagem e representação dos dados é um componente essencial para uma ferramenta de mineração de dados, de modo que a camada de modelo do MVC implementa a maior parte do sistema. Para manipulação de dados, optou-se pelo uso de bibliotecas da linguagem de programação Python específicas para essa tarefa, com foco no desempenho e compatibilidade com outros módulos. Para representação e realização de cálculos envolvendo matrizes utilizou-se a biblioteca Numpy[7] (*Numeric Python*), um módulo para computação científica de alto desempenho escrito em C/C++. Este módulo traz suporte para representação de matrizes multidimensionais por meio do tipo de dados *numpy.array*, o qual implementa métodos para operações matemáticas de matrizes, como *array.transpose*() para transposição (inversão das linhas e colunas).

Para a realização de consultas e operações complexas sobre as matrizes de dados, empregou-se a biblioteca Pandas[8] (*Python Data Analysis Library*), a qual encapsula estruturas do NumPy, fornecendo ferramentas de manipulação e análise de alto desempenho. A biblioteca Pandas adiciona à linguagem de programação Python o tipo de dados *pandas.DataFrame*, que permite manipular de forma fácil tabelas de dados e séries temporais, e adiciona novos recursos às matrizes do NumPy, como *labels*, índices, agrupamentos, junções e filtragem de dados. Como os dados representados nos DataFrames são textuais, provenientes de discussões *online*, utilizou-se o módulo NLTK[9] (*Natural Language Toolkit*) para o seu processamento. Este módulo implementa diversas técnicas para manipulação de dados textuais, como tokenização, produção de *n-grams*, identificação de classe gramatical (*parts-of-speech*) e listas de *stopwords* comuns.

[6] https://www.w3.org/Protocols/
[7] https://numpy.org/
[8] https://pandas.pydata.org/
[9] https://www.nltk.org/

Para modelagem de tópicos, que tem como objetivo identificar os principais assuntos semânticos abordados nos textos da discussão virtual, foi utilizada a biblioteca Scikit-learn[10], que possui implementações dos modelos NMF e LDA. Esta biblioteca Python possui uma API consistente e simples de manipular – além de ampla documentação técnica e guia teórico sobre os algoritmos e suas implementações[11].

A ferramenta foi desenvolvida para funcionar com textos de discussões *online* em inglês ou em português. O idioma é identificado de forma automática, analisando-se a predominância de preposições em português ou em inglês no conjunto de todos os textos. Enquanto a etapa de modelagem de tópicos é independente do idioma – por modelar a co-ocorrência de termos independentemente da sua semântica –, a análise de sentimentos requer bibliotecas próprias para cada idioma. Para o idioma inglês, foi utilizada a biblioteca VADER[12], um léxico produzido a partir da avaliação de dez avaliadores humanos independentes, com foco na análise de textos de mídias sociais. De modo a manter a consistência do código, para textos em português é utilizada a biblioteca de Léxico para Inferência Adaptada (LeIA)[13], um *fork* do léxico VADER, adaptado para textos em português.

De modo a possibilitar que a ferramenta possa ser instanciada com diferentes modelos de mineração de dados e evitar o retrabalho com codificação para alterar esses modelos ou integrar novos modelos, é utilizada programação orientada a objetos e sua técnica de polimorfismo. Ela permite que se manipule de forma homogênea instâncias de classes com diferentes comportamentos, por meio da referência a uma classe abstrata que define uma interface a ser seguida.

O uso do polimorfismo torna a ferramenta independente dos objetos que representam os módulos de mineração de dados e suas implementações, contanto que eles sigam a mesma API de chamada. Por exemplo, seguindo o padrão da biblioteca *scikit-learn*, os módulos de modelagem de tópicos da ferramenta devem implementar o método $.fit\_transform(representacao\_vetorial)$, o qual recebe os textos vetorizados, treina o modelo e retorna os tópicos identificados. Apesar de apresentarem a mesma assinatura, instâncias dos modelos NMF e LDA comportam-se de maneira diferente para cada chamada, mas retornam os dados dos tópicos identificados em um mesmo formato, de modo que a ferramenta possa processá-los.

Uma vez processada a requisição do usuário pela camada de modelo, a camada de controle do *back-end* retorna o conteúdo da resposta para ser renderizada na camada de visão. Esta resposta utiliza a Notação de Objetos JavaScript[14], do inglês *JavaScript*

[10] https://scikit-learn.org
[11] https://scikit-learn.org/stable/modules/decomposition.html
[12] https://github.com/cjhutto/vaderSentiment
[13] https://github.com/rafjaa/LeIA
[14] https://www.json.org

Object Notation (JSON), um formato padronizado com foco na interoperabilidade de dados, que possui analisadores (*parsers*) nativos na maioria das linguagens de programação. Os dados no formato JSON são retornados via HTTP para a interface de usuário, que os analisa e renderiza os componentes gráficos com os resultados da mineração de dados e as visualizações do *front-end*.

3.2.3 *Front-end*

Constituindo a camada de visão do MVC, o *front-end* da ferramenta foi desenvolvido na forma de uma página *web*, de modo a ser independente de dispositivo e sistema operacional, demandando do usuário apenas acesso a um navegador de internet. Utilizou-se as tecnologias HTML5[15] e CSS3[16] para a estruturação e estilização visual da página, e a linguagem JavaScript para o tratamento de eventos e a comunicação com o servidor. Para o estilo de design da página, foi utilizado o padrão *Material Design*, uma linguagem de design padronizada e bem documentada proposta pela Google, que sintetiza os princípios clássicos do bom design com a inovação da tecnologia e da ciência (MEW, 2015).

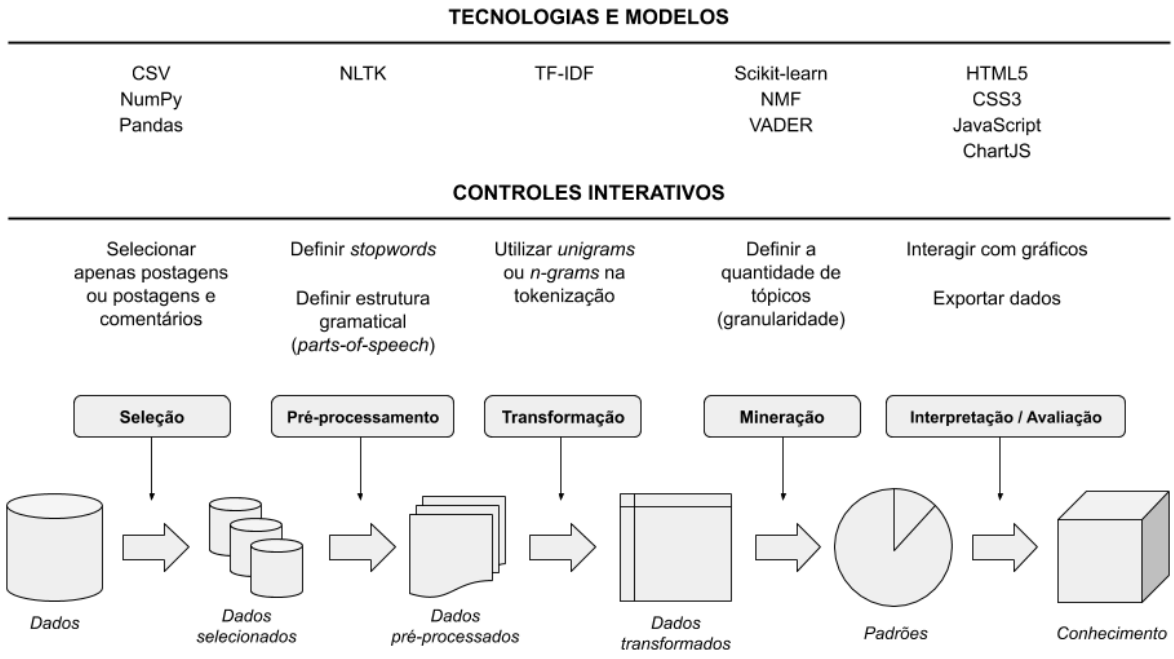
As requisições da camada de visão no *front-end* para a camada de controle no *back-end* utilizam a técnica *Asynchronous JavaScript and XML* (AJAX). Nela, a linguagem de programação JavaScript do navegador realiza requisições HTTP assíncronas, trocando mensagens com o servidor e atualizando a interface com os dados recebidos. Por não necessitar da atualização da página a cada requisição AJAX, é possível trazer mais dinamismo e interação à interface com o usuário – abstraindo o fato de que a maioria dos processamentos ocorrem em um servidor remoto.

Os dados recebidos nas requisições HTTP via AJAX estão abstraídos no formato JSON, que por ser o formato nativo das estruturas de dados da linguagem de programação JavaScript, podem ser processados de forma simples para atualizar as informações da página. Os dados em JSON também são utilizados como entrada para a biblioteca JavaScript ChartJS[17], a qual os utiliza para renderização dos gráficos que sumarizam visualmente os resultados.

As tecnologias, modelos e controles interativos utilizados em cada etapa do processo de KDD na ferramenta desenvolvida são sumarizados na Figura 15. Sua instanciação para avaliação com usuários e estudo de caso empregou o processo de vetorização TF-IDF, o modelo de segmentação de tópicos NMF e o léxico para análise de sentimentos VADER. Essa escolha se justifica por serem modelos simples, intuitivos e de baixo custo computacional, possibilitando uma execução rápida dos algoritmos e uma melhor resposta da ferramenta durante seu processo de uso.

[15] https://html.spec.whatwg.org/multipage/
[16] https://www.w3.org/TR/css3-roadmap/
[17] https://www.chartjs.org/

Figura 15 – Tecnologias, modelos e controles interativos relacionados a cada etapa do processo de KDD na ferramenta desenvolvida



Fonte: autor

3.3 Processo de uso

Uma das principais limitações observadas em ferramentas para análise de discussões *online*, as quais possuem teor predominantemente qualitativo, consiste na ausência de propostas de processos metodológicos para as análises, limitando seu rigor e reprodutibilidade. Dessa forma, torna-se relevante a definição de um processo de uso que possibilite estabelecer etapas claras para condução da análise de discussões virtuais, garantindo que sejam produzidos resultados fundamentados nos dados capazes de responder às questões de pesquisa.

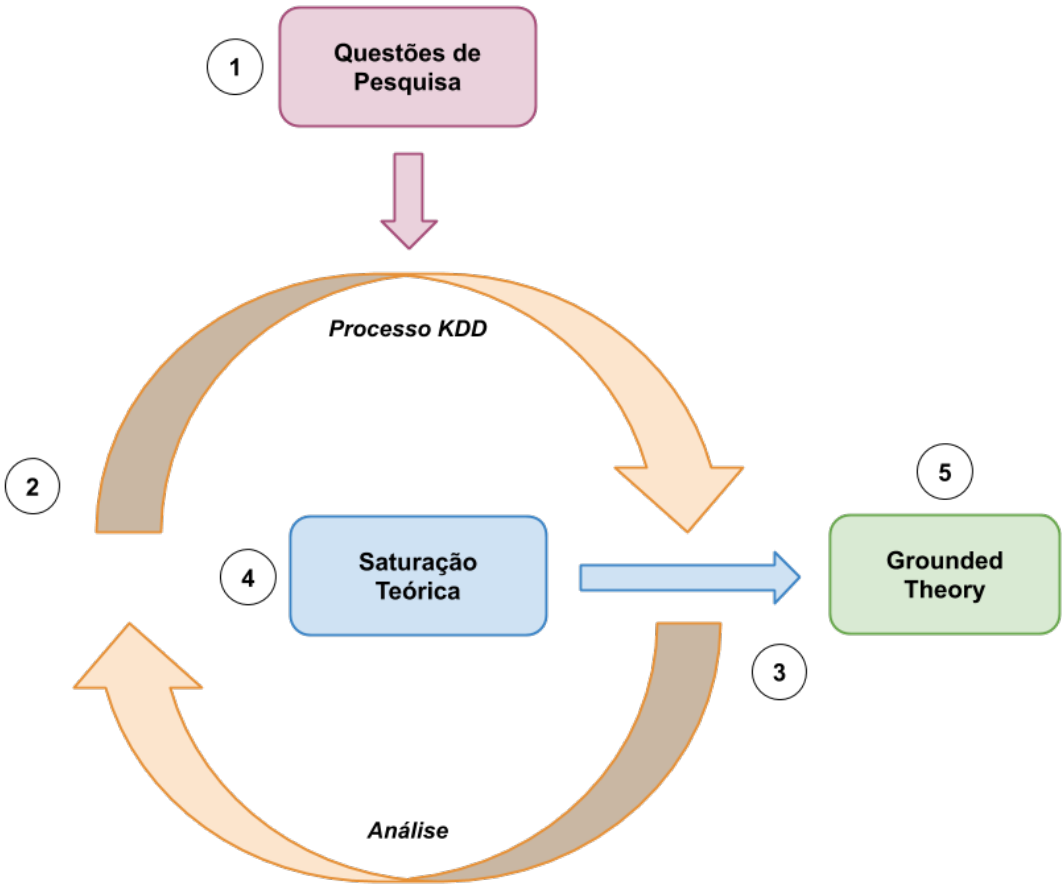
Devido ao caráter interativo, exploratório e orientado a dados das análises de discussões *online* centradas no pesquisador com a ferramenta proposta, seu processo de uso baseia-se na Teoria Fundamentada em Dados, do inglês *Grounded Theory* (GT). Este método fornece uma abordagem metodológica para pesquisas qualitativas, e seu uso em pesquisas de Sistemas de Informação vem aumentando gradualmente ao longo dos anos, à medida que a pesquisa qualitativa em geral se torna mais prevalente (URQUHART; FERNÁNDEZ, 2016).

Assim como o processo de KDD, a Teoria Fundamentada em Dados caracteriza-se por um processo circular, onde o pesquisador busca significados nos dados por meio de sua coleta, codificação e análise comparativa contínuas. A comparação constante de instâncias de dados a serem categorizados em busca de significado é a técnica de condução do seu

processo de análise, sendo executada iterativamente até que se atinja a saturação teórica – momento onde as categorias identificadas (tópicos) começam a se estabilizar e novas análises não trazem novos resultados.

Este processo de investigação integra-se naturalmente ao fluxo de uso da ferramenta desenvolvida para responder às questões de pesquisa propostas nas análises, podendo ser observadas diversas similaridades entre a Teoria Fundamentada em Dados e a mineração de dados – principalmente envolvendo técnicas de modelagem de tópicos. Como observam MULLER et al. (2016), ambos processos buscam se fundamentar nos dados, nos quais começam e retornam de forma iterativa, e só terminam após a interpretação e construção de teorias em busca de um novo conhecimento – requerendo crucialmente interpretação e julgamento humano em cada etapa. A Figura 16 ilustra o processo iterativo de análise dos dados e sua ordem de execução, realizado por meio do uso da ferramenta desenvolvida.

Figura 16 – Processo de uso da ferramenta



Fonte: autor

O processo de uso proposto integra o ciclo de KDD à Teoria Fundamentada em Dados, de modo que as etapas do processo de descoberta do conhecimento com participação do usuário especialista são executadas sucessivas vezes – até o momento em que não seja possível produzir conhecimentos capazes de responder às questões de pesquisa. Complementarmente, a combinação do processo de KDD com a Teoria Fundamentada em

Dados deve ser pautada pelas práticas de pesquisa, coleta e análise de dados estabelecidas pela Netnografia, de modo a assegurar um rigor metodológico e preservar a privacidade dos participantes da discussão analisada. O fluxo típico de uma análise de discussão *online* aplicando o processo de uso proposto é exemplificado a seguir.

A análise se inicia a partir de uma demanda de pesquisa a ser realizada, investigando determinado campo e seu contexto. Por exemplo, suponha que a empresa que desenvolve um *software* tem interesse na caracterização dos recursos do mesmo pelos seus usuários, de modo a direcionar os esforços de seus desenvolvedores. Avaliações presenciais com usuários demandam tempo e possuem limitações quanto a quantidade e variedade de perfis dos participantes. Dessa forma, a empresa opta por analisar a opinião de um grande volume de usuários a partir de seus comentários sobre o *software* em discussões *online*. Para isso, o primeiro passo será a definição, por parte dos pesquisadores da empresa, de questões de pesquisa pertinentes. Questões de pesquisa relevantes para a empresa poderiam ser: "Q1: Quais são os recursos mais populares do *software*?" e "Q2: Quais são os recursos mais criticados pelos usuários do *software*?".

Definidas as questões de pesquisa, o próximo passo consiste na identificação e seleção do ambiente de discussão que será analisado, avaliando seu potencial para responder às questões propostas. No contexto do exemplo da empresa de *software*, este ambiente pode ser um fórum sobre programas de computador, os comentários nas páginas do *software* em mídias sociais ou trocas de mensagens de usuários do *software* em grupos de mensagens instantâneas de desenvolvedores como o Slack[18]. Caso discussões de diferentes fontes sejam relevantes e complementares para a análise, elas poderão ser coletadas em conjunto, de modo a produzir um *dataset* mais rico e completo.

A ferramenta desenvolvida recebe como entrada as discussões coletadas, persistidas em um arquivo no formato *Comma-Separated Values* (CSV). A coleta dos dados pode ser realizada por meio de Interfaces de Programação de Aplicativos, do inglês Application Programming Interfaces (APIs) ou coletores *web* (*web crawlers*). APIs são interfaces de chamada remota que permitem que um aplicativo recupere dados de uma plataforma *online*, como uma rede social. O uso de APIs representa uma solução mais simples para coleta de dados, uma vez que costumam ser bem documentadas e utilizar padrões de comunicação e representação de dados interoperáveis. Entretanto, algumas APIs podem impor limitações de acesso às consultas, como limite máximo de chamadas e restrições nos dados retornados. Como alternativa, podem ser programados coletores *web* que simulam uma navegação nas páginas da plataforma *online*, coletando o conteúdo das mesmas à medida que são acessadas – buscando burlar as limitações das APIs.

Coletores são mais complexos de se implementar, uma vez que devem ser programados especificamente para cada tipo de página que precisarão navegar e identificar os dados

[18] https://slack.com

a serem coletados, muitas vezes em meio a códigos pouco estruturados e padronizados. Além disso, deve-se observar a questão legal da política de uso da plataforma cujos dados estão sendo coletados. Após a coleta dos dados da discussão em que será conduzida a pesquisa, deve-se realizar um processo de anonimização das informações pessoais de seus participantes. Esta etapa é importante para proteger sua privacidade, garantindo o anonimato e a confidencialidade dos membros da discussão.

Com os dados das discussões *online* de interesse identificados, coletados e anonimizados, inicia-se a sua análise interativa, por meio da ferramenta proposta. Nela, o pesquisador pode configurar interativamente parâmetros de cada etapa do processo de KDD, iniciando pela etapa de seleção, e obtendo ao final da etapa de interpretação/avaliação o conjunto de assuntos identificados e suas estatísticas. Este resultado parcial é avaliado pelo pesquisador, que identifica e nomeia os assuntos pertinentes para responderem às questões de pesquisa propostas. Este processo se repete iterativamente, com o pesquisador explorando as configurações das etapas de KDD.

No exemplo da empresa de *software*, uma primeira execução do processo poderia ajustar a etapa de seleção, restringindo a análise somente às postagens e excluindo-se os comentários às mesmas, de modo a identificar os assuntos gerais postados. Em uma segunda iteração, os comentários às postagens poderiam ser incluídos na análise, de modo a aprofundar a descoberta de assuntos mais específicos. Já em uma terceira execução, a etapa de pré-processamento poderia ser ajustada para processar apenas substantivos e adjetivos dos textos, identificando assuntos referentes a recursos do *software* e a opinião dos usuários.

O processo de análise finaliza quando, após múltiplas execuções do processo de KDD, não são mais identificados assuntos relevantes para a pesquisa, atingindo o estágio de saturação teórica da Teoria Fundamentada em Dados. Nesta etapa final, os novos conhecimentos produzidos são utilizados para responder às questões de pesquisa propostas, estabelecendo uma narrativa ou teoria fundamentada nos dados.

No próximo capítulo é realizada a avaliação da ferramenta desenvolvida e de seu processo de uso com dados de discussões virtuais do mundo real. A avaliação de usabilidade e aceitação é realizada por meio de um experimento com usuários, que empregaram a ferramenta e seu processo na análise de um fórum de discussão voltado para iniciantes em programação. Complementarmente, é conduzido um estudo de caso exploratório pelo autor, analisando uma discussão *online* sobre neurociência, demonstrando a aplicabilidade da Topic Insights, e buscando responder às questões de pesquisa propostas.

4 Processo de avaliação

4.1 Avaliação com usuários

A avaliação do artefato proposto e desenvolvido (ferramenta e seu respectivo processo de uso) está prevista no Ciclo do *Design* do método DSR, onde é avaliado se os objetivos para a solução do problema foram atendidos. O artefato foi avaliado com usuários por meio do método de avaliação de usabilidade *System Usability Scale* (SUS) e do modelo de aceitação de tecnologia *Unified Theory of Acceptance and Use of Technology* 2 (UTAUT 2). Os resultados das duas avaliações são relacionados por meio da triangulação dos métodos, possibilitando uma discussão geral em relação à avaliação do artefato. A triangulação consiste na combinação de métodos diferentes para analisar o mesmo fenômeno, de modo a fundamentar a construção de teorias e consolidar as conclusões a respeito do que está sendo investigado (DENZIN; LINCOLN, 2008). Ela possibilita validar e ampliar as interpretações da avaliação, adotando diferentes perspectivas e visões, por meio da combinação de diferentes tipos de dados sob a mesma abordagem teórica para a produção de mais conhecimento do que seria possível com base em uma só perspectiva (FLICK, 2012).

Para condução do experimento, foi submetido um projeto para apreciação pelo comitê de ética em pesquisas, por meio do sistema eletrônico da Plataforma Brasil, que foi avaliado e aprovado pelo Comitê de Ética em Pesquisa Envolvendo Seres Humanos das Unidades Educacionais de São João del-Rei (CEPSJ), da Universidade Federal de São João del-Rei (UFSJ), sendo CAAE 11970919.2.0000.5151 o código do processo registrado na Plataforma Brasil (Apêndice A). Com a pesquisa aprovada, foi reunido um grupo de participantes para avaliação, consistindo em 20 alunos da graduação de Tecnologia em Sistemas para Internet do Instituto Federal do Sudeste de Minas Gerais, *Campus* Barbacena, onde o autor desta dissertação é docente. A graduação de Tecnologia em Sistemas para Internet é um curso superior de curta duração (3 anos) na modalidade tecnólogo, com enfoque no desenvolvimento de sistemas baseados em tecnologias *web*. Nos três períodos iniciais os alunos estudam, dentre outras disciplinas, Lógica de Programação, Estrutura de Dados, Banco de Dados e Redes de Computadores – todas com uma abordagem teórica aliada ao desenvolvimento de projetos práticos. Nos demais períodos, são abordadas disciplinas voltadas para o desenvolvimento e a integração de sistemas e plataformas, como *Web Services*, Sistemas Distribuídos, Serviços e Gerência de Redes de Computadores e Programação Móvel.

O critério de inclusão dos participantes foi definido como ter cursado ou estar cursando a disciplina de Lógica de Programação ofertada no primeiro período do curso –

uma vez que a discussão virtual a ser analisada no experimento foi coletada de um fórum voltado para iniciantes em programação[1], com aproximadamente 3.000 registros. Dessa forma, os participantes da avaliação podem ser considerados especialistas no domínio das discussões a serem analisadas por meio da ferramenta e seu processo de uso. Além disso, todos os participantes gozam de maioridade e plena capacidade de ação.

Antes da avaliação os participantes leram, concordaram e assinaram um Termo de Consentimento Livre e Esclarecido (TCLE), observando que sua participação seria voluntária e não obrigatória, e com informações sobre as medidas tomadas para resguardar sua privacidade e sua integridade (Apêndice B). O experimento foi realizado no Laboratório de Redes de Computadores do Instituto Federal do Sudeste de Minas Gerais, Campus Barbacena, e a participação de cada um consistiu em três etapas, com duração máxima total de 30 minutos. Foram explicados aos participantes alguns aspectos em relação ao contexto da avaliação, como o propósito da ferramenta e seu respectivo processo de uso desenvolvidos como parte da pesquisa de mestrado apresentada na presente dissertação, bem como que a mesma é de uso livre, não possuindo custos para sua utilização.

Na Etapa 1, com até 5 minutos de duração, a ferramenta foi apresentada para os participantes do experimento, descrevendo seu processo de uso e suas funcionalidades, envolvendo: selecionar e carregar os dados de uma discussão *online*, selecionar interativamente os parâmetros de análise do processo de Descoberta de Conhecimento em Bases de Dados (KDD), navegar e interpretar os gráficos e indicadores produzidos pela análise e exportar os resultados. Esta etapa consistiu em uma apresentação básica do uso da ferramenta e seu processo, não sendo proporcionado mais nenhum auxílio técnico pelo aplicador da avaliação nas etapas subsequentes.

A segunda etapa, com até 15 minutos de duração, consistiu na apresentação aos participantes de questões de pesquisa a serem respondidas por meio da análise interativa do fórum de discussão sobre programação, utilizando a ferramenta e seu processo de uso. A tarefa envolvia carregar a base de dados, iniciar a análise configurando livremente o valor dos parâmetros, explorar e interpretar os resultados e exportar os dados gerados, identificando a partir dos gráficos e indicadores: 1) três assuntos de destaque; 2) um assunto predominantemente negativo; 3) um assunto predominantemente positivo; 4) um assunto recorrente e 5) um assunto emergente.

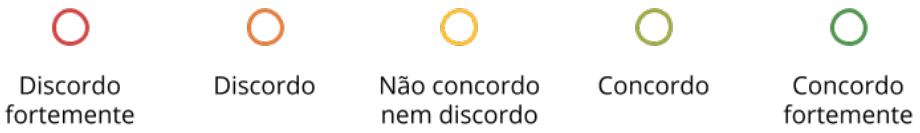
Na Etapa 3, com até 10 minutos de duração, os participantes responderam aos questionários SUS e UTAUT 2, em um formulário anônimo *online* desenvolvido na ferramenta Google Forms, finalizando o experimento. A seguir são apresentados os métodos SUS e UTAUT2, bem como a análise dos resultados da avaliação da ferramenta e seu processo de uso com usuários.

[1] www.clubedohardware.com.br/forums/forum/181-programação-iniciantes

4.1.1 Avaliação de usabilidade

O método utilizado para esta avaliação foi o *System Usability Scale*, um método padronizado amplamente adotado em pesquisas científicas para a avaliação de usabilidade e aceitação de produtos e serviços (SAURO, 2011). Ele consiste em um instrumento de avaliação do tipo questionário, com dez questões na escala de Likert (Figura 17), cujas respostas variam entre "1 – Discordo fortemente" a "5 – Concordo fortemente" (BROOKE et al., 1996): 1) Acho que gostaria de utilizar a ferramenta com frequência; 2) Achei a ferramenta mais complexa do que o necessário; 3) Achei a ferramenta fácil de usar; 4) Acho que precisaria de ajuda de um técnico para conseguir usar a ferramenta; 5) Considero que as várias funcionalidades da ferramenta estão bem integradas; 6) Acho que a ferramenta apresenta muitas inconsistências; 7) Suponho que a maioria das pessoas aprenderia a utilizar a ferramenta rapidamente; 8) Considerei a ferramenta muito complicada de utilizar; 9) Eu me senti confiante ao usar a ferramenta; 10) Eu precisei aprender várias coisas novas antes de conseguir usar a ferramenta.

Figura 17 – Faixa de respostas às questões do formulário SUS



Fonte: autor

O SUS permite quantificar quão bem os usuários conseguem interagir com um dado produto ou serviço, medindo a usabilidade e suas dimensões de eficácia, eficiência e satisfação em determinado contexto de uso (STANDARDIZATION, 2018). A eficácia reflete a habilidade do usuário em efetuar uma tarefa atingindo seus objetivos, enquanto a eficiência descreve os recursos gastos pelo usuário para realizar a tarefa – geralmente o recurso tempo, quando avaliados sistemas de informação. Já satisfação relaciona-se à percepção do usuário de quão bem o produto atendeu às suas necessidades e objetivos.

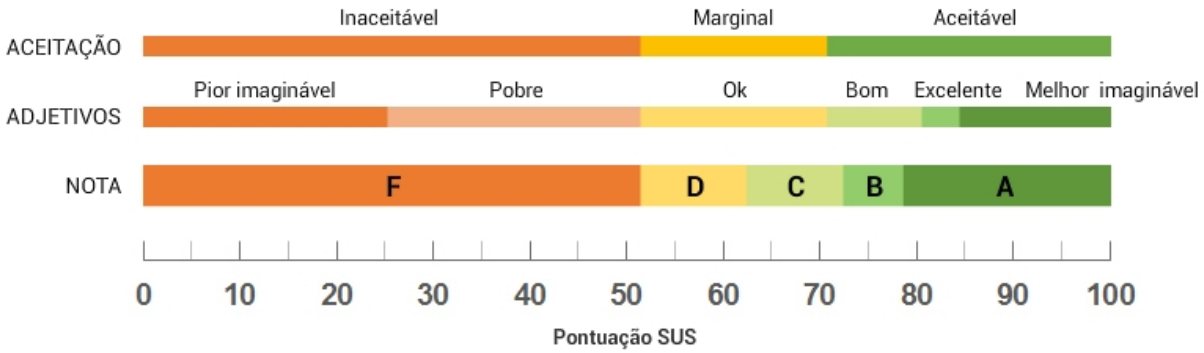
Além de ser um instrumento disponível gratuitamente, sem taxas de uso ou de tabulação, este método tem como vantagens o fato de ser agnóstico em relação à tecnologia. Desta forma, ele pode ser utilizado para avaliar uma variedade de produtos e serviços, incluindo *sites*, *hardware*, sistemas de interação, sistemas de votação, *apps*, sistemas médicos, dentre outros (KORTUM; BANGOR, 2013).

Outra vantagem na adoção deste método consiste no mesmo possibilitar calcular o índice numérico da usabilidade geral do sistema. Cada uma das dez questões alterna entre perguntas positivas e negativas sobre o produto a ser avaliado. Desta forma, para se calcular o índice de usabilidade, deve-se subtrair o valor 1 da pontuação das respostas às questões ímpares, e subtrair o valor das respostas pares do valor 5. Após este procedimento,

os valores de todas as perguntas devem ser somados e multiplicados por 2,5, de modo que os resultados possíveis fiquem normalizados entre 0 e 100.

Este formato de resultado numérico proporciona uma linguagem comum e intuitiva para que desenvolvedores do produto avaliado descrevam e discutam seus atributos de usabilidade, e possam comparar suas versões (KORTUM; BANGOR, 2013). Segundo a escala do SUS, valores inferiores a 50 pontos indicam usabilidade muito ruim, entre 51 e 64 indicam usabilidade ruim, entre 65 e 74 indicam usabilidade neutra, valores iguais ou superiores a 75 pontos indicam uma boa usabilidade e valores iguais ou superiores a 80 pontos indicam uma usabilidade muito boa. BANGOR; KORTUM; MILLER (2009) apresentam um meta-estudo onde validam a confiabilidade do método SUS analisando centenas de estudos realizados, e propõem uma escala com faixas de aceitação, bem como adjetivos e notas para auxiliar na interpretação das pontuações (Figura 18).

Figura 18 – Faixas de aceitação para auxiliar na interpretação das pontuações do SUS



Fonte: elaborada pelo autor com base em Bangor et al. (2009)

Além da avaliação da nota numérica final do SUS, pode ser realizada a análise das estatísticas descritivas como média e desvio padrão das respostas dos participantes para cada questão – possibilitando uma compreensão mais detalhada da percepção dos usuários. Também é possível associar conjuntos de questões específicas do SUS a algumas Heurísticas de Nielsen (NIELSEN; MOLICH, 1990), as quais constituem princípios gerais de *design* amplamente adotados no desenvolvimento e avaliação de sistemas. As heurísticas de Nielsen relacionadas às questões do SUS são: flexibilidade e eficiência de uso (questões 5, 6 e 8), reconhecimento em vez de memorização (questão 2) e prevenção de erros (questão 6) (TENÓRIO et al., 2010). Já em relação às dimensões da usabilidade definidas pela STANDARDIZATION (2018), a facilidade de aprendizagem é contemplada pelas questões 3, 4, 7 e 10, enquanto a satisfação relaciona-se às questões 1, 4 e 9.

Os dados coletados foram tabulados e suas estatísticas descritivas são apresentadas na Tabela 2. O resultado final normalizado da avaliação da ferramenta e seu processo de uso por meio do *System Usability Scale* foi de 84,5 pontos, indicando uma alta taxa de aceitação na escala proposta por BANGOR; KORTUM; MILLER (2009). Obtendo a

nota “A” da escala, e adjetivo “excelente”, este resultado positivo da avaliação indica a facilidade de uso da ferramenta e seu processo na análise interativa de discussões *online*.

Tabela 2 – Estatísticas descritivas das respostas ao questionário SUS, onde a coluna DP representa o desvio padrão

Questão	**Média**	**Mínimo**	**Máximo**	**DP**
1) Acho que gostaria de utilizar a ferramenta com frequência	4,20	3	5	0,70
2) Achei a ferramenta mais complexa do que o necessário	1,45	1	3	0,60
3) Achei a ferramenta fácil de usar	4,25	3	5	0,72
4) Acho que precisaria de ajuda de um técnico para conseguir usar a ferramenta	2,10	1	4	0,97
5) Considero que as várias funcionalidades da ferramenta estão bem integradas	4,60	4	5	0,50
6) Acho que a ferramenta apresenta muitas inconsistências	1,15	1	2	0,37
7) Suponho que a maioria das pessoas aprenderia a utilizar a ferramenta rapidamente	4,05	3	5	0,76
8) Considerei a ferramenta muito complicada de utilizar	1,35	1	2	0,49
9) Eu me senti confiante ao usar a ferramenta	4,20	3	5	0,70
10) Eu precisei aprender várias coisas novas antes de conseguir usar a ferramenta	1,45	1	3	0,60

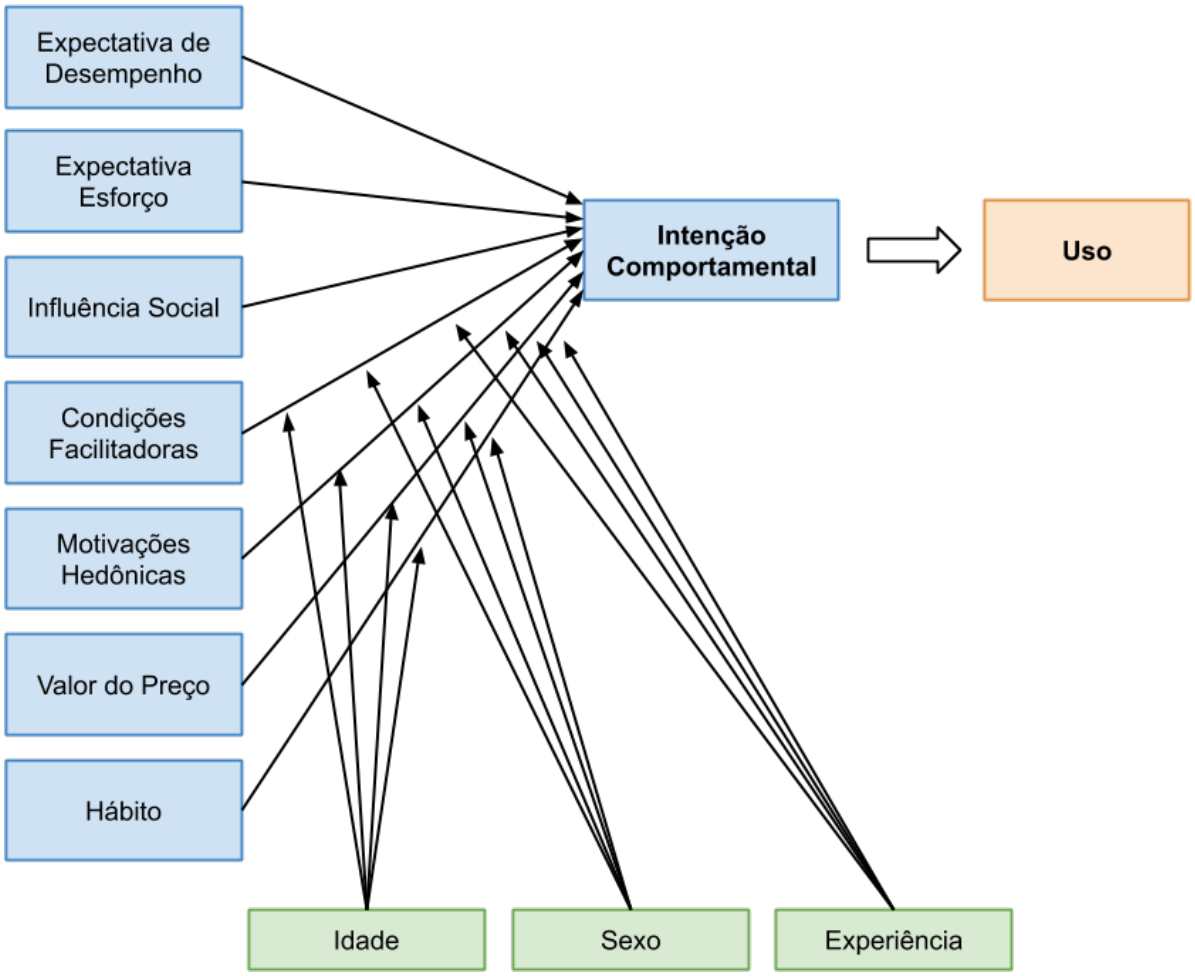
4.1.2 Avaliação de aceitação

Para a avaliação de aceitação da ferramenta e do processo de uso desenvolvidos, foi utilizado o modelo de aceitação de tecnologia UTAUT. Este modelo teórico unificado foi proposto por VENKATESH et al. (2003), analisando e integrando elementos de oito modelos e teorias que ajudam a explicar fatores relacionados à aceitação e ao uso de tecnologias. Para sua publicação, o modelo UTAUT foi validado empiricamente, utilizando-se dados de quatro organizações durante um período de seis meses, onde superou os oito modelos individuais. Na presente pesquisa, foi utilizada a versão 2 do modelo UTAUT (VENKATESH; THONG; XU, 2012), a qual aprimora a proposta inicial do modelo, possibilitando analisar o impacto de novos aspectos na aceitação, como prazer de uso, preço e hábito.

A partir da necessidade de se compreender o comportamento do usuário diante da tecnologia, o modelo UTAUT 2 permite identificar fatores que levam um usuário a utilizar uma tecnologia em diferentes contextos, e explicar suas intenções de uso. Para isso, ele se baseia em sete construtos e três moderadores (Figura 19). Os construtos constituem os aspectos capazes de influenciar no uso da tecnologia analisada: expectativa de desempenho,

expectativa de esforço, influência social, condições facilitadoras, motivação hedônica, valor do preço e hábito. Já os moderadores são variáveis com potencial para alterar a intensidade do impacto dos construtos: idade, gênero e experiência.

Figura 19 – Modelo UTAUT 2



Fonte: elaborada pelo autor com base em Venkatesh et al. (2012)

Para a presente pesquisa, foram aplicados os construtos: expectativa de desempenho, expectativa de esforço, condições facilitadoras, motivação hedônica e valor do preço. O construto relacionado ao hábito não foi utilizado no modelo por demandar que a tecnologia avaliada já esteja em uso prévio – o que não é o caso da ferramenta e processo de uso desenvolvidos na pesquisa. Da mesma forma, o construto de influência social também não foi empregado, uma vez que o artefato avaliado não está presente no dia-a-dia dos usuários de forma a ter um impacto social em relação ao seu uso. Os moderadores de idade, gênero e experiência também não foram aplicados no modelo, uma vez que não representam modificadores relevantes para o contexto da análise.

A aplicação do UTAUT 2 utilizou um instrumento de avaliação do tipo questionário, composto por 21 questões na escala de Likert, em uma escala de sete pontos: 1) Discordo Totalmente; 2) Discordo em grande parte; 3) Discordo; 4) Neutro; 5) Concordo; 6) Concordo

em grande parte; 7) Concordo Totalmente. Ao contrário do SUS, no questionário UTAUT 2 todas as questões são afirmações onde quanto maior o valor de concordância da resposta, melhor o resultado da avaliação do artefato. As questões estão distribuídas de acordo com cada um dos construtos empregados e com a intenção comportamental (Tabela 3), sendo baseadas na validação e adaptação transcultural do modelo UTAUT 2 para o contexto e idioma brasileiro, realizado por NISHI et al. (2017).

Tabela 3 – Construtos e questões do questionário UTAUT 2

Construto	**Questões**
Expectativa de desempenho: grau em que o usuário acredita que a utilização do artefato avaliado o auxilia para atingir um desempenho adequado nas tarefas realizadas com o mesmo.	D1: Eu acho a ferramenta útil no meu dia-a-dia. D2: Usar a ferramenta aumenta minhas chances de conseguir coisas que são importantes para mim. D3: Usar a ferramenta me ajuda a realizar as coisas mais rapidamente. D4: O uso da ferramenta aumenta a minha produtividade.
Expectativa de esforço: grau de esforço que o usuário espera demandar para realizar tarefas com o artefato, sendo um esforço menor demandado por uma capacidade do mesmo em facilitar a execução das tarefas.	E1: Acho a ferramenta fácil de usar. E2: É fácil para mim ficar habilidoso(a) no uso da ferramenta. E3: Minha interação com a ferramenta é clara e compreensível. E4: Aprender a usar a ferramenta é fácil para mim.
Condições facilitadoras: grau em que o usuário acredita que a infraestrutura técnica e organizacional de seu ambiente podem facilitar o uso do artefato.	C1: A ferramenta é compatível com outras tecnologias que eu uso. C2: Eu tenho os recursos financeiros necessários para usar a ferramenta. C3: Posso obter ajuda de outros quando tenho dificuldades em usar a ferramenta. C4: Eu tenho o conhecimento necessário para usar a ferramenta.
Motivações hedônicas: grau de divertimento ou prazer obtido ao utilizar o artefato	M1: Usar a ferramenta é divertido. M2: Usar a ferramenta é agradável. M3: Usar a ferramenta é muito prazeroso.
Valor do preço: grau que o usuário espera que o custo do artefato impacte em sua adoção.	V1: A ferramenta está a um preço razoável. V2: A ferramenta tem um bom custo benefício. V3: Considerando o valor atual, a ferramenta possui um bom preço de mercado.'
Intenção comportamental: grau esperado em que a rotina de uso do artefato possa impactar na sua aprendizagem e utilização	I1: Eu pretendo continuar utilizando a ferramenta no futuro. I2: Sempre tentarei utilizar a ferramenta no meu dia-a-dia. I3: Eu pretendo continuar a usar a ferramenta frequentemente.

As estatísticas descritivas das respostas ao questionário UTAUT 2 são apresentadas na Tabela 4. A partir de cada construto, são propostas hipóteses para avaliação da aceitação

e intenção de uso da ferramenta e do seu processo de uso:

H1: A expectativa de desempenho afeta positivamente a intenção de uso da ferramenta e seu processo.

H2: A expectativa de esforço afeta positivamente a intenção de uso da ferramenta e seu processo.

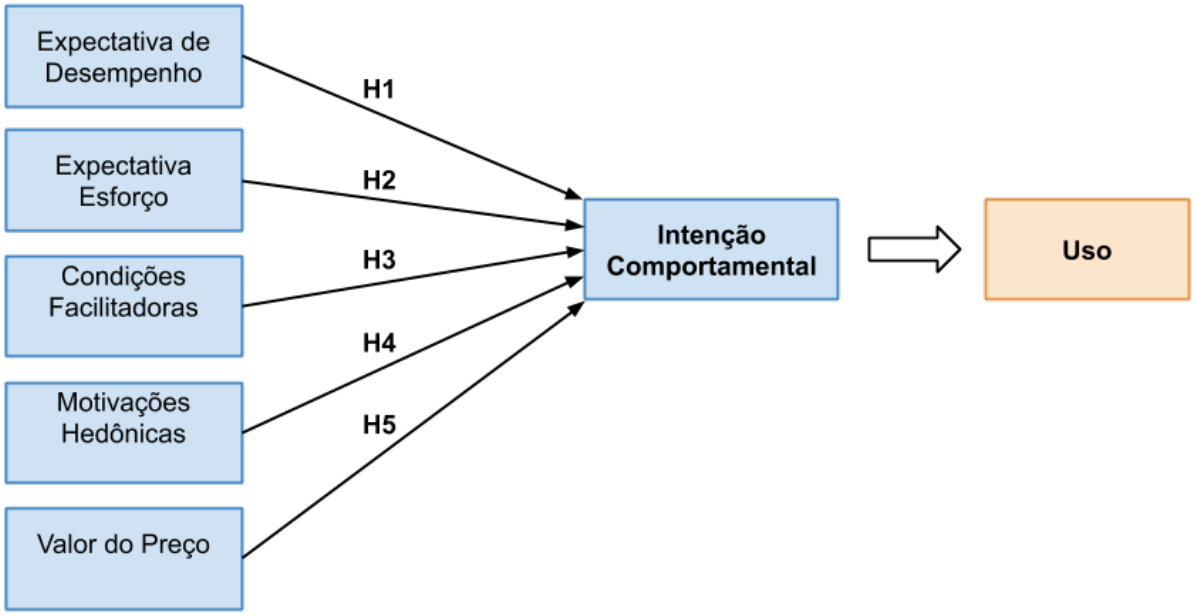
H3: As condições facilitadoras afetam positivamente a intenção de uso da ferramenta e seu processo.

H4: As motivações hedônicas afetam positivamente a intenção de uso da ferramenta e seu processo.

H5: O valor do preço afeta positivamente a intenção de uso da ferramenta e seu processo.

Após a compilação dos construtos e hipóteses, foi desenvolvido o modelo de avaliação de aceitação de tecnologia proposto nessa pesquisa, com cinco hipóteses derivadas do modelo UTAUT 2 (Figura 20).

Figura 20 – Hipóteses derivadas do modelo UTAUT 2



Fonte: autor

Antes da análise das hipóteses, foi realizado o teste de normalidade Shapiro-Wilk (GONZÁLEZ-ESTRADA; COSMES, 2019), constatando que os dados apresentam distribuição não-normal, rejeitando a hipótese nula H0 de normalidade ao nível de significância de 5% (todos os valores da variável estatística W ficaram acima de 0,05, conforme apresentado na Tabela 5). Dessa forma, os dados foram normalizados, sendo reescalados individualmente de modo que cada conjunto de respostas ao questionário apresente distribuição normal para sua adequada análise.

Tabela 4 – Estatísticas descritivas das respostas ao questionário UTAUT 2, onde a coluna DP representa o desvio padrão

Questão	**Média**	**Mínimo**	**Máximo**	**DP**
1) Eu acho a ferramenta útil no meu dia-a-dia.	5,80	3	7	1,32
2) Usar a ferramenta aumenta minhas chances de conseguir coisas que são importantes para mim.	5,70	4	7	0,98
3) Usar a ferramenta me ajuda a realizar as coisas mais rapidamente.	6,50	4	7	1,00
4) O uso da ferramenta aumenta a minha produtividade.	6,30	3	7	1,08
5) Acho a ferramenta fácil de usar.	6,00	3	7	1,12
6) É fácil para mim ficar habilidoso(a) no uso da ferramenta.	6,45	5	7	0,76
7) Minha interação com a ferramenta é clara e compreensível.	6,10	4	7	0,91
8) Aprender a usar a ferramenta é fácil para mim.	6,45	5	7	0,76
9) A ferramenta é compatível com outras tecnologias que eu uso.	5,75	3	7	1,21
10) Eu tenho os recursos financeiros necessários para usar a ferramenta.	6,70	4	7	0,80
11) Posso obter ajuda de outros quando tenho dificuldades em usar a ferramenta.	5,80	4	7	0,95
12) Eu tenho o conhecimento necessário para usar a ferramenta.	6,20	4	7	1,15
13) Usar a ferramenta é divertido.	5,80	4	7	1,01
14) Usar a ferramenta é agradável.	6,40	5	7	0,60
15) Usar a ferramenta é muito prazeroso.	5,65	4	7	0,93
16) A ferramenta está a um preço razoável.	6,85	4	7	0,67
17) A ferramenta tem um bom custo benefício.	6,85	4	7	0,67
18) Considerando o valor atual, a ferramenta possui um bom preço de mercado.	6,80	4	7	0,7
19) Eu pretendo continuar utilizando a ferramenta no futuro.	6,20	5	7	0,70
20) Sempre tentarei utilizar a ferramenta no meu dia-a-dia.	5,80	4	7	0,77
21) Eu pretendo continuar a usar a ferramenta frequentemente.	5,80	3	7	1,06

Tabela 5 – Resultado do teste de normalidade Shapiro-Wilk, onde *W* representa o resultado do teste e *H0* a hipótese de que os dados possuem uma distribuição normal

Questão	*W*	**H0**
1) Eu acho a ferramenta útil no meu dia-a-dia.	0,80	FALSA
2) Usar a ferramenta aumenta minhas chances de conseguir coisas que são importantes para mim.	0,88	FALSA
3) Usar a ferramenta me ajuda a realizar as coisas mais rapidamente.	0,56	FALSA
4) O uso da ferramenta aumenta a minha produtividade.	0,66	FALSA
5) Acho a ferramenta fácil de usar.	0,80	FALSA
6) É fácil para mim ficar habilidoso(a) no uso da ferramenta.	0,71	FALSA
7) Minha interação com a ferramenta é clara e compreensível.	0,84	FALSA
8) Aprender a usar a ferramenta é fácil para mim.	0,71	FALSA
9) A ferramenta é compatível com outras tecnologias que eu uso.	0,87	FALSA
10) Eu tenho os recursos financeiros necessários para usar a ferramenta.	0,44	FALSA
11) Posso obter ajuda de outros quando tenho dificuldades em usar a ferramenta.	0,88	FALSA
12) Eu tenho o conhecimento necessário para usar a ferramenta.	0,70	FALSA
13) Usar a ferramenta é divertido.	0,85	FALSA
14) Usar a ferramenta é agradável.	0,74	FALSA
15) Usar a ferramenta é muito prazeroso.	0,89	FALSA
16) A ferramenta está a um preço razoável.	0,24	FALSA
17) A ferramenta tem um bom custo benefício.	0,24	FALSA
18) Considerando o valor atual, a ferramenta possui um bom preço de mercado.	0,33	FALSA
19) Eu pretendo continuar utilizando a ferramenta no futuro.	0,80	FALSA
20) Sempre tentarei utilizar a ferramenta no meu dia-a-dia.	0,85	FALSA
21) Eu pretendo continuar a usar a ferramenta frequentemente.	0,77	FALSA

A validação das hipóteses foi realizada avaliando a capacidade de cada construto explicar ou não a intenção de uso da ferramenta e seu processo, por meio da técnica de regressão *Partial Least Square* (PLS), comumente adotada nesse tipo de pesquisa (CHAO, 2019). Esta técnica reduz os dados de entrada (respostas às questões) ao menor conjunto de componentes não correlacionados possível e aplica regressão para analisar a capacidade

dos mesmos preverem os dados de intenção de uso. Esta redução de dimensionalidade é importante para avaliar a dimensionalidade de cada construto, removendo possíveis correlações desnecessárias entre múltiplas questões de cada construto que possam prejudicar a análise (HAIR et al., 2009).

A análise de regressão PLS foi validada utilizando-se a técnica de validação cruzada *10-folds*. Nesta técnica, os dados de entrada são divididos em dez partes, e a cada parte (*fold*) iterada, é realizado o treino com as outras nove, e a validação com a mesma – obtendo-se dez resultados para validação com diferentes conjuntos de dados. Utilizou-se a métrica de avaliação Mean Absolute Error (MAE), a qual computa, para cada conjunto de previsões, a média absoluta dos erros entre cada valor previsto e o valor esperado. Como resultado da validação, o erro médio para cada construto apresentou-se próximo de zero ($< 0{,}004$), assim como o desvio padrão dos conjuntos de previsões ($< 0{,}02$) – indicando que as hipóteses H1, H2, H3, H4 e H5 podem ser consideradas válidas (Tabela 6).

Tabela 6 – Resultados da regressão PLS com 10 *folds*, onde a coluna MAE representa o erro médio absoluto da regressão e DP o seu desvio padrão

Hipótese	**MAE**	**DP**	**Resultado**
H1: A expectativa de desempenho afeta positivamente a intenção de uso da ferramenta e seu processo	0,0023	0,0164	VÁLIDA
H2: A expectativa de esforço afeta positivamente a intenção de uso da ferramenta e seu processo	0,0045	0,0264	VÁLIDA
H3: As condições facilitadoras afetam positivamente a intenção de uso da ferramenta e seu processo	0,0028	0,0178	VÁLIDA
H4: As motivações hedônicas afetam positivamente a intenção de uso da ferramenta e seu processo	0,0027	0,0161	VÁLIDA
H5: O valor do preço afeta positivamente a intenção de uso da ferramenta e seu processo	0,0021	0,0135	VÁLIDA

4.1.3 Discussão

A avaliação do artefato desenvolvido (ferramenta e respectivo processo de uso) é mista, combinando dois métodos quantitativos com propósitos específicos. O SUS tem como objetivo avaliar a usabilidade do artefato, tendo como resultado final um número entre 0 e 100 que representa um indicador normalizado da usabilidade geral. Já o UTAUT 2 possibilita obter visões complementares, analisando sua utilidade e aceitação, por meio da identificação dos construtos que impactam na sua intenção de uso. Dessa forma, os métodos de avaliação empregados em conjunto são quantitativos convergentes, com ênfase igual e coleta simultânea.

A nota final normalizada do artefato pelo método SUS foi de 84,5 pontos em 100. Na escala de avaliação proposta por (BANGOR; KORTUM; MILLER, 2009), este valor

indica a alta taxa de usabilidade do artefato, categorizando-o com o adjetivo excelente, nota A. Com base nesse indicador, pode-se afirmar que a ferramenta e o processo de análise desenvolvidos configuram-se como um artefato fácil de se aprender e se utilizar na condução de pesquisas em discussões *online* centradas no pesquisador, com controle das etapas do processo de KDD e metodologia de análise bem definida. Esta conclusão é reforçada pelo resultado da análise das estatísticas descritivas do questionário UTAUT 2, onde todas as questões tiveram nota máxima 7 e mínima 3, e menor valor médio de questão equivalente a 5,65. Como neste questionário todas as afirmações são de pontos positivos em relação ao artefato analisado, estes resultados numéricos indicam uma alta taxa de aceitação e intenção de uso.

A investigação das respostas dos participantes em relação ao questionário SUS (Tabela 2) também permite traçar um panorama geral em relação à satisfação de uso, facilidade de aprendizagem e a aspectos das Heurísticas de Nielsen (TENÓRIO et al., 2010) apresentados pela ferramenta e seu processo de uso. As três questões que compõem a heurística de Nielsen "Flexibilidade e eficiência de uso" (5, 6 e 8) apresentam os menores desvios-padrões nas respostas do questionário, indicando que a maioria dos participantes da avaliação tiveram percepções fortemente semelhantes nesse aspecto. A baixa concordância para as afirmações "Acho que a ferramenta apresenta muitas inconsistências" e "Considerei a ferramenta muito complicada de utilizar" e alta concordância para a afirmação "Considero que as várias funcionalidades da ferramenta estão bem integradas" indicam a flexibilidade e eficiência gerais no uso da ferramenta e de seu método. Atender a essa heurística é importante para possibilitar que diferentes perfis de pesquisadores possam utilizar a ferramenta de forma objetiva e funcional, sem ficarem confusos em relação a seus recursos na condução das análises.

A média de resposta 1,45 à afirmação "Achei a ferramenta mais complexa do que o necessário" indicam que a ferramenta e seu processo de uso apresentam facilidade de memorização pelos usuários, atendendo à heurística de Nielsen de "reconhecimento em vez de memorização". Essa característica é importante em um sistema de informação para garantir que o usuário possa utilizar a ferramenta por meio de seu respectivo processo de análise sem precisar memorizar várias informações ao longo de seu uso.

Com média de respostas 1,15 e maior valor de concordância equivalente a 2 para a afirmação "Acho que a ferramenta apresenta muitas inconsistências", observa-se que a ferramenta atende à heurística de Nielsen de "prevenção de erros". Isso indica que os recursos da ferramenta estão bem integrados a seu processo de uso, de modo a conduzir bem sua utilização pelo usuário e prevenir os possíveis erros que ele possa cometer.

O aspecto de facilidade de aprendizagem, representada pelas questões 3, 4, 7, e 10, possui uma boa aceitação média dos usuários, mas apresenta maior desvio padrão em algumas das afirmações que a compõem – indicando pontos a que podem ser aprimorados.

Uma questão a ser observada é a necessidade de um técnico para auxiliar no uso da ferramenta (questão 4). Enquanto a média das respostas, com valor 2,10, indica discordância à afirmação, as respostas individuais vão da faixa de valor 1 (forte discordância) à 4 (concordância), indicando que existem usuários que tiveram dificuldade para utilizar a ferramenta e seu processo sozinhos. Dessa forma, uma possibilidade de aprimoramento do uso da ferramenta por profissionais que não são da área de mineração de dados consiste em uma melhoria na documentação de uso e no treinamento inicial, para que os usuários não dependam de muita ajuda durante a realização de suas análises.

A satisfação do usuário pode ser avaliada pelas respostas às questões 1, 4 e 9. A questão 1, "Acho que gostaria de utilizar a ferramenta com frequência", possui média de resposta 4,20 e menor concordância dentro da faixa neutra, com valor 3. Por sua vez, a questão 4, "Acho que precisaria de ajuda de um técnico para conseguir usar a ferramenta", possui uma maior variabilidade de respostas – inclusive, com o maior desvio padrão de todas as dez questões, com sua concordância indo de 1 a 4. Apesar da resposta média ser 2,10, apresenta-se como uma oportunidade de melhoria do uso da ferramenta um tutorial ou manual de uso, de modo a reduzir a necessidade de um profissional da área de mineração de dados auxiliando o pesquisador. Já para a questão 9, "Eu me senti confiante ao usar a ferramenta", observa-se bons resultados, com média 4,20.

Complementando os aspectos de usabilidade que configuram a facilidade de uso do artefato desenvolvido, o modelo de aceitação UTAUT 2 permite mapear individualmente quais aspectos estão relacionados à sua intenção de uso. Foram avaliadas cinco hipóteses por meio dos construtos Expectativa de desempenho, Expectativa de esforço, Condições facilitadoras, Motivações hedônicas e Valor do preço. A análise da validade de cada hipótese, por meio da técnica de regressão PLS, permite investigar se elas possuem ou não impacto na intenção de uso do artefato desenvolvido.

A hipótese H1, "A expectativa de desempenho afeta positivamente a intenção de uso da ferramenta e seu processo", apresentou relação válida com a intenção de uso, indicando que este construto é relevante para analisar a aceitação do artefato. Este construto relaciona-se ao grau em que o usuário acredita que a utilização do artefato avaliado o auxilia para atingir um desempenho adequado nas tarefas realizadas com o mesmo. A segunda hipótese, H2, "A expectativa de esforço afeta positivamente a intenção de uso da ferramenta e seu processo" também foi identificada como válida. Ela se relaciona ao grau de esforço que o usuário espera demandar para realizar tarefas com o artefato, sendo um esforço menor demandado por uma capacidade do mesmo em facilitar a execução das tarefas. Com as hipóteses H1 e H2 se mostrando válidas, pode-se afirmar que os participantes da avaliação consideraram a ferramenta e seu processo de uso capazes de contribuir significativamente com a análise de discussões *online*, de forma fácil e clara para se utilizar – indo ao encontro das conclusões de usabilidade identificadas na avaliação SUS.

A validade da hipótese H3, “As condições facilitadoras afetam positivamente a intenção de uso da ferramenta e seu processo”, também foi comprovada pela regressão PLS, indicando que compatibilidade tecnológica, recursos financeiros, ajuda de uso e conhecimentos prévios são condições capazes de facilitar o uso do artefato desenvolvido. Este construto reforça as observações realizadas na análise do questionário SUS, indicando que o desenvolvimento de uma documentação ou manual de uso da ferramenta e seu processo podem ser relevantes para melhorar a usabilidade e aceitação do artefato.

A hipótese H4, “As motivações hedônicas afetam positivamente a intenção de uso da ferramenta e seu processo”, também apresentou relação válida com a intenção de uso, indicando que o grau de divertimento ou prazer obtido ao utilizar a ferramenta e seu processo relaciona-se diretamente com a aceitação da mesma. Os valores médios das respostas às questões que pertencem a esse construto indicam que seu uso, no geral, é agradável, divertido e prazeroso – contribuindo para um maior envolvimento emocional do usuário com o artefato.

A última hipótese proposta na avaliação com o modelo UTAUT 2 é a H5, “O valor do preço afeta positivamente a intenção de uso da ferramenta e seu processo”. Esta hipótese relaciona-se ao grau que o usuário espera que o custo do artefato impacte em sua adoção, e teve sua relação com a intenção de uso do artefato validada por meio da regressão PLS. Como foi informado aos participantes da avaliação, o artefato desenvolvido nesta dissertação é de uso livre e gratuito, não dependendo de licenças ou outros tipos de custos. Este aspecto mostrou-se relevante para o interesse na adoção do artefato, com as questões relacionadas ao construto obtendo os maiores valores médios de resposta, bem como menores valores de desvio padrão.

4.2 Estudo de caso

Seguindo o arcabouço do método *Design Science Research* (DSR), o Ciclo do Rigor forneceu como entrada para o Ciclo do *Design* fundamentações científicas – revisão da literatura, processo de análise e método de avaliação – para o desenvolvimento e a avaliação da ferramenta e seu processo de uso. Após estes terem sido projetados e avaliados, o Ciclo do Design deve retornar para o Ciclo do Rigor novas adições à base de conhecimento de seu contexto, na forma de contribuições científicas. Neste contexto, esta seção apresenta um estudo de caso exploratório conduzido pelo autor, aplicando a ferramenta e seu processo de uso na análise de uma discussão *online* do mundo real, visando demonstrar sua utilidade e contribuir com novos conhecimentos sobre a mesma.

Para o estudo de caso foi selecionado um fórum de discussão *online* sobre neurociência do site Reddit[2] – uma plataforma de fóruns de discussão sobre temas diversos

[2] https://www.reddit.com/

popular na Internet, com potencial para análises e obtenção de *insights* na área de pesquisa em saúde (CHOUDHURY; DE, 2014). O tema neurociência foi escolhido por abordar um assunto que faz interseção com diversas áreas do conhecimento, possuindo tópicos de discussão ricos em conteúdo multidisciplinar, com grande volume de dados, e diversos perfis de participantes (DAS et al., 2014). Deve-se destacar que não houve a participação de um especialista no domínio da discussão virtual (neurocientista) durante a condução da análise, sendo seu papel realizado pelo autor para ilustrar a aplicação real do artefato desenvolvido.

4.2.1 Pesquisas relacionadas

Diferentes técnicas de mineração de dados vêm sendo aplicadas em pesquisas envolvendo a análise de discussões *online* na área de saúde, abrangendo diversos temas, como: vício em drogas (KIM et al., 2017), câncer (CHO et al., 2018) e uso de medicamentos (ABDELLAOUI et al., 2018). Em meio à grande variedade de técnicas de mineração de dados aplicáveis, FAN; GORDON (2014) apresentam uma visão geral do processo de análise de mídias sociais, descrevendo as principais técnicas de mineração de dados aplicáveis ao contexto: modelagem de tópicos, análise de sentimentos e análise visual.

Técnicas de modelagem de tópicos possibilitam a sumarização automática de assuntos (tópicos) em grandes conjuntos de textos, atacando o desafio principal das pesquisas em discussões *online*, que consiste na impossibilidade de leitura de todo o conteúdo para análise dos assuntos discutidos. Nesta técnica de mineração de dados não-supervisionada, as postagens da discussão são organizadas automaticamente em grupos semanticamente relacionados (LIU, 2012), que são identificados pelos seus termos mais relevantes, os quais caracterizam o assunto principal de cada conjunto. Essa técnica é explorada no contexto da saúde por ABDELLAOUI et al. (2018), que a aplicam a mensagens de fóruns públicos mencionando determinados medicamentos, buscando segmentar os resultados em tópicos semânticos, e identificar para análise aqueles que se relacionem ao não cumprimento das instruções prescritas pelo médico.

Na análise de sentimentos, a polaridade das opiniões, emoções e sentimentos manifestados pelos usuários é avaliada automaticamente, por meio da análise computacional da linguagem escrita (SONG et al., 2009). ZHENG; LI; FARZAN (2018) demonstram o potencial desta técnica no contexto da saúde, aplicando-a para análise de sentimentos na exploração das dinâmicas e impactos de grupos de apoio em discussões *online* sobre doenças – conseguindo segmentar automaticamente milhares de postagens como positivas ou negativas para investigação, de uma forma que seria impossível manualmente.

Os resultados gerados pela combinação das técnicas de modelagem de tópicos e análise de sentimentos possibilitam a descoberta automática de tópicos relevantes para análise em grandes volumes de dados. Entretanto, para que seu potencial seja plenamente

explorado pelos pesquisadores, precisam ser apresentados em formatos compreensíveis e intuitivos. CHEN; ZHU; CONWAY (2015) exploram o uso de visualizações gráficas do resultado da modelagem de tópicos aplicada a uma discussão *online* sobre questões de saúde envolvidas no uso de cigarros eletrônicos, propondo que representações visuais dos tópicos possibilitam sua melhor compreensão e comparação.

4.2.2 Materiais e métodos

Com uma comunidade ativa, popularidade crescente e dados das discussões publicamente acessíveis (WENINGER; ZHU; HAN, 2013), a plataforma Reddit apresenta-se como um espaço virtual relevante para discussões e uma potencial fonte de dados para estudos. O presente estudo de caso consiste na análise da comunidade de discussão "Neuroscience: your brain on Reddit"[3], utilizando a ferramenta e o processo de uso desenvolvidos. A comunidade de discussão possui no presente momento mais de 50 mil membros cadastrados, com variados perfis, como médicos, pesquisadores, entusiastas e pacientes. Suas discussões abrangem os mais variados temas relacionados à neurociência – possibilitando um vasto panorama sobre a área.

Na plataforma Reddit, um fórum sobre determinado assunto é chamado *subreddit*, e seus usuários inscritos podem adicionar postagens, incluindo textos, imagens e *links*. Postagens podem receber comentários dos usuários, criando uma estrutura hierárquica de discussão – com a possibilidade de comentários em resposta a outros comentários. Os usuários também podem votar positivamente ou negativamente uma postagem ou comentário, produzindo *scores* que permitem avaliar sua popularidade e relevância.

A interface web do Reddit permite filtrar as postagens de destaque de um *subreddit* em três categorias baseadas no sistema de pontuação gerado pelos votos dos usuários:

- *Top*: postagens com a maior quantidade de votos positivos da comunidade em todo seu período de existência, indicando assuntos que despertaram o interesse dos usuários e foram amplamente visualizados e discutidos;

- *Hot*: postagens que receberam grande quantidade de votos positivos e comentários, em um curto período de tempo, indicando assuntos emergentes que estão captando o interesse da comunidade;

- *Controversial*: postagens que receberam muitos comentários e votos, mas com quantidades semelhantes de votos positivos e negativos, indicando assuntos polêmicos ou controversos.

[3] https://www.reddit.com/r/neuro/

Para a criação do *dataset* utilizado neste estudo de caso, foi desenvolvido um *web crawler* em Python para coletar todas as postagens e seus respectivos comentários de cada uma dessas três categorias de destaque da comunidade sobre neurociência, totalizando 22.154 registos. Como questões de pesquisa sobre a comunidade de neurociência analisada, buscou-se investigar a caracterização dos principais assuntos discutidos, sua polaridade e seu interesse ao longo do tempo, sendo formuladas da seguinte maneira:

Q1: Como se caracterizam os assuntos discutidos frequentemente na comunidade?

Q2: Quais são as discussões emergentes e quais estão perdendo interesse na comunidade ao longo do tempo?

Q3: Quais são e como se caracterizam os assuntos nos quais há uma forte predominância de sentimentos?

Para a modelagem de tópicos, que consiste na identificação dos principais assuntos semânticos discutidos na comunidade, a ferramenta de análise foi instanciada com a técnica de Fatoração de Matriz Não-Negativa (NMF), a qual no geral produz tópicos coerentes para *datasets* variados (O'CALLAGHAN et al., 2015). Os textos de cada tópico tiveram sua polaridade analisada por meio da ferramenta VADER, por ser um léxico otimizado para analisar sentimentos de textos de mídias sociais no idioma inglês. Para cada texto, a ferramenta calcula a valência e a intensidade de seus termos com base em um dicionário de pesos, gerando valores numéricos para os sentimentos positivo, negativo e neutro. Por se tratar de uma comunidade com discussões predominantemente técnicas, como esperado o sentimento de maior destaque em cada tópico foi o neutro. Dessa forma, para as análises da polaridade este sentimento foi desconsiderado, comparando-se apenas o valor dos sentimentos positivo e negativo.

Os *datasets* produzidos foram analisados com a ferramenta interativa de mineração de dados desenvolvida, utilizando seu processo de uso baseado na abordagem metodológica da Teoria Fundamentada em Dados e pautado no processo sistemático de pesquisa da Netnografia. Por meio da produção de uma série de agrupamentos exploratórios de tópicos semânticos, foram investigados padrões de discussões na comunidade, ajustando de forma iterativa os parâmetros de mineração do processo de KDD. Os resultados produzidos a cada iteração foram analisados em relação à sua capacidade de resposta às questões de pesquisa, e a busca foi refinada à procura de novos padrões de discussões e seus relacionamentos, até o ponto onde não eram produzidos novos conhecimentos relevantes – atingindo a etapa de saturação teórica.

4.2.3 Conjunto de dados

Para a presente pesquisa, foram produzidos três *datasets* no formato CSV, coletados do fórum de discussão sobre neurociência da plataforma Reddit, utilizando os filtros *"top"*,

"hot" e *"controversial"*. O *dataset top.csv* contém as postagens e comentários com maior quantidade de votos positivos da comunidade ao longo de todo o seu tempo de duração, representando discussões duradouras e consideradas altamente relevantes pelos membros. Ele possui 9.146 linhas, sendo 970 postagens e 8.176 comentários, produzidos por 3.033 usuários. A Tabela 7 descreve algumas estatísticas complementares do *dataset*:

Tabela 7 – Estatísticas descritivas do *dataset top.csv*

	Média	**DP**	**Mínimo**	**Máximo**
Comentários/Postagens	9,46	7,29	1	29
Postagens/Autor	0,32	1,55	0	53
Cometários/Autor	2,69	13,94	0	735
Postagens/Dia	23,66	43,02	0	153
Comentários/Dia	199,41	412,57	0	1.786
Palavras/Postagem	16,53	9,11	4	54
Palavras/Comentário	49,22	67,99	1	979

O segundo *dataset* coletado, *hot.csv*, é composto pelas postagens que tiveram elevadas taxas de votos positivos e comentários em curto período de tempo, representando postagens emergentes que despertam interesse rápido na comunidade. Este *dataset* possui 5.661 linhas, contendo 979 postagens e 4.682 comentários, com 1.711 usuários envolvidos. A Tabela 8 descreve algumas estatísticas complementares do *dataset*:

Tabela 8 – Estatísticas descritivas do *dataset hot.csv*

	Média	**DP**	**Mínimo**	**Máximo**
Comentários/Postagens	6,28	5,92	1	35
Postagens/Autor	0,57	2,29	0	58
Cometários/Autor	2,74	5,73	0	115
Postagens/Dia	23,46	45,46	0	142
Comentários/Dia	126,54	214,59	5	772
Palavras/Postagem	13,20	7,75	3	50
Palavras/Comentário	61,05	83,15	1	962

O terceiro *dataset*, *controversial.csv*, contém as postagens com grande interação da comunidade, mas com uma quantidade equilibrada de votos positivos e negativos, ou seja, consideradas mais controversas. São 7.347 linhas, das quais 978 são postagens e 6.369 são comentários. O número de usuários envolvidos é de 2.661. A Tabela 9 descreve algumas estatísticas complementares do *dataset*:

Apesar de provenientes da mesma comunidade virtual de discussão, os três *datasets* possuem diferenças significativas em relação às estatísticas de suas postagens e comentários. Nos registros do *dataset hot.csv*, tendem a existir mais postagens e comentários por mesmo autor, indicando que determinados usuários acompanham com maior frequência as discussões principais e são mais ativos na comunidade. Já o *dataset controversial.csv*, por seu caráter polêmico, possui em média mais comentários por postagem e por dia do que o

Tabela 9 – Estatísticas descritivas do *dataset controversial.csv*

	Média	**DP**	**Mínimo**	**Máximo**
Comentários/Postagens	7,63	6,56	1	31
Postagens/Autor	0,37	1,54	0	70
Cometários/Autor	2,39	10,90	0	537
Postagens/Dia	22,23	40,18	0	149
Comentários/Dia	144,75	259,61	0	920
Palavras/Postagem	13,31	7,83	3	63
Palavras/Comentário	62,93	85,43	1	1.242

de discussões *hot.csv*, mas ainda assim, menos do que o das discussões com mais votos positivos, *top.csv*. Como representa as discussões que se desenvolveram mais rapidamente na comunidade, o *dataset hot.csv* possui comentários todos os dias, ao contrário do *dataset top.csv* e até mesmo do *controversial.csv*, que também possui postagens com grande interação em um curto período de tempo – mas ainda assim possui dias sem comentários.

Em relação ao tamanho dos textos, pode-se observar que as discussões com maior quantidade de votos positivos (*top.csv*) possuem em média postagens com consideravelmente mais palavras, indicando que postagens que descrevem e contextualizam melhor seu propósito tendem a promover melhores discussões. Além disso, essas discussões possuem no geral comentários mais curtos e objetivos, com menos palavra em relação aos comentários dos outros dois *datasets*. Também é interessante observar que a postagem e o comentário mais longos fazem parte do *dataset controversial.csv*, e ambos *datasets* possuem comentários com apenas uma palavra – na maioria das amostras, um agradecimento.

Como os três *datasets* são provenientes da mesma comunidade, existem postagens e comentários em comum entre os mesmos, como, por exemplo, postagens que têm alta quantidade de votos positivos e que os mesmos foram obtidos em um curto período de tempo. A maior interseção de dados encontra-se entre as discussões emergentes em um curto período de tempo e as controversas, compartilhando 1.174 registros.

4.2.4 Análise exploratória

A presente seção apresenta a análise exploratória das discussões *online* sobre neurociência da plataforma Reddit, buscando responder às questões de pesquisa propostas. Para cada *dataset*, o processo interativo de análise fundamentado em dados foi executado, ajustando-se os parâmetros do processo de KDD iterativamente na ferramenta desenvolvida. Após atingir a saturação teórica, onde não foi possível identificar mais tópicos novos e relevantes para pesquisa, obteve-se um conjunto de assuntos semânticos para cada *dataset*.

As Tabelas 10, 11, e 12 apresentam os dez tópicos semânticos mais significativos gerados pelo processo de uso da ferramenta e selecionados para a análise final, a partir dos *datasets top.csv*, *hot.csv* e *controversial.csv*. São apresentados os principais termos

que representam cada tópico, a descrição sucinta do tópico gerada manualmente pelo autor, o volume de postagens e comentários do tópico ao longo do tempo e a polaridade predominante em relação aos sentimentos positivo e negativo.

Q1: Como se caracterizam os assuntos discutidos frequentemente na comunidade?

A exploração e caracterização dos assuntos discutidos com maior frequência na comunidade possibilita traçar um panorama geral da discussão *online*. Para esta pesquisa, considerou-se como assuntos frequentes: 1) Aqueles que aparecem em mais de um dos *datasets* analisados, compartilhando termos descritivos de seus tópicos e representando um mesmo assunto geral; 2) Sub-tópicos de um mesmo assunto geral, presentes em um ou mais *datasets* e 3) Assuntos exclusivos de um único *dataset*, mas que se mostram constantes ou recorrentes no gráfico de sua linha do tempo.

O primeiro assunto frequente observado na pesquisa relaciona-se à doença cerebral degenerativa Alzheimer. Discussões sobre a mesma estão presentes em todos os três *datasets* analisados, todas com um volume crescente de postagens ao longo do tempo na comunidade. Nos *datasets* de discussões com mais votos positivos (*top.csv*) e de discussões de rápido interesse (*hot.csv*), as interações dos usuários caracterizam-se por assuntos mais técnicos (termos *"amyloid", "protein"* e *"neurobiology"*), outras doenças relacionadas (termos *"parkinsons"* e *"huntingtons"*), tratamento (termos *"symptoms", "drug"* e *"restores"*) e prevenção (termo *"prevent"*). Já no *dataset* de discussões polêmicas (*controversial.csv*), o assunto Alzheimer é discutido com menos profundidade técnica, com termos de destaque relacionados a hipóteses e opiniões pessoais dos participantes, como *"hypothesis"* e *"think"*.

Outro tópico frequente são os comentários de agradecimentos, que foram identificados pela ferramenta como assuntos semânticos característicos dos *datasets top.csv* e *hot.csv*. Estes assuntos têm uma distribuição relativamente constante na linha do tempo da comunidade, e grande teor positivo, com termos como *"amazing", "wow", "fantastic", "interesting", "cool"* e *"helpful"*. Pelos temos que descrevem os tópicos, também podemos identificar pelo que os usuários estão agradecendo: compartilhamentos, postagens, respostas, ajudas e conselhos. Também é interessante observar que este tópico de agradecimentos não aparece no *dataset controversial.csv*, indicando que em discussões controversas, há menos colaboração e empatia entre os usuários – reduzindo o potencial da comunidade.

Um assunto recorrente e que recebe grande engajamento da comunidade relaciona-se a percepções estranhas de tempo, identificado no *dataset hot.csv* como dois tópicos. No primeiro, observa-se discussões específicas sobre o efeito nomeado pelos participantes como *"fast feeling"* (termos *"feeling", "fast"* e *"neuro"*), onde vários usuários compartilham experiências de terem vivenciado a passagem do tempo de forma mais acelerada do

Tabela 10 – Análise dos tópicos do *dataset top.csv*

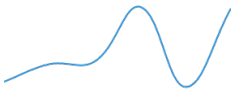
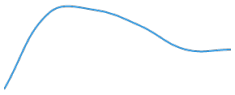
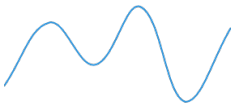
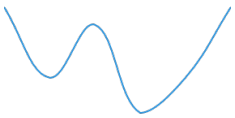
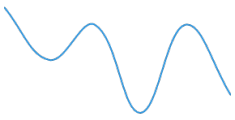
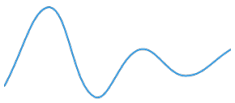
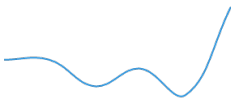
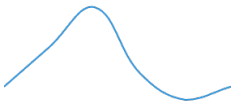
Termos	**Descrição**	**Atividade ao longo do tempo**	**Sentimento Predominante**
Brain, human, cells, activity, project, neurons, regions, changes, part, new	Cérebro humano		Positivo
thanks, sharing, lot, amazing, check, posting, much, post, wow, fantastic	Agradecimentos		Positivo
Research, am, phd, doing, lab, school, labs, funding, years, animal	Pesquisa em Neurociência		Positivo
Study, explained, brainpost, humans, new, psychiatry, stimulation, effects, rewards, rats	Discussões sobre estudos publicados		Positivo
Memory, learning, work, loss, enhances, alters, view, does, dopamine, restores	Estudos sobre memória		Positivo
Alzeheimers, disease, amyloid, protein, disorder, huntingtons, parkinsons, drug, restores, prevent	Alzheimer		Positivo
Mri, gif, data, xpost, neuro, look, based, imaging, animated, showing	Imagem de ressonância magnética (MRI)		Positivo
Neuroscience, society, computacional, disorders, books, carees, school, university, professor, students	Área acadêmica		Positivo
Neuroscientist, books, life, interview, thoughts, prof, own, slice, scientist, drugs	Neurocientistas		Positivo
Good, book, molecular, free, online, computational, neurobiology, basic, introductory, suggestions	Sugestões de livros		Positivo

Tabela 11 – Análise dos tópicos do *dataset hot.csv*

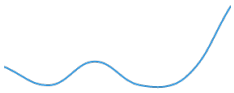
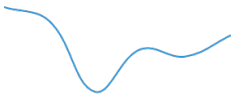
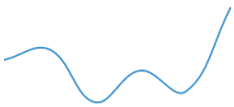
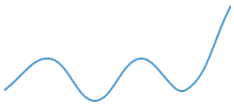
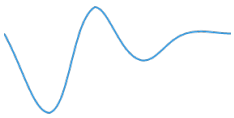
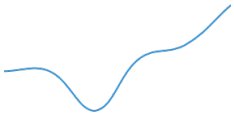
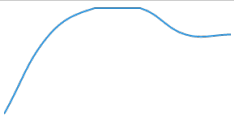
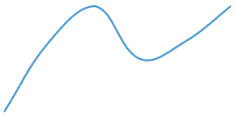
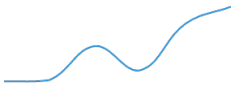
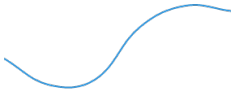
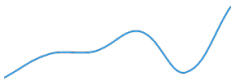
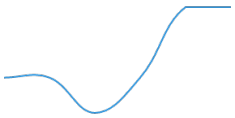
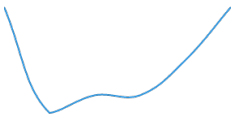
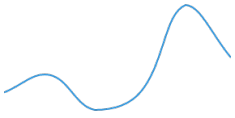
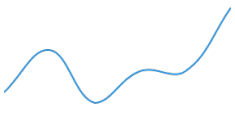
Termos	**Descrição**	**Atividade ao longo do tempo**	**Sentimento Predominante**
Thanks, interesting, cool, answer, helpful, replay, advice, lot, man, yeah	Agradecimentos		Positivo
Brain, damage, injury, blood, cells, oxygen, hypoxia, cause, activity, areas	Danos cerebrais		Positivo
New, study, explained, brainpost, humans, psychiatry, molecular, mice, rats, journal	Novos estudos publicados		Positivo
Dopamine, seretonin, receptors, effects, release, drugs, reward, reuptake, long, cause	Estudos sobre dopamina e serotonina		Positivo
Neuroscience, computational, programs, phd, neuro, resources, server, perceptual, abstracts, lab	Área acadêmica		Positivo
Feeling, fast, neuro, nerve, caffeine, odd, getting, minutes, thing, understand	Sensações do tipo “fast feelings”		Positivo
Disease, alzheimers, parkinsons, patients, protein, neurobiology, amyloid, roundtable, dollar, symptoms	Doença de Alzheimer e doença de Parkinson		Positivo
Time, weird, perception, memories, symptoms, sensation, periods, change, taking, short	Percepções estranhas do tempo		Negativo
Mind, effect, stress, thought, behavior, body, fast, state, neuro, analysis	Efeitos da fadiga		Positivo
Help, mouse, need, neuralgia, club, mri, behaviour, journal, report, cant	Ajuda com estudos envolvendo camundongos		Positivo

Tabela 12 – Análise dos tópicos do *dataset controversial.csv*

Termos	**Descrição**	**Atividade ao longo do tempo**	**Sentimento Predominante**
New, eeg, meditation, neuroscientists, activity, company, psychostimulants, transcendental, optogenetics, cells	Eletroencefalografia da meditação		Positivo
Neuroscience, psychology, field, phd, cognitive, degree, masters, computational, learn, program	Carreira em neurociência		Positivo
Doctor, see, go, talk, symptoms, ask, neurologist, medical, advice, said	Busca por conselhos médicos		Negativo
Memory, working, memories, recall, longterm, information, problems, loss, hippocampus	Funcionamento da memória		Positivo
Help, need, please, am, someone, find, understand, symptoms, try, sounds	Busca por ajuda		Positivo
Science, article, read, post, love, link, cognitive, interesting, title, results	Discussões sobre artigos científicos		Positivo
Dissease, alzheimers, prevent, problem, hypothesis, stress, think, pdf, effects, exercise	Prevenção do Alzheimer		Positivo
Consciousness, theory, know, neuro, serotonin, light, quantum, solving, state, beliefs	Teorias sobre a consciência		Negativo
Mind, control, ways, has, reading, efect, technologies, blown, conscious, drug	Controle mental		Negativo
Mri, scan, tell, look, got, results, head, has, neurologist, ct	Pedidos de ajuda com resultados de ressonância magnética		Negativo

que o normal. Há um viés positivo na polaridade das discussões, com os participantes demonstrando interesse em compreender o efeito, mas não o percebendo como algo ruim ou prejudicial. Já o segundo tópico do assunto é mais abrangente em relação aos fenômenos discutidos (termos *"time", "weird"* e *"perception"*), e apresenta uma polaridade predominantemente negativa na análise de sentimentos, com os usuários descrevendo os fenômenos como sintomas de algum problema.

A área acadêmica em neurociência é outro assunto identificado de discussão constante na comunidade, aparecendo como tópico nos três *datasets*. Em *top.csv*, o assunto relaciona-se principalmente a estudar na área (termos *"school", "university", "professor"* e *"students"*), enquanto em *hot.csv* as discussões têm um foco maior na área computacional (termos *"computational", "programs", "server"* e *"resources"*). Já no *dataset* de discussões controversas, há menos discussões ao longo do tempo sobre o assunto, e seu foco está mais relacionado ao campo da psicologia, com os termos *"psychology", "field"* e *"cognitive"*.

Relacionado aos assuntos sobre a área acadêmica, outro tópico frequente na comunidade refere-se a discussões sobre publicações científicas – o qual aparece na análise dos três *datasets*. Entretanto, enquanto nos *datasets top.csv* e *hot.csv* os temos dos tópicos indicam que o foco é a explicação de novos estudos (termos *"new", "study"* e *"explained"*), no *dataset controversial.csv* a discussão é menos profunda, como observado nos termos *"read", "title", "interesting"* e *"love"*.

Sub-tópicos de pesquisa em neurociência também aparecem como assuntos de discussão frequentes em todos os *datasets*, abrangendo temas como: cérebro, memória, dopamina, serotonina, efeito do estresse e pesquisa em laboratório com ratos – todas discussões com teor positivo. Em relação ao tópico de pesquisas com ratos identificado no *dataset hot.csv*, é interessante notar que além de um tópico que costuma ter respostas em curto período de tempo, seu gráfico de ocorrências ao longo do tempo mostra que é o tópico mais constante em toda a comunidade de discussão. Outros assuntos frequentes envolvendo pedidos de ajuda relacionam-se à avaliação de sintomas (termos *"help", "need", "please", "understand"* e *"symptoms"*), análise de resultados de imagens de ressonância magnética (termos *"mri", "scan", "tell"* e *"look"*) e resposta à busca de conselhos médicos (termos *"doctor", "see", "go", "talk"* e *"symptoms"*) – todos presentes no *dataset controversial.csv*.

Q2: Quais são as discussões emergentes e quais estão perdendo interesse na comunidade ao longo do tempo?

A análise da evolução temporal dos principais assuntos discutidos na comunidade permite identificar e caracterizar aqueles que estão perdendo ou ganhando interesse – possibilitando direcionar pesquisas para os temas de interesse emergente. Para responder a essa questão, foi realizada a análise do volume de comentários e postagens de cada

tópico ao longo do tempo, por meio da interpretação do gráfico de *sparkline* gerado. Para a seleção dos assuntos, exclui-se aqueles cujos gráficos apresentam ocorrências constantes ou recorrentes ao longo do tempo.

As discussões sobre a doença de Alzheimer figuram não só como um dos assuntos mais discutidos na comunidade analisada, como também um assunto de interesse crescente. Este resultado é coerente com estatísticas que indicam o crescimento de casos da doença ao longo dos anos (MATTHEWS et al., 2019), o que promove o interesse das pessoas e também reforça a necessidade de pesquisas e discussões científicas sobre a doença.

Outros assuntos que vêm ganhando interesse crescente relacionam-se a discussões sobre estudos em neurociência, com a presença deste tópico específico (termos *"study"* e *"explained"*), e sub-tópicos como: danos cerebrais, dopamina e serotonina, efeitos do estresse, efeitos da meditação e teoria da consciência. Observa-se também alguns assuntos emergentes com manifestações menos técnicas e mais pessoais, como relatos de percepção estranha do tempo, e pedidos de ajuda, conselhos e discussão sobre a carreira na neurociência.

Além do interesse crescente da comunidade em relação a tópicos específicos sobre neurociência, é interessante observar que há discussões emergentes com grande quantidade de votos positivos (*dataset top.csv*) sobre neurocientistas de destaque. Elas abrangem diversos aspectos, como seus livros publicados, sua vida, entrevistas realizadas e pensamentos (termos *"neuroscientist", "books", "life", "interview"* e *"thoughts"*).

Em relação ao volume de postagens ao longo do tempo, observa-se uma menor quantidade de assuntos que estão perdendo o interesse do que aqueles que vêm ganhando destaque. Tópicos de discussão relacionados ao compartilhamento de livros gratuitos para *download* (termos *"good", "book", "free"* e *"online"*) apresentam uma queda clara no volume de discussões da comunidade ao longo do tempo. Analisando-se amostras das postagens, percebe-se que este comportamento se deve a políticas contra pirataria e de conscientização dos usuários.

Discussões compartilhando imagens e dados de ressonância magnética (termos *"mri", "gif"* e *"data"*) mostraram-se recorrentes por um longo período, mas vêm dando indicações de perda de interesse da comunidade. Já as discussões relacionadas a controle mental (termos *"mind", "control"* e *"ways"*) apresentaram um pico de interesse recente da comunidade, mas agora vêm perdendo seu volume de postagens, reforçado pelas várias críticas dos usuários em relação a este tema polêmico.

Q3: Quais são e como se caracterizam os assuntos nos quais há uma forte predominância de sentimentos?

Como uma comunidade de discussões predominantemente técnicas, a polaridade principal de todas as postagens na análise de sentimentos é neutra. Entretanto, analisando-

se apenas os sentimentos positivos e negativos, observa-se que seus valores variam, permitindo assim identificar o sentimento da comunidade em relação a cada assunto.

Todos os assuntos do *dataset* com as postagens mais bem votadas (*top.csv*) apresentam valor positivo na análise de sentimentos, indicando que discussões bem avaliadas caracterizam-se por uma polaridade positiva e amistosa. Já no *dataset hot.csv*, que abstrai os assuntos bem votados e com interações imediatas da comunidade, encontramos apenas um assunto de polaridade negativa. Neste assunto, onde os usuários descrevem percepções de tempo estranhas (termos *"time", "weird"* e *"perception"*), o motivo do valor negativo na análise de sentimentos relaciona-se mais à forma como os usuários descrevem suas percepções do que a reação da comunidade em si.

Como esperado, o *dataset* de discussões controversas (*controversial.csv*) possui a maior ocorrência de assuntos com sentimentos negativos. Por seu caráter polêmico, nesses assuntos, a polaridade negativa das discussões é ocasionada pela forma como a comunidade reage a determinadas ideias e comportamentos nos tópicos. Destaca-se o assunto sobre controle mental (termos *"mind", "control"* e *"ways"*), já mencionado como tópico negativo que vem perdendo o interesse e outros três assuntos que, pelo contrário, vêm aparecendo cada vez mais nas discussões. Dois deles estão relacionados: pedidos de ajuda com resultados de exames de ressonância magnética (termos *"mri", "scan", "tell"* e *"look"*), e resposta à busca de conselhos médicos (termos *"doctor", "see", "go", "talk"* e *"symptoms"*) – ambos voltados para a crítica de que determinadas questões seriam mais pertinentes para um profissional médico e não a uma discussão *online* pública.

Finalmente, o terceiro assunto com polaridade negativa, presente no *dataset controversial.csv*, identifica discussões relacionadas a teorias sobre a consciência (termos *"consciousness"* e *"theory"*), que por seu caráter fascinante mas especulativo, produz longas discussões envolvendo crenças (termos *"know"* e *"beliefs"*) e até mesmo física quântica (termo *"quantum"*).

4.2.5 Discussão

O estudo de caso exploratório teve como objetivo demonstrar como a ferramenta e seu processo de análise abrem novas possibilidades de investigações científicas dirigidas a dados no contexto das mídias sociais, analisando uma comunidade de discussão *online* sobre o tema neurociência. Por meio dela, foram utilizadas técnicas de mineração e visualização de dados para identificar e caracterizar assuntos semânticos relevantes em meio a milhares de postagens e comentários – o que seria inviável por meio da sua leitura e análise manual.

A ferramenta foi aplicada a três *datasets* coletados da comunidade a ser analisada, cada um referente a um aspecto das discussões: mais bem votadas, com grande interação em um curto período de tempo e controversas. Os assuntos identificados abrangem diversos

tópicos em neurociência, que vão de discussões técnicas como análise de artigos recentes a pedidos de conselhos e especulações sobre o funcionamento da consciência. Foi empregado o processo de uso proposto nesta dissertação, baseado na Teoria Fundamentada em Dados, para responder às questões de pesquisa propostas, por meio da análise sistemática e rigorosa dos dados.

A primeira questão de pesquisa buscou investigar os principais assuntos discutidos com mais frequência na comunidade. O assunto de destaque refere-se à doença de Alzheimer, sendo discutido nos três *datasets* analisados, com diversos enfoques: prevenção, diagnóstico, tratamento e doenças relacionadas. Agradecimentos e pedidos de ajuda também são tópicos recorrentes, seja em relação à avaliação de sintomas, análise de resultados de exame e conselhos médicos.

Outra manifestação pessoal identificada foi o assunto sobre percepções estranhas de tempo, com os participantes compartilhando experiências de efeitos estranhos que presenciaram, como a passagem acelerada do tempo (*"Fast Feeling"*). A área acadêmica e carreira em neurociência também são temas que despertam bastante interesse da comunidade, com discussões relacionadas a universidades, pesquisas de laboratório e análises de artigos publicados. Outros tópicos relacionados identificados foram: cérebro, memória, dopamina, serotonina, efeito do estresse e pesquisa em laboratório com ratos.

O objetivo da segunda questão de pesquisa consistiu na identificação e caracterização dos assuntos emergentes e os que vêm perdendo o interesse da comunidade ao longo do tempo. Identificou-se que discussões sobre estudos em neurociência vêm apresentando interesse crescente da comunidade, com tópicos como carreira, Alzheimer, danos cerebrais, dopamina e serotonina, efeitos do estresse, efeitos da meditação e teoria da consciência. Conselhos, pedidos de ajuda e relatos de experiências também apresentando um volume crescente de postagens, indicando a abertura da comunidade para discussões mais pessoais e menos técnicas. Foram identificados apenas três assuntos que vêm perdendo o interesse da comunidade: compartilhamento de livros gratuitos para *download*, imagens e dados de ressonância magnética e controle mental.

A terceira questão de pesquisa buscou investigar os assuntos que possuem uma predominância de sentimentos, com polaridade positiva ou negativa. Percebeu-se que a maioria dos assuntos identificados têm uma polaridade positiva, indicando que as interações na comunidade são predominantemente amistosas. A maioria dos assuntos negativos emergiram das postagens com muita interação mas proporção equivalente de votos positivos e negativos, já que se caracterizam por serem controversas. No geral são assuntos pouco técnicos ou especulativos, como técnicas de controle mental, ajuda com resultados de ressonância magnética, respostas à busca de conselhos médicos e teorias sobre a consciência.

A consistência dos assuntos identificados pela análise e sua capacidade de responder

de forma satisfatória às questões de pesquisa produzindo novos conhecimentos indica a utilidade da ferramenta e do processo de uso propostos na presente dissertação, bem como sua relevância para a pesquisa em discussões *online*. No próximo capítulo, de conclusão, são discutidas as contribuições e limitações da dissertação, sendo também propostas sugestões de trabalhos futuros para aperfeiçoar e estender a pesquisa.

5 Conclusão

O presente trabalho investigou o desafio de se efetuar análises interativas de discussões *online* combinando técnicas de mineração de dados, com um processo de análise centrado no pesquisador e fundamentado em teorias metodologicamente rigorosas e reprodutíveis. Por se tratar de uma pesquisa envolvendo inovação tecnológica, foi adotado o método *Design Science Research*, o qual se propõe a guiar de forma científica o desenvolvimento de artefatos tecnológicos, por meio de um método cientificamente válido e metodologicamente reconhecido.

Para a especificação e avaliação do artefato e seu processo de uso no Ciclo do *Design* do DSR, os ciclos de Relevância e de Rigor envolveram a investigação da literatura científica relacionada a análises de discussões *online*, mapeando suas técnicas, recursos e limitações. Foram identificados trabalhos que tratam da análise de discussões *online* por meio de técnicas de mineração de dados, dentre as quais se destacaram a modelagem de tópicos, a análise de sentimentos e a visualização de dados. Apesar de muitos trabalhos possuírem um componente interativo em suas análises, identificou-se que os mesmos não envolviam o usuário pesquisador especialista no domínio do problema em todas as etapas do processo de Descoberta de Conhecimento em Bases de Dados (KDD), limitando as possibilidades das análises. Além disso, observou-se a ausência de especificações de processos de uso para conduzir as análises, o que limita sua reprodutibilidade.

Uma vez definido o problema e os requisitos para sua solução, foi realizada a especificação e o desenvolvimento de um artefato (ferramenta interativa e respectivo processo de uso) para mineração de dados de discussões *online*, capaz de incluir o pesquisador em todas as etapas do processo de KDD. Constituindo a contribuição tecnológica deste trabalho de dissertação, o desenvolvimento da ferramenta fundamentou-se nos trabalhos relacionados e nas conjecturas teóricas do contexto do problema, de modo a fundamentar a pesquisa e garantir sua relevância e rigor.

Para o processo de análise de discussões *online* centrada no usuário por meio da ferramenta desenvolvida, foi especificado um fluxo de uso pautado nas práticas da Netnografia e da Teoria Fundamentada em Dados, sendo integrado a todas as etapas de KDD, de modo a possibilitar a condução de pesquisas éticas, rigorosas e reprodutíveis. O processo se baseia na coleta e análise rigorosa e sistemática dos dados, para a construção de teorias fundamentadas que respondam às questões de pesquisa propostas. Este método integra-se ao uso da ferramenta na forma de um ciclo, onde são efetuadas sucessivas análises ajustando os parâmetros das etapas do processo de KDD, até que se atinja a saturação teórica – momento onde não são produzidos novos conhecimentos para se responder às

questões de pesquisa propostas.

A relevância e utilidade da ferramenta e seu processo foram avaliados com usuários, que após uma apresentação do funcionamento do artefato, realizaram a análise interativa de uma discussão *online* sobre programação de computadores, e responderam aos questionários SUS e UTAUT 2 para avaliação de sua usabilidade e aceitação. Os resultados indicam excelente usabilidade e alta aceitação dos usuários, que conseguiram utilizar a ferramenta e seu processo de uso para analisar o fórum de discussão apresentado e responder às questões de pesquisa propostas. Observou-se também que a intenção de uso e adoção da ferramenta e de seu processo está relacionada a cinco construtos: expectativa de desempenho, expectativa de esforço, condições facilitadoras, motivações hedônicas e preço.

Após o desenvolvimento e avaliação do artefato da pesquisa (ferramenta e respectivo processo de uso), foi realizado um estudo de caso exploratório, consistindo na análise de uma discussão *online* sobre neurociência, demonstrando a aplicabilidade da pesquisa desenvolvida. A consistência dos assuntos identificados pela análise e sua capacidade de responder de forma satisfatória às questões de pesquisa indicam a utilidade da ferramenta e do processo de uso desenvolvidos. Os resultados produzidos também contribuem para uma maior compreensão sobre os assuntos discutidos na área de neurociência, possibilitando caracterizar os tópicos de relevância, seu interesse ao longo do tempo e a percepção dos participantes da discussão. Como limitação do estudo de caso, deve-se observar que não houve a participação de um especialista no domínio da discussão analisada, sendo seu papel realizado pelo autor para ilustrar a aplicação real do artefato desenvolvido.

Os resultados obtidos com a pesquisa permitem corroborar as conjecturas propostas e apresentadas no quadro teórico do método DSR, produzindo novos conhecimentos descritivos e prescritivos a partir da inter-relação dos aspectos sociais e tecnológicos que compõem o problema investigado. A contribuição descritiva relaciona-se à compreensão do problema e às teorias que viabilizam sua solução. Ela consiste na identificação de que a análise de discussões *online* suportadas por técnicas de mineração de dados pode apresentar maior profundidade e rigor científico com o envolvimento do pesquisador em todas as etapas de KDD, e com a definição de um processo de análise consistente pautado em teorias científicas. Já a contribuição prescritiva relaciona-se ao resultado prático da pesquisa, representado pela ferramenta desenvolvida, a qual possui uma combinação de recursos interativos superior às existentes na literatura, mostrando-se viável para a análise de discussões *online*, e contribuindo com avanços práticos na área.

Para a avaliação e o estudo de caso, o artefato desenvolvido foi instanciado com a técnica de vetorização TF-IDF, modelagem de tópicos por fatoração de matrizes (NMF) e léxico para análise de sentimentos VADER. Como trabalho futuro, pretende-se instanciar o artefato com outras técnicas, como vetorização com *word embeddings*, modelagem de tópicos probabilística (LDA) e análise de sentimentos baseada em aprendizado de máquina.

O objetivo é investigar comparativamente o impacto destas técnicas no resultado das análises interativas, bem como sua percepção pelos usuários – possibilitando avaliar se sua variação é tão significativa quanto as opções interativas de ajustes de parâmetros de KDD.

A partir dos resultados da avaliação de usabilidade e aceitação, pretende-se também desenvolver uma documentação de uso para a ferramenta e seu processo, bem como investigar a aplicação de técnicas colaborativas de gamificação, do inglês *gamification*, ao artefato. A gamificação consiste no emprego de dinâmicas e mecânicas de jogos a aplicações que pertencem a outros contextos, com o objetivo de motivar e engajar os usuários. Para estimular a produção coletiva de conhecimento científico revisado e de qualidade por meio do artefato, pretende-se explorar elementos de jogos colaborativos, como placares e metas. Por meio da adição de um recurso que possibilite aos usuários votarem nos tópicos identificados e rotulados por cada um, é possível criar placares dos usuários com os tópicos mais bem avaliados ou com a maior quantidade de tópicos identificados e bem avaliados. Já o recurso de metas, bem como o de medalhas virtuais (*badges*), pode ser implementado por meio da definição de objetivos de dificuldade crescente a serem conquistados, como por exemplo: identificar dez tópicos, exportar os resultados de uma análise, obter cinco votos positivos em um tópico identificado ou rotular o tópico de outro usuário com um texto que o mesmo considere mais descritivo do que o proposto inicialmente.

Resultados parciais desta pesquisa de dissertação foram publicados pelo autor nos eventos internacionais IEEE 20^{th} International Conference on Information Reuse and Integration for Data Science e II Latin American Workshop on Computational Neuroscience. Na conferência IEEE IRI 2019, o trabalho "*Combining Data Mining Techniques for Evolutionary Analysis of Programming Languages*" – um dos candidatos ao "*Best student paper award*" do evento – foca no desafio de se analisar como uma comunidade de desenvolvedores que utiliza determinada linguagem de programação reage às evoluções da mesma, como introdução de novos recursos e modificação de sintaxe. Para isso, foram combinadas técnicas de modelagem de tópicos, análise de sentimentos e visualização de dados em um estudo longitudinal, capazes de analisar em paralelo a documentação técnica e os fóruns de discussão da linguagem. O segundo trabalho, apresentado no *workshop* II LAWCN 2019, "*Interactive analysis of the discussion from a virtual community on neuroscience*", combina técnicas de mineração de dados para a análise exploratória de uma comunidade na área de saúde, apresentando o fluxo de uso baseado na teoria fundamentada em dados integrada ao processo de KDD para a condução de uma análise metodologicamente rigorosa e reprodutível. O trabalho foi selecionado entre os melhores trabalhos da conferência para publicação no *Springer's Communications in Computer and Information Science*, sendo o conteúdo deste trabalho a base para a seção de estudo de caso da presente dissertação.

Referências

ABDELLAOUI, R. et al. Detection of cases of noncompliance to drug treatment in patient forum posts: topic model approach. *Journal of medical Internet research*, JMIR Publications Inc., Toronto, Canada, v. 20, n. 3, p. e85, 2018. Citado na página 73.

ALENCAR, A. B.; OLIVEIRA, M. C. F. de; PAULOVICH, F. V. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 2, n. 6, p. 476–492, 2012. Citado na página 48.

ALMEIDA, R. J. de A. et al. Combining data mining techniques for evolutionary analysis of programming languages. In: IEEE. *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. Los Angeles, California, USA, 2019. p. 1–8. Citado na página 34.

ANANDARAJAN, M.; HILL, C.; NOLAN, T. Modeling text sentiment: Learning and lexicon models. In: *Practical Text Analytics*. USA: Springer, 2019. p. 151–164. Citado na página 46.

BANGOR, A.; KORTUM, P.; MILLER, J. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, Usability Professionals' Association, v. 4, n. 3, p. 114–123, 2009. Citado 2 vezes nas páginas 62 e 69.

BARTL, M.; KANNAN, V. K.; STOCKINGER, H. A review and analysis of literature on netnography research. *International Journal of Technology Marketing*, Inderscience Publishers (IEL), v. 11, n. 2, p. 165–196, 2016. Citado na página 21.

BASKERVILLE, R. et al. Design science research contributions: finding a balance between artifact and theory. *Journal of the Association for Information Systems*, v. 19, n. 5, p. 3, 2018. Citado na página 28.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado na página 44.

BROOKE, J. et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, London–, v. 189, n. 194, p. 4–7, 1996. Citado na página 61.

ÇAĞDAŞ, V.; STUBKJÆR, E. Design research for cadastral systems. *Computers, Environment and Urban Systems*, Elsevier, v. 35, n. 1, p. 77–87, 2011. Citado na página 31.

CASS, S. The top programming languages 2019. *IEEE Spectrum*, 2019. Citado na página 50.

CHAO, C.-M. Factors determining the behavioral intention to use mobile learning: An application and extension of the utaut model. *Frontiers in psychology*, Frontiers, v. 10, p. 1652, 2019. Citado na página 68.

CHEN, A. T.; ZHU, S.-H.; CONWAY, M. What online communities can tell us about electronic cigarettes and hookah use: a study using text mining and visualization techniques. *Journal of medical Internet research*, JMIR Publications Inc., Toronto, Canada, v. 17, n. 9, p. e220, 2015. Citado na página 74.

CHEN, Y. et al. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, Elsevier, v. 163, p. 1–13, 2019. Citado na página 44.

CHO, H. et al. Visual cancer communication on social media: An examination of content and effects of# melanomasucks. *Journal of medical Internet research*, JMIR Publications Inc., Toronto, Canada, v. 20, n. 9, p. e10501, 2018. Citado na página 73.

CHOUDHURY, M. D.; DE, S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: *Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, Michigan, USA: Association for the Advancement of Artificial Intelligence, 2014. Citado na página 73.

CICHOCKI, A.; PHAN, A.-H. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, The Institute of Electronics, Information and Communication Engineers, v. 92, n. 3, p. 708–721, 2009. Citado na página 46.

DAS, S. et al. Pain research forum: application of scientific social media frameworks in neuroscience. *Frontiers in neuroinformatics*, Frontiers, v. 8, p. 21, 2014. Citado na página 73.

DEBUSE, J. et al. Building the kdd roadmap. In: *Industrial Knowledge Management*. London: Springer, 2001. p. 179–196. Citado na página 23.

DENZIN, N. K.; LINCOLN, Y. S. Introduction: The discipline and practice of qualitative research. SAGE Publications, Inc, 2008. Citado na página 59.

DRESCH, A.; LACERDA, D. P.; JÚNIOR, J. A. V. A. *Design science research: método de pesquisa para avanço da ciência e tecnologia*. Porto Alegre, Brasil: Bookman Editora, 2015. Citado na página 27.

FAN, W.; GORDON, M. D. The power of social media analytics. *Commun. Acm*, v. 57, n. 6, p. 74–81, 2014. Citado 4 vezes nas páginas 21, 22, 33 e 73.

FAYYAD, U. et al. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM, v. 39, n. 11, p. 27–34, 1996. Citado na página 22.

FLICK, U. *Introdução à metodologia de pesquisa: um guia para iniciantes*. Porto Alegre, Rio Grande do Sul, Brasil: Penso Editora, 2012. Citado na página 59.

FULLER, R. B. A comprehensive anticipatory design science. *Royal Architectural Institute of Canada*, v. 34, 1957. Citado na página 27.

GEIST, I. A framework for data mining and kdd. In: ACM. *Proceedings of the 2002 ACM symposium on Applied computing*. Magdeburg, Germany, 2002. p. 508–513. Citado na página 23.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, v. 1, n. 12, p. 2009, 2009. Citado na página 47.

GONZÁLEZ-ESTRADA, E.; COSMES, W. Shapiro–wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 89, n. 17, p. 3258–3272, 2019. Citado na página 66.

GREGORY, S. A. Design science. In: *The design method*. Boston, Massachusetts, USA: Springer, 1966. p. 323–330. Citado na página 27.

HAIR, J. F. et al. *Análise multivariada de dados*. Porto Alegre, Rio Grande do Sul, Brasil: Bookman Editora, 2009. Citado na página 69.

HEVNER, A. R. A three cycle view of design science research. *Scandinavian journal of information systems*, v. 19, n. 2, p. 4, 2007. Citado na página 30.

HEVNER, A. R. et al. Design science in information systems research. *MIS quarterly*, JSTOR, p. 75–105, 2004. Citado 3 vezes nas páginas 27, 29 e 31.

HOQUE, E.; ABID, E. Visual exploration of topic controversy in online conversations. In: SPRINGER. *International Conference on Human-Computer Interaction*. Cham, Germany, 2019. p. 369–373. Citado na página 36.

HOQUE, E.; CARENINI, G. Interactive topic modeling for exploring asynchronous online conversations: Design and evaluation of convisit. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, ACM, v. 6, n. 1, p. 7, 2016. Citado na página 35.

HOQUE, E.; CARENINI, G. Interactive topic hierarchy revision for exploring a collection of online conversations. *Information Visualization*, SAGE Publications Sage UK: London, England, p. 1473871618757228, 2018. Citado na página 35.

HU, Y. et al. Interactive topic modeling. *Machine learning*, Springer, v. 95, n. 3, p. 423–469, 2014. Citado na página 35.

HUTTO, C. J.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *Eighth international AAAI conference on weblogs and social media*. Ann Arbor, Michigan, USA: Association for the Advancement of Artificial Intelligence, 2014. Citado na página 47.

JR, J. F. N.; CHEN, M.; PURDIN, T. D. Systems development in information systems research. *Journal of management information systems*, Taylor & Francis, v. 7, n. 3, p. 89–106, 1990. Citado na página 27.

KIM, S. J. et al. Scaling up research on drug abuse and addiction through social media big data. *Journal of medical Internet research*, JMIR Publications Inc., Toronto, Canada, v. 19, n. 10, p. e353, 2017. Citado na página 73.

KORTUM, P. T.; BANGOR, A. Usability ratings for everyday products measured with the system usability scale. *International Journal of Human-Computer Interaction*, Taylor & Francis, v. 29, n. 2, p. 67–76, 2013. Citado 2 vezes nas páginas 61 e 62.

KOZINETS, R. V. The field behind the screen: Using netnography for marketing research in online communities. *Journal of marketing research*, SAGE Publications Sage CA: Los Angeles, CA, v. 39, n. 1, p. 61–72, 2002. Citado na página 21.

KOZINETS, R. V. *Netnography: Redefined*. 2. ed. Newbury Park, California, USA: SAGE Publications Ltd, 2015. Citado 2 vezes nas páginas 21 e 22.

LANDTHALER, J. et al. Extending thesauri using word embeddings and the intersection method. In: *ASAIL@ ICAIL*. London, UK: CEUR-WS, 2017. Citado na página 42.

LAZER, D. et al. Computational social science. *Science*, American Association for the Advancement of Science, v. 323, n. 5915, p. 721–723, 2009. Citado na página 21.

LIU, B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012. Citado 2 vezes nas páginas 44 e 73.

LUIZ, W. et al. A feature-oriented sentiment rating for mobile app reviews. In: INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE. *Proceedings of the 2018 World Wide Web Conference*. Lyon, France, 2018. p. 1909–1918. Citado na página 34.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. *Decision support systems*, Elsevier, v. 15, n. 4, p. 251–266, 1995. Citado na página 29.

MARCH, S. T.; STOREY, V. C. Design science in the information systems discipline: an introduction to the special issue on design science research. *MIS quarterly*, v. 32, n. 4, p. 725–730, 2008. Citado na página 29.

MATTHEWS, K. A. et al. Racial and ethnic estimates of alzheimer's disease and related dementias in the united states (2015–2060) in adults aged $\geq$ 65 years. *Alzheimer's & Dementia*, Elsevier, v. 15, n. 1, p. 17–24, 2019. Citado na página 83.

MEW, K. *Learning Material Design*. Birmingham, UK: Packt Publishing Ltd, 2015. Citado na página 53.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. Stateline, Nevada, USA: NIPS, 2013. p. 3111–3119. Citado na página 42.

MULLER, M. et al. Machine learning and grounded theory method: convergence, divergence, and combination. In: ACM. *Proceedings of the 19th International Conference on Supporting Group Work*. Sanibel Island, Florida, USA: ACM, 2016. p. 3–8. Citado na página 55.

NABLI, H.; DJEMAA, R. B.; AMOR, I. A. B. Efficient cloud service discovery approach based on lda topic modeling. *Journal of Systems and Software*, Elsevier, v. 146, p. 233–248, 2018. Citado na página 45.

NIELSEN, J.; MOLICH, R. Heuristic evaluation of user interfaces. In: ACM. *Proceedings of the SIGCHI conference on Human factors in computing systems*. USA: ACM, 1990. p. 249–256. Citado na página 62.

NISHI, J. M. et al. *A (re) construção do modelo UTAUT 2 em contexto brasileiro*. Tese (Doutorado) — Universidade Federal de Santa Maria, 2017. Citado na página 65.

O'CALLAGHAN, D. et al. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, Elsevier, v. 42, n. 13, p. 5645–5657, 2015. Citado na página 75.

O'DONOHOE, S. *Netnography: Doing ethnographic research online*. United Kingdom: Taylor & Francis, 2010. Citado na página 21.

PEFFERS, K. et al. A design science research methodology for information systems research. *Journal of management information systems*, Taylor & Francis, v. 24, n. 3, p. 45–77, 2007. Citado 2 vezes nas páginas 27 e 31.

PIMENTEL, M. Design science research: método de pesquisa científica com desenvolvimento de artefatos. In: SBC. *XIII Simpósio Brasileiro de Sistemas Colaborativos*. São Paulo, SP, Brasil, 2016. Citado 2 vezes nas páginas 28 e 32.

RIBEIRO, F. N. et al. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, SpringerOpen, v. 5, n. 1, p. 23, 2016. Citado na página 47.

SAURO, J. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Denver, Colorado, USA: Measuring Usability LLC, 2011. Citado na página 61.

SHNEIDERMAN, B. et al. Realizing the value of social media requires innovative computing research. *Communications of the ACM*, ACM, v. 54, n. 9, p. 34–37, 2011. Citado na página 22.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. Rio de Janeiro, Brasil: Elsevier Brasil, 2017. Citado na página 34.

SIMON, H. A. The sciences of the artificial (vol. 136). *MIT press, CambridgeSteyaert C, Katz J (2004) Reclaiming the space of entrepreneurship in society: geographical, discursive and social dimensions. Entrep Reg Dev*, v. 16, n. 3, p. 179196Vaghely, 1996. Citado na página 27.

SKEPPSTEDT, M. et al. Topics2themes: Computer-assisted argument extraction by visual analysis of important topics. In: *Proceedings of the 3rd Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources at LREC*. Potsdam, Germany: Applied Computational Linguistics, 2018. v. 18. Citado na página 36.

SONG, Y. et al. Topic and keyword re-ranking for lda-based topic modeling. In: ACM. *Proceedings of the 18th ACM conference on Information and knowledge management*. USA, 2009. p. 1757–1760. Citado na página 73.

STANDARDIZATION, I. O. for. *ISO 9241: Ergonomics of human-system interaction-Pt. 11: Usability: Definitions and concepts*. Vernier, Geneva, Switzerland: ISO, 2018. Citado 2 vezes nas páginas 61 e 62.

TAKEDA, H.; VEERKAMP, P.; YOSHIKAWA, H. Modeling design process. *AI magazine*, v. 11, n. 4, p. 37–37, 1990. Citado na página 27.

TENÓRIO, J. M. et al. Desenvolvimento e avaliação de um protocolo eletrônico para atendimento e monitoramento do paciente com doença celíaca. *Revista de Informática Teórica e Aplicada*, v. 17, n. 2, p. 210–220, 2010. Citado 2 vezes nas páginas 62 e 70.

TIRRONEN, V.; WEBER, M. Sparkline histograms for comparing evolutionary optimization methods. In: *IJCCI (ICEC)*. Jyvaskylaa, Finland: University of Jyvaskylaa, 2010. p. 269–274. Citado na página 48.

URQUHART, C.; FERNÁNDEZ, W. Using grounded theory method in information systems: the researcher as blank slate and other myths. In: *Enacting Research Methods in Information Systems: Volume 1*. Cham, Germany: Springer, 2016. p. 129–156. Citado na página 54.

VENKATESH, V. et al. User acceptance of information technology: Toward a unified view. *MIS quarterly*, JSTOR, p. 425–478, 2003. Citado na página 63.

VENKATESH, V.; THONG, J. Y.; XU, X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly*, v. 36, n. 1, p. 157–178, 2012. Citado na página 63.

VIJAYARANI, S.; ILAMATHI, M. J.; NITHYA, M. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, v. 5, n. 1, p. 7–16, 2015. Citado na página 39.

WANG, J. et al. Interactive topic model with enhanced interpretability. In: *IUI Workshops*. Los Angeles, USA: ACM, 2019. Citado na página 36.

WENINGER, T.; ZHU, X. A.; HAN, J. An exploration of discussion threads in social news sites: A case study of the reddit community. In: IEEE. *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. Niagara Falls, Canada, 2013. p. 579–583. Citado na página 74.

WIERINGA, R. J. *Design science methodology for information systems and software engineering*. Berlin: Springer, 2014. Citado na página 28.

YANG, Y.; YAO, Q.; QU, H. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics*, Elsevier, v. 1, n. 1, p. 40–47, 2017. Citado na página 36.

ZEMMOURI, E. M. et al. Goal driven approach to model interaction between viewpoints of a multi-view kdd process. In: IEEE. *2012 Next Generation Networks and Services (NGNS)*. New York, USA, 2012. p. 1–6. Citado na página 23.

ZHENG, K.; LI, A.; FARZAN, R. Exploration of online health support groups through the lens of sentiment analysis. In: SPRINGER. *International Conference on Information*. San Francisco, California, USA, 2018. p. 145–151. Citado na página 73.

ZOU, C.; HOU, D. Lda analyzer: A tool for exploring topic models. In: IEEE. *2014 IEEE International Conference on Software Maintenance and Evolution*. British Columbia, Canada, 2014. p. 593–596. Citado na página 36.

Apêndices

APÊNDICE A – Parecer do Comitê de Ética em Pesquisa (CEP)

Parecer do Comitê de Ética em Pesquisa (CEP) no qual é aprovada a condução da pesquisa de avaliação de aceitação e usabilidade da ferramenta e do processo de uso desenvolvidos.

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Avaliação de aceitação e usabilidade de ferramenta para apoio à análise interativa de comunidades virtuais

Pesquisador: Dárlinton Barbosa Feres Carvalho

Área Temática:

Versão: 2

CAAE: 11970919.2.0000.5151

Instituição Proponente: Universidade Federal de São João Del Rei-UFSJ/MG

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 3.381.044

Apresentação do Projeto:

O projeto em análise será coordenado pelo Prof. Dárlinton Barbosa Feres Carvalho do Departamento das Ciências da Computação e pelo prof. Rafael José de Alencar Almeida do Instituto Federal do Sudeste de Minas Gerais campus Barbacena. Segundo os propositores, o crescente volume de informações disponíveis publicamente em ambientes para discussão virtual, como fóruns e sites de redes sociais. Entretanto, sua análise manual apresenta-se inviável devido ao grande tempo consumido no processo, demandando sistemas de informação capazes de apoiar as pesquisas por meio de técnicas de mineração de dados. Sendo assim, foi desenvolvida uma ferramenta de mineração de dados para apoio à análise interativa de fóruns virtuais para pesquisadores de qualquer área do conhecimento. Desta forma, o experimento proposto consiste em avaliar a aceitação e usabilidade de tal ferramenta por usuários reais. Para tal, o avaliado passará por um pequeno treinamento no uso da ferramenta e a realização das atividades. Logo após, os participantes preencherão um questionário UTAUT2 – “Unified Theory of Acceptance and use of Technology2” e um questionário SUS – “System Usability Scale”, para avaliar, respectivamente, a aceitação e a usabilidade da ferramenta. O modelo UTAUT2, integra elementos de diversos modelos e teorias que ajudam a explicar a aceitação da tecnologia. Ele consiste em um questionário quantitativo que utiliza a Escala de Likert, onde o usuário indica sua intenção de uso em relação a diversos aspectos do sistema avaliado, possibilitando compreender o comportamento de usuários diante de uma tecnologia, e avaliar

assim a aceitação dos mesmos em relação a um sistema. Complementarmente, o modelo SUS tem como objetivo avaliar de forma qualitativa a usabilidade (facilidade de uso) de um sistema, consistindo em um questionário (também baseado na Escala de Likert) sobre a percepção do usuário em relação ao sistema, que terá como resultado final uma nota da usabilidade do sistema, normalizada em uma escala de 0 a 100.

Objetivo da Pesquisa:

De acordo com o pesquisar o propósito do presente estudo será ¨Avaliar a aceitação e a usabilidade da ferramenta de análise de comunidades virtuais, por meio de um experimento com usuários, utilizando os métodos UTAUT2 e SUS.¨

Avaliação dos Riscos e Benefícios:

De acordo com pesquisador, por se tratar de experimento envolvendo o uso de um software, há poucos riscos envolvidos durante sua condução. Nota-se, no entanto, o risco de desgaste mental (estresse) durante a realização das atividades utilizando a ferramenta de software. Para mitigar esse risco, todos os participantes serão orientados no início do experimento de que poderão interromper as atividades a qualquer momento, não tendo obrigação de participar de todas as atividades ou ir até o final. Não haverá benefício material ou financeiro para a participação no experimento, havendo apenas com benefício indireto a participação de uma aula interativa em um ambiente virtual e conhecimento de novas tecnologias. Foi especificado o grupo de estudo sendo solucionada tal pendência anterior, não havendo mais considerações neste item.

Comentários e Considerações sobre a Pesquisa:

De acordo com o referencial teórico avaliado este comitê entende que tal estudo possui relevância em sua área específica.

Considerações sobre os Termos de apresentação obrigatória:

Foram apresentados a folha de rosto com as devidas assinaturas. os termos de responsabilidade do pesquisador, bem como, do seu auxiliar. Além disso, foi apresentado o termo do diretor do IFET Sudeste campus Barbacena referendando a execução e disponibilidade de instalações para a execução do mesmo. Quanto ao TCLE, todas as solicitações apresentadas no parecer anterior foram atendidas. À título de sugestão este CEP recomenda a troca de identificação para apenas números do sujeito tendo em vista a extrema especificidade amostral o que poderia acarretar fácil identificação.

Recomendações:

Sugere-se apenas a substituição da forma de identificação do voluntário para números tendo em vista o tamanho amostral restrito e de propiciar fácil identificação.

Conclusões ou Pendências e Lista de Inadequações:

À luz deste novo encaminhamento este CEP entende que o projeto intitulado ¨Avaliação de aceitação e usabilidade de ferramenta para apoio à análise interativa de comunidades virtuais¨encontra-se apto para a sua execução quanto aos preceitos éticos legislados de acordo com o Conselho Nacional de Pesquisa em sua Resolução 466/2012 e da Norma Operacional 001/2013. Este CEP reitera a necessidade da realização do responsável do relatório final da pesquisa a ser encaminhado ao CEPSJ ao final do prazo execução da mesma.

Considerações Finais a critério do CEP:

O protocolo de pesquisa “Avaliação de aceitação e usabilidade de ferramenta para apoio à análise interativa de comunidades virtuais” e documentações apresentadas se encontram em consonância com os princípios éticos em pesquisa envolvendo seres humanos nos termos da Resolução 466/2012; 510/2016 e Norma operacional 001/2013. Somos, portanto, de parecer favorável à aprovação do referido protocolo e documentações. Informamos que relatórios parcial, e final da pesquisa devem notificados por meio da Plataforma Brasil e os resultados obtidos publicados e/ou encaminhados às instituições colaboradoras, aos órgãos e entidades representantes da sociedade

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1324203.pdf	18/05/2019 17:01:59		Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_____v2.DOC	18/05/2019 17:00:55	Dárlinton Barbosa Feres Carvalho	Aceito
Declaração de Instituição e Infraestrutura	diretor_geral.pdf	15/04/2019 14:58:56	Dárlinton Barbosa Feres Carvalho	Aceito
Projeto Detalhado / Brochura Investigador	projeto_de_pesquisa_plataforma_brasil.odt	15/04/2019 14:58:31	Dárlinton Barbosa Feres Carvalho	Aceito
Declaração de Pesquisadores	termo_rafael.pdf	15/04/2019 14:56:57	Dárlinton Barbosa Feres Carvalho	Aceito
Declaração de	termoresponsabilidade.pdf	11/04/2019	Dárlinton Barbosa	Aceito

Pesquisadores	termoresponsabilidade.pdf	10:54:48	Feres Carvalho	Aceito
Folha de Rosto	folhaderosto.pdf	11/04/2019 10:53:33	Dárlinton Barbosa Feres Carvalho	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

SAO JOAO DEL REI, 10 de Junho de 2019

Assinado por:
Jacqueline Domingues Tibúrcio
(Coordenador(a))

APÊNDICE B – Termo de Consentimento Livre e Esclarecido (TCLE)

Termo de Consentimento Livre e Esclarecido (TCLE) aprovado pelo Comitê de Ética em Pesquisa Envolvendo Seres Humanos das Unidades Educacionais de São João del-Rei (CEPSJ), da Universidade Federal de São João del-Rei (UFSJ). O termo foi lido e assinado por todos os participantes que aceitaram participar da pesquisa de avaliação de aceitação e usabilidade da ferramenta e do processo de uso desenvolvidos.

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Prezado participante,

Você está sendo convidado(a) a participar da pesquisa "AVALIAÇÃO DE ACEITAÇÃO E USABILIDADE DE FERRAMENTA PARA APOIO À ANÁLISE INTERATIVA DE COMUNIDADES VIRTUAIS", desenvolvida por RAFAEL JOSÉ DE ALENCAR ALMEIDA sob orientação do Prof. Dr. DÁRLINTON BARBOSA FERES CARVALHO, do Departamento de Ciência da Computação da Universidade Federal de São João del-Rei.

Você terá direito a ser esclarecido sobre o que desejar em qualquer momento da pesquisa. Não haverá nenhum custo por sua participação neste estudo, assim como não haverá também remuneração financeira por sua participação nesse estudo. Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa Envolvendo Seres Humanos das Unidades Educacionais de São João del-Rei (CEPSJ) da Universidade Federal de São João del-Rei (UFSJ), que tem a finalidade de proteger eticamente o participante. Portanto, caso ocorra dano decorrente de sua participação na pesquisa, por parte do pesquisador e das instituições envolvidas nas diferentes fases da pesquisa, conforme Resolução CNS nº 466/2012, reconhecemos seu direito a ressarcimento ou indenização de acordo com as leis vigentes no país.

O objetivo da pesquisa é avaliar a aceitação e a usabilidade da ferramenta de análise de comunidades virtuais, por meio de um experimento com usuários, utilizando os métodos UTAUT2 e SUS, pois os resultados podem beneficiar promovendo melhor compreensão acerca do artefato tecnológico criado. Destaca-se ainda um benefício direto para o participante de receber um treinamento sobre a realização de análises de comunidades virtuais utilizando uma ferramenta de apoio na interpretação dos resultados.

O convite a sua participação se deve aos seus conhecimentos de Lógica de Programação, por estar cursando ou pois já tenha cursado esta disciplina, além de gozar de maioridade e plena capacidade de ação. O pesquisador lhe endereçará o material via e-mail para a avaliação dos instrumentos de avaliação do SUS e do UTAUT2. Estando ciente disso, poderá aceitar ou recusar este convite no momento em que desejar, sem que sua decisão lhe traga quaisquer tipos de prejuízo.

A sua participação é voluntária, isto é, **ela não é obrigatória**, possuindo plena autonomia para decidir se quer ou não participar, bem como retirar seu consentimento ou sua participação a qualquer momento. Sua recusa não trará nenhum prejuízo em sua relação com os pesquisadores, ou com a instituição. Contudo, ela é muito importante para a execução da pesquisa.

Serão garantidas a confidencialidade e a privacidade das suas informações prestadas, sendo toda a avaliação realizada no local de exame. Na presente pesquisa, você será identificado pelas iniciais de seu nome, bem como, um número que será intransferível. Qualquer dado que possa identificá-lo será omitido na divulgação dos resultados da pesquisa, e o material será armazenado em local seguro. A qualquer momento, durante a pesquisa, ou posteriormente, você poderá solicitar do pesquisador informações sobre sua participação e/ou sobre a pesquisa, o que poderá ser feito através dos meios de contato explicitados neste documento.

Os riscos para os participantes da pesquisa são mínimos e estão relacionados ao desgaste mental (estresse) durante a realização das atividades utilizando a ferramenta de software. Caso isso ocorra, serão oferecidas pausas na realização do experimento para recuperação bem como apoio e encaminhamento médico caso seja necessário. Vale ainda ressaltar que o tempo estimado para a realização das atividades propostas é inferior a 1h de modo a minimizar este risco.

O cenário foi construído focando nos objetivos deste estudo, a fim de se enquadrar na situação de uso em que será simulado um estudo com o objetivo de aproximar o máximo possível o uso profissional da ferramenta. Dessa maneira, os participantes serão observados pelo(a) facilitador(a) no ato da utilização da ferramenta, bem como utilizar-se-á dos instrumentos de avaliação (questionários) do SUS e UTATU2, cujo objetivo é avaliar sistematicamente o artefato tecnológico utilizado pelos participantes. A sua participação consistirá em:

Etapa 1 (até 5 minutos de duração): a ferramenta será apresentada para os participantes do experimento, descrevendo seu modo de uso e suas funcionalidades, envolvendo:

- Selecionar e carregar os dados de uma comunidade virtual
- Selecionar os parâmetros de análise (processo KDD)
- Navegar e interpretar os gráficos e indicadores produzidos pela análise
- Exportar os resultados

Etapa 2 (até 15 minutos de duração): serão propostas aos participantes as seguintes tarefas:

- Carregar a base de dados da comunidade sobre programação
- Iniciar a análise, configurando livremente o valor dos parâmetros
- Explorar e interpretar os resultados da análise, identificando:
  - Três assuntos de destaque discutidos no fórum
  - Um assunto predominantemente negativo
  - Um assunto predominantemente positivo
  - Um assunto recorrente
  - Um assunto emergente (recente)
- Exportar os dados gerados

Etapa 3 (até 5 minutos de duração): os participantes deverão responder ao questionário UTAUT2 para avaliação da aceitação da ferramenta.

Etapa 4 (até 5 minutos de duração): os participantes deverão responder o questionário SUS (System Usability Scale), para avaliação da usabilidade da ferramenta.

Um pequeno treinamento de uso da ferramenta de análise de comunidades virtuais (5 minutos) e realizarão atividades propostas que envolvam o uso da ferramenta para se analisar uma comunidade virtual sobre lógica de programação (15 minutos). Após a realização das atividades, os participantes preencherão um questionário UTAUT2 (5 minutos) e um questionário SUS (5 minutos). E que poderá ser interrompido por solicitação do participante.

Ao final da pesquisa, o projeto, o protocolo e os relatórios serão mantidos em arquivo pelo CEPSJ, por um período de 5 anos após o encerramento do estudo, podendo esse arquivamento processar-se em meio digital, conforme Resolução 466/12 e orientações do CEPSJ.

Os resultados gerais poderão ser divulgados de forma agregada em palestras dirigidas ao público participante, artigos científicos e em dissertações ou teses acadêmicas. Os resultados de forma individual poderão ser repassados aos participantes estando a equipe de pesquisadores à disposição para eventuais esclarecimentos.

Poderão ser feitas algumas imagens durante a realização dos procedimentos neste estudo para fazer parte dos dados para estudo ou divulgadas em periódicos e reuniões científicas.

Por favor, sinta-se à vontade para fazer qualquer pergunta sobre este estudo. Se outras perguntas surgirem mais tarde, poderás entrar em contato com os pesquisadores. Em caso de dúvida quanto à condução ética do estudo, entre em contato com o Comitê de Ética em Pesquisa da CEPSJ. O Comitê de Ética é a instância que tem por objetivo defender os interesses dos participantes da pesquisa em sua integridade e dignidade e para contribuir no desenvolvimento da pesquisa dentro de padrões éticos. Dessa forma o comitê tem o papel de avaliar e monitorar o andamento do projeto de modo que a pesquisa respeite os princípios éticos de proteção aos direitos humanos, da dignidade, da autonomia, da não maleficência, da confidencialidade e da privacidade.

Tel e Fax - (0XX) 32- 3379- 5598 e-mail: cepsj@ufsj.edu.br
Endereço: Praça Dom Helvécio, 74, Bairro, Dom Bosco, São João del-Rei, MG, CEP: 36301-160, Campus Dom Bosco

Se desejar, consulte ainda a Comissão Nacional de Ética em Pesquisa (Conep): Tel: (61) 3315-5878 / (61) 3315-5879 e-mail: conep@saude.gov.br

Contato com o pesquisador a responsável: Prof. Dr. Dárlinton Carvalho
Email: darlinton@ufsj.edu.br **Telefone:** (32) 3379-4935

Declaro que entendi os objetivos e condições de minha participação na pesquisa e concordo em participar. Declaro que este documento foi elaborado em duas vias, rubricadas em todas as suas páginas e assinadas, ao seu término, pelo convidado a participar da pesquisa, assim como pelo pesquisador responsável, ou pela(s) pessoa(s) por ele delegada(s).

Barbacena, _____ de _____ de _____.

_____	_____
Nome do Participante	**Assinatura do Participante**
_____	_____
Nome do Pesquisador	**Assinatura do Pesquisador**