Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Explorando postagens em rede social descentralizada e acadêmica: mecanismo para identificação de trending topics e conexão com temas de disciplinas curriculares

Amanda Augusto da Silva
Hugo Pereira Rezende
Italo Marcos Brandão

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Orientadora
Prof.a Dr.a Germana Menezes da Nóbrega

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Explorando postagens em rede social descentralizada e acadêmica: mecanismo para identificação de trending topics e conexão com temas de disciplinas curriculares

Amanda Augusto da Silva
Hugo Pereira Rezende
Italo Marcos Brandão

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Prof.a Dr.a Germana Menezes da Nóbrega (Orientadora)
CIC/UnB

Prof. Dr. Jorge Henrique Cabral Fernandes
CIC/UnB

Prof. Dr. Edison Ishikawa
CIC/UnB

Prof. Dr. Jorge Henrique Cabral Fernandes
Coordenador do Curso de Computação — Licenciatura

Brasília, 28 de Abril de 2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

AAU923e

Augusto da Silva, Amanda
Explorando postagens em rede social descentralizada e acadêmica: mecanismo para identificação de trending topics e conexão com temas de disciplinas curriculares / Amanda Augusto da Silva, Hugo Pereira Rezende, Italo Marcos Brandão; orientador Germana Menezes da Nóbrega. -- Brasília, 2023.
89 p.

Monografia (Graduação - Computação Licenciatura) -- Universidade de Brasília, 2023.

1. educação formal e informal. 2. trending topics. 3. análise de trending topics. 4. redes sociais descentralizadas e educação. 5. processamento de linguagem natural. I. Italo Marcos Brandão, Hugo Pereira Rezende,. II. Menezes da Nóbrega, Germana, orient. III. Título.

Dedicatória

Amanda Silva

Eu dedico este trabalho primeiramente aos meus pais, Elizabeth e Vilmar, que sempre me incentivaram aos estudos, e ao meu irmão, Murilo, que me deu suporte em todas as minhas escolhas. Dedico também ao Alexandre, que esteve presente em todas as etapas da minha graduação e me apoiou nesse período de conclusão. Aos meus amigos que estiveram sempre ao meu lado nos momentos felizes e também de dificuldades, o meu muito obrigada, em especial para minha amiga Kelly que sempre me apoiou em todos os momentos.

Hugo Rezende

Eu Dedico esse trabalho a Deus, aos meus pais, Julio e Rita, meus irmãos Pedro e Victor, minhas avós Nilda e Natalice, a todos os amigos que fizeram parte da minha vida durante esses longos anos de graduação.

Italo Marcos

Eu dedico este trabalho aos meus pais, Aparecido e Ariene, e ao meu irmão Gabriel, pela compreensão e suporte durante toda a jornada de graduação.

Agradecimentos

Agradecemos primeiramente à professora e orientadora Dra. Germana Menezes da Nóbrega, por ter nos auxiliado ao longo deste projeto e contribuído de forma significativa para a conclusão deste. Agradeçemos também à Universidade de Brasília, que nos proporcionou todas as oportuniades e estrutura para nossa formação, e nos permitiu contato com vários professores e colegas que fizeram parte da nossa evolução pessoal e acadêmica. Por fim, agradedcemos à todos que tenham contribuído de uma forma direta ou indireta para que tenhamos chegado nesse momento.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Este trabalho tem como objetivo abordar a integração do ambiente formal e informal de ensino do departamento de Ciência da Computação da Universidade de Brasília, por meio da proposta de implementação de uma ferramenta geração de *trending topics* na rede social CICFriend, a fim de explorar as possibilidades de aprendizagem e engajamento dos alunos durante o processo de ensino-aprendizado, utilizando os assuntos mais comentados pela comunidade e promovendo uma educação mais contextualizada e relevante. Além disso, disponibilizar um protótipo funcional de geração de *trending topics* a partir de postagens de uma rede social utilizando Processamento de Linguagem Natural.

Palavras-chave: formal, informal, redes sociais descentralizadas, *trending topics*, educação, processamento de linguagem natural

Abstract

This work aims to address the integration of formal and informal teaching environment of the Computer Science department of the University of Brasilia, through the proposal of implementation of a tool to generate trending topics in the social network CICFriend, in order to explore the possibilities of learning and engagement of students during the teaching-learning process, using the most commented subjects by the community and promoting a more contextualized and relevant education. In addition, to provide a functional prototype for generating trending topics from social network posts using Natural Language Processing.

Keywords: formal, informal, decentralized social media, trending topics, education, social networks, natural language processing

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

CIC Departamento de Ciência da Computação.

DCN Diretrizes Curriculares Nacionais.

OSN Online Social Network.

RSD Rede Social Descentralizada.

RSDs Redes Sociais Descentralizadas.

RSO Rede Social Online.

RSOs Redes Sociais Online.

SBC Sociedade Brasileira De Computação.

TICs Tecnologias da Informação e Comunicação.

TTs *Trending Topics*.

UnB Universidade de Brasília.

Capítulo 1

Introdução

Este trabalho visa investigar os benefícios e desafios potenciais da implementação de um *add-on* para disponibilização de uma *feature* de *trending topics* no CiCFriend, uma instância da rede social descentralizada Friendica, que está implantada no ambiente smartUnB.ECOS. A proposta do *add-on* utiliza um algoritmo de processamento de linguagem natural para extrair palavras e tópicos de dados gerados pelos usuários através dos posts, comentários e compartilhamentos. O Algoritmo analisa esses dados para identificar padrões e frequência nas postagens dos usuários da plataforma que podem ser relevantes para a comunidade universitária, relacionando-os aos conteúdos, competências e tópicos das disciplinas.

1.1 Contextualização

A tecnologia evoluiu rapidamente nas últimas décadas, e esse progresso provocou grandes alterações no estilo de vida da sociedade. Pode-se observar o aumento do uso de dispositivos móveis como celulares e *tablets*, além do aumento do uso da Internet, que vem mudando a forma com que as pessoas conversam, se informam e participam da sociedade[1]. Atualmente as pessoas passam mais tempo conectadas à Internet que desconectadas dela, dessa forma, isso faz com que elas busquem meios para que as relações interpessoais passem a acontecer também *online*.

Os dados publicados pela Ookla[2] indicam que os usuários de Internet no Brasil poderiam esperar as seguintes velocidades de conexão no início de 2022: a velocidade média de conexão móvel via redes celulares foi de 22.60 Mbps, já a velocidade média de conexão à Internet fixa foi de 83.25 Mbps. A análise revela que a velocidade média no Brasil para a conexão móvel aumentou em 3.42 Mbps (+17.8%) nos 12 meses anteriores ao início

[1] `https://datareportal.com/reports/digital-2022-global-overview-report`
[2] `https://www.ookla.com/articles/global-index-market-analyses-q1-2022#brazil`

de 2022. Por outro lado, os dados da Ookla mostram que as velocidades de conexão à Internet fixa no Brasil aumentaram em 39.31 Mbps (+89.5%) durante o mesmo período.

A análise da Kepios[3] revela que os usuários de mídias sociais no Brasil aumentaram em 21 milhões (+14,3%) entre 2021 e 2022, além de mostrar que em janeiro de 2022, havia 171,5 milhões de usuários de mídias sociais no Brasil. O número de usuários no início de 2022 foi equivalente a 79,9% da população total, mas é importante lembrar que este número pode não representar indivíduos únicos.

Desse modo, o cotidiano das pessoas em todo o mundo está cada vez mais ligado às redes sociais. Segundo a pesquisa[3], foram gastas 12,5 trilhões de horas *online*, e esse número é um novo marco na adoção da Internet. Também tiveram novos recordes de tempo de uso e números de usuários de mídias sociais. Os serviços de *microblogging*, como uns dos serviços representativos das Redes Sociais Online (RSOs), proporcionam uma abordagem conveniente para todos ao redor do mundo para ler notícias, enviar mensagens e trocar opiniões em relação aos serviços tradicionais de mídia, como TV ou jornal.

O papel das Redes Sociais Descentralizadas (RSDs) tem evoluído bastante junto com as novas tecnologias como o aumento da velocidade de conexão em dispositivos móveis a partir do 3G, 4G e atualmente o 5G. Isso vem causando mudanças significativas em vários âmbitos da sociedade, e consequentemente, no campo educacional. De acordo com [1], as Tecnologias da Informação e Comunicação (TICs) proveem recursos para favorecer e enriquecer as aplicações e processos da área de educação.

No contexto pós-pandêmico (COVID-19), é notória a necessidade de buscar novas formas de auxiliar a aprendizagem e inovar os ecossistemas virtuais de interação entre os alunos, que foram fortemente prejudicados com as soluções temporárias do ensino remoto emergencial. É importante ressaltar que estudos confirmaram um aumento significativo de perturbações psicológicas como ansiedade, depressão e estresse entre os estudantes universitários no período pandêmico comparativamente a períodos anteriores [2].

Dessa maneira, torna-se interessante propiciar uma plataforma de integração para o ambiente universitário, servindo para fomentar a comunicação e colaboração de docentes e discentes. Uma solução para tal plataforma se apresenta na forma de uma RSD, levando em conta sua natureza de promover autonomia e transparência aos usuários da rede, sobre como e quais dados pessoais estão sendo utilizados, além de possuírem código aberto, o que permite os usuários configurarem e controlarem a rede de acordo com suas preferências. Em [3], é defendido que as RSDs têm o potencial de proporcionar um melhor ambiente dentro do qual os usuários podem ter mais controle sobre sua privacidade, e a propriedade e disseminação de suas informações.

[3] `https://datareportal.com/reports/digital-2022-global-overview-report`

Dado esse cenário, uma área bastante visada é a de análise das tendências temáticas nas mídias sociais, as quais têm se tornado cada vez mais importantes nos últimos anos, pois permitem uma compreensão mais profunda das conversas e sentimentos em torno de tópicos específicos. Este tipo de análise pode ser utilizada para diversos fins, incluindo marketing e relações públicas [4], bem como pesquisa política e social [5]. Um estudo de Kwak [6] descobriu que os *Trending Topics* (TTs) no Twitter são capazes de prever com precisão eventos do mundo real, tais como flutuações na bolsa de valores e vendas de bilheteria. Além disso, um estudo de Bollen [7] descobriu que a análise de sentimentos desses TTs pode ser usada para prever o estado de espírito da população em geral. Como as redes sociais continuam a desempenhar um papel maior na sociedade, a capacidade de compreender e analisar os tópicos em tendências só se torna mais crucial.

Nesse contexto, o presente trabalho está inserido no ambiente smartUnB.ECOS, um projeto de ecossistema educacional digital para campus universitário que busca prover a interoperabilidade de ferramentas de comunicação e de educação a fim de fomentar a socialização e a aprendizagem [8], conforme Figura 1.1. Nesse projeto, juntamente com trabalhos anteriores [9, 10], foi implantada a rede social descentralizada Friendica para a comunidade do Departamento de Ciência da Computação (CIC) da Universidade de Brasília (UnB).

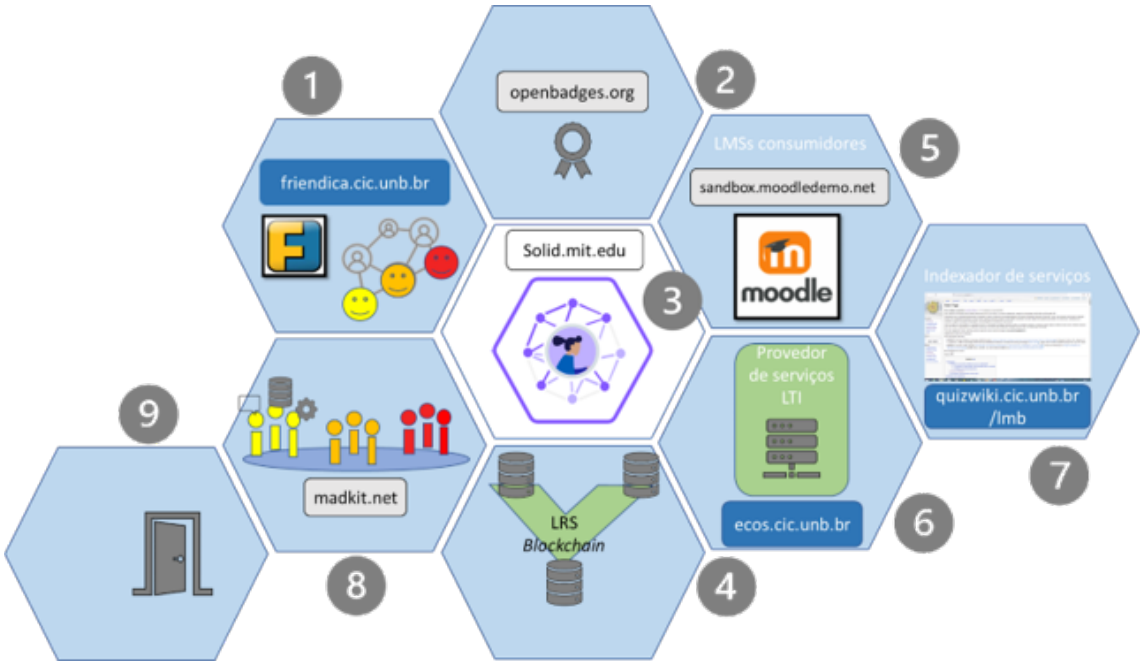

Figura 1.1: smartUnB.ECOS [11]

1.2 Motivação e Justificativa

Um dos objetivos do processo de educação do ensino superior é o preparo dos alunos para a vida, no entanto, muitas vezes percebe-se que no ensino formal há uma falta de interação dos conteúdos aprendidos em sala de aula com o mundo real, e até uma falta de prática sobre a aplicabilidade desses na vida dos estudantes e no mercado de trabalho. Em contrapartida do contexto formal, as redes sociais são utilizadas como um campo informal que registra conversas e debates, tornando-se uma fonte relevante de aprendizado e troca de informações. Além disso, as redes sociais têm trazido benefícios para o ambiente acadêmico, como a facilidade de acesso a conteúdos, a possibilidade de conexão com outros estudantes e pesquisadores e a oportunidade de discussão e colaboração entre os usuários. Esses benefícios podem ser melhor aproveitados com o uso das RSDs, devido a possibilidade de personalização e extensibilidade de suas funcionalidades a partir de *add-ons*.

Tendo em vista os trabalhos realizados anteriormente em [9, 12], os quais apresentaram resultados positivos em relação a novas funcionalidades a fim de aumentar o engajamento da comunidade do CIC na rede social CiCFriend, e de modo a enriquecer o debate entre alunos e professores com conteúdos atuais que possam ser conectados aos tópicos sugeridos pelas matérias nas matrizes curriculares, o presente trabalho visa investigar os benefícios da implementação de um *add-on* de *trending topics* no CiCFriend para apresentação dos temas mais discutidos pelos usuários da comunidade acadêmica. O uso de *trending topics* como fonte de *insights* em diversas áreas foi previamente estudado na literatura, mas pouco investigado no âmbito educacional. Além disso, o uso de uma rede social descentralizada como plataforma para esta funcionalidade ainda não foi amplamente explorado.

1.3 Objetivos

1.3.1 Objetivo Geral

No âmbito do projeto smarUnB.ECOS, e o estudo feito anteriormente em [12], que aborda as definições de aprendizagem formal e informal, o presente trabalho tem como objetivo geral contribuir com uma integração do ambiente de ensino formal com o informal do CIC, onde o primeiro pode ser considerado como as matérias e disciplinas do departamento, e o segundo a rede social CICFriend. Essa integração pode ser realizada através da proposta de implementação de uma ferramenta de geração de *trending topics*, onde a comunidade do departamento teria um acesso mais rápido e facilitado aos assuntos mais comentados na rede, incorporando-os aos estudos dos alunos em sala de aula.

1.3.2 Objetivos Específicos

- Prover protótipos de tela da ferramenta de *trending topics* no CICFriend, apresentando as funcionalidades propostas a partir desse *add-on*;
- Prover protótipo de mecanismo de geração de *trending topics*;
- Propor uma integração dos tópicos do *trending topics* com os conteúdos abordados nas disciplinas do CIC.

1.4 Pressupostos

- O *add-on* do Friendica proposto neste trabalho poderá ser utilizado para futura implantação no CICFriend;
- O ecossistema smartUnB.ECOS utiliza redes sociais descentralizadas como componente. Este trabalho visa apresentar uma ideia de nova funcionalidade à RSD como forma de fomentar o uso e o engajamento dos alunos e professores na rede social;
- O *add-on* sugerido poderá ser disponibilizado para todos os tipos de contas da rede social Friendica para melhor aproveitamento e estímulo à trocas sobre os temas em tendências do cotidiano, a fim de relacioná-los às discussões durante o processo de ensino-aprendizagem;
- Existe espaço e capacidade de processamento nos servidores utilizados pela Universidade de Brasília para a adição da aplicação sugerida.

1.5 Estrutura do Trabalho

Este trabalho está estruturado em 6 capítulos, incluindo este capítulo de introdução, onde é realizada a contextualização do problema a ser apresentado ao longo do documento, seguido da motivação e justificativa, objetivos e as premissas, suposições e hipóteses assumidas para o desenvolvimento da pesquisa.

Em seguida, é desenvolvido o capítulo de temáticas e trabalhos relacionados, onde são apresentadas descrições de todo o referencial teórico utilizado como base para a compreensão dos aspectos do campo em que o trabalho está inserido.

No capítulo 3, têm-se os fundamentos, onde são expostos os tópicos fundamentais para o entendimento do estudo realizado, seguido pelo capítulo 4, o qual detalha a proposta de implementação da *feature* de *trending topics*, complementado pelo capítulo 5, que apresenta a arquitetura de uma aplicação de demonstração.

Por fim, é apresentado o capítulo de conclusão, relatando conclusões, objetivos alcançados e sugerindo possíveis caminhos a serem tomados futuramente.

Capítulo 2

Temáticas e trabalhos relacionados

O objetivo deste capítulo é fornecer uma visão geral da literatura existente sobre o uso da análise de *trending topics* em redes sociais como fonte de informação, assim como destacar os desafios e os resultados da implementação de redes sociais em ambientes educacionais. Além disso, apresenta estudos anteriores que são relevantes para as pesquisas atuais.

2.1 Redes Sociais

As redes sociais se tornaram um grande exemplo de globalização, conectando todo o mundo com nenhum tipo de limite geográfico. Com todo o avanço dos últimos anos das Tecnologias da Informação e Comunicação (TICs), houve um impacto profundo no mundo moderno, e isso vem sendo alvo de análises de diferentes dimensões. Os impactos das Redes Sociais Online(RSOs) na sociedade são discutidos por vários campos de conhecimento, como comentado por [13], desde as ciências sociais até a matemática ou biologia, afirmam os autores:

> O estudo das redes sociais vem acompanhado de transformações nos campos paradigmáticos, nas técnicas de investigação, bem como na recorrência a tradicionais campos de pesquisa e na criação de novos universos de construção do conhecimento.

No trabalho [14], uma RSO é definida como uma plataforma online que (1) fornece serviços para um usuário construir um perfil público e declarar explicitamente a conexão entre seu perfil e de outros usuários; (2) permite que um usuário compartilhe informações e conteúdo com os usuários escolhidos ou com o público. Além disso, os autores afirmam que as RSOs tiveram um impacto significante nas mudanças de paradigmas em aspectos sócio-econômicos e técnicos da colaboração e interação, comparável ao causado pela implantação da *World Wide Web*(WWW) na década de 1990.

Para muitos usuários, as RSOs se tornaram uma parte indispensável de suas vidas [15], e levando em consideração os avanços nas tecnologias da *internet*, os sistemas entregam

cada vez mais possibilidades e atividades para o dia a dia das pessoas [16]. Como afirmado em [13], do ponto de vista da tecnologia as aplicações estão cada vez mais complexas, porém mais simples para o usuário interagir. Esses benefícios estão diretamente ligados com o número crescente de usuários de RSOs, mostrado na Figura 2.1[1], e como citado por [17]:

> A facilidade de acesso às informações, as diversas ferramentas disponíveis gratuitamente e a rápida interação proporcionada fizeram com que o número de usuários das redes sociais digitais fosse maximizado.

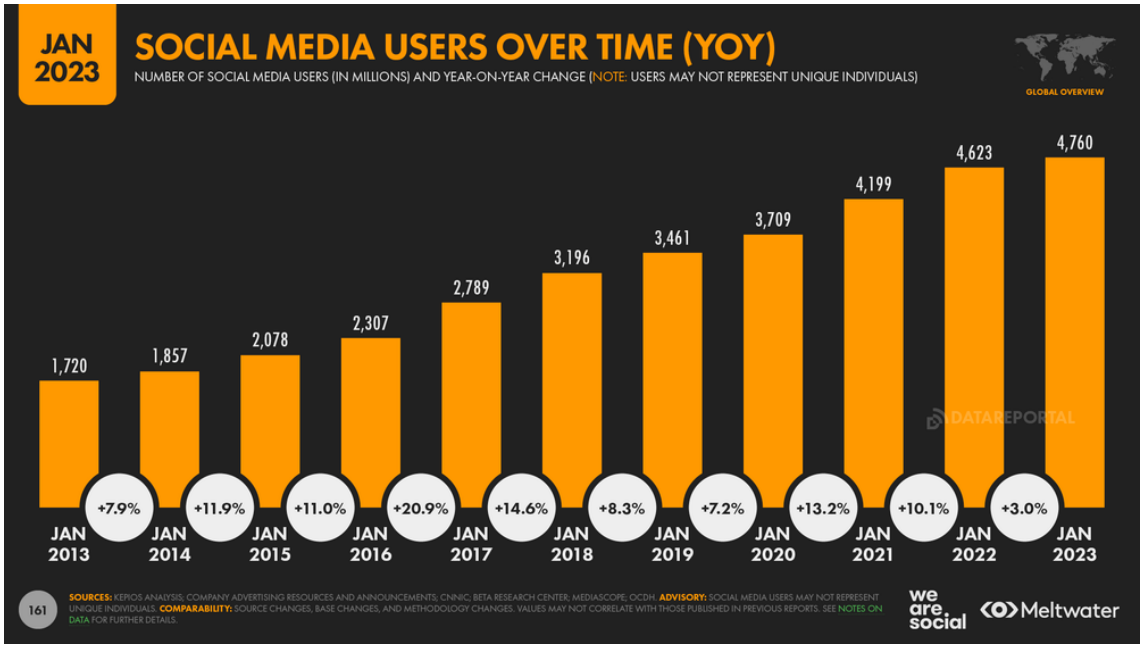

Figura 2.1: Gráfico que mostra o aumento dos usuários de mídias sociais ao longo dos anos [18].

A pesquisa de Kepios[1] mostrou que atualmente existem 4,76 bilhões de usuários de mídia social em todo o mundo, o que equivale a pouco menos de 60% da população global total. Além disso, as redes sociais favoritas dos usuários ativos entre 16 e 64 anos ao redor do mundo são o WhatsApp, Instagram e Facebook, com os percentuais de favoritismo de 15,8%, 14,3% e 14,2% respectivamente. Por outro lado, a mesma pesquisa indica que o Facebook ainda é a rede social mais acessada do mundo em relação ao tempo gasto de uso diário na plataforma, com 19 horas e 43 minutos, seguido do Instagram com 12 horas.

Nesse contexto, é observado o quanto a importância e relevância das redes sociais foram evoluindo ao longo do tempo, tornando- se parte da vida das pessoas diariamente. A autora de [19], afirma que

[1] `https://datareportal.com/reports/digital-2023-global-overview-report`

> [...] a lógica da Internet como plataforma de rede social facilita às pessoas a oportunidade de se associarem a outros com quem partilham interesses, encontrar novas fontes de informação e publicação de conteúdo e opinião.

2.2 Utilização de Redes Sociais Na Educação

Mesmo com o crescente uso de plataformas de mídias sociais, como Facebook, LinkedIn, etc., na Educação Superior como ferramentas de aprendizado, a adoção delas como instrumentos formais de ensino e aprendizagem pelos educadores ainda é bastante limitada e sujeita a muitas restrições [20], além de questões de privacidade e segurança, e o potencial de distração e diminuição da produtividade [21]. Apesar disso, a prática da implementação das redes sociais para o auxílio da educação é defendida por vários autores.

Atualmente, há um crescente número de pesquisas que abordam a influência das redes sociais nas universidades [22, 23, 24]. Em [25] os autores defendem que o uso de tecnologias como Twitter e blogs podem, em conjunto, ser um incentivo para permitir que tanto estudantes quanto professores participem ativa e instantaneamente e se comuniquem uns com os outros em atividades educacionais sem depender apenas do contato pessoal e dentro de sala de aula.

As oportunidades oferecidas por essas plataformas na educação não se limitam a aumentar as interações entre os estudantes e os instrutores, mas também se estendem para apoiar o aprendizado baseado em evidências, como a Aprendizagem Cooperativa (AC) apresentada em [26], que também afirma que:

> Muitos pesquisadores (Antil et al., 1998; Cohen Lotan, 1997; Dyson et al., 2004; Dyson Casey, 2012; Perkins, 1999) fizeram as conexões entre a AC e a teoria do construtivismo social. Os construtivistas sociais acreditam que o aprendizado é um processo social e só pode ser alcançado através de ensino recíproco, colaboração entre pares, aprendizagem cognitiva, instrução baseada em problemas, instrução ancorada e outros métodos que envolvem interações com outros (Shunk, 2000). O princípio desta estrutura teórica está nas descobertas de um aprendiz engajado, ativo e criativo (Rovegno Dolly, 2006).

Além disso, a pesquisa apresentada em [21], defende que a mídia social pode ser usada para apoiar uma ampla gama de atividades de aprendizagem(Figura 2.2), incluindo comunicação, colaboração e criação de conteúdo. Também foi observado que essas redes podem servir de apoio para abordagens de ensino e aprendizagem, como construtivismo, aprendizagem baseada em problemas e aprendizagem combinada. Os autores destacam que as redes sociais favorecem uma série de objetivos educacionais, como promover o engajamento, aumentar a motivação e melhorar o desempenho dos alunos.

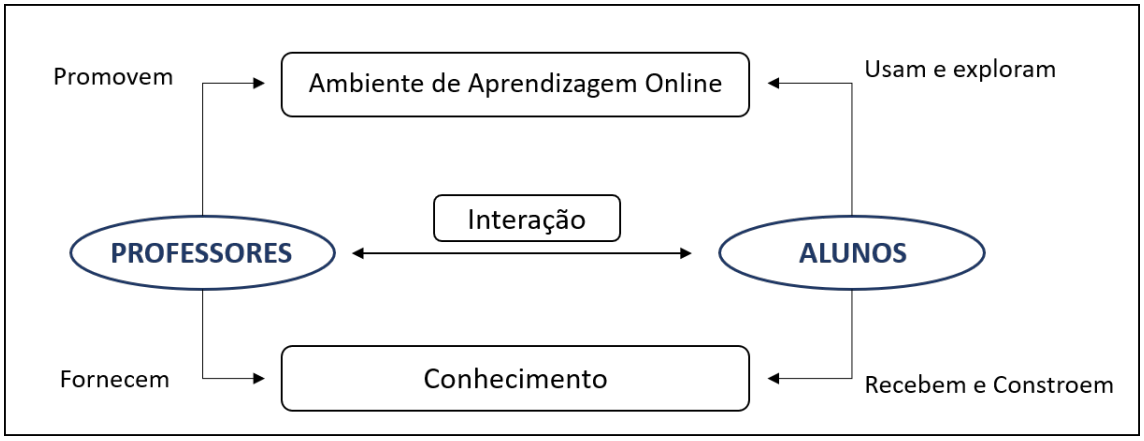

Figura 2.2: Relação de professores e estudantes com ambiente online (Figura adaptada de [21]).

Trabalhos como [8, 27, 9] evidenciam os benefícios da incorporação de redes sociais na educação, nesse contexto de expandir os meios de integração entre os agentes que participam do processo de ensino-aprendizagem. A integração de Online Social Network (OSN) no ambiente educacional traz várias vantagens tais como a criação de uma comunidade fomentando o engajamento dos alunos, aumentando as realizações, facilitando a gestão da informação e promovendo o compartilhamento de conhecimentos entre eles.

2.3 Caracterização de Disciplinas Curriculares

O papel das RSOs na educação vem sendo mais significativo com o passar dos anos. Conforme abordado em [28], uma rede social não é apenas uma ferramenta, é um lugar tão real e natural como qualquer lugar da vida onde interações sociais formais/informais acontecem. O autor afirma que, ainda assim, os contextos formais de educação superior ainda estão, em sua maioria, presos em sistemas de gestão de aprendizagem institucionais fechados, enquanto um novo mundo de conexões sociais cresce e se desenvolve fora das escolas.

Dito isso, através das diretrizes formais que regulam o ensino superior é possível extrair informações para que sejam aplicadas em ambientes informais, como por exemplo as competências e eixos de formação explicados a seguir, os quais podem ser utilizados como base para construir meios de ensino informais que relacionem essas habilidades a situações da vida real, formando um pensamento crítico e criativo.

No Brasil, os cursos de graduação na área da computação precisam ser realizados de acordo com as Diretrizes Curriculares Nacionais (DCN) [29], ou DCN16, do Ministério da Educação (MEC). Essas diretrizes direcionam as instituições a oferecerem aos egressos uma formação sólida para enfrentar os desafios seja no mercado de trabalho ou na área

acadêmica. Elas servem de base para as universidades realizarem cursos na área de tecnologia da informação e criação de projetos pedagógicos, e também para projetarem seus cursos com uma identidade e matriz curricular própria [29].

A Sociedade Brasileira De Computação (SBC)[2] é uma organização sem fins lucrativos que traz estudantes, professores, pesquisadores e profissionais da área da computação. A SBC possui alguns objetivos tais como:

- Fomentar o acesso à informação através da tecnologia.
- Promover inclusão digital.
- Encorajar a pesquisa e o ensino de computação no Brasil.
- Contribuir para o treinamento de profissionais de computação com responsabilidade social.

Os membros da SBC também são responsáveis por discutir como os cursos de graduação devem ser realizados, através de comitês que devem acontecer pelo menos de dez em dez anos. Eles também têm a responsabilidade de criar currículos e diretrizes e discutir as formas de avaliar estes cursos pelo MEC.

Além disso, a SBC foi responsável pela criação do documento dos Referenciais de Formação (RFs), os quais realizam uma organização das competências e habilidades apresentadas pela DCN, conforme mostrado na Figura 2.3, servindo de auxílio e referência para os coordenadores de curso de graduação na elaboração dos Planos Políticos Pedagógicos (PPCs) de todos os cursos da área de computação. Esses PPCs devem, obrigatoriamente, estar em conformidade com a DCN vigente [30]. As etapas da estrutura de organização realizada pelos RFs, foi descrita em [31] como:

> [...]o perfil esperado para o egresso determina o objetivo geral do curso, decomposto em diferentes eixos de formação. Os eixos de formação objetivam capacitar o egresso em competências genéricas. Para alcançar cada competência, são relacionadas diversas competências derivadas, que determinam a necessidade de serem desenvolvidas em conteúdos específicos.

[2] `sbc.org.br/institucional-3/sobre`

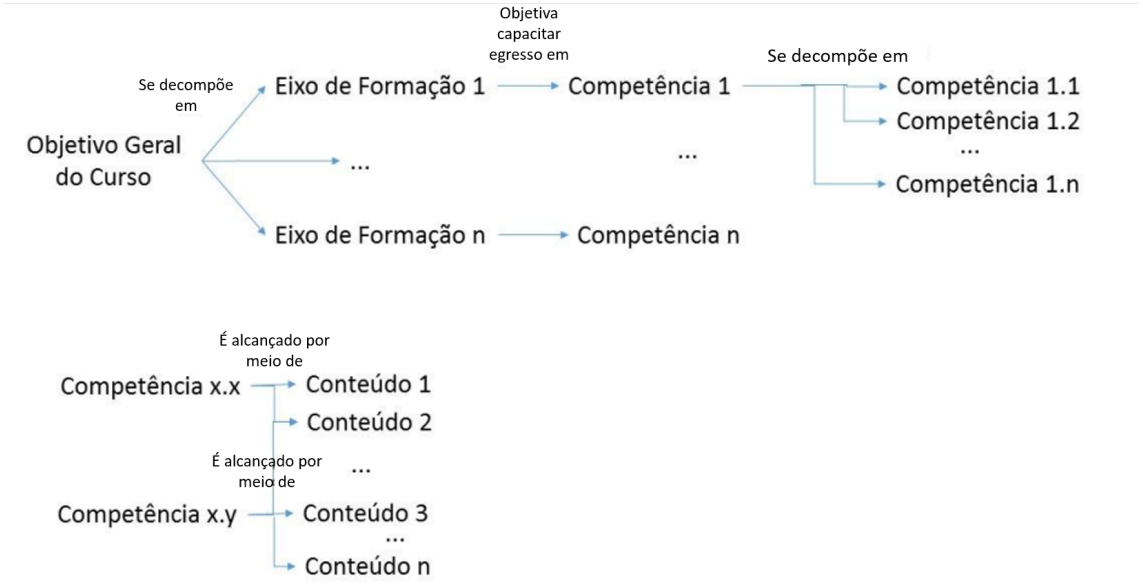

Figura 2.3: Estrutura conceitual dos Referenciais de Formação em Computação[31]

Os PPCs devem ser elaborados e atualizados pelas Instituições de Ensino Superior (IES), associando os conteúdos curriculares às competências presentes na DCN. Entretanto, compete às IES definir uma estratégia de como usar essas informações e referências na elaboração do PPC, em especial, na descrição dos componentes curriculares do curso.

A partir disso, cabe a cada curso das IESs realizar a criação das ementas das disciplinas do curso, descrevendo seus conteúdos. É importante que esses conteúdos apresentem relação com os conteúdos, competências e eixos do RF.

2.4 *Trending topics*

Trending topics em serviços de *microblogging* se referem a quando um tópico específico ou uma *hashtag* é mencionada ou discutido por uma grande quantidade de usuários dentro de um curto período de tempo, eles são gerados e atualizados conforme os eventos de entrada de informação. Como as plataformas de mídia social se tornaram uma fonte primária de notícias e informações, os TTs são frequentemente associados a notícias de última hora ou eventos que têm um impacto no mundo real. Desse modo, os tópicos se tornaram uma ferramenta de fácil acesso e com grande eficiência em descobrir o que está acontecendo e o que as pessoas pensam sobre os assuntos do momento [32, 33].

Nesse contexto, a identificação desses tópicos em redes de *microblogging* se tornou uma área de interesse crescente entre pesquisadores e indústrias. Além disso, a capacidade de detectar e prever *trending topics* em tempo real pode ter um valor científico, social e comercial significativo [34, 35]. Por exemplo, o conhecimento antecipado de desastres na-

turais ou provocados pelo homem pode proporcionar às equipes de resgate mais eficiência em suas ações [36].

Esse processo de identificação e previsão de tendências nas plataformas de *microblogging* normalmente envolve três etapas [32]: coleta do máximo de dados relevantes da plataforma, identificação de padrões e tendências dentro dos dados e uso dessas informações para prever a população futura de certos tópicos. A eficácia dessas etapas depende muito da quantidade e qualidade dos dados coletados na primeira etapa. Se a fonte de dados em determinados momentos for considerada tendenciosa, um conjunto de dados maior pode ser necessário para eliminar o viés e fornecer resultados precisos de detecção de tendências em todos os períodos de tempo [32].

Essa funcionalidade(*feature*) de *trending topics* oferece uma maneira potencialmente útil de ajudar usuários a obterem convenientemente uma impressão rápida e concisa dos temas mais comentados naquele momento.

Quando se fala em *trending topics*, o Twitter é a rede social mais famosa nesse aspecto, pois com seus aproximados 368,4 milhões de usuários ativos em 2023 [3], o usuário consegue ter informações sobre os assuntos mais falados no mundo em um determinado momento. As palavras-chave ou *tags* mencionadas pelos usuários são compiladas em tempo real, formando uma lista dos tópicos, levando em consideração o número de publicações contendo esses tópicos, a quantidade de compartilhamentos, o fator "novidade", ou seja, se o assunto começou a ser comentado recentemente, entre outros critérios que não são divulgados pela rede social[4]. Além disso, o Twitter possui duas páginas de tendências, uma personalizada de acordo com os interesses do usuário e outra com as tendências a partir da geolocalização, a qual pode ser alterada.

Além do Twitter, o Google Trends[5], uma ferramenta desenvolvida pelo Google Notícias, permite acompanhar a evolução do número de buscas por uma determinada palavra-chave ou termo de busca naquele momento ou em um passado recente. Os dados são apresentados de diferentes formas na ferramenta, como listas, gráficos de linha e gráficos em forma do mapa de uma região. Na página inicial é mostrada uma lista das pesquisas que estão em alta, além de apresentar os gráficos de notícias e estatísticas recentes. A ferramenta permite que o usuário faça uma busca com qualquer intervalo de tempo a partir do ano 2004, apesar de ter sido criado em 2006.

[3] `https://www.websiterating.com/pt/research/twitter-statistics/#references`
[4] `https://resultadosdigitais.com.br/marketing/trending-topics/`
[5] `https://trends.google.com.br/trends/?geo=BR`

2.5 Análise de *trendings topics* em redes sociais

Comunidades de pesquisa, formadas por membros de diversas áreas acadêmicas, têm investigado o uso potencial dos dados criados e armazenados através de tecnologias ou plataformas de redes sociais, para desenvolver novas percepções ou obter *insights* em diferentes áreas, como por exemplo previsões de preço de ações, prevenção de epidemias, monitoramento precoce de eventos, previsões eleitorais, gestão de crises humanitárias, relações públicas, difusão de informações e opiniões públicas [32, 37].

Nesse contexto, na pesquisa realizada em [38], é explorado o papel dos *trending topics* na formação do engajamento com as notícias nas redes sociais. Os autores conduzem um experimento na plataforma chinesa de *microblogging Weibo*, no qual manipulam a visibilidade dos *trendings topics* e medem os efeitos no engajamento dos usuários com o conteúdo de notícias. O estudo conclui que esses tópicos desempenham um papel significativo no controle, ou no processo de filtragem e divulgação de informações nas mídias sociais. Quando eles ficam mais visíveis, os usuários ficam mais propensos a se envolverem com o conteúdo de notícias, sugerindo que os tópicos em relevância servem como uma ferramenta poderosa para chamar a atenção para certas informações. Os autores também afirmam que os *trending topics* podem servir como uma dica para os usuários encontrarem informações relevantes. Os resultados deste estudo destacam a importância da análise dessa funcionalidade na compreensão da dinâmica do engajamento de notícias nas mídias sociais e o papel das plataformas de OSN em moldar a quais informações os usuários são expostos. Também é enfatizada a necessidade de mais pesquisas sobre os mecanismos dos *trending topics* e seu impacto no comportamento dos usuários, bem como as implicações para as sociedades democráticas, onde a função de *gatekeeper*(moderador) das mídias sociais está se tornando cada vez mais discutida.

Já no estudo apresentado em [5], que tem como objetivo identificar e quantificar os vieses econômicos e culturais presentes nos *trending topics* das redes sociais, os autores usam um conjunto de dados dos *trending topics* no Twitter da Espanha e do México para analisar as diferenças nos tipos de tópicos em cada país. Eles descobriram que os tópicos econômicos eram mais propensos a tendência na Espanha, enquanto os tópicos culturais eram mais propensos à tendência no México. Os autores também analisaram os tipos de usuários que conduziam esses termos em relevância e descobriram que, na Espanha, eles eram direcionados por usuários com níveis mais altos de poder econômico, enquanto no México, eram direcionados por usuários com níveis mais altos de poder cultural. O estudo destaca a importância de entender os tópicos tendenciosos das mídias sociais e o impacto potencial que eles podem ter na formação da opinião pública e na tomada de decisões. Os autores sugerem que as plataformas de OSN devem trabalhar para minimizar esses vieses, a fim de promover *trending topics* mais diversos e representativos.

Essas características dos tópicos em tendência nas redes sociais podem ser utilizadas não apenas para educação, mas também para negócios, *marketing*, relações públicas, entre outros. Além disso, as pesquisas apresentadas indicam que é importante acompanhar os tópicos em tendência nas OSN, pois eles podem ser uma valiosa fonte de informação para análise de dados em diferentes campos.

2.6 A problemática das Redes Sociais Centralizadas e os estudos da Descentralização

As redes sociais centralizadas possuem um agente central que detém todos os dados registrados no servidor, ou seja, toda a informação passa pelo nó central para, então, poder ser distribuída para os demais nós, como ilustra a Figura 2.4. Esse poder dado ao operador do sistema gera efeitos colaterais indesejados, porém a importância das OSNs para a comunicação interpessoal diária da população, coloca esses fornecedores em uma posição de *gatekeeper* para partes da vida social de seus usuários. Devido a esta dependência, os usuários tendem a aceitar esses efeitos colaterais mencionados e até condições desvantajosas de uso, uma vez que os fornecedores podem excluir perfis das OSNs acabando com esse meio de comunicação para esses usuários [39].

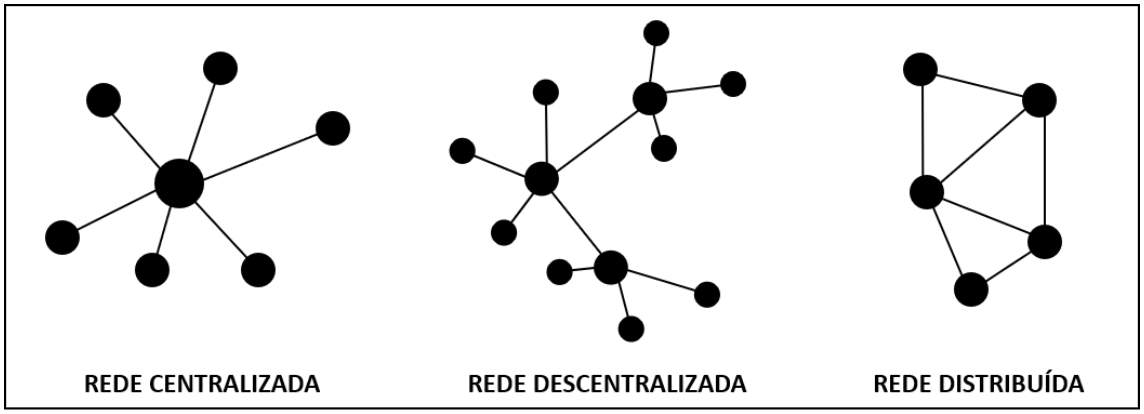

Figura 2.4: Topologia de redes de comunicação (Figura adaptada de [40]).

A partir desses pontos, são evidenciados alguns dos problemas enfrentados pelos usuários de redes sociais centralizadas, como apresentados pela pesquisa [41], onde a autora fala dos diferentes tipos de censura que ocorrem nas plataformas atuais: a censura específica de conteúdo com respeito a diferentes regras e tradições em diferentes países, bem como a censura específica de pessoas, o que significa proibir subconjuntos da população de acessar a rede. Além disso, é comentado que algumas dessas censuras são feitas pelo mecanismo de *shadowban*, ou seja, o conteúdo não é distribuído para a rede, se tornando

invisível aos outros usuários, sem que seja de fato excluído. Essas censuras, em sua maioria, estão relacionadas com os mecanismos de controle de conteúdo e manipulação de discursos [42].

A manipulação midiática é abordada pelos autores de [43], onde eles citam dois casos relacionados à política. O primeiro, nas eleições dos Estados Unidos de 2016, onde algumas empresas do Texas formataram informações com o intuito de aumentar a visibilidade de conteúdos e publicidades benéficas ao ex-presidente Donald Trump, na rede Facebook. Já o segundo caso ocorreu na França, na eleição de 2017, envolvendo o Google e também o Facebook, os quais intervieram nos conteúdos supostamente para prevenir notícias falsas(*fake news*) em plano eleitoral.

Empresas detentoras das plataformas podem rentabilizar os dados dos usuários compartilhando os interesses destes com outras empresas [44]. A existência de um agente central que detém todos os dados registrados no servidor, combinado com incentivos à monetização, pode causar sérios problemas à privacidade dos usuários e da sociedade de um modo geral [45].

Em contrapartida, tem-se as estruturas das redes sociais descentralizadas, baseadas na arquitetura distribuída, na qual possui um conjunto de redes trocando informações entre si (Figura 2.4), como exemplo da Friendica e Diáspora que foram propostas como alternativa às principais OSN centralizadas dominantes nos dias de hoje [46]. De acordo com o trabalho [47], os autores sugerem que as redes sociais descentralizadas vieram para devolver aos usuários o controle sobre os seus dados no que se refere ao respeito à privacidade, propriedade e divulgação de informações.

Em [39] também podemos encontrar uma definição dos componentes de uma RSD:

Usuário: Pessoa ou organização que possui um identificador e um perfil de usuário;

Perfil de usuário: Representação digital de um usuário, contendo todos os itens de dados de propriedade do usuário;

ID do usuário: Identificador único para cada usuário, sendo parte do sistema;

Conteúdo: Item de dados que é armazenado ou compartilhado dentro da OSN;

Conexão: Afiliação ou familiaridade declarada entre usuários (por exemplo, solicitação de amizade ou seguir alguém(*follow*);

Nó: Dispositivo de rede utilizado pelos usuários para se conectar à rede social;

Servidor: Dispositivo de rede que suporta de forma confiável a prestação de serviço.

Pode-se observar que as RSD apresentam vários benefícios aos usuários, como exemplo os perfis autogerenciados independentes, controles de dados e sistemas de gerenciamento de conteúdo, interoperabilidade e flexibilidade, ferramentas de recompensa, e também a arquitetura de hospedar vários servidores em uma única plataforma [8, 12].

2.7 Redes Sociais Descentralizadas e a Educação

Uma das áreas de pesquisa no campo das redes sociais descentralizadas tem sido o uso delas como ferramenta para negócios e educação, apesar de não possuir um largo número de estudos. Por exemplo, a autora de [9] examinou o potencial das redes sociais descentralizadas para promover o engajamento e a colaboração dos estudantes em um ambiente universitário. O estudo descobriu que redes sociais descentralizadas podem proporcionar aos estudantes um senso de comunidade e pertencimento, o que pode levar a um maior engajamento e participação nas discussões em sala de aula. Além de ter conseguido demonstrar que a seriedade do projeto e a contínua atualização dos recursos disponíveis, favorece a popularidade e envolvimento da comunidade do Friendica.

Outro estudo, apresentado em [12], mais focado na utilização da RSD, teve embasamento teórico em aprendizagem informal, não-formal e formal. O autor investigou o uso de recompensas(*badges*) digitais como uma ferramenta para motivar os estudantes a se engajarem em redes sociais descentralizadas na "tentativa de suprir a necessidade de integração e fomentação das práticas sociais de aprendizagem por meio de uma técnica de gamificação". O estudo concluiu que o uso desses *badges* pode ser uma forma eficaz de incentivar a participação, além de afirmar que estruturas de aprendizagem formais podem não ser tão eficientes para os alunos quanto às estratégias de cunho social.

Os autores de [10], tiveram como objetivo a implantação de uma rede social no ambiente acadêmico. Durante o estudo, verificaram que as RSDs eram uma alternativa interessante e com relevância literária, devido a trabalhos anteriores. Eles também examinaram a percepção inicial dos usuários em relação à adoção da rede social descentralizada. O estudo descobriu que os estudantes geralmente têm percepções positivas das RSDs e que elas podem ser uma ferramenta eficaz para promover o engajamento e a colaboração dos estudantes.

2.8 BraSNAM

Nesta seção, é apresentada a busca de estudos sobre o uso de trending topics voltado ao auxílio da educação. Para obter artigos e trabalhos que estudem essa utilização, foi sele-

cionado o Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)[6], um congresso organizado pela Sociedade Brasileira de Computação que acontece anualmente desde 2012, onde são discutidos e abordados tópicos relacionados a análise de redes sociais. Também foi limitado o espaço de tempo considerado relevante, determinado como 2020-2022.

Para realizar a pesquisa, foram definidas as seguintes combinações de busca de palavras-chave: "trending topics" e "educação", "trending topics" e "ensino", "trending topics" e "escola". A partir desses termos de pesquisa, foi montada a seguinte tabela com a quantidade de artigos relevantes encontrados.

Tabela 2.1: Quantidade de trabalhos relacionados encontrados na BraSNAM

	2020	**2021**	**2022**
BraSNAM	0	0	0

Para exemplificar a lacuna encontrada em trabalhos relacionados ao tema de pesquisa, foi escolhido um trabalho publicado no BraSNAM, um workshop se tornou um importante evento que reúne pesquisadores para discutir métodos de análise, tendências e fenômenos que ocorrem nas redes sociais. Desde o começo do congresso é possível visualizar um aumento no número de membros participantes, conforme análise realizada no trabalho [48], que vem ganhando relevância a cada ano também devido ao crescimento do uso de redes sociais pela sociedade.

No estudo realizado no artigo [48], foi feita uma análise dos 10 anos do BraSNAM, onde é possível visualizar os temas mais discutidos baseados nas palavras-chave do artigo. Dentre os 230 trabalhos publicados foram identificadas 469 palavras-chave, com 804 conexões entre elas e 45 componentes conectados. Para melhorar a legibilidade, o autor aplicou um filtro baseado no grau do nó, onde são apresentados apenas os nós com grau maior que 9, e apenas os componentes conectados. A partir da Figura 2.5 é possível observar os temas mais discutidos e a relação entre eles.

[6] `https://csbc.sbc.org.br/2022/brasnam/`

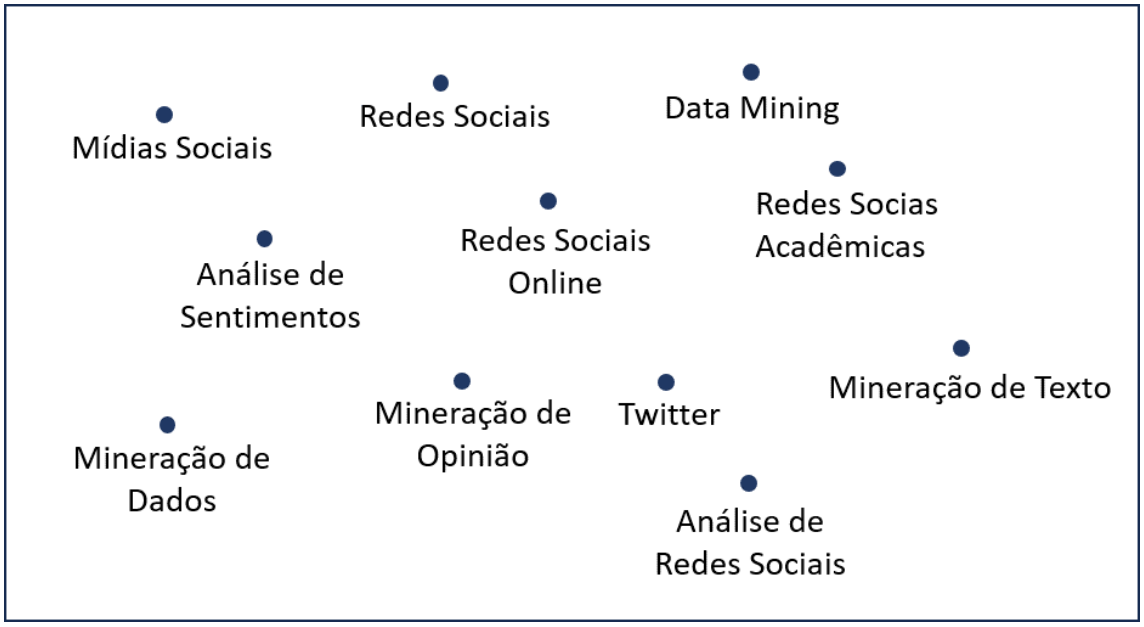

Figura 2.5: Relação de temas citados nos artigos da BraSNAM (Figura adaptada de [48]).

Na análise das publicações dos eventos da BraSNAM[7], e na Figura 2.5, é possível observar o grande número de trabalhos que abordam temas que estão diretamente conectados com este trabalho, como: redes sociais, midias sociais, redes sociais online, data mining, análise de redes sociais, mineração de texto, entre outros. Apesar de terem sidos encontrados temas relacionados, a partir das buscas realizadas e da análise desse artigo, foi observado nesta pesquisa uma lacuna em relação a trabalhos que relacionem o uso de trending topics voltado à educação.

[7] `https://sol.sbc.org.br/index.php/brasnam/index`

Capítulo 3

Fundamentação para a proposta

Neste capítulo, são expostos os tópicos fundamentais para o entendimento do estudo realizado neste trabalho, abordando a revisão de algumas partes da estrutura do Friendica, o estudo do processamento de linguagem natural, e por fim uma análise da identificação automática de *trending topics*.

3.1 Friendica: perspectivas usuário e desenvolvedor

Nesta seção abordamos alguns pontos específicos do Friendica, iniciando com a visão geral da rede social.

3.1.1 Visão geral

O Friendica é uma alternativa de rede social com arquitetura descentralizada sem autoridade ou propriedade central e de código totalmente aberto[1]. A plataforma realiza a integração da comunicação social conectando-se a outros projetos independentes e serviços cooperativos, podendo ser instâncias públicas ou privadas. O usuário possui a autonomia de configurar sua rede de acordo com suas preferências, como a funcionalidade de mostrar diferentes perfis para diferentes pessoas/comunidades, já que a rede permite que seja criada mais de uma conta por e-mail, e possui diversos tipo de conta (pessoal, fórum, notícia, organizações). Além disso, o software é focado em recursos de privacidade e segurança, com várias opções detalhadas de controle de relacionamento, tipos de interações, entre outras configurações, e também visa uma rede aberta e livre das corporações de coletas e vendas de dados pessoais.

[1] `https://friendi.ca/`

O código-fonte e *add-ons* do Friendica ficam hospedados no GitHub[2], eles são mantidos e atualizados por desenvolvedores da comunidade. A plataforma também conta com fóruns específicos para aqueles que desejam configurar suas instâncias, criar novos *add-ons* ou resolver problemas de implantação, além de possuir fóruns de suporte e problemas gerais da plataforma.

O Friendica não possui um servidor central, mas centenas de servidores públicos espalhados pelo mundo[3], como a Figura 3.1, que mostra uma página da instância venera.social, um servidor localizado na Finlândia mantido por uma comunidade local [4]. Além disso, diferenciando-se das redes sociais famosas, que não interagem com outros softwares de OSNs, um usuário do Friendica consegue se interligar com qualquer pessoa do Twitter ou Facebook, além de conseguir postar e receber conteúdos do Tumblr e Wordpress, entre outros sistemas.

A rede Friendica possui várias das funcionalidades presentes nas principais redes sociais, como a marcação de perfis através do "@", comentários em postagens, menção de *tags* com "#", mensagens privadas, entre outros. Um perfil pode ser facilmente excluído ou migrado para outro servidor.

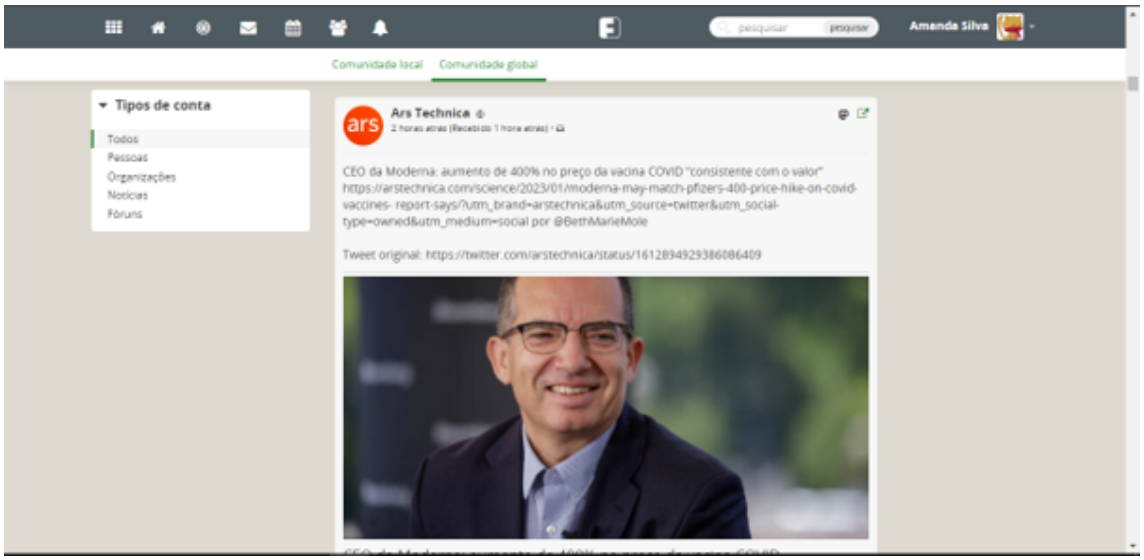

Figura 3.1: Nó venera.social (Imagem de tela do perfil da usuária Amanda.)

3.1.2 Contas do tipo fórum

Os fóruns são uma das alternativas de tipo conta que a rede social oferece, podendo ser utilizados para várias funcionalidades como fóruns de discussão, contas de celebridades,

[2] `https://github.com/friendica/friendica`
[3] `https://dir.friendica.social/servers`
[4] `https://venera.social/`

canais de anúncios, canais de notícias ou páginas de organizações, dependendo de como o usuário deseja interagir com outras pessoas [5] . Para criar esse tipo de conta, o usuário deve registrar uma conta adicional a partir de uma conta normal pré-existente, e ela terá que possuir um *nickname* (nome de usuário) exclusivo, igual todas as contas criadas no Friendica. O usuário que realizar a criação do fórum será o administrador da conta, porém podem ser adicionados novos administradores posteriormente.

As contas tipo fórum podem ser públicas, permitindo que qualquer pessoa se torne amigo/seguidor do fórum sem a necessidade de aprovação, ou privados, onde é necessária a solicitação de vinculação prévia. Esse tipo de perfil conta com todas as funcionalidades de um normal, porém possui uma configuração exclusiva de recompartilhamento(*reshare*), onde todas as publicações que contenham a marcação do *nickname* da conta, através dos símbolos "@" ou "!", feitas por usuários do fórum, são automaticamente repostadas por ele.

Essa funcionalidade de repostar *posts* que contenham a marcação do fórum, pode ser relevante dentro do contexto acadêmico, uma vez que as contas desse tipo podem ser utilizadas para criação de comunidades. Desse modo, mesmo que dois perfis não tenham amizade, mas sigam o fórum, eles verão todos os *posts* dessa *tag*.

Além dessas funcionalidades, a partir do *add-on* desenvolvido no trabalho [12] os fóruns são contas que ganharam a opção de entrega de *badges*, ou seja, o usuário ganha uma conquista conforme sua atuação dentro da plataforma, segundo as configurações presentes em todos esses tipos de conta. De acordo com [6], "um *badge* digital é um registro *online* de uma dessas conquistas, monitorado por uma comunidade em que o beneficiário tenha interagido e obtido o emblema, bem como o trabalho feito para obtê-lo." Os *badges* são controlados pelos administradores do fórum, sendo deles a responsabilidade de realizar a confirmação das entregas para os usuários.

3.1.3 *Tags* em perfil

As *tags* (ou etiquetas) no Friendica são caracterizadas pelo caractere *hash*(#) acompanhado da palavra ou termo que se tornará uma *tag* nesse contexto. Ela cria um *link* entre os posts que a utilizam[7], possibilitando uma busca generalizada para o termo em questão. Por exemplo, o termo '#captcha' fornecerá um *link* de busca para todos os posts que possuem '#captcha' em seu conteúdo.

Elas geralmente têm um mínimo de três caracteres de comprimento[7]. Termos de busca mais curtos podem não render resultado de busca, embora isso dependa da configuração

[5] `https://wiki.friendi.ca/docs/forums`
[6] `https://support.mozilla.org/pt-BR/kb/o-que-e-um-badge`
[7] `https://wiki.friendi.ca/docs/tags-and-mentions`

do banco de dados. A exceção presente na formação das etiquetas por letras é o uso do caractere hífen(-) para separar os caracteres, portanto não é possível criar uma cujo alvo contenha um sublinhado, por isso elas são sempre formadas por palavras únicas e sem espaços. As "*topical tags*"também não são ligadas se forem formadas por números, por exemplo, "#1". Se o usuário desejar usar uma *hashtag* numérica, é preciso adicionar algum texto descritivo, como "#2022WorldCup".

No Friendica, há a possibilidade de adicionar *tags*, ou palavras-chave, ao perfil dos usuários. Nas configurações é possível visualizar duas entradas, a de etiquetas públicas e a de privadas, onde o usuário consegue escrever várias palavras relevantes sobre suas preferências. Na primeira, esses termos serão utilizados para que o sistema realize a indicação de amigos em potenciais, ou melhor, perfis que possuem a mesma linha de interesses com base nas palavras-chave. Na segunda entrada, podem ser colocadas outras *tags*, que nunca são mostradas para outros usuários, mas serão utilizadas pelo sistema para a busca de perfis semelhantes.

Ainda existe um diretório de perfis organizado por essas *tags* públicas, onde é possível pesquisar por termos e visualizá-las, além de possuir a visualização de países e linguagens mais populares, mostrado na Figura 3.2, encontrada no diretório do Friendica[8]. Também é possível filtrar os perfis por: todos, pessoas, organizações, fóruns e novos, e por essa funcionalidade o usuário é capaz de encontrar perfis de acordo com suas predileções.

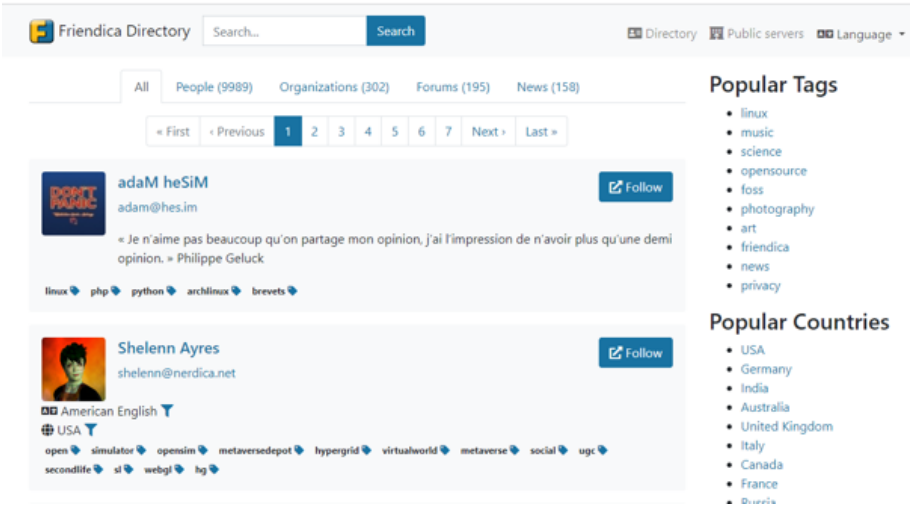

Figura 3.2: Imagem de tela da página do Diretório do Friendica.

Essa funcionalidade pode ser relevante para a criação de filtros e comunidades que buscam conteúdos específicos. Os fóruns podem conter as *tags* referentes a alguns tópicos e a partir dessa relação serem indexadas à página de busca pelos *trending topics*.

[8] `https://dir.friendica.social/`

3.1.4 *Add-ons*

O Friendica possui algumas extensões padrões do sistema, mostrados na Figura 3.3, como por exemplo os conectores de postagem cruzadas, que permitem que os usuários se conectem com outros sistemas pré-estabelecidos e realizem postagens através do Friendica nessa outra plataforma, e possui também conectores bidirecionais, possibilitando posts, retransmissão e leitura da outra rede. Essas extensões, quando habilitadas, podem ser formatadas pelo usuário nas configurações do perfil, marcando um *checkbox* nas funcionalidades adicionais desejadas. Os Add-ons são aplicações inteiramente integradas ao sistema com o intuito de expandir as funcionalidades e de fácil habilitação/desabilitação pelos gestores do sistema [9].

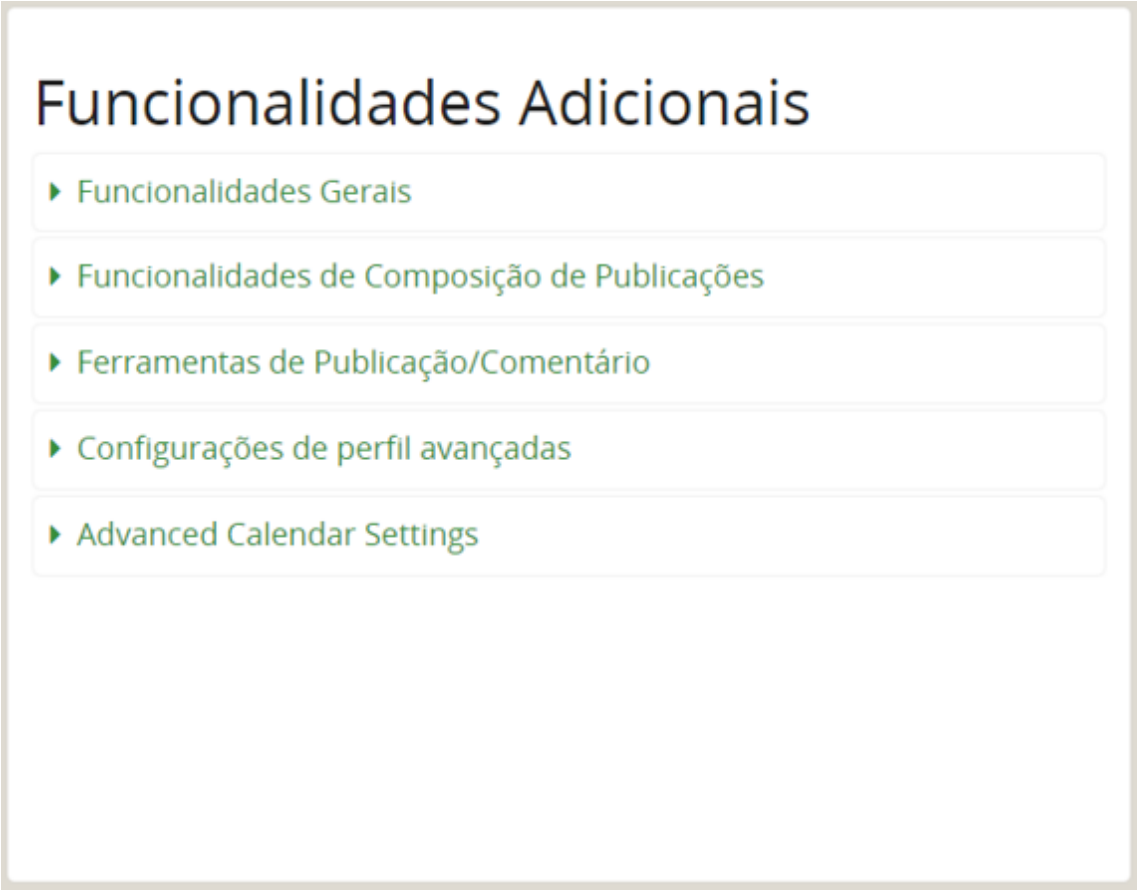

Figura 3.3: Imagem de tela perfil do perfil da usuária Amanda mostrando os Add-ons padrões.

Na página de documentação oficial do Friendica[8] são mostradas algumas regras para a implementação de complementos, os quais precisam ser desenvolvidos em PHP e JavaScript, conforme o código fonte do sistema. Os *add-ons* disponíveis também estão

[9] `https://wiki.friendi.ca/docs/addons?s[]=addon`

armazenados no GitHub, no repositório oficial do Friendica, para que todos os usuários tenham acesso às extensões de funcionalidades oferecidas pela comunidade. Para que as pessoas tenham acesso a essas extensões, elas precisam ser adicionadas a uma instância e habilitadas por um administrador.Os add-ons normalmente são desenvolvidos pelos gerenciadores do sistema, porém na rede social Friendica, qualquer usuário pode criar um add-on, o qual pode ser habilitado por um administrador .

3.1.5 *Trending Tags*

Durante a pesquisa realizada neste trabalho, foi possível identificar que o Friendica apresenta uma funcionalidade de "trending tags" em sua estrutura. Na instância do CiCFriend essa *feature* ainda não é utilizada como padrão pela comunidade. É possível verificar a habilitação dela na aba de configuração conforme a Figura 3.4.

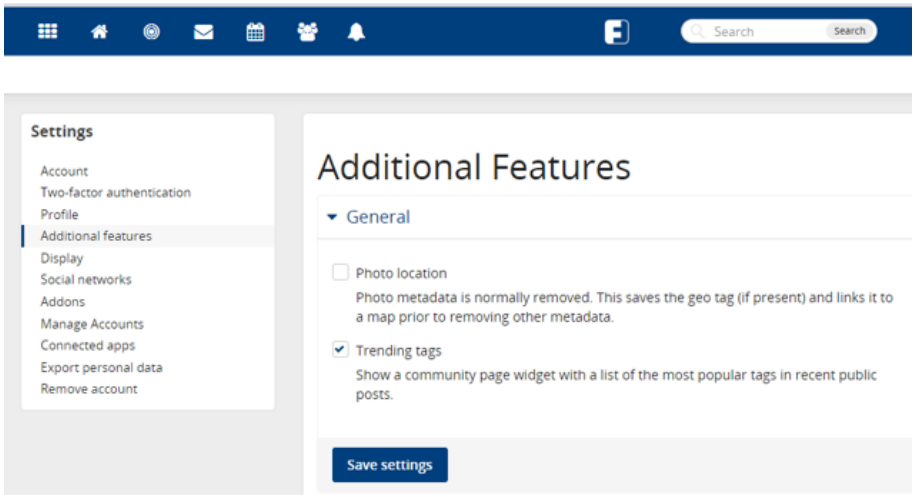

Figura 3.4: Imagem de tela das configurações para habilitar o *trending tag* do perfil da usuária Amanda.

O funcionamento dessa *feature* depende da utilização do uso de tags nos posts, evidenciadas pelo "" antes de palavras, ou seja, para uma publicação ser vista nos assuntos populares, é necessário o uso da "". Essa dependência do uso de tags provoca uma grande limitação na geração dos tópicos de tendência, já que posts que não contenham uma tag específica, mesmo comentando do mesmo termo, não serão contabilizados para a lista. Além disso, conforme citado anteriormente, as tags possuem algumas regras de escrita, como por exemplo, em caso do uso de mais de uma palavra na tag, não é permitido incluir espaço entre elas. Essas limitações diferenciam os *trending tags* dos *trending topics*, já que no segundo é realizada a mineração de todo o texto de uma postagem, utilizando Processamento de Linguagem Natural para identificar os tópicos de tendência, os quais podem ser palavras únicas, frases curtas, termos específicos, tags, entre outros.

A lista das *trending tags* geradas é apresentada na aba esquerda da página da "Comunidade Local", conforme mostrado na Figura 3.5. Essa lista mostra as *tags* utilizadas nas últimas 24 horas. Também é possível selecionar algum item presente na lista e visualizar os posts que utilizaram essa tag específica.

As informações do funcionamento dessa *feature* foram obtidas através de testes realizados na plataforma, pois não foram encontrados documentos das especificações no diretório oficial do Friendica. A partir desses testes, foi observado que a atualização da lista não é realizada instantaneamente, e também não foi possível identificar um limite de itens na lista.

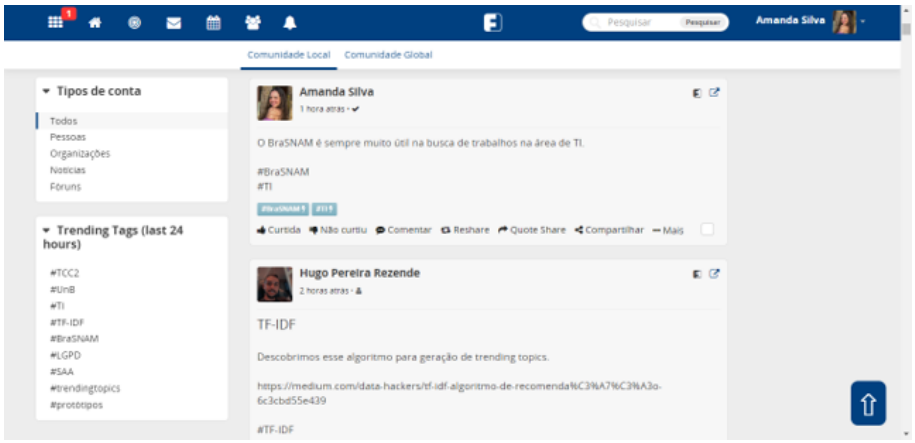

Figura 3.5: Imagem de tela do *trending tag* do perfil da usuária Amanda.

3.1.6 *Tag Cloud*

Uma outra funcionalidade que o Friendica também possui, é a de geração de *Tag Cloud*. Essa *feature* oferece a visualização das *tags* utilizadas por um usuário, em forma de nuvem, e fica localizada na página do perfil pessoal. Quando habilitada (Figura 3.6), essa função permite que seja visualizada a *Tag Cloud* de outros usuários, que sejam amigos ou possuam perfis públicos. Essa nuvem é disposta com *tags* de diferentes cores e tamanhos, organizadas de acordo com a quantidade de vezes que o usuário utilizou as *tags*, conforme mostrado na Figura 3.7.

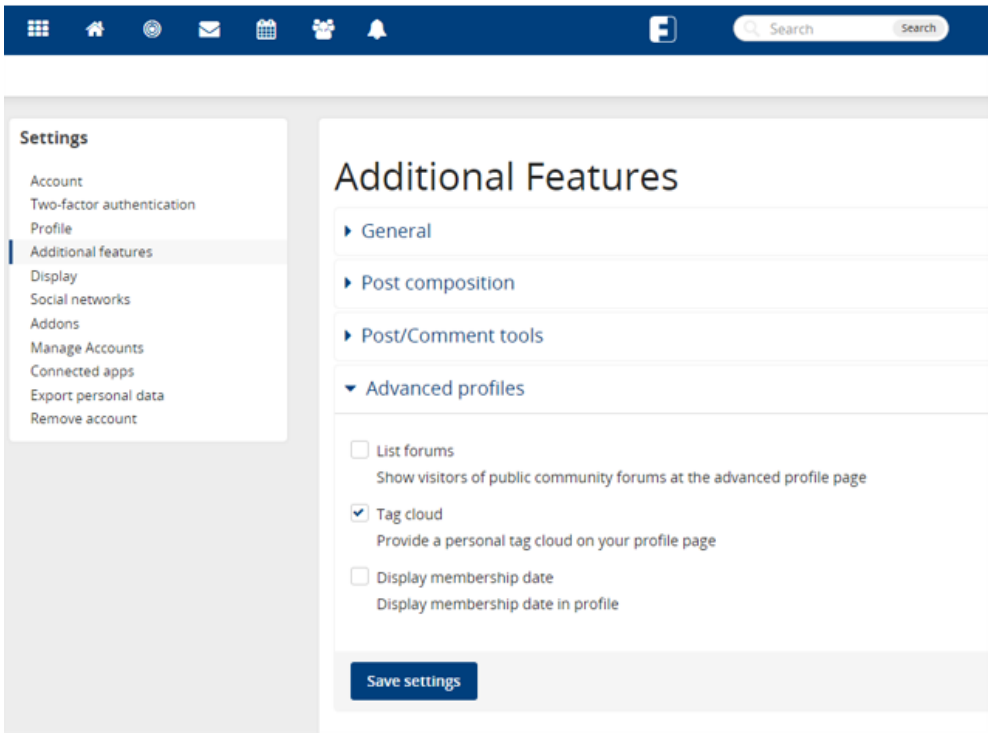

Figura 3.6: Imagem de tela das configurações para habilitar a *Tag Cloud.*

Essa *feature* pode ser utilizada para apresentar um resumo, baseado na utilização de *tags*, dos temas mais falados por um usuário, podendo também demonstrar interesses, dúvidas ou domínio de certos termos. As informações do funcionamento dessa *feature* foram obtidas através de testes realizados na plataforma, pois não foram encontrados documentos das especificações no diretório oficial do Friendica.

Figura 3.7: Imagem de tela da *Tag Cloud* do perfil da usuária Amanda.

3.2 Identificação automática de *trending topics* em redes sociais

Nesta seção elucidamos possíveis soluções para a funcionalidade de identificação automática de *trending topics*. Introduziremos o conceito de "Processamento de Linguagem Natural" e motivar seu uso, trabalhar a fundamentação teórica por trás de opções de algoritmos que podem servir para a implementação de *trending topics* em redes sociais, e por fim selecionar o algoritmo adequado para o nosso caso. As informações utilizadas para embasar os conceitos desta seção foram extraídas dos capítulos 1, 2 e 3 do livro *Practical Natural Language Processing* [49].

3.2.1 Processamento de Linguagem Natural

Um desafio recorrente para a computação é o de compreender a linguagem humana, permitindo analisar, extrair, e até gerar novas informações baseando-se em conteúdo gerado por seres humanos, visto que é recorrente no dia-a-dia ferramentas como assistentes de voz, análises de textos, filtros de spam, entre outros [49].

Processamento de Linguagem Natural ou PLN, (do inglês Natural Language Processing ou NLP) é uma vertente de Inteligência Artificial (IA) focada na interação entre linguagem humana e computadores. Envolve o uso de técnicas computacionais para analisar, entender e gerar textos que mimetizam a comunicação humana, seja na forma de texto ou voz. É uma abordagem utilizada para desenvolver uma gama de soluções em vários segmentos, tais como tradução, sumarização, classificação e representação de textos, análise de sentimento, detecção de spam, entre outras [49].

No contexto de uma RSD, o Processamento de Linguagem Natural apresenta-se como uma possível ferramenta, pois permite analisar os dados textuais gerados pelos usuários e extrair informações e padrões do conteúdo [49]. Essa extração de informações pode ser utilizada para identificar tópicos populares e assuntos em alta na RSD, aumentando o engajamento, visibilidade e alcance de publicações que têm relevância para a rede, além de fomentar a interação dos usuários.

Existem diversos algoritmos de PLN que podem ser aplicados para implementar a funcionalidade de *trending topics* [49], desde algoritmos mais complexos e usados em grandes corporações como o *Latent Semantic Analysis* (LSA) e *Word2Vec*, ou algoritmos mais simples baseados em modelos como o *Bag of Words* ou medidas estatísticas como o TF-IDF (*Term Frequency-Inverse Document Frequency*) [49]. Tais algoritmos são utilizados em vários casos do mundo real como detecção de spam, extração de textos, classificação de *tickets* de suporte, etc.

Voltando-se para o tópico de *trending topics*, houve uma quantidade significativa de pesquisas sobre o uso de algoritmos de PLN para análise de dados gerados por usuários em redes sociais (por exemplo [50, 51] e sobre o uso de técnicas de visualização de dados para apresentar tendências. Estes estudos fornecem *insights* e técnicas valiosas que podem ser usadas na proposta do *add-on* para geração de tópicos em tendências no CICFriend.

Em suma, o uso de PLN se apresenta como uma solução viável para a geração de *trending topics* da plataforma CICFriend por sua capacidade de interpretar textos convencionais e gerar conteúdo em linguagem natural, que se faz essencial levando em conta que a RSD é utilizada por pessoas e seu conteúdo deve favorecer a experiência do usuário.

3.2.2 Aplicando Processamento de Linguagem Natural para a implementação de *trending topics*

Em meio ao grande volume de usuários recorrentes e o alto volume de publicações a todo momento, além do desafio de compreender e processar a linguagem humana, temos que as técnicas de PLN se apresentam como uma possível solução para a identificação de *trending topics* em tempo real na plataforma CICFriend, dada a capacidade de lidar com inferência de contexto, definir relevância das palavras e outras funcionalidades [52], que será mostrada adiante. Além de que, no contexto de *trending topics* pode-se encontrar textos ambíguos, informais, com erros de digitação e até dependentes de contexto, como algum evento importante ou uma tendência da época, e para lidar com isso são necessárias técnicas de pré-processamento e normalização do texto, também presentes no escopo da PLN.

Levando em conta todo o processo de manipular um texto, desde a morfologia até a identificação de contexto, temos uma sequência de passos recorrentes nas soluções de PLN para extrair informações de um texto e poder executar a tarefa desejada, seja fazer análise de sentimento de um texto ou gerar recomendações baseadas em informações prévias. É notável que alguns passos do processo envolvem uma noção subjetiva e inerente ao ser humano, como identificar o contexto das publicações, por exemplo. Logo, existem limitações naturais e intangíveis às máquinas, e esse fator deve ser levado em conta na seleção do algoritmo adequado para a identificação de *trending topics*.

As etapas do processo de identificação de *trending topics* serão descritas a seguir, desde a etapa de pré-processamento do texto até opções de algoritmos.

Pré-processamento do Texto

Antes de lidar com o texto para executar a tarefa desejada com PLN, é necessário efetuar uma normalização do texto, com o objetivo de formatá-lo em um formato mais compre-

ensível para a máquina. Esse é um passo necessário pois os dados vêm em linguagem informal, com erros de digitação, sem estrutura definida e com informações que podem não ser úteis para a tarefa, que acabam por poluir o texto e afetar o desempenho.

A seguir será apresentada uma sequência de passos de pré-processamento e normalização do texto chamada *pipeline*, e ignorar essa etapa pode interferir na *performance* do algoritmo e na qualidade dos resultados gerados. É uma prática comum entre cientistas de dados despender boa parte do tempo para limpar e normalizar os dados, antes de executar a operação desejada.

A Figura 3.8 mostra as etapas recorrentes em um processo de *pipeline*.

Figura 3.8: Representação dos passos de pré-processamento.

Conjunto de Textos

A primeira etapa de pré-processamento consiste em reunir o conjunto de textos utilizados na tarefa desejada. Para essa etapa é necessário extrair os textos de alguma fonte, como arquivos de texto, HTML, PDF, entre outros. No caso de textos extraídos da *web*, a técnica mais comum é a de *web scraping* (raspagem de dados), que é uma forma de mineração de texto aplicada especificamente para páginas HTML, removendo as *tags* de formatação HTML para extrair o conteúdo da página. Isso pode ser feito usando bibliotecas em Python, como o Beautiful Soup[10] ou Scrapy [11].

Detecção de Linguagem e Tradução

O processo de detecção de linguagem e tradução se trata de uma etapa opcional de pré-processamento de textos, mas apesar disso é conveniente em muitos casos. A detecção de linguagem consiste em identificar o idioma em que o texto foi escrito, e isso se faz essencial em algumas aplicações de PLN pois pode acontecer do corpo de texto conter palavras de diferentes idiomas, como por exemplo em uma rede social, onde os usuários podem postar em vários idiomas.

Logo, identificar o idioma pode ser um passo necessário para preparar o texto e evitar inconsistências no texto, pois muitos algoritmos e modelos são específicos para um idioma

[10] `https://www.crummy.com/software/BeautifulSoup/`
[11] `https://scrapy.org/`

e dessa forma, um modelo treinado em um idioma específico não seria adequado para determinada tarefa de PLN em outro idioma.

Remoção de *stop words*

Um texto não é composto apenas por suas palavras-chave e termos principais, os termos responsáveis por definir o significado e ideia principal do texto. Naturalmente, a interpretação do texto foca nos termos-chave e ignora o resto das palavras como artigos, preposições, verbos, etc. Tais termos carregam apenas função estrutural no texto em sua linguagem natural para compreensão humana, mas não apresentam significado e não contribuem para dar sentido ao texto. Esses termos são chamados de *stop words* [49], e para um melhor processamento do texto pelo algoritmo de PLN, é necessária uma etapa de pré-processamento para removê-los. A Figura 3.9 mostra o antes e depois do processo de remoção de *stop words* aplicado a um texto. Para o processo de remoção, foi utilizada uma ferramenta disponível na internet [12].

A primeira etapa de pré-processamento consiste em iterar o texto para remover as *stop words*, utilizando uma lista pré-definida de palavras e uma estrutura de repetição (laços *for* ou *while*) no algoritmo e buscando cada palavra do texto na lista de *stop words*. No geral, essa lista é composta de artigos, preposições e até verbos que podem ser removidos do texto sem perda semântica, porém vale ressaltar que alguns termos podem aparecer em contextos específicos e por conta disso, devem ser mantidos no texto para preservar o contexto. Por exemplo, no termo "tênis da Jordan", a preposição "da" relaciona os termos "tênis" e "Jordan", atribuindo o nome ao tênis no contexto de marca e produto. Removendo a preposição, os termos perdem relação e o contexto é afetado, parecendo se tratar de dois termos diferentes e sem relação, quando de fato um complementa o significado do outro. Existem implementações prontas para cada idioma que são capazes de interpretar e lidar com casos de uso, como a bilbioteca NLTK [13].

[12] `https://tools.fromdev.com/remove-stopwords-online.html`
[13] `https://www.nltk.org/`

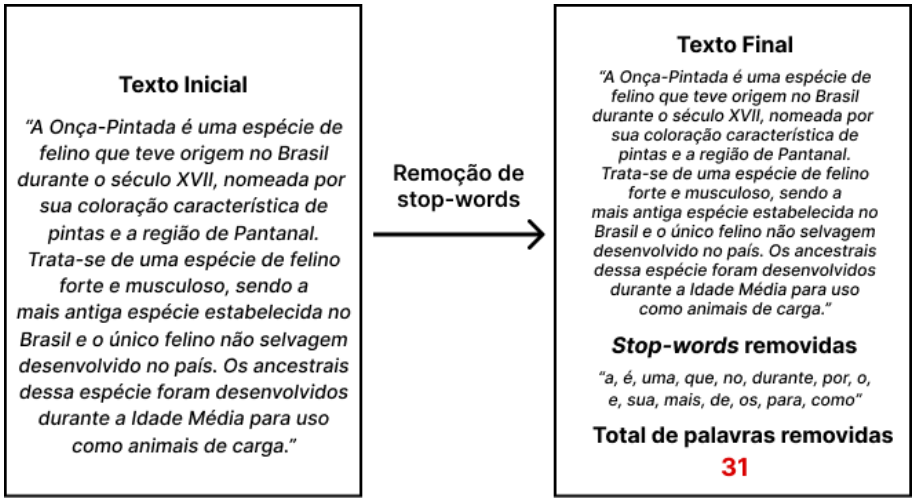

Figura 3.9: Comparação do antes e depois de um texto após a remoção de stop words

Separação de frases e *tokenização* de palavras

Os algoritmos de PLN não processam o texto como um longo bloco, mas sim palavra por palavra, e essa etapa consiste em separar os textos em frases (segmentação de sentenças) e subsequentemente separar frases em palavras (*tokenização*). Dessa forma, o texto fica mais fácil de ser processado, além de permitir tarefas adicionais como correção do texto, atribuição de funções gramaticais (para atribuir relacionamento entre as palavras), etc [49].

De forma intuitiva, poderíamos dividir os textos em frases ao separar as frases pelos sinais de pontuação, por ponto-final ou ponto de interrogação, por exemplo. Porém, apesar de funcionar bem para perguntas e frases curtas, essa abordagem falha em casos especiais como no uso do ponto-final em abreviações como "Sr." ou "Sra.", e reticências "...".

Existem dois tipos principais de algoritmos para segmentação de frases, algoritmos baseados em regras (*rule-based*) e algoritmos com aprendizado de máquina (*machine learning*) [49]. Os algoritmos *rule-based* usam um conjunto predefinido de regras para separar as frases de acordo com sinais de pontuação, espaços em branco e outras regras para cobrir casos especiais como vocativos e palavras hifenizadas. Esses algoritmos também estão presentes nas árvores sintáticas gramaticais, como mostra a Figura 3.10. Já os algoritmos *machine learning* comumente usam modelos estatísticos ou redes neurais para aprender os padrões do texto e automaticamente delimitar as sentenças [53], ou seja, é fornecido um conjunto de vários textos e frases e é informado qual é o modelo necessário, qual o resultado esperado, então o algoritmo detecta os padrões de frases e aprende a delimitar sentenças de textos posteriores por conta própria.

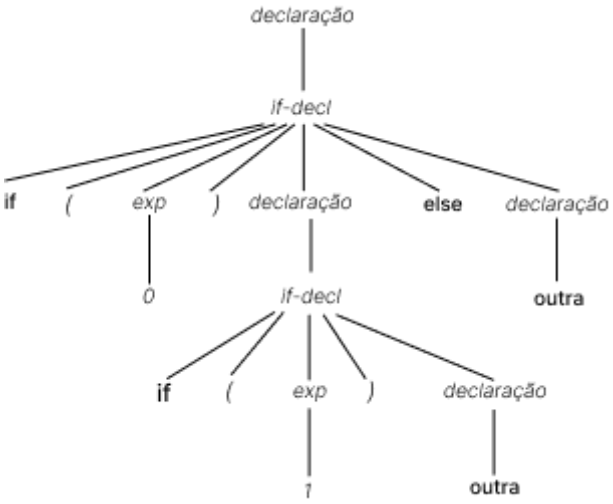

Figura 3.10: Exemplo de árvore sintática, representando o algoritmo *rule-based*

Após a separação do texto em frases, temos a subdivisão das frases em palavras. Assim como na separação de frases, também podemos separar as palavras pelos sinais de pontuação. Porém, também existem casos especiais como em palavras com apóstrofo, hífen, cifrões, e outros caracteres especiais.

Apesar da etapa de segmentação de frases ser trivial na separação de palavras, a etapa de *tokenização* depende bastante do idioma utilizado. Isso se dá por conta de características específicas de cada linguagem, como por exemplo, a presença do "ç" no português, ou a falta de acentos na língua inglesa. Existem soluções como o *spaCy* [14], que permitem definir regras específicas por linguagem, cobrindo os casos de inflexões, prefixos, sufixos e morfologia complexa.

Dependendo da ferramenta utilizada pode ser necessário definir o formato de entrada do arquivo. No caso de uma carta ou e-mail, por exemplo, existem formatos específicos para capturar os dados e é possível usar uma expressão regular para tal. Para o contexto de *trending topics*, existem implementações específicas para lidar com *tweets* e publicações de redes sociais.

Lidando com o contexto

As publicações dos usuários em uma rede social podem envolver eventos aleatórios como uma tendência, assuntos do momento, publicações anteriores, e até a bagagem cultural da sociedade, formando assim o contexto das publicações e até da própria rede social.

Assim como entender e processar a linguagem humana com PLN é um desafio, um outro desafio recorrente é o de relevar o contexto nos algoritmos de PLN. Não há um meio de transferir o conhecimento e contexto dos usuários para as máquinas, e isso se

[14] `https://spacy.io/`

apresenta como um problema pois não há como gerar conteúdo relevante para o usuário sem levar em conta o próprio contexto dos usuários. Manipular o contexto é um tópico delicado em PLN pois, ele influencia no significado das palavras e das frases em um texto. O significado de uma palavra ou frase pode mudar dependendo do contexto em que ela é utilizada, tornando difícil para os algoritmos entenderem de forma precisa o significado do texto [49].

Um dos maiores problemas com contexto semântico em PLN é que geralmente ele se apresenta de forma implícita e/ou subjetiva [49], e não de forma clara e direta, sendo necessária a interpretação do texto para extrair seu verdadeiro significado. Por exemplo, o termo "Minhocão" pode ter dois significados diferentes dependendo da localidade do usuário. Se for residente de Brasília, "Minhocão" remete ao prédio do Instituto de Ciências Central (ICC) da Universidade de Brasília (UnB), porém, em São Paulo o termo faz referência à via expressa Elevado Presidente João Goulart, no centro da cidade. Logo, sem a informação adicional da cidade em questão, o termo se torna ambíguo.

Outro problema com o contexto é que as palavras podem estar relacionadas com assuntos e áreas de conhecimento específicas. Por exemplo, a palavra "operação" pode se referir a uma operação bancária, uma cirurgia, uma expressão matemática, e vários outros significados de acordo com a área de conhecimento. Logo, sem conhecimento do assunto, a informação fica incompleta, afetando o significado e compreensão da mesma.

Existem várias soluções em NLP para lidar com o contexto de palavras e textos. Os algoritmos mais refinados como o *fastText* e *Word2Vec* (criados por *Facebook* e *Google*, respectivamente) utilizam da técnica de "*word-embeddings*" (palavras incorporadas), que consiste em criar um vetor numérico para representar as palavras, de forma que palavras similares têm valores numéricos similares. A vantagem de tal técnica é permitir que o algoritmo capture a relação das palavras, formando assim uma noção do contexto. Por exemplo, ao percorrer um ou vários textos e notando que alguns substantivos próprios vêm acompanhados de substantivos como "país", "nação", "pátria" ou "estado", o algoritmo pode perceber que tais substantivos próprios se referem a países, formando assim um contexto que posteriormente o permite identificar outros países citados no texto. Resumindo, existem opções de algoritmos que podem capturar os relacionamentos entre palavras, formando uma noção de contexto semântico e permitindo identificar o significado implícito do texto, ultrapassando ambiguidades e outras possíveis limitações da compreensão do texto por parte da máquina.

3.2.3 Opções de modelos e algoritmos de Processamento em Linguagem Natural para a implementação dos *trending topics*

A seguir abordaremos opções de algoritmos que podem solucionar a questão dos *trending topics* em uma rede social. Para selecionar o algoritmo adequado, devemos seguir alguns critérios como:

- O algoritmo deve ser capaz de identificar o contexto das palavras, ou alguma forma de identificar e atribuir relevância a elas;
- O algoritmo deve ser capaz de rodar no ambiente da plataforma CICFriend, levando em conta suas limitações computacionais;
- O algoritmo deve retornar a relevância das palavras de forma ordenada e decrescente, seguindo o formato de *trending topics*;
- O algoritmo deve ser capaz de acompanhar quais tópicos se mantêm relevantes conforme novas publicações são feitas

Glossário

Antes de prosseguir, é necessário explicar alguns termos recorrentes do contexto de algoritmos em PLN e ML. As definições a seguir foram retiradas do livro *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* [53].

- Modelo: um *software* com a tarefa de fazer predições e reconhecer padrões, utilizando técnicas de IA para "aprender" utilizando dados. De maneira superficial, o objetivo do campo de IA é automatizar tarefas e mimetizar o pensamento humano, então é comum dizer que o modelo está sendo "treinado" para executar tal tarefa. Em outras palavras, o modelo recebe um conjunto de dados, identifica padrões e a partir daí pode ser executado por conta própria e retornar os resultados desejados.
- *Training Set*, *Training Data*, conjunto de treinamento ou conjunto de teste: o conjunto de dados inicial utilizado para treinar o modelo. É essencial que o *training set* seja compatível com o resultado que esperamos, que contenha dados que simulem o cenário real. Dessa forma, o modelo tem mais chances de operar retornando os resultados esperados, alinhados com o cenário real, permitindo-o ser executado em produção.
- *Overfitting*: quando um modelo se torna enviesado sobre os dados do conjunto de teste. Com o *training set* retorna os resultados esperados, mas com dados reais (em produção) retorna resultados inconsistentes.

- Esparsidade: a presença de vários valores "vazios" ou sem representação/significado em um grande conjunto de dados. Geralmente é visto como falta de otimização do algoritmo, e a consequência direta é o desperdício de recursos computacionais como armazenamento e memória, afetando o desempenho do algoritmo.
- Dimensionalidade: o número de atributos ou características relevantes utilizados para representar um conjunto de dados. Quanto mais atributos, maior deve ser o tratamento dos dados, pré-processamento, quantidade de passos no algoritmo, e tempo decorrido nas etapas de treinamento do modelo.
- *Out of Vocabulary (OOV) Problem*: situação ocorrida quando o modelo encontra um termo fora do seu vocabulário (cenário de testes para treinar o modelo), e como consequência de não saber lidar com o novo termo, retorna resultados imprecisos e não-esperados.
- Redes neurais: as redes neurais representam um tipo de algoritmos em ML que se baseiam na estrutura do cérebro humano, com o objetivo de analisar e processar dados complexos. São capazes de detectar padrões mais implícitos no conjunto de dados.

Modelos de Espaço Vetorial (*Vector Space Models*)

O Modelo de Espaço Vetorial (do inglês *Vector Space Model*, ou VSM) [49], é um modelo algébrico, uma representação matemática para as unidades de texto (caracteres, fonemas, palavras e frases) pois para processar o texto precisamos tê-los representados em forma numérica.

Na prática, cada texto passa a ser representado por meio de um vetor numérico, onde a dimensão do vetor representa alguma característica específica do documento como o contexto ou o relacionamento entre palavras, e cada elemento representa uma palavra em si. É importante destacar a dimensão pois, textos (ou até as palavras apenas) que têm significados similares são representados por vetores com tamanho parecido, além da similaridade na distribuição dos valores.

A característica de representar os textos como vetores de números permite que os modelos tenham representações simples, tornando simples a implementação de algoritmos baseados em VSM para diversas tarefas de NLP, incluindo tarefas que não precisam trabalhar com memória (aprendizado de informações prévias) e formação de contexto [49]. Outra vantagem é que vetores são estruturas básicas presentes em diversas linguagens de programação, tanto é que existem diversas bibliotecas com implementações de algoritmos

aplicando VSM, como a scikit-learn[15].Existem 3 principais meios de implementar um VSM, que são: *One-Hot Encoding*, *Bag of Words* e *Bag of N-Grams*.

Problemas dos modelos VSM

A vantagem de utilizar VSMs também gera seu maior gargalo. Ao representar os textos por meio de vetores com valores 0s e 1s, os textos perdem suas características semânticas, nuances de cada linguagem e contexto, resultando em uma representação reducionista. Logo, ao mesmo tempo que transpor um texto para um modelo VSM permite aplicar operações matemáticas nos vetores (como utilizar distância euclidiana), o eventual processamento do texto fica limitado às propriedades do próprio vetor, como termos com frequências parecidas e a semelhança entre dois vetores. Abaixo iremos abordar três problemas que surgem na utilização de modelos VSM.

Perda do contexto semântico do texto

Como consequência de reduzir o texto a valores numéricos, os modelos VSM acabam removendo relações e funções importantes das palavras, gerando assim um problema para identificar o contexto e significado do texto. Isso ocorre pois as dimensões consideradas em cada modelo (frequência das palavras, posição no texto) não é determinante para a identificação do contexto.

Por exemplo, a Uber utiliza o modelo *Bag of Words* para separar *tickets* de suporte [49], pois pela representação vetorial de uma frase e levando em conta a frequência das palavras, é possível separar os textos que tenham palavras similares e encaminhar para o setor de atendimento adequado. Porém, suponhamos que apareça um *ticket* de suporte onde um passageiro reclama sobre "limpeza do banco", e em outro *ticket* um outro passageiro escreve "parar no banco". Ambos os textos teriam representações vetoriais parecidas, pois são frases com o mesmo total de palavras e a palavra "banco" tem a mesma frequência em ambos, porém, no primeiro caso o passageiro pode estar falando sobre o banco do carro, e no segundo texto o passageiro pode estar se referindo à uma agência bancária como o destino de sua viagem.

Logo, um modelo VSM não consegue captar a diferença do significado e diferença entre as palavras, e levar em conta apenas as dimensões e características dos vetores não é o suficiente para determinar a relevância das palavras. Essa limitação dos modelos VSM é crucial para a implementação da funcionalidade de *trending topics*, onde é necessário determinar a relevância das palavras.

[15] `https://scikit-learn.org/stable/`

Esparsidade

Por conta da representação dos textos ou palavras em forma de vetores, é natural perceber que quanto maior o texto, maior será o tamanho do vetor. Além de afetar o processamento por conta do tamanho do vetor em memória, pode ocorrer de um vetor ter inúmeros valores vazios (iguais a zero), como é o caso do modelo *One-Hot Encoding*, onde cada palavra é representada por um vetor preenchido com 0s e apenas um valor 1 de acordo com a posição da palavra no texto. Esse é o problema da esparsidade, que ocorre como uma consequência natural dos modelos VSM que utilizam representações vetoriais.

Especificamente no caso de uma rede social temos que o problema da esparsidade surge naturalmente, pois conforme o número de publicações aumenta, além da possibilidade de ter publicações com textos longos, o número de vetores (e por consequência vetores esparsos) irá aumentar. Logo, a esparsidade é uma limitação em relação aos modelos VSM para a implementação da funcionalidade de *trending topics*.

Out-of-Value Problem

Nos modelos VSM as representações vetoriais são diretamente baseadas no texto e em suas palavras, que define a característica de cada representação e também o tamanho do vetor. Por conta da limitação do modelo VSM em representar palavras de acordo com o vocabulário do texto, quando surgem palavras que não estavam presentes no texto, não há uma forma de representá-las e elas são tratadas como fora-de-vocabulário, o que pode levar a uma série de problemas.

Esse problema é conhecido como *Out-of-Value problem* ou *out-of-vocabulary* (do inglês, "sem valor" ou "fora do vocabulário"), e suas implicações em PLN são bem claras, no sentido que novas palavras não podem ser representadas em um modelo VSM, o que pode levar a informações importantes sendo perdidas. Isso é problemático para várias aplicações, como em análise de sentimento por exemplo, ou qualquer nicho que lide com um volume recorrente de textos.

No contexto de redes sociais, onde podemos ter um grande volume de publicações de novos assuntos a todo momento e o conteúdo de cada publicação é imprevisível, é inviável ter um vocabulário pré-definido para o modelo VSM, porque além do trabalho necessário para criar o vocabulário ainda surgiria o problema da esparsidade. Logo, como o modelo fica limitado aos dados de treinamento, em produção não saberá lidar com uma palavra fora do vocabulário, tornando-o inviável para a funcionalidade de *trending topics*.

3.2.4 O modelo TF-IDF e sua aplicação para a implementação de *trending topics*

As representações apresentadas anterioremente compartilham de diversos problemas que podem afetar consideravelmente a implementação de *trending topics* em uma rede social, problemas que surgem conforme o corpo de texto (publicações) aumenta. O problema principal é que não há diferença entre as representações das palavras, todas são tratadas de forma igual em representações numéricas que desconsideram os demais atributos das palavras e sua representação semântica. Com isso temos uma gama de problemas como a esparsidade, que afeta a *performance*, o *OOV problem* que torna o algoritmo inviável para um corpo de texto variável como é o caso de uma RSD, e o problema de contexto e atribuição de relevância das palavras, função crucial para definir os *trending topics*.

O TF-IDF ou "*term frequency–inverse document frequency*" é um modelo estatístico utilizado para atribuir relevância à palavras, calculando o peso de uma palavra de acordo com sua frequência em um texto e contrabalanceando esse valor com a frequência da palavra em outros textos de um mesmo grupo [49]. Basicamente, o TF-IDF quantifica a importância de uma palavra em relação às outras palavras no documento, e a relevância de um termo aumenta de acordo com a frequência em um texto, mas diminui de acordo com a proporção de frequência da palavra em outros documentos do corpo. Dessa forma, é possível medir a relevância de uma palavra de forma ponderada, onde um termo com frequência baixa (ou média) no mesmo texto, e alta frequência no decorrer de vários textos se apresenta relevante. O termo "TF-IDF" é representado por duas fórmulas, sendo TF e IDF respectivamente, e TF-IDF correspondendo ao produto de ambas.

Uma vantagem do TF-IDF é o fato de não necessitar de um grande conjunto de dados para treinamento, como os modelos de redes neurais, que é um fator crucial para situações onde a língua ou o vocabulário são limitados [49]. A pesquisa realizada em [54] aponta que o trabalho para aplicar análise de sentimento sobre opiniões públicas em redes sociais geralmente foca na língua inglesa, com poucos estudos voltados para línguas menores, e propõe treinar classificadores utilizando o TF-IDF para analisar tweets no idioma marati. O modelo foi utilizado para a predição do resultado de uma eleição em 2019, e obteve um resultado melhor que os classificadores *state-of-the-art* da época.

A fórmula TF representa a frequência do termo em um dado texto. Existem diversos textos no conjunto e cada um tem um comprimento diferente, então é natural que um certo termo tenha uma frequência maior se o texto for mais longo. Conforme mostra a Equação 3.1, o valor TF é dado pela divisão do número de ocorrências de um determinado termo em uma frase pelo número total de termos da respectiva frase[49].

$$TF(t, f) = \frac{Número\ de\ ocorrências\ do\ termo\ t\ na\ frase\ f}{Número\ total\ de\ termos\ na\ frase\ f} \tag{3.1}$$

A fórmula IDF mede a importância do termo no decorrer do conjunto de textos. O poder disso está no fato de que vários termos comuns como preposições, artigos e verbos (*stop words*) podem ter uma frequência alta, e por conta disso atuam como termos estruturais para a linguagem natural, mas que no geral não acrescentam ao significado do texto. Conforme mostra a Equação 3.2, a fórmula é dada pelo logaritmo na base e do total de frases em um determinado conjunto de frases dividido pelo número de frases que incluem um determinado termo.

$$IDF(t) = \log \mathrm{e} \frac{Número\ total\ de\ frases\ no\ conjunto}{Número\ de\ frases\ que\ incluem\ o\ termo\ t} \tag{3.2}$$

A ideia por trás do TF-IDF é favorecer palavras que tenham uma frequência média e que estejam presentes em vários textos diferentes, em vez de ter alta frequência em apenas um texto. Fazendo uma relação com *trending topics*, temos que se um determinado termo está presente nas publicações de vários usuários, mesmo que tenha uma frequência baixa na própria publicação, pode ser considerado relevante, e o algoritmo traduz isso pelo balanço do fator IDF. Podemos correlacionar o funcionamento do algoritmo com uma curva de sino, onde termos muito raros (baixa frequência), ou termos muito comuns (frequência alta) são punidos pelo algoritmo e não são considerados relevantes. A Figura 3.11 exibe o formato da curva de sino, relacionando frequência e relevância e das palavras.

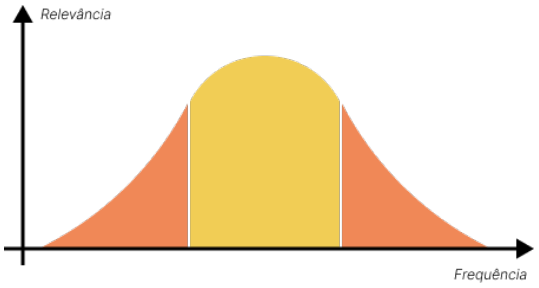

Figura 3.11: Representação de uma “curva de sino”

Para o conjunto de frases que definimos, temos a Tabela 3.1 com os devidos valores TF, IDF, e TF-IDF para cada palavra. Pelo TF-IDF se tratar de um modelo estatístico, temos uma representação numérica do peso da palavra em relação ao texto. Quanto maior o valor do termo em relação aos outros, mais relevante ele é. Pode acontecer de existir termos com o mesmo valor TF-IDF, e isso ocorre pois os termos têm exatamente a mesma frequência no conjunto de frases.

Como o TF-IDF é uma medida estatística e não gera representações vetoriais como os modelos VSM, não enfrenta o problema de esparsidade, e pelo mesmo motivo, o algo-

ritmo não fica limitado aos termos do vocabulário utilizados para formar a representação vetorial, resolvendo também o *OOV problem*. Conforme ilustrado na Figura 3.11 sobre a relação da fórmula com a curva de sino, o algoritmo é capaz de atribuir relevância pois aplica peso aos termos, punindo termos muito raros ou muito comuns, e balanceando o peso de acordo com sua presença no decorrer de outros textos.

Tabela 3.1: Tabela do cálculo TF-IDF para os termos do conjunto

Palavra	**TF**	**IDF**	**TF-IDF**
gato	0,33	0,414	0,136
cachorro	0,17	1	0,17
homem	0,33	0,414	0,136
persegue	0,17	1	0,17
morde	0,17	0,414	0,07

Observando a Tabela 3.1 podemos ver que os pesos são retornados em forma numérica, ou seja, podemos ordená-los de forma decrescente para distinguir o termo mais relevante do menos relevante. Sendo assim, o TF-IDF cumpre os critérios definidos para a seleção de um algoritmo para a funcionalidade de *trending topics*, pois é capaz de atribuir relevância às palavras, lidar com o problema de esparsidade conforme o texto escala, ordenar os termos de acordo com a relevância (em valor numérico) das palavras, e conseguir acompanhar conforme um termo se mantém relevante ou não, por apresentar um fator de balanço (IDF) de acordo com a frequência do termo em vários textos.

Para executar um teste local do algoritmo TF-IDF, foi gerado um texto arbitrário com a ferramenta *ChatGPT*[16], cujo tema envolve a relação entre inteligência artificial, geração de novos empregos e o efeito disso na população. Após a aplicação do TF-IDF no texto abaixo, foram obtidos os resultados exibidos na Tabela 3.2, que apresenta as palavras mais relevantes extraídas do texto.

```
Com a evolução da tecnologia e da inteligência artificial (IA),
muitos questionamentos surgem sobre o futuro do trabalho e da
economia. Alguns temem que a crescente automação e IA resultem
em demissões em massa, levando a um cenário apocalíptico de
desemprego generalizado e até mesmo no fim do mundo como conhecemos.

Porém, outros argumentam que a IA não é apenas uma ameaça aos
```

[16] `chat.openai.com/`

```
empregos existentes, mas também uma oportunidade para novas
ocupações e uma nova revolução industrial. Com a adoção da
tecnologia de ponta pelas empresas, novas oportunidades surgirão
para aqueles com habilidades em áreas como programação, análise de
dados e design de sistemas.

Ao mesmo tempo, é importante considerar como o governo pode ajudar
aqueles que são impactados pelas mudanças trazidas pela IA. Medidas
como fornecer treinamento gratuito para novas habilidades, a criação
de novos programas de emprego e a expansão da rede de seguridade
social podem ser cruciais para aqueles que são afetados pela automação
e a IA.
```

Tabela 3.2: Tabela do cálculo TF-IDF para os termos do texto de demonstração

Termo	**TF-IDF**
IA	0,020
novas	0,012
tecnologia	0,008
automação	0,008
habilidades	0,008
evolução	0,004
inteligência	0,004
artificial	0,004
questionamentos	0,004
futuro	0,004

3.2.5 Uma alternativa aos modelos VSM: representações distribuídas

Para resolver os problemas citados anteriormente nos modelos VSM, foram criados métodos para lidar com representações de baixa dimensão, permitindo manipular longos textos sem lidar com a esparsidade e dimensionalidade. Esses métodos utilizam redes neurais para criar representações densas de informações, evitando lidar com representações geométricas e os problemas de desempenho apresentados por elas [49].

As redes neurais são capazes de identificar e atribuir relacionamento entre as palavras, sendo capazes de capturar o significado e contexto delas. Assim como os modelos de VSM, as redes neurais também geram representações vetoriais das palavras, de forma que

palavras que aparecem em contextos parecidos terão representações parecidas. Porém, diferente deles, conseguem lidar com a esparsidade e *OOV problem*.

De forma superficial, a rede neural trabalha analisando uma palavra e ao mesmo tempo analisa as palavras vizinhas, para identificar o relacionamento entre eles e posteriormente ser capaz de identificar o mesmo padrão. Por exemplo, ao notar que alguns substantivos próprios vêm acompanhados do mesmo conjunto de substantivos como "*país*", "*nação*", "*pátria*" ou "*estado*", e sabendo que os substantivos próprios se referem a países, a rede neural percebe que tais substantivos próprios se referem a países, formando assim o contexto e posteriormente será capaz de identificar outros países citados no texto, sempre que houver um nome acompanhado por algum dos termos sinônimo de "*nação*", "*país*", etc.

A rede neural gera um único vetor contendo todas as palavras, logo, não tem campos vazios ou vetores preenchidos com inúmeros 0s como nos modelos de VSM que vimos anteriormente. Adicionalmente, pela capacidade de formar contexto e inferir significado, quando surge alguma palavra nova no texto, que não estava no vocabulário do *training set*, a rede neural é capaz de inferir o significado dela baseado no contexto das palavras anteriores (incluindo o *training set*).

Logo, o modelo de representações distribuídas que funcionam com redes neurais são capazes de lidar com problemas anteriores que os modelos de VSM não conseguem, como problemas de contexto, preservação semântica das palavras, relevância (por meio de frequência em contextos diferentes), esparsidade e *OOV problem*. Porém, como veremos adiante, esse modelo não parece ser uma solução viável para o caso específico de *trending topics* na rede social CICFriend.

1. *Word2Vec*

O *Word2Vec* da Google [55], é um algoritmo de representação distribuída, cujo funcionamento consiste em capturar relações semânticas e sintáticas de um texto, usando uma representação vetorial e usando uma rede neural para processar e fazer predições das palavras. Logo, é capaz de inferir se uma palavra pertence a um contexto, e dado um contexto pode inferir uma palavra.

Uma grande vantagem desse algoritmo é o fato dele capturar o significado das palavras de uma forma que os modelos de VSM não conseguem, por criar representações densas (sem campos vazios, evitando esparsidade) e tendo um modelo de predição. Essa funcionalidade permite que o algoritmo seja utilizado para sistemas de recomendação como músicas ou filmes, onde baseando-se em recomendações e interesse prévio do usuário, é possível inferir outras obras que o usuário possa gostar.

Com isso, o *Word2Vec* se apresenta como uma boa solução para os *trending topics*, pois consegue medir a relação entre palavras e definir quais palavras aparecem sobre o mesmo contexto ou não, além de permitir inferir tópicos relacionados com os *trending topics*, com o modelo de predição.

Para o algoritmo funcionar de forma precisa quanto à predição de palavras e formação de contexto, a rede neural necessita de um grande conjunto de dados para treinamento para fazer comparações entre as palavras, extrair as nuances entre elas, identificar relacionamentos e assim criar representações mais precisas. No geral, são necessários conjuntos com centenas de milhares de palavras para obter uma rede neual não-enviesada e precisa [52].

2. *fastText*

O *fastText* [56], é uma biblioteca de classificação de textos desenvolvida pelo *Facebook*. Assim como o *Word2Vec*, o *fastText* é capaz de capturar relações semânticas de um texto, por meio de uma rede neural e representando os dados por meio de um vetor denso.

No geral, o *fastText* tem funcionamento parecido com o *Word2Vec* e atende casos de uso parecidos, também resolvendo os problemas de contexto e esparsidade. Porém, seu diferencial é a capacidade de subdividir as palavras morfologicamente, expandindo a capacidade de relacionar palavras, formar contexto e inferir novas palavras. Dessa maneira, o *fastText* tem mais possibilidades de relacionamento entre palavras e uma inferência potente comparada ao *Word2Vec*, lidando melhor com o *OOV problem*.

3.3 Princípios de usabilidade na implementação do componente de *trending topics*

Quando se fala em criar uma aplicação, uma interface, ou até mesmo uma simples *feature*, sempre deve ser levado em conta como tal funcionalidade se aplica ao usuário [57], como ele irá interpretar, interagir, e até mesmo seu sentimento em relação à aplicação como um todo. Essa necessidade culminou no surgimento da área de UX (do inglês, *user experience*), que pode ser traduzida como "experiência do usuário" ou até "*design* voltado ao usuário".

Os princípios de UX são cruciais no desenvolvimento de interfaces pois cobrem todo o processo de interação do usuário com a aplicação, incluindo desde o projeto e navegação entre telas, até acessibilidade, facilidade de uso e tomadas de decisão [58]. Ignorar os princípios de UX pode afetar negativamente a experiência do usuário [57], podendo reduzir o engajamento e satisfação com a aplicação ou até fazer o usuário buscar outras aplicações.

No contexto de *trending topics*, os assuntos em alta ficam organizados de forma que estejam alinhados com a experiência do usuário, evitando uma sobrecarga de informações ou conflito com outras *features* da plataforma, permitindo assim propiciar uma experiência mais proveitosa e aumentar o engajamento do usuário com a aplicação, focando em trazer informações relevantes para o usuário. Para alcançar tais efeitos, é necessário aplicar conceitos de UX não tanto com o objetivo de trazer uma experiência positiva para o usuário, como para evitar confusão por parte dele e até que o mesmo desista de usar a RSD. Iremos elaborar conceitos essenciais na concepção do *layout* da interface de *trending topics*, incluindo a organização dos tópicos em formato de lista, o posicionamento da lista à esquerda da tela, e a hierarquia de informações.

3.3.1 Lei de Hick

A Lei de Hick é um princípio da psicologia que relaciona o "número de escolhas possíveis" em uma interface com o tempo decorrido para o usuário tomar uma decisão. De forma resumida, com menos opções disponíveis o usuário pode se situar rapidamente na aplicação e tomar as ações desejadas [59].

Matematicamente, é dito que o tempo necessário para um usuário fazer uma ação cresce de forma logarítmica conforme o número de opções aumenta [60]. Intuitivamente, isso ocorre pois é necessário que o usuário percorra a tela e interprete as possibilidades para não só escolher o que deseja fazer, como para evitar uma ação errada. Esse princípio é relevante no campo de UX pois implica diretamente em como um usuário irá interagir com uma interface, principalmente por destacar a importância de favorecer ter um conjunto limitado de opções em vez de sobrecarregar o usuário com informações.

Na prática, uma interface deve ter um número limitado de elementos, para manter um tempo de decisão baixo no lado do usuário [60]. Por exemplo, em *websites* é recomendado não ter menus com longas listas de categorias, e evitar também que essas categorias abram uma lista de subcategorias [60], pois a seleção de um item em um menu envolve leitura de frases, busca, compreensão, etc [60].

A importância desse conceito se dá pois uma interface com muitas opções pode confundir o usuário (principalmente usuários iniciantes), e por consequência pode reduzir o engajamento ou até fazer o usuário desistir de utilizar a aplicação. Em [61], um estudo conduzido com adolescentes concluiu que no geral, usuários nessa faixa etária constumam navegar rapidamente pelos *websites*, o que pode dificultar no processo de achar o que estão procurando. Por conta disso, preferem informações pictoriais em vez de grandes blocos de informação, apoiando a Lei de Hick.

Em suma, a Lei de Hick se faz essencial por permitir que sejam apresentados ao usuário apenas os tópicos de seu interesse, para evitar uma sobrecarga mental e melhorar

a experiência de uso do usuário [60]. Adiante veremos como aplicar isso no contexto de *trending topics*, dispondo os *trending topics* no formato de listas.

3.3.2 Organizando os *trending topics* no formato de listas

As listas são uma escolha de *design* comum [62] para apresentar informações de forma clara e organizada, e isso se dá por conta do formato vertical em que as informações são apresentadas, seguindo o padrão de orientação de leitura. Listas costumam ser utilizadas em vários contextos diferentes como em menus, exibição de produtos, ou na exibição de *posts* em uma rede social. A grande vantagem de usar esse formato é que, por se tratar de elementos em forma sequencial, o usuário consegue se situar rapidamente, navegar pela lista e encontrar o que deseja evitando informações irrelevantes, por exemplo. Para tal, é necessário que haja um parâmetro de organização na lista, seja ordem alfabética, decrescente ou cronológica [62].

No contexto de *trending topics*, o formato de lista é efetivo para ajudar o usuário a identificar os tópicos por ordem de relevância, do tema mais comentado ao menos comentado, e dessa forma ele pode se situar rapidamente sobre os assuntos em alta à medida que a lista é atualizada com novos tópicos. Outro ponto relevante ao aplicar listas no *design* é que, como as informações estão de forma organizada, o usuário pode achar rapidamente o que precisa, possibilitando-o focar em conteúdo relevante e aumentando o engajamento na plataforma. Isso nos leva ao conceito de *information foraging*.

3.3.3 *Information Foraging*

Information Foraging (coleta de informações) é um conceito cognitivo que representa o processo de buscar informações que satisfaçam uma necessidade ou desejo [63]. No campo de UX, se aplica à forma que os usuários interagem com sistemas digitais e *websites*, incluindo redes sociais. No contexto de redes sociais, ocorre quando o usuário procura por conteúdo de seu interesse, e nisso fica evidente a importância dos *trending topics* [63].

Um dos principais princípios por trás do *information foraging* é que os usuários são motivados pela necessidade de informações e que o sistema deve facilitar o acesso a elas [63], permitindo os usuários acharem o que procuram. Levando em conta que redes sociais são feitas por e para pessoas, é essencial que o conteúdo e assuntos presentes sejam de interesse dos usuários, fomentando o engajamento e uso da rede [61].

No contexto de redes sociais, ter uma seção de *trending topics* pode direcionar o usuário para os assuntos em alta, permitindo-o ficar a par dos assuntos relevantes e interagir com os outros usuários, fazendo-o ficar mais tempo na rede social. Levando em conta os princípios anteriores, é possível criar uma experiência eficiente e intuitiva para os usuários,

maximizando engajamento e satisfação. Porém, uma pergunta que surge naturalmente é "o que acontece após o usuário interagir com a lista de *trending topics*?". Para responder essa pergunta apresentaremos o princípio de *progressive disclosure*.

3.3.4 *Progressive Disclosure*

Progressive Disclosure é um princípio de UX que envolve apresentar informações ao usuário de forma gradual, por demanda, conforme ele se familiariza com a aplicação [64]. Essa técnica possibilita o usuário focar nas *features* principais da aplicação, reduzindo a sobrecarga de informações e facilitando o entendimento. Desta forma, um usuário iniciante não fica confuso e consegue aprender mais facilmente como usar a aplicação [64].

Esse conceito visa reduzir a complexidade de uso ao separar as informações de forma hierárquica e evitar poluição visual [57], além de separar as funcionalidades da aplicação, visto que as *features* mais utilizadas e/ou mais importantes ficam em evidência, e *features* mais específicas ficam "escondidas", de modo que o usuário intencionado consegue encontrá-las e utilizá-las. Na prática, as funções primárias de um aplicativo devem aparecer já na tela inicial do usuário, e após alguma outra ação do usuário, as funcionalidades secundárias são exibidas. Caso esse princípio seja ignorado, funcionalidades primárias são mescladas com funcionalidades secundárias, poluindo a interface e afetando a experiência de usuários iniciantes e até mesmo usuários avançados [64].

No contexto de *trending topics* é essencial aplicar tal conceito pois não é interessante que sejam exibidas para o usuário todas as publicações possíveis (incluindo as que não são relevantes para ele). Isso não só causaria poluição visual e aumentaria o tempo de decisão do usuário (Lei de Hick), como impediria ele de focar no que é de seu interesse, quando na verdade a aplicação deve fornecer isso para ele (*information foraging*). Ao separar os assuntos em *trending topics*, o usuário pode visualizar os assuntos que têm interesse, e ao clicar em algum tópico, é redirecionado para uma listagem que inclui posts do tal assunto.

Logo, a prática de *progressive disclosure* é essencial para estabelecer a hierarquia de funcionalidades na aplicação, diminuir a tomada de decisões do usuário, levar informações relevantes ao momento atual, reduzir a poluição visual e evitar confusão no uso da aplicação.

3.3.5 Posicionamento dos *trending topics* no *layout*

O sentido de leitura ocidental costuma partir da esquerda para a direita, em textos, listas, tabelas, etc [57]. Portanto, levando em conta que esse é o padrão de navegação em *websites* e outras aplicações, seja *mobile* ou *desktop* [57], o posicionamento do componente de listagem do *trending topics* deve ficar à esquerda da tela, como justificaremos a seguir.

Em [65] foi conduzida uma pesquisa para entender a distribuição de atenção do usuário em monitores, usando a tecnologia de *eyetracking*, que consiste em utilizar um dispositivo que captura a direção que o usuário olha para mapear com a respectiva região na tela. A pesquisa concluiu que os usuários passam cerca de 69% do tempo olhando para o lado esquerdo da tela, mais que o dobro do tempo gasto no lado oposto, que foi de 30%. A pesquisa correlacionou o tempo gasto em cada região da tela (com as regiões divididas em *pixels* de 0 a 1100, representando a tela inteira), e representou a distribuição de atenção por regiões da tela com os valores organizados de forma crescente conforme os *pixels* da tela. Os resultados indicam que as 5 primeiras faixas de valores (de 100 a 500 *pixels*) representam o lado esquerdo da tela, e que essas faixas contêm os maiores percentuais de tempo de visualização, implicando que o usuário passa mais tempo olhando para o lado esquerdo da tela.

Logo, concluindo que o usuário passa boa parte do seu tempo de navegação focando no lado esquerdo da tela, é viável posicionar o componente de *trending topics* à esquerda da tela, pois por se tratarem de tópicos relevantes ao usuário, faz sentido posicioná-lo na sua região de mais atenção.

3.4 Referencial tecnológico

Nesta seção iremos abordar a fundamentação técnica para implementar uma aplicação de *trending topics* utilizando o modelo TF-IDF. As ferramentas sugeridas compõem uma aplicação cliente-servidor em formato de demonstração, no modelo da plataforma CIC-Friend.

3.4.1 Aplicação-cliente - React

No contexto da arquitetura cliente-servidor, o cliente desempenha o papel responsável pelas interações do usuário com o sistema. No caso de uma aplicação de *trending topics*, deve permitir que o usuário faça e veja publicações, além de conferir em tempo real as atualizações de *trending topics*. Na aplicação de demonstração, foi utilizada a biblioteca de *front-end* React[17] para implementar o lado do cliente.

React[18] é uma biblioteca Javascript desenvolvida pelo Facebook e utilizada para construir interfaces de usuário (UI) em aplicações *web*. O código escrito em *React* funciona com uma sintaxe chamada JSX, que é semelhante ao HTML e sua estrutura de *tags*, que permite escrever código Javascript e HTML juntos, criando componentes reutilizáveis que podem ser combinados para construir interfaces de usuário complexas. Além disso, o

[17] `https://react.dev/`
[18] `https://react.dev/`

React possui um sistema de "reconciliação" que permite atualizar a interface do usuário de forma eficiente quando ocorrem mudanças nos dados e interações do usuário. O *React* é amplamente utilizado no desenvolvimento web moderno, pois permite criar interfaces de usuário escaláveis, de alta performance e fáceis de manter.

Com a popularização do *React*, muitos desenvolvedores começaram a utilizá-lo como uma alternativa ao modelo tradicional de desenvolvimento web, que envolve a manipulação direta do DOM e a atualização do HTML e CSS a partir de código JavaScript, por ter um conjunto de funções (por padrão) que reduzem escrita do código e manutenção por parte dos desenvolvedores. A abordagem do *React*, por outro lado, é baseada em componentes, o que significa que a interface do usuário é dividida em partes menores e independentes, cada uma sendo representada por um componente. Isso torna a construção e a manutenção de interfaces de usuário complexas muito mais gerenciável, escalável e eficiente. Tudo isso faz com que o *React* seja uma escolha popular para o desenvolvimento de aplicativos web modernos e dinâmicos.

3.4.2 Aplicação-servidor - NodeJS

Apesar do usuário ter contato apenas com o lado do cliente, com a interface, não significa que não esteja acontecendo nada por baixo dos panos. Nesse sentido, como já explicado antes, a aplicação se divide em cliente e servidor, e o servidor funciona como uma aplicação separada. Para sua implementação, é necessário um *framework* de *back-end*, e no contexto de nossa aplicação usaremos o *Node.js*.

Node.js[19] é uma plataforma de desenvolvimento de software construída em cima do motor V8 do Google Chrome, ou seja, permite executar código Javascript fora do navegador. O *Node.js* usado principalmente para construir aplicativos de servidor da *web* e APIs REST, mas também pode ser usado para criar aplicativos de linha de comando e até mesmo aplicativos *desktop*. Assim como o *React* é apenas uma biblioteca e necessita de um *framework* para complementar seu funcionamento, o *Node.js* também conta com uma gama de *frameworks* para a construção de sistemas, como o Express.JS, AdonisJS, etc.

Além disso, o Node.js é altamente escalável e pode lidar com um grande volume de solicitações de maneira eficiente, o que o torna uma escolha popular para o desenvolvimento de aplicativos web em tempo real.

[19] `https://nodejs.org/en`

3.4.3 Framework para aplicação cliente-servidor - Next.js

Como já diz o nome, para construir aplicações cliente-servidor, é necessário ter uma aplicação para cliente e outra para o servidor, seja em aplicações separadas ou na mesma aplicação, onde nesse caso existem *frameworks fullstack* como *Ruby on Rails*[20] ou Django[21]. Para a aplicação de demonstração foi utilizado o *frameworkNext.js*[22], um *framework fullstack* que utiliza o *React* para renderizar a aplicação do cliente.

Next.js[23] é um framework de desenvolvimento *web* de código aberto construído sobre o *React*. Ele permite que o desenvolvedor crie aplicativos *web* escaláveis complementando o ecossistema do *React* com uma ampla gama de recursos integrados, incluindo renderização do lado do servidor (SSR), geração de sites estáticos (SSG), e um recurso chamado *API Routes*, que simula um servidor *web*, permitindo a aplicação ter um *back-end* próprio.

Por se tratar de um *framework* e ter funcionalidades prontas, o *Next.js* permite que os desenvolvedores criem aplicações complexas sem ter que se preocupar com tarefas comuns de configuração, como roteamento, gerenciamento de estado, inicialização do servidor, otimização de imagens, etc. Além disso, como o *Next.js* utiliza *React*, qualquer biblioteca de terceiros criada para *React*, também funciona com o *Next.js*.

Como mencionado anteriormente, o *Next.js* conta com uma funcionalidade chamada *API Routes*, que atua como um servidor próprio da aplicação. Esse servidor é implementado em *Node.js*, ou seja, ao mesmo tempo que utilizamos *Next.js* para a aplicação no lado do cliente (interface *web*), o *Next.js* também pode lidar com o lado do servidor, sendo uma aplicação completa incluindo a arquitetura cliente-servidor no mesmo projeto (com exceção do banco de dados).

3.4.4 Banco de dados - FaunaDB

O último componente da arquitetura cliente-servidor da aplicação é o banco de dados, que no caso da aplicação *demo* fica responsável pelo armazenamento de postagens e usuários da aplicação. Porém, um banco de dados não é uma aplicação a parte e deve ser executado em um servidor para estar acessível para o servidor da aplicação. Nesse sentido, iremos introduzir o conceito de computação *serverless* (do inglês, "sem servidor").

Servidores são um dos principais componentes de qualquer aplicativo moderno, no entanto, configurá-los e mantê-los pode ser uma tarefa complexa e demorada para muitos desenvolvedores, pois exige um conhecimento técnico a parte do desenvolvimento, envolvendo infraestrutura, redes, endereçamento IP, etc. O modelo *serverless* permite

[20] `https://pt.wikipedia.org/wiki/Ruby_on_Rails`
[21] `https://pt.wikipedia.org/wiki/Django_(framework_web)`
[22] `https://nextjs.org/`
[23] `https://nextjs.org/`

que os desenvolvedores se concentrem exclusivamente em escrever código, sem precisar se preocupar com a configuração ou a manutenção de servidores, onde toda a configuração é feita de forma automática por provedores da nuvem. Na prática, o servidor tem um funcionamento *on-demand*, onde em vez de ficar no ar de forma indefinida, é executado apenas quando a aplicação recebe algum *request*, fazendo com que o desenvolvedor pague apenas pelos momentos em que a aplicação realmente foi utilizada. Isso pode ser uma grande vantagem para aplicativos com demandas variáveis de tráfego, onde os recursos do servidor precisam se adaptar rapidamente às mudanças de uso.

Para ter um banco de dados em produção, é necessário que ele esteja instalado em algum servidor, que também exige configuração e manutenção por si só. Uma alternativa então é usar o modelo *serverless* para executar o banco de dados, de forma que um banco de dados *serverless* é um serviço gerenciado que fornece todas as funcionalidades de um banco de dados tradicional, sem que o desenvolvedor precise se preocupar com a configuração ou a manutenção de servidores de banco de dados. Como opção para a aplicação de demonstração, foi utilizado o *FaunaDB* [24], que é um banco de dados *serverless* com recursos como escalabilidade automática, alta disponibilidade e segurança integrada. Além disso, o *FaunaDB* oferece um modelo de dados flexível e uma API rica em recursos, permitindo que os desenvolvedores criem aplicativos escaláveis e flexíveis com facilidade.

FaunaDB é um banco de dados distribuído, seguro, e *serverless*. Oferece escalabilidade horizontal automática, fácil integração com outras ferramentas e serviços e uma *query language* própria denominada FQL. Ele é projetado para suportar aplicativos modernos, que exigem uma infraestrutura de banco de dados escalável e de alto desempenho, que possa suportar a carga de tráfego em constante crescimento e fornecer uma experiência de usuário sem interrupções. O FaunaDB permite que os desenvolvedores criem aplicativos altamente disponíveis e escaláveis, sem se preocupar com o gerenciamento e a manutenção da infraestrutura do banco de dados, além de oferecer segurança, conformidade e privacidade de dados. Ele é baseado em uma arquitetura de microsserviços distribuídos, que garante uma alta disponibilidade e uma latência extremamente baixa. Com sua API GraphQL nativa, o FaunaDB também torna a integração com aplicativos web e móveis uma tarefa fácil e intuitiva.

3.4.5 Hospedagem - Vercel

Para ter a aplicação em produção, é utilizado um ambiente de hospedagem. Para a aplicação de demonstração foi utilizada a *Vercel*[25].

[24] `https://fauna.com/home`
[25] `https://vercel.com/`

Vercel[26] é uma plataforma de hospedagem e computação em nuvem voltada para a distribuição de aplicativos *web*, e oferece uma série de recursos e funcionalidades que podem tornar o processo de desenvolvimento mais fácil e mais rápido, como CI/CD, *Edge Network* e *serverless functions*. No quesito de hospedagem, a *Vercel* oferece um plano gratuito, tornando-a uma ótima escolha para subir a aplicação sem custo.

[26] `https://vercel.com/`

Capítulo 4

Proposta de implementação

Neste capítulo será detalhada a proposta de implementação da *feature* de *trending topics*. A proposta foi dividida nas etapas de visão geral e estruturação dos casos de uso, seguido da prototipação da ferramenta.

4.1 Visão geral da dinâmica provida pela funcionalidade proposta

A rede social CiCFriend, instância do Friendica, é uma RSD utilizada pelos alunos e professores do Departamento CIC para compartilharem ideias e informações a partir de suas funcionalidades, e uma delas é a realização de *posts*. Esses *posts* são publicados nos perfis dos usuários, e quando possuem a marcação de um fórum eles também são mostrados a todos os membros deste fórum. A partir dos *posts* os usuários podem compartilhar seus pensamentos, dúvidas e opiniões. A comunidade pode interagir com esses *posts* através de *likes*(gostar), *dislikes*(não gostar), comentários e compartilhamentos.

A funcionalidade sugerida neste trabalho, apresentada por protótipos neste capítulo, tem como finalidade a implementação de uma aba que mostra aos usuários da rede social CiCFriend os tópicos em relevância do momento, ou seja, os conteúdos que estão sendo mais comentados pelos usuários da rede. Esses tópicos serão gerados a partir dos *posts* dos usuários, e serão apresentados em forma de uma lista ordenada, conforme mostrado na ilustração do funcionamento do *add-on* (Figura 4.1).

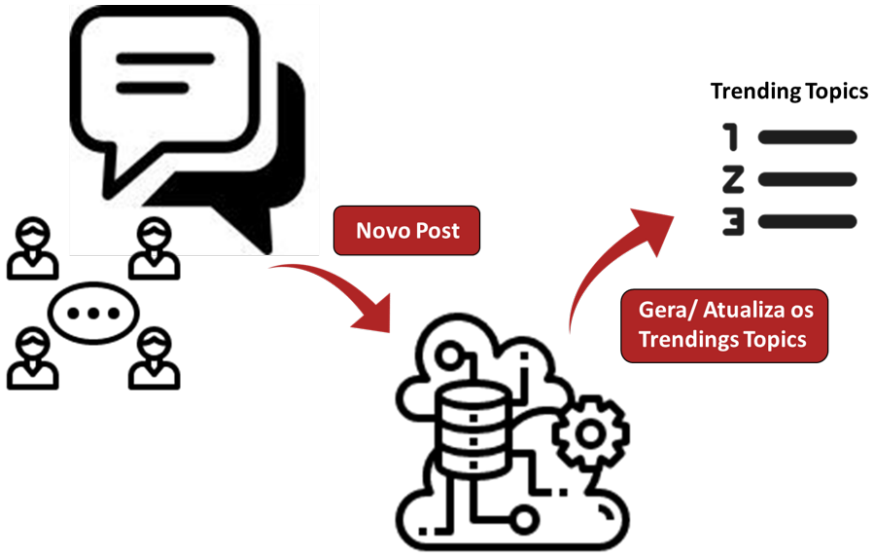

Figura 4.1: Ilustração da geração dos *trending topics*

O usuário, ao acessar a rede social, terá rápido acesso aos temas e assuntos mais conversados nos últimos dias. Também será possível acessar os *posts*, acompanhar as discussões e participar dos debates e colaborar com a comunidade.

A Figura 5.1 ilustra as possibilidades de interação do *add-on* proposto, iniciando-se pela visualização da lista dos *trending topics*. A partir dessa lista, o usuário pode selecionar um dos tópicos de sua escolha, onde ele terá a visualização da lista dos *posts* que contém esse tópico específico. Ainda na página desses *posts*, também será apresentada a lista de fóruns relacionados com aquele tema. O relacionamento entre os fóruns e os tópicos irá ocorrer a partir das *tags* públicas.

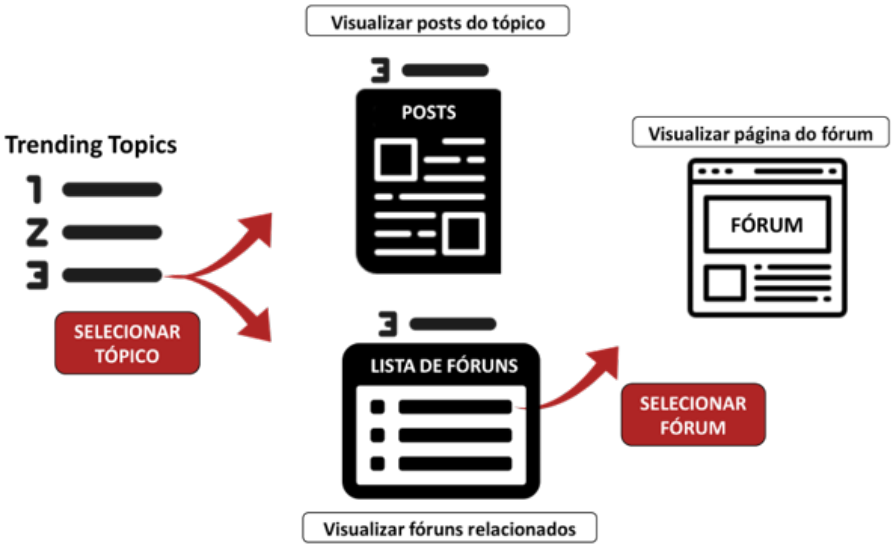

Figura 4.2: Ilustração das funcionalidades a partir do *trending topics*

As contas do tipo fórum serão utilizadas para representar uma comunidade que tenha um interesse em comum. Esse interesse pode ser uma linguagem de programação, uma tecnologia ou uma disciplina do curso. Para o caso de um fórum de disciplina em específico, as *tags* presentes nesse perfil poderão ser configuradas de acordo com os seus assuntos específicos. Os assuntos serão retirados dos conteúdos programáticos, das ementas adotadas pelo CIC, e das bases curriculares da SBC. Desse modo, os temas em evidência poderão ser utilizados como gancho de assuntos ou objeto de análise dentro das aulas e fóruns das matérias.

4.2 Casos de uso

Os casos de uso foram modelados na ferramenta *LucidChart*[1] e estão disponíveis no repositório do Git juntamente com a implementação.

Como primeiro passo da apresentação da proposta, foram definidos casos de uso para a *feature* de *trending topics*. Nas Figuras 4.3 e 4.4, são apresentados os diagramas de casos de uso, onde temos como atores os gestores da instância e os usuários do CiCFriend, que podem se tornar participantes de fóruns específicos, criados por qualquer membro da comunidade, seja ele um professor, aluno ou gestor da instância. Para criar uma conta do tipo fórum o usuário que já possui conta deverá acessar a página de gerenciamento de contas e adicionar a conta do tipo fórum. O usuário criador do fórum se torna automaticamente o administrador do mesmo.

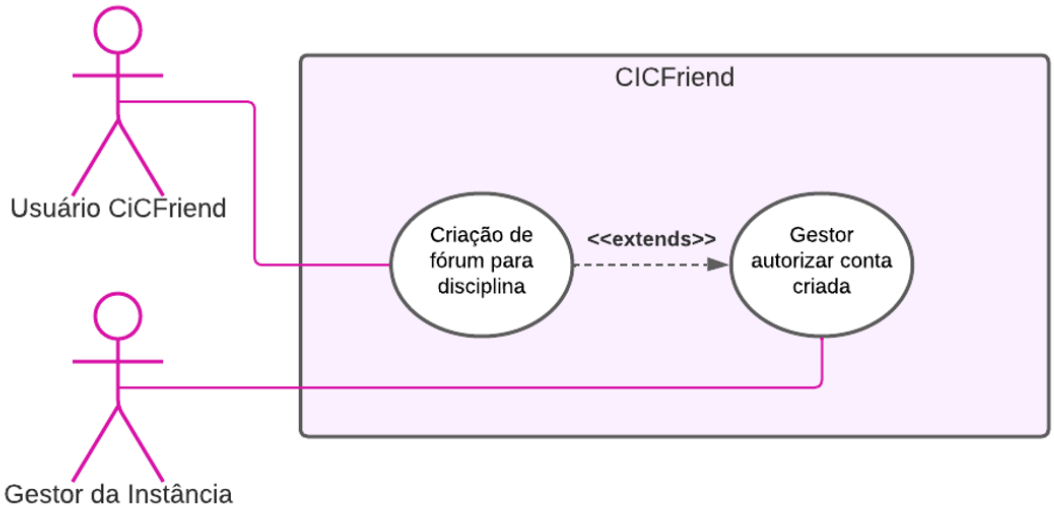

Figura 4.3: Diagrama de Uso 1 - Criação de Conta Tipo Fórum

[1] `https://www.lucidchart.com/pages/pt`

Na Figura 4.3, é apresentado o diagrama de casos de uso que representa uma ação de criação de fórum para uma disciplina, que pode ser realizada por qualquer usuário do CiCFriend. Esses fóruns serão utilizados para interligar os *trending topics* às matérias do CIC.

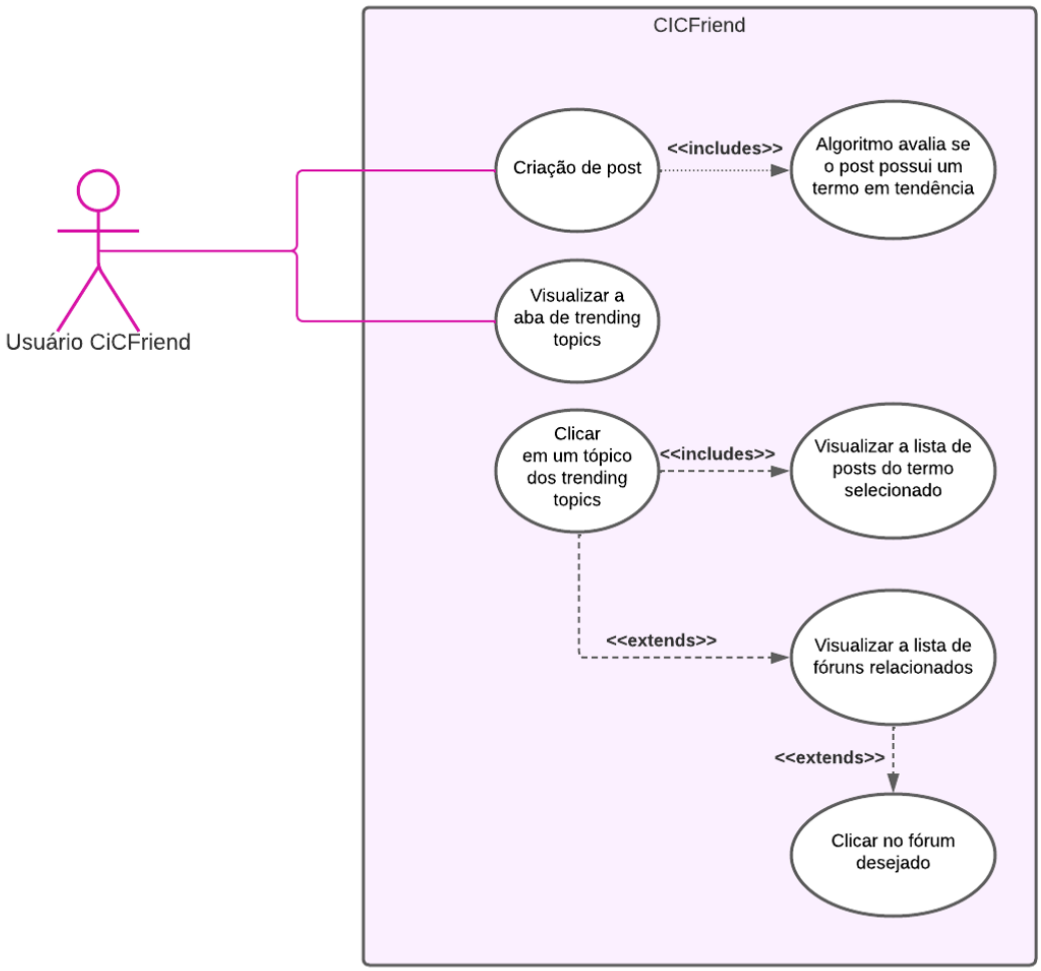

Figura 4.4: Diagrama de Uso 2 - *Trending Topics*

Já na Figura 4.4, são mostradas as ações que podem ocorrer voltadas aos *trending topics*. Essas ações têm início a partir da ação base que é a criação de postagens para a formação da lista de termos em tendência.

Caso de Uso 1	**Criação de Fórum para Disciplina**
Descrição:	Usuário realiza a criação da conta fórum para uma disciplina.
Atores:	Usuários do CICFriend
Pré-condição:	Ator cadastrado na rede Friendica
Cenário:	1. O ator entra na tela de gerenciar contas e clica na opção “registrar conta adicional”. 2. O ator escolhe um nickname para seu fórum de acordo com a disciplina. 3. O ator informa sua senha para a validação da conta. 4. O ator entra na sua conta novas através da aba de contas. 5. Ator entra na aba “contas”, presente no menu superior direito, aba “tipos avançados de contas”, e muda o tipo da conta para fórum comunitário.
Extensões:	4.1 Sistema requer autorização de gestor para criação de novas contas. (a) Gestor autoriza nova conta criada.

Ao criar um fórum para a disciplina, o gestor da conta deve acessar as configurações de perfil, na aba “Diversos”, campo “Palavras-Chave Públicas”, e realizar a inserção dos termos chaves daquela disciplina, para que eles sejam relacionados aos termos dos *trending topics*.

Caso de Uso 2	**Criação de postagem**
Descrição:	Usuário realiza publicação de uma postagem no feed.
Atores:	Usuários do CICFriend
Pré-condição:	Ator cadastrado na rede Friendica
Cenário:	1. Ator realiza publicação da postagem a partir da sua página ou da página de Conversas dos seus amigos. 2. O sistema armazena a postagem realizada no feed para utilização na geração de *trending topics*. -> Caso de uso: Algoritmo avalia se o post possui um termo em tendência. 3. É avaliado pelo algoritmo se aquele post possui um termo em tendência naquele momento. -> Caso de uso: Algoritmo avalia se o post possui um termo em tendência. 4. Se o post conter um termo em tendência naquele momento, ele será mostrado na lista de *posts* do tópico em questão. -> Caso de uso: Algoritmo avalia se o post possui um termo em tendência.

Caso de Uso 3	**Clicar em um tópico do *trending topic***
Descrição:	Usuário visualiza a lista de tópicos e clica em algum de sua preferência.
Atores:	Usuários do CICFriend
Pré-condição:	*Trending Topics* gerado
Cenário:	1. Usuário acessa a aba da Comunidade Local e visualiza a aba de *Trending Topics* na lateral esquerda da página. 2. Usuário clica em algum tópico dos *trending topics* de acordo com sua preferência. 3. Usuário é redirecionado para a página de *posts* que utilizaram o tópico escolhido. 4. Usuário visualiza a lista de *posts* do tópico escolhido -> Caso de uso: Visualizar a lista de *posts* do termo selecionado.
Extensões:	3.1 Usuário pode visualizar a lista de fóruns relacionados, se houver pelo menos um. (a) Usuário pode clicar em um fórum presente na lista para entrar na página dele.

4.3 Mapeamento dos *trending topics* para as Disciplinas do CIC

Uma das funcionalidades do *add-on* de *trending topics* apresentado nessa proposta, é a de realizar a integração dos tópicos em tendência com os conteúdos das disciplinas do CIC, conforme apresentado na Seção 2.3. As possibilidades de conectar publicações ao contexto universitário são diversas, expandindo a barreira do conhecimento e tornando a experiência de navegação mais holística.

Para a realização do mapeamento das palavras-chave, serão utilizadas como base principal as ementas das disciplinas do CIC. Reiterando a descrição apresentada na Seção 2.3, a criação das ementas dos cursos de computação precisam estar de acordo com as Diretrizes Curriculares Nacionais, as quais apresentam uma base descritiva para todas as IES.

Nas ementas das disciplinas é possível ter acesso aos temas de conteúdos que serão abordados pelos professores durante o curso. Esses temas serão referências primárias para elaboração de palavras-chave utilizadas no perfil do fórum, conforme explicado na Seção 4.4, para o relacionamento com os assuntos mais comentados da plataforma (Figura 4.5), levando em consideração que no trabalho [9] se convencionou que cada oferta de disciplina do CIC pode ter seu perfil associado a uma conta do tipo fórum.

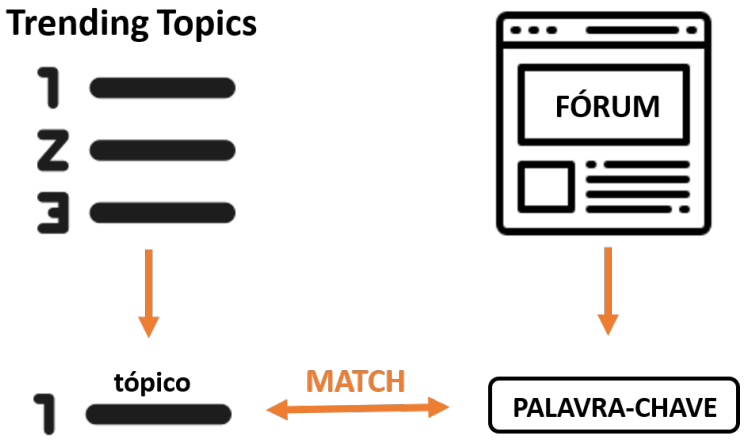

Figura 4.5: Mapeamento dos *trending topics* para as contas tipo fórum

Assim como as competências presentes no referencial de formação proposto pela SBC podem ser desenvolvidas em diversos conteúdos, os temas de uma disciplina da computação também podem ser abordados de diversas formas, ou introduzidos por diversos assuntos. Desse modo, os professores podem a partir da ementa escolher entre tópicos a serem discutidos ou apresentados em sala de aula durante o processo de ensino aprendizagem.

Essa decomposição dos tópicos poderá ser realizada utilizando outras fontes de temas, como por exemplo, os planos de aula, os assuntos abordados em sala de aula, novas tecnologias e até mesmo o que os professores e alunos considerarem importante. Com base nisso, a fim de ilustrar como será obtida a associação dos *trending topics* com as disciplinas, foram realizados exemplos de uma seleção de palavras-chave de duas matérias do CIC, mostrando também como uma palavra-chave pode ser associada a mais de uma matéria.

Na Figura 4.6, obtida no portal de gestão de atividades acadêmicas utilizado pela UnB (SIGAA[2]), pode-se observar alguns dados da matéria Algoritmos e Programação de Computadores (APC), que é um componente curricular obrigatório do Curso de Computação - Licenciatura. Nesses dados, está presente a ementa/descrição do componente.

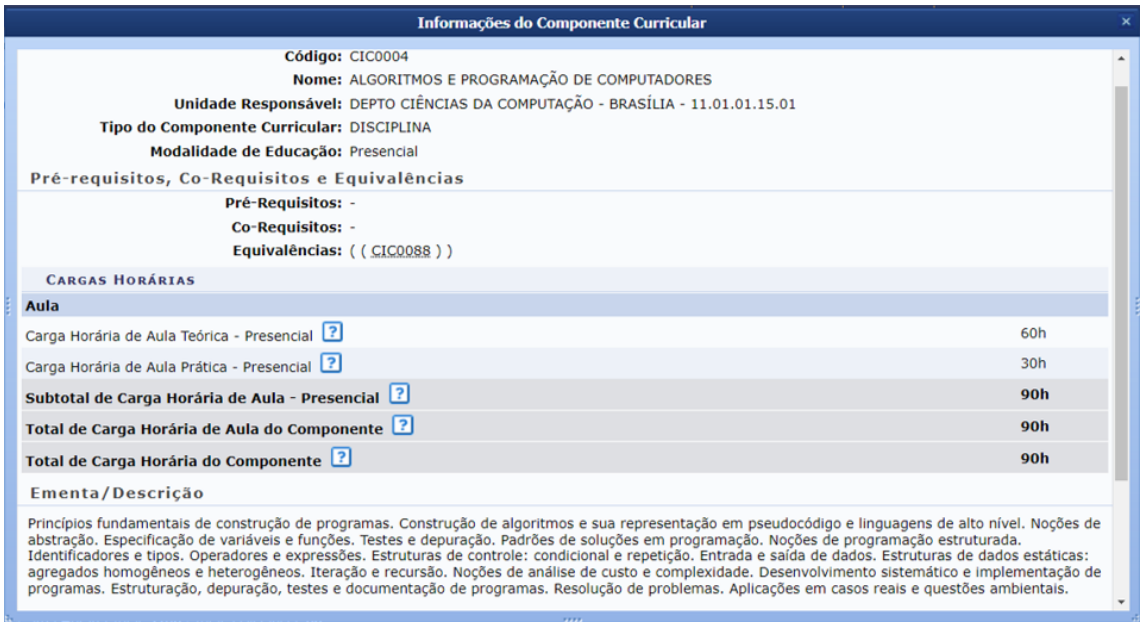

Figura 4.6: Ementa da matéria APC no SIGAA

A partir do texto da ementa da matéria APC, foram identificas e filtradas as seguintes palavras-chave: programas, algoritmos, pseudocódigos, linguagens, abstração, variáveis, funções, teste, depuração, programação estruturada, identificadores, tipos, operadores, estrutura de controle, entrada de dados, saída de dados, complexidade, desenvolvimento, implementação, programas.

Essas palavras-chave ainda podem ser utilizadas para formar uma cadeia de termos associados, levando em consideração as fontes de temas sugeridas acima, para que sejam selecionadas as palavras-chave dessa matéria. Por exemplo, pode-se associar “linguagens”,

[2] `https://sigaa.unb.br/sigaa/`

no contexto da ementa de APC, com a linguagem “python” ou “C++”, entre outros, e a ligação entre os termos seria realizada conforme o exemplo apresentado na Figura 4.7.

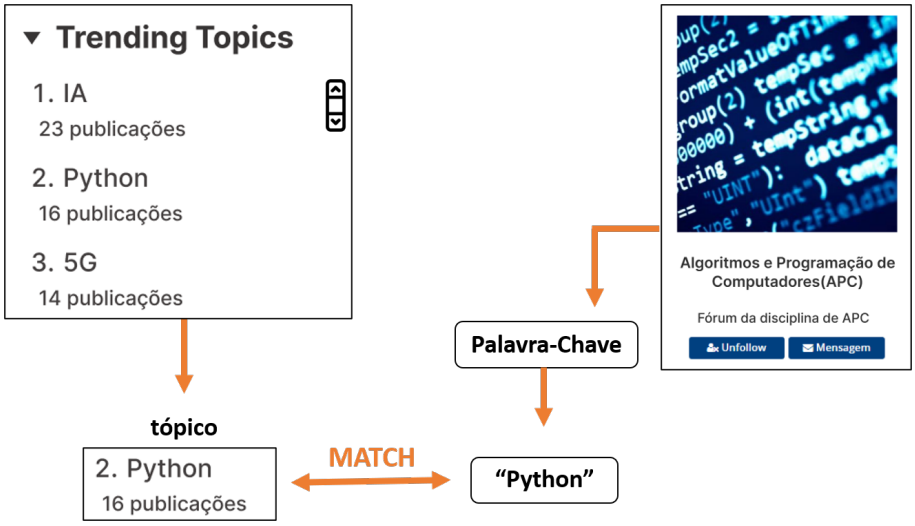

Figura 4.7: Exemplo de ligação entre um tópico e uma palavra-chave utilizando a matéria APC

Para outro exemplo, foi utilizada a disciplina de Sistemas de Informação(SI) (Figura 4.8), que a partir da ementa disponibilizada no SIGAA é possível observar os seguintes temas e assuntos abordados durante o curso: Fundamentos da teoria geral de sistemas, teoria da informação: conceito de informação, conceito de dados, representação de dados e de conhecimento sistemas de informação: fases e etapas documentação, prototipação, modelagem conceitual: abstração, modelo entidade-relacionamento, análise funcional, administração, de dados estudo de caso. Tais tópicos também ainda poderão ser destrinchados em cadeias de termos relacionados.

Figura 4.8: Ementa da matéria SI no SIGAA

A partir desses assuntos conseguimos observar que existem ligações entre os assuntos da disciplina de APC e SI, como por exemplo, as aplicações de algoritmos estudadas na primeira estão relacionadas aos Fundamentos da Teoria Geral de Sistemas (TGS), abordados em SI, pois compreender os conceitos da TGS pode ajudar a melhorar a concepção e implementação de algoritmos eficientes em sistemas computacionais. Além disso, percebe-se que as matérias usadas como exemplos compartilham de palavras-chave iguais, como no caso dos tópicos: dados, abstração e programas. Ou seja, a partir dessas relações, será possível visualizar as disciplinas que possuem contextos conectados.

Na direção de configurar essa integração, será necessário o mapeamento de palavras-chave que caracterizam uma disciplina específica, as quais deverão ser adicionadas ao fórum criado para essa matéria. Na etapa de adição dessas palavras-chave ao perfil do fórum, o administrador da conta deverá seguir as instruções apresentadas na seção 4.4 (Interfaces).

4.4 Interfaces e interações resultantes

Nesta seção, serão apresentados os protótipos do *add-on* de *trending topics* aplicado à interface do CiCFriend, conforme os casos de uso modelados. Para a prototipagem, foram criadas telas baseadas no *layout* do CiCFriend. Essas telas apresentam situações

com usuários hipotéticos, e publicações baseadas em tópicos presentes em disciplinas do CIC e que também são discutidos pela sociedade, a fim de exemplificar a usabilidade da ferramenta e as configurações necessárias para a aplicação do mapeamento de disciplinas curriculares.

Na Figura 4.9, ilustramos os primeiros passos da criação de uma conta do tipo fórum. O usuário abre as configurações, na opção “Gerenciar Contas”, e seleciona “Registre uma conta adicional”.

Na tela sequencial (4.10), são mostrados os campos de informações necessárias para a criação do fórum. Nota-se que um dos campos é a solicitação de aprovação para o admistrador do CiCFriend.

Por fim, como mostrado na Figura 4.11, é necessário concordar com os termos de uso e Política de Dados e clicar na opção “Registrar”.

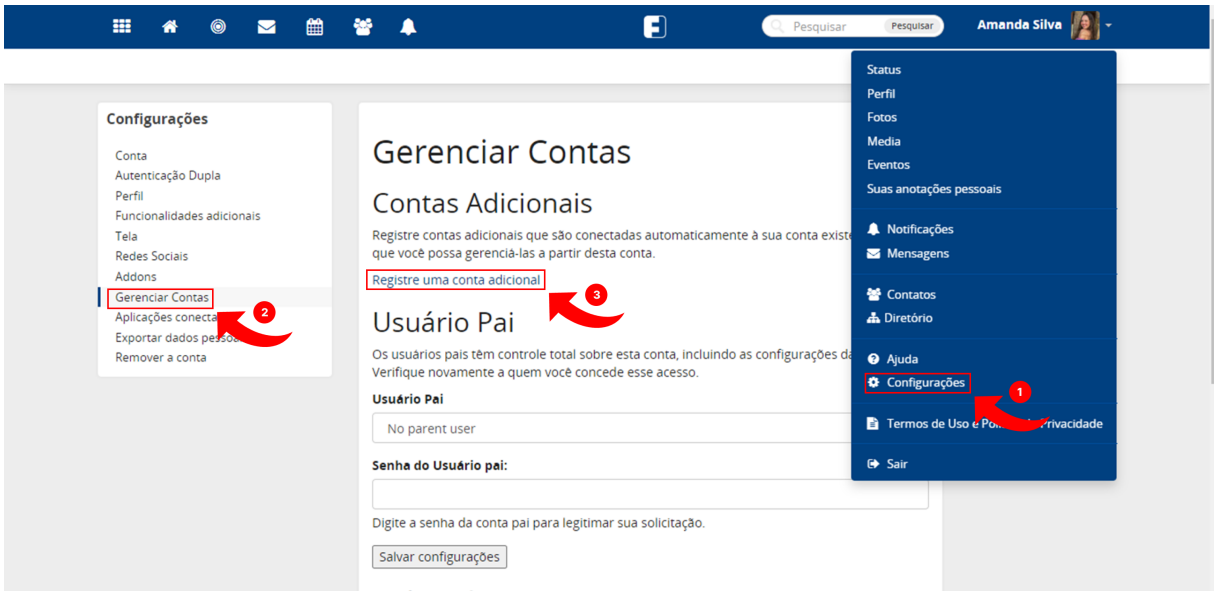

Figura 4.9: Acessando as configurações para criação de uma conta fórum

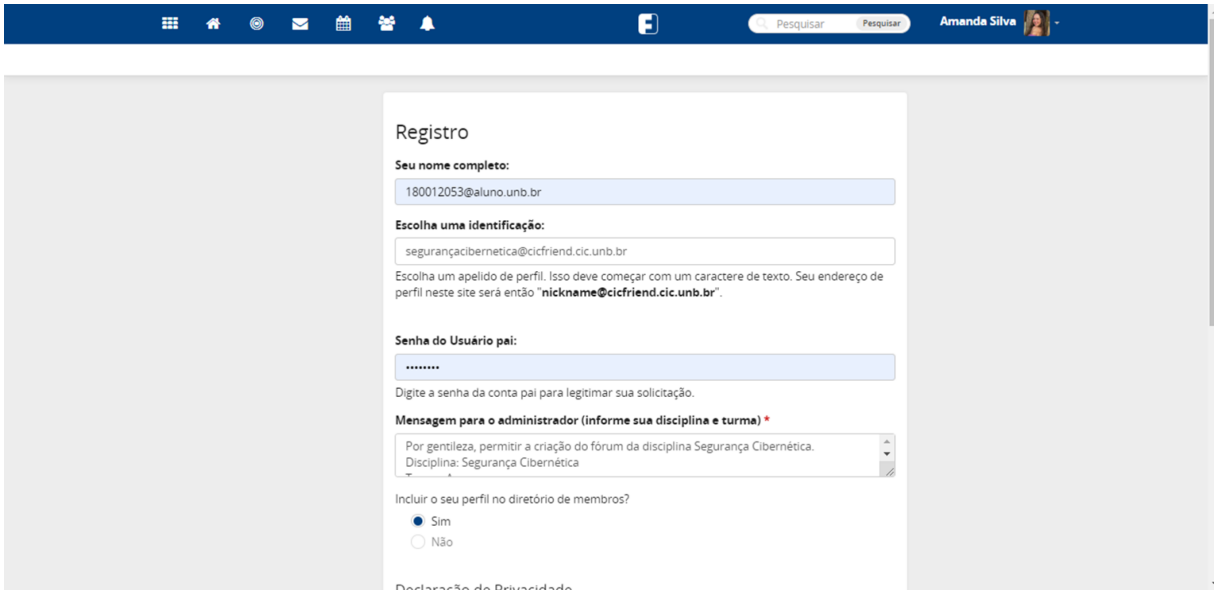

Figura 4.10: Preenchendo as informações para criação de uma conta fórum

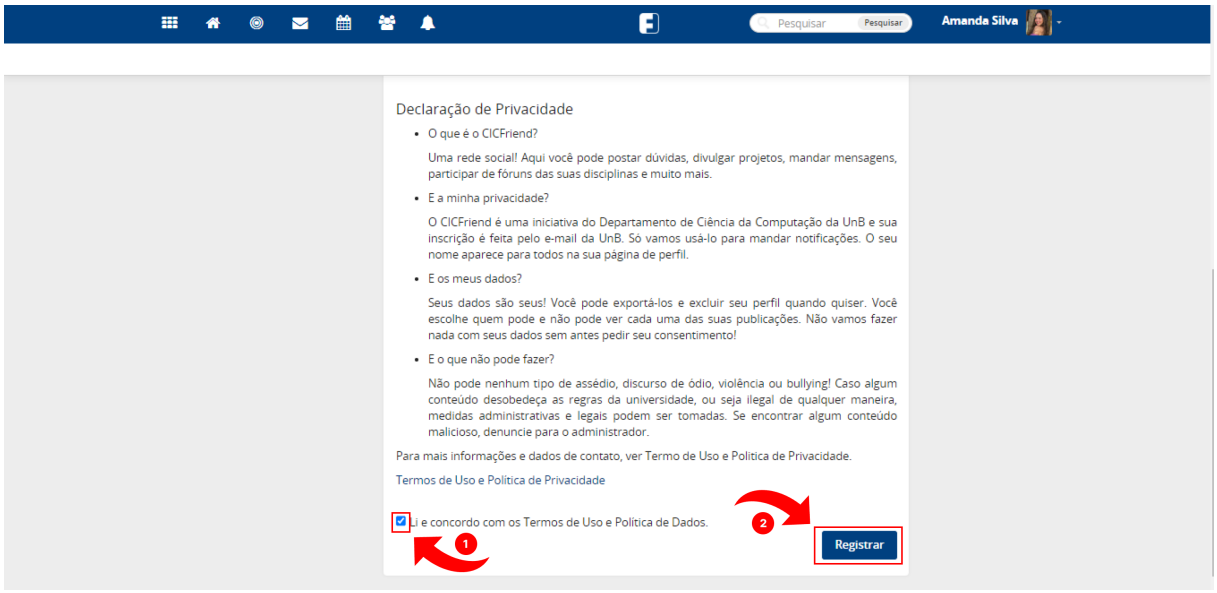

Figura 4.11: Aceitando os termos de uso e registrando a conta fórum

Para configurar as *tags* públicas do fórum da disciplina é importante que o administrador adicione palavras-chave relevantes ao tema da matéria, pois essas poderão ser utilizadas para relacionar os fóruns aos tópicos do *trending topics*. Além disso, elas também sugerem esse fórum àqueles que têm interesse nesses tópicos.

Na Figura 4.12, o administrador do fórum da disciplina acessa as configurações. Já na próxima tela (Figura 4.13), é acessada a opção "Perfil" e depois "Miscelânia", onde contém o campo "Palavras-Chave Públicas", os quais são incluídos os tópicos relacionados com o conteúdo.

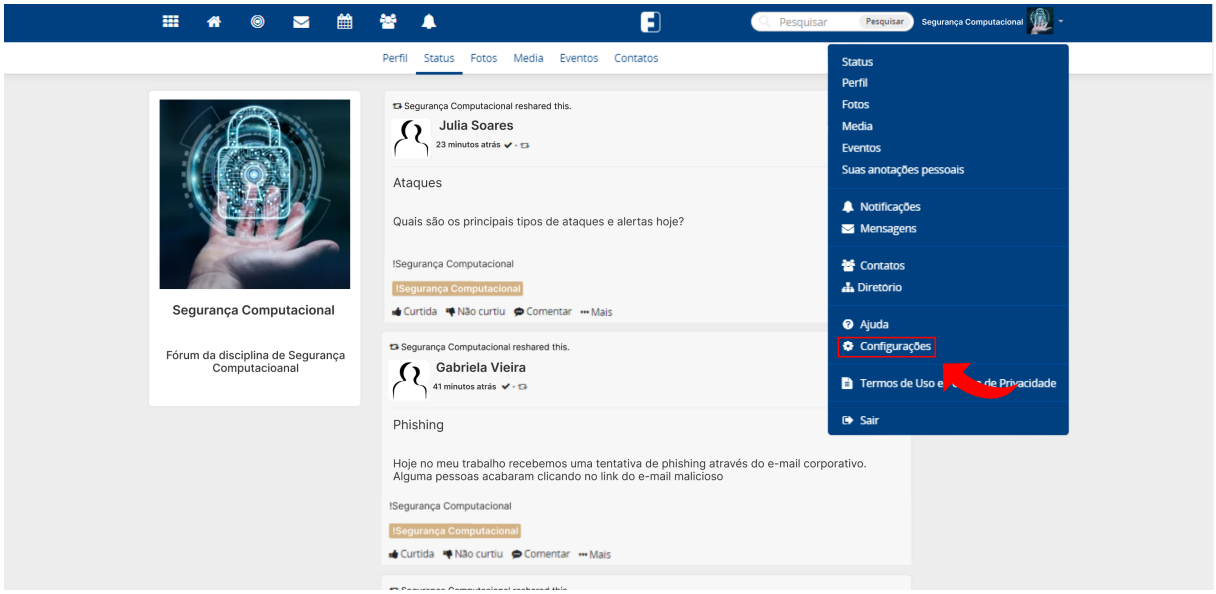

Figura 4.12: Entrando nas configurações da conta fórum

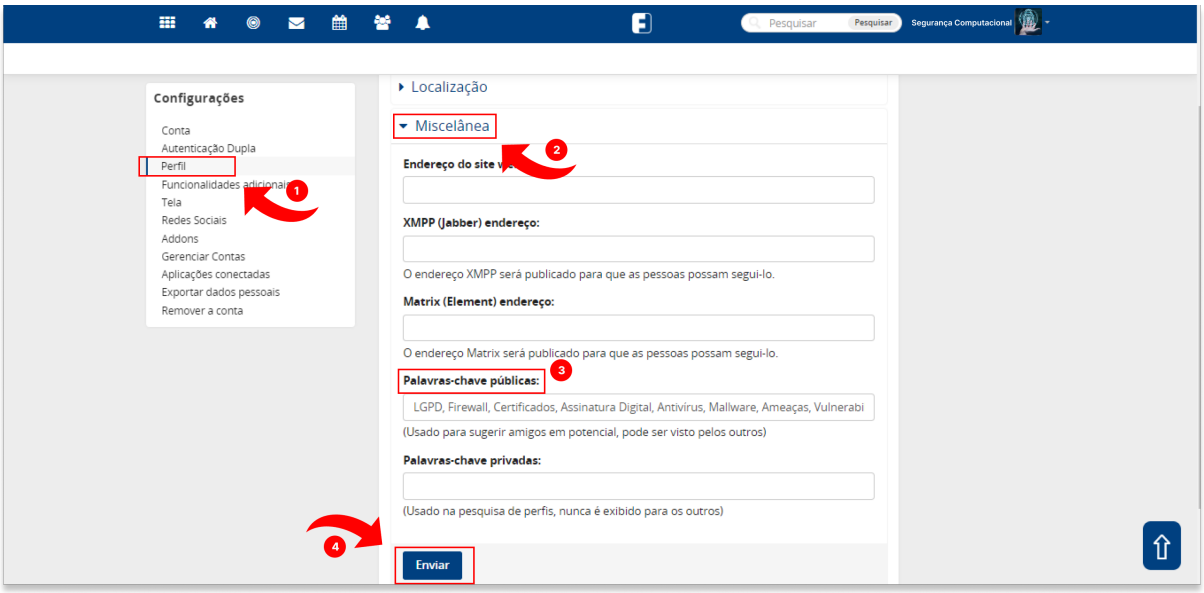

Figura 4.13: Registrando as tags no perfil da conta fórum

Na tela apresentada na Figura 4.14, é mostrado o procedimento de criação de uma publicação no CICFriend, onde o usuário seleciona o ícone indicado em "1", para que apareça o campo de criação de *post*, o qual deve ser preenchido conforme a vontade do usuário. Em seguida, basta clicar em "Compartilhar".

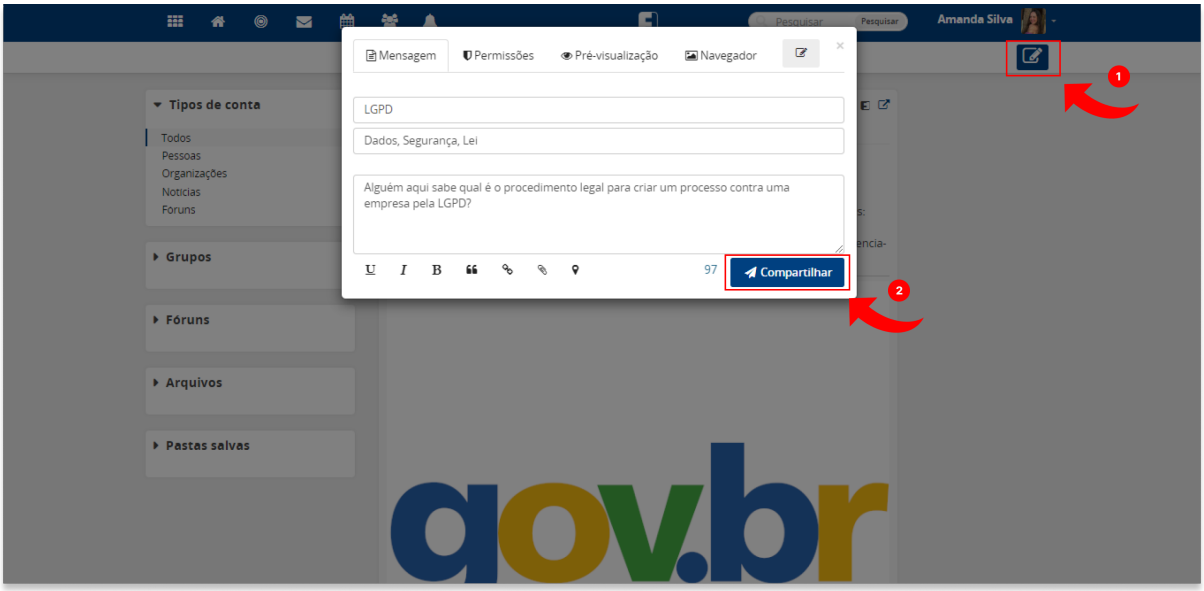

Figura 4.14: Criando uma publicação

Na aba de publicações da "Comunidade Local", acessada pelo passo a passo mostrado na tela da Figura 4.15, os *trending topics* são apresentados na barra esquerda. Eles estão dispostos de maneira ordenada, de acordo com a frequência de utilização, em uma lista de 10 itens. A quantidade de itens da lista pode ser personalizada conforme o escolhido em uma futura implementação. Na indicação "3", o usuário clica em um dos tópicos da lista, que o direcionará para a tela da Figura 4.16.

Nessa tela, é exibida a lista de *posts* que utilizaram o tópico selecionado, ordenada do mais recente ao mais antigo, além de expor os fóruns relacionados na barra esquerda. Ao clicar em um fórum da lista, será feito o redirecionamento à página do mesmo, apresentando também as postagens em que ele foi marcado (Figura 4.17).

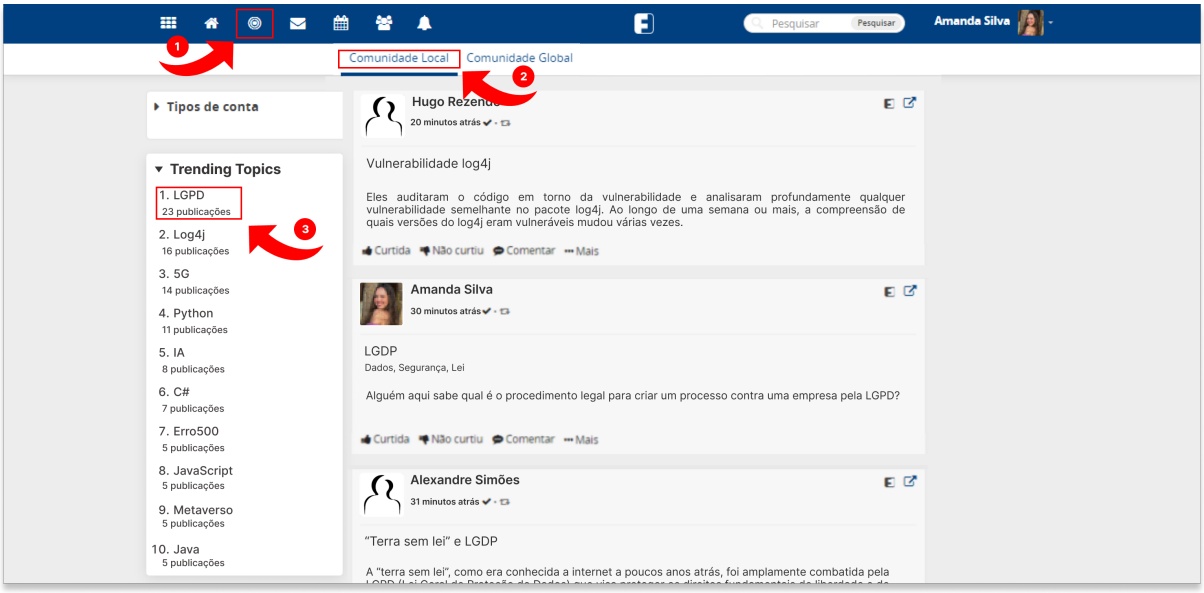

Figura 4.15: Acessando a comunidade local para visualizar os *trending topics*

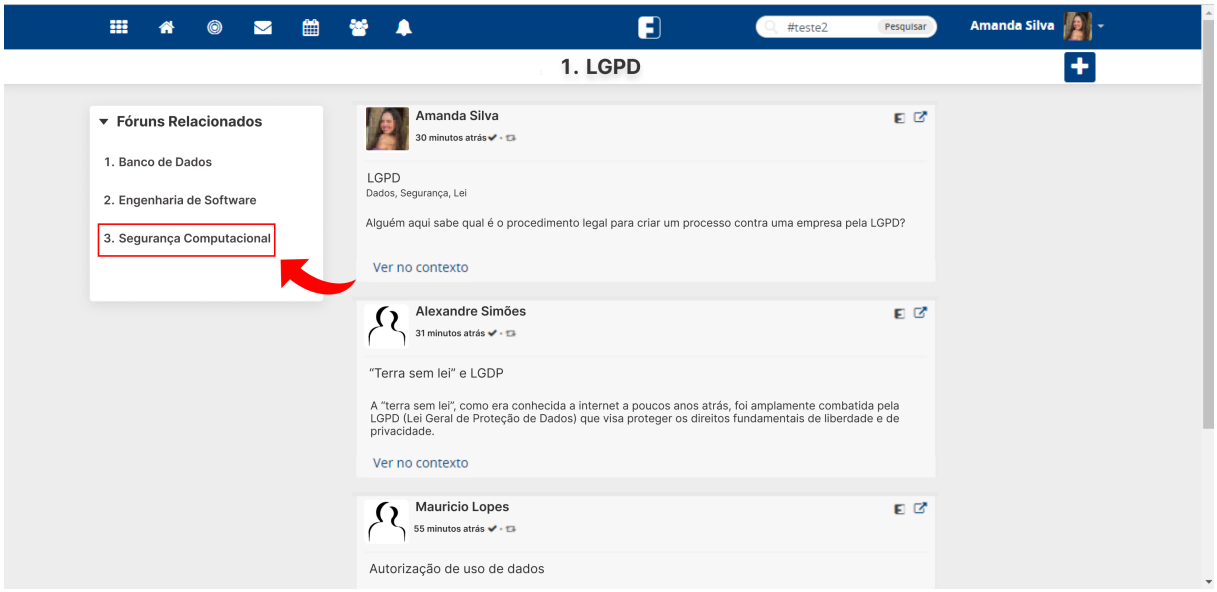

Figura 4.16: Visualizando a lista de *posts* do tópico selecionado

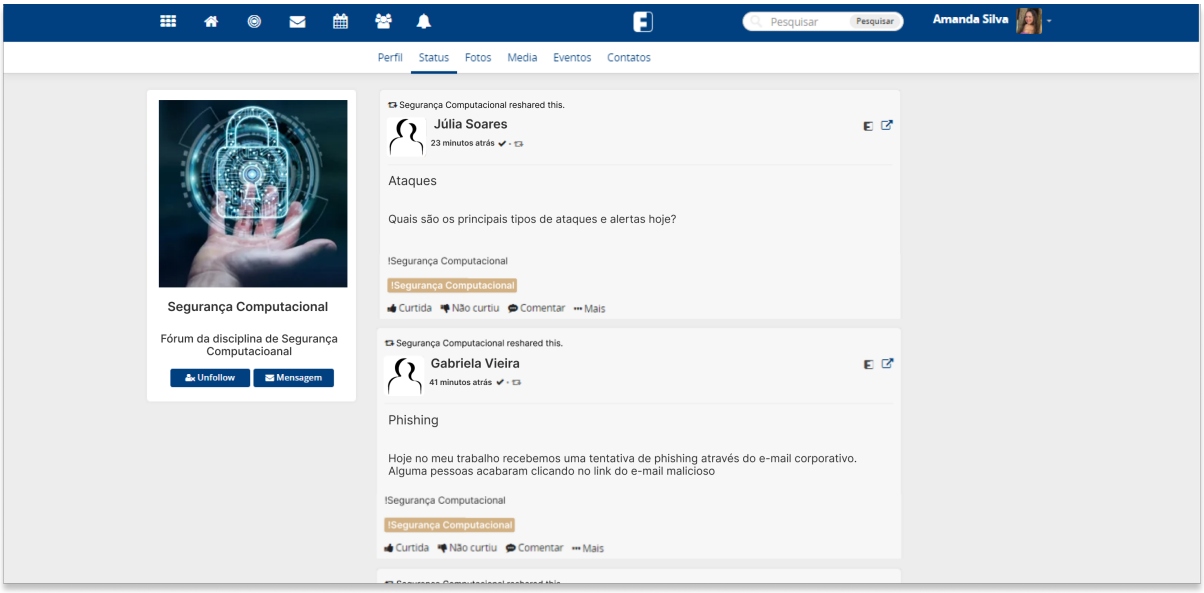

Figura 4.17: Visualizando a conta do fórum selecionado

4.5 Aplicando PLN para a implementação de *trending topics* na CICFriend

Dada a visão geral da ferramenta e os casos de uso citados anteriormente, é possível ter uma breve ideia do funcionamento da *feature* de *trending topics* por parte do usuário. Nesta seção, iremos abordar a funcionalidade na perspectiva do sistema, elucidando como as técnicas de PLN e o algoritmo selecionado se aplicam no contexto de *trending topics*. Iremos cobrir desde os passos de pré-processamento do algoritmo, a necessidade de cada passo do processo, e como o conjunto de passos resulta na geração dos *trending topics*. A Figura 4.18 detalha os passos de pré-processamento utilizados no processo de criação de *trending topics* no contexto da rede CICFriend.

Figura 4.18: *Pipeline* utilizado da geração de *trending topics*

4.5.1 Pré-processamento das publicações dos usuários

O primeiro passo é o pré-processamento das publicações. Sem o tratamento e normalização do texto, o algoritmo e seu resultado são afetados, e consequentemente, a experiência do usuário. Levando em conta que a língua portuguesa é rica em símbolos, acentos e flexões

de palavras, esse passo se faz ainda mais necessário. Iremos cobrir os passos obrigatórios e opcionais do processo.

Detecção de Linguagem

O primeiro passo consiste em detectar a linguagem utilizada pelo usuário antes de fazer qualquer processamento. Isso se faz necessário pois, como dito anteriormente, cada linguagem tem suas nuances, e para poder tratá-las, é necessário determinar a linguagem utilizada. Passos subsequentes como a remoção de *stop-words* ou *tokenização* são crucialmente afetados por essa etapa. Além de que, levando em conta que o conteúdo da publicação pode se referir a diversos assuntos, incluindo notícias internacionais, é de se esperar termos na língua inglesa, por exemplo.

Após detectar a linguagem da publicação usando uma bilbioteca como a NLTK[3], inicia-se o processo de remoção de *stop-words* para filtrar o conteúdo relevante das publicações dos usuários.

Remoção de *stop-words*

O processo de remoção de *stop-words* no contexto de *trending topics* serve para filtrar as palavras que apresentam relevância semântica de palavras estruturais como artigos, preposições, etc. Por conta da necessidade de identificar as palavras para poder separá-las, é fundamental o passo anterior de detecção da linguagem utilizada na publicação.

Esse processo se faz necessário pois, é a primeira etapa em que há uma separação de palavras possivelmente relevantes para a publicação, e consequentemente palavras que podem ser *trending topics*. Com a remoção de *stop-words* o texto se torna mais conciso e fácil de ser processado, garantindo mais *performance* no algoritmo e precisão no resultado gerado.

Essa etapa é crucial pois as etapas seguintes irão processar o conjunto de publicações, e para evitar que o algoritmo itere palavras irrelevantes para o contexto, diminuindo a *performance* e usando mais processamento, a remoção de palavras sem importância semântica é essencial. Ao final do processo é esperado ter um grupo de palavras que tenham significado para o usuário, nos deixando um passo mais perto de gerar *trending topics* para o usuário. Após a remoção de *stop-words*, entra o processo de segmentação de sentenças.

[3] `https://www.nltk.org/`

Segmentação de sentenças

Essa etapa consiste na subdivisão do texto em partes menores, em sentenças e até palavras. Esse passo é necessário pois os algoritmos de PLN operam com sentenças, e não com textos inteiros. Dessa maneira, é possível o algoritmo identificar a estrutura do texto e entender o relacionamento entre as frases.

A importância desse passo se dá ao fato de que, dividindo os textos em frases, o algoritmo consegue analisá-las de forma isolada, podendo dispor mais processamento em cada passo e gerando mais precisão nos resultados. No caso de *trending topics*, essa etapa contribui para a identificação dos relacionamentos entre as frases e diminui a quantidade de termos irrelevantes no geral.

Esse processo é uma consequência natural à remoção de *stop-words*, dado que após a remoção de artigos, preposições e outros termos estruturais, o algoritmo fica mais propenso a extrair informações relevantes da publicação, e por consequência, gerar *trending topics* que estejam alinhados com o contexto dos usuários.

Após a etapa de separação do texto em frases, vem a etapa de divisão das frases em palavras, no passo de *tokenização*.

Tokenização

Naturalmente, após a divisão do conteúdo da publicação em textos, vem a divisão em palavras para análise do algoritmo no processo chamado de *tokenização*. Com a divisão em palavras, o algoritmo pode processar melhor as informações, identificar padrões e até fazer predições.

Esse processo é crucial pois além de favorecer a limpeza e normalização do texto, além da *performance* e qualidade dos resultados, pode permitir que etapas adicionais sejam aplicadas ao processo, como correção de erros de digitação, busca de sinônimos (*lematização*), colocar as palavras em caixa baixa, etc. No contexto de *trending topics* esse processo é vital pois, as redes sociais envolvem um ambiente informal, então podem haver erros de digitação, abreviações, etc. Também é possível aplicar um algoritmo de análise de sentimento para identificar o efeito ou sensação causada por cada palavra, que pode ser relevante para a geração de *trending topics*.

Após os processos de detecção de linguagem, remoção de *stop-words*, e divisão em palavras e frases que atuam como um processo de filtragem dos termos, essa etapa do processo deve conter as palavras mais relevantes para a geração de *trending topics*. Essa etapa de pré-processamento permite tratar cada palavra de forma unitária, e por conta disso, essa é a última etapa de tratamento do texto antes de utilizar o algoritmo.

Lowercasing

Lowercasing corresponde ao passo de normalizar as palavras em caixa baixa, transformando caracteres maiúsculos e minúsculos. É um passo opcional, não afeta a remoção de palavras relevantes e nem a performance do algoritmo, porém é viável para eliminar palavras duplicadas no texto, visto que a mesma palavra em caixa alta e caixa baixa conta como duas palavras no vocabulário.

Em alguns casos pode reduzir consideravelmente o total de palavras em uma publicação, mas por se tratar de um passo extra, adiciona mais processamento na *pipeline* e pode não gerar resultados notáveis. Por conta disso, é um passo opcional e vem por último na etapa de pré-processamento.

4.5.2 Aplicação do algoritmo TF-IDF

Após a etapa de tratamento e pré-processamento do texto, o conteúdo da publicação está pronto para ser passado ao algoritmo para a geração de *trending topics*.

Conforme explicado anteriormente, o modelo TF-IDF compara a frequência de cada palavra no decorrer de uma frase ou texto com sua frequência no decorrer de todo o conjunto de publicações. No caso, o algoritmo pode ser aplicado para uma única publicação, onde irá comparar a frequência dos termos em uma frase com a frequência no mesmo texto e retornar os termos mais relevantes da publicação de um usuário, e também pode ser aplicado para um grupo de publicações, onde aplica o processo de forma recursiva e após retornar os termos mais relevantes de uma publicação, compara a frequência dos mesmos no decorrer das outras publicações.

Como o modelo TF-IDF é a etapa final do processo, ele é executado após a filtragem de *stop-words*, normalização do texto e lematização. O produto final do algoritmo TF-IDF é o valor númerico das palavras, correspondendo à relevância de cada palavra de forma decrescente, se adequando ao formato de *trending topics*. Posteriormente, dada a relevância de cada palavra e formação de contexto após processar o conteúdo das publicações, é possível aplicar modelos de *machine learning* para correlacionar os *trending topics* das publicações do CICFriend com disciplinas do CIC, como por exemplo o algoritmo *word2vec* [55]. Dessa forma, é possível formar uma base de publicações relacionadas que apontem para as mesmas disciplinas, fomentando discussões extracurriculares sobre os conteúdos, notícias e experiências relacionadas à universidade.

O algoritmo foi testado em uma versão de demonstração na plataforma CICFriend. O conteúdo das publicações utilizadas foi retirado de notícias na internet, para fins de teste. A Figura 4.19 exibe os *trending topics* em uma aplicação de demonstração[4] com

[4] `https://friendica-demo.vercel.app/`

funcionamento similar ao CICFriend. As publicações utilizadas envolvem falhas de segurança no aplicativo Nubank[5], e a Figura 4.20 apresenta os *trending topics* de publicações envolvendo a ferramenta de inteligência artificial *ChatGPT* [6].

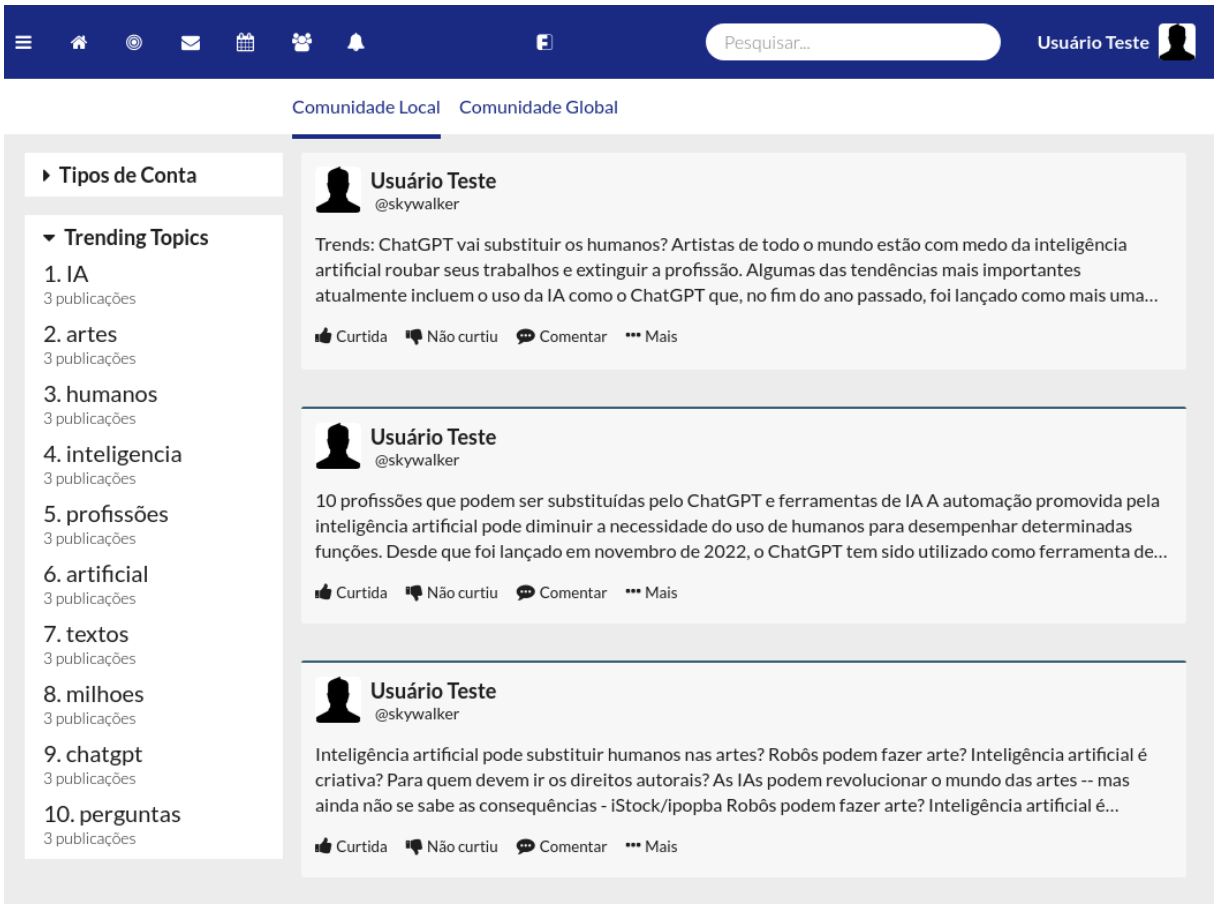

Figura 4.19: *Trending topics* em versão *demo* do CICFriend sobre a ferramenta ChatGPT

[5] `https://tecnoblog.net/noticias/2022/05/19/nubank%2Dtinha%2Dfalha%2Dde%2Dseguranca%2Dque%2Dfacilitava%2Droubo%2Dde%2Ddinheiro%2Dusando%2Do%2Dgmail/`

[6] `https://revistapegn.globo.com/tecnologia/noticia/2023/02/10%2Dprofissoes%2Dque%2Dpodem%2Dser%2Dsubstituidas%2Dpelo%2Dchatgpt%2De%2Dferramentas%2Dde%2Dia.ghtml`

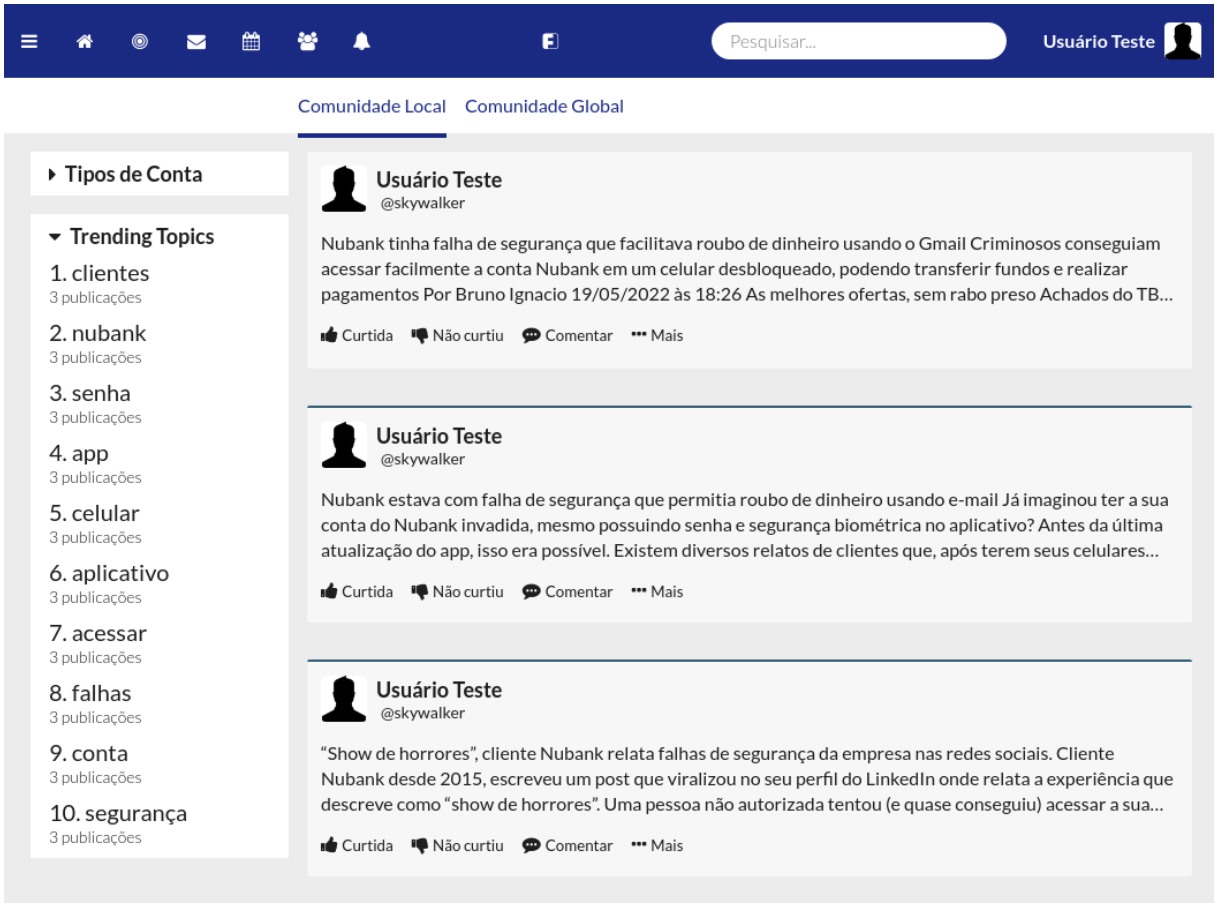

Figura 4.20: *Trending topics* em versão *demo* do CICFriend sobre uma falha de segurança do Nubank

4.6 Considerações Finais

Neste capítulo, foi apresentada a proposta de implementação de um *add-on* de geração de *trending topics* para o CICFriend. A partir da pesquisa de trabalhos relacionados, foi possível evidenciar a importância dos *trending topics* na promoção de discussões relevantes e no acesso à informações em redes sociais. Acredita-se que a implementação da ferramenta proposta irá gerar benefícios significativos para a comunidade do CIC, permitindo a identificação mais rápida e eficiente de tópicos relevantes e a promoção da interação dos ambientes formais e informais do contexto acadêmico.

Primeiramente, foram desenvolvidos protótipos de tela para a criação do modelo das funcionalidades possíveis a partir dos *trending topics*. Em seguida, foi descrita a arquitetura proposta para a ferramenta, que inclui a utilização de algoritmos de processamento de linguagem natural e aprendizado de máquina para identificar e classificar os tópicos mais relevantes na rede social acadêmica. Ainda que não tenha sido implantada, foi im-

plementada uma aplicação de demonstração de geração de trending topics conforme a arquitetura sugerida no trabalho.

Capítulo 5

Aplicação de demonstração

Com o intuito de fornecer uma visualização e experimentação da funcionalidade de *trending topics*, foi implementada uma aplicação clone do Friendica, focando na visualização e geração automática de *trending topics*, e criação de postagens dos usuários. O objetivo deste capítulo é explicar a arquitetura utilizada para a construção da aplicação de demonstração. A aplicação é de uso livre e está acessível na internet [1].

5.1 Arquitetura da aplicação

A aplicação segue o modelo *client-server* (cliente-servidor), que é o modelo comum para aplicações da *web*[2]. Essa estrutura consiste na comunicação de diversas máquinas (clientes) com uma aplicação central (servidor), onde o cliente tem acesso a uma interface de usuário (UI) para utilizar a aplicação e ter interação (envio e recebimento de dados) com o servidor, enquanto que o servidor é a parte que processa as solicitações do cliente. A Figura 5.1 exibe um diagrama de uma estrutura cliente-servidor padrão, com um cliente acessando a aplicação pelo computador e se comunicando com o servidor hospedado na nuvem.

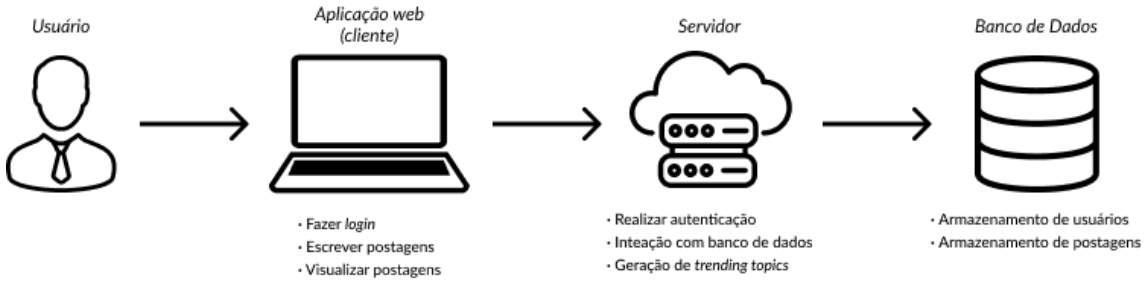

Figura 5.1: Diagrama de arquitetura cliente-servidor para a aplicação *demo*

[1] `https://friendica-demo.vercel.app/`
[2] `https://pt.wikipedia.org/wiki/Modelo_cliente%E2%80%93servidor`

O cliente não tem acesso direto ao servidor e suas funcionalidades, apenas à aplicação, que é executada em um navegador *web*, enviando solicitações e recebendo respostas ao servidor por requisições *HTTP*. Aplicações da *web* são construída utilizando linguagens de marcação (HTML), estilização (CSS) e linguagens de *script* (JavaScript), e as três em conjunto formam a página *web* que o usuário tem acesso. No diagrama da Figura 5.1 o cliente pode fazer login, escrever e visualizar postagens, e todas essas operações dependem de interação com o servidor.

O servidor é a parte da aplicação *web* que processa as solicitações do cliente, cuidando de receber as requisições do cliente, tratar os dados e enviar uma resposta. Ele é responsável por gerenciar a lógica do negócio, recuperar e armazenar dados em um banco de dados, e devolver uma resposta apropriada ao cliente. O servidor é construído em uma linguagem de programação do lado do servidor, como Node.js, Ruby on Rails, Django ou Java. No diagrama da Figura 5.1 o servidor recebe as requisições do cliente, realiza autenticação dos dados para permitir o *login*, se comunica com o banco de dados para gerenciar postagens, e cuida de gerar *trending topics* baseado nas publicações, que seria uma regra de negócio da aplicação.

5.2 Implementação da aplicação de demonstração

Levando em consideração o referencial tecnológico apresentado no capítulo 3, temos um conjunto de ferramentas modernas para implementar um sistema cliente-servidor para a aplicação de demonstração, levando em conta a geração de *trending topics*, interação de usuários, e outras funcionalidades citadas previamente.

5.2.1 Aplicação do cliente

Para a implementação do cliente (interface visual do usuário), foi utilizado o *framework NextJS*, para a construção de componentes reutilizáveis em tela e permitir atualizar a interface de acordo com as interações do usuário. No caso, quando o usuário faz uma nova postagem, a lista de postagens é atualizada com a nova publicação e a lista de *trending topics* é atualizada após processar a nova postagem e gerar novos *trending topics* usando as palavras da publicação.

Dentre as funcionalidades possíveis, o usuário pode visualizar publicações prévias (incluindo as feitas por outros usuários), visualizar os *trending topics* e se cadastrasr na aplicação. Para fazer login o usuário deve utilizar uma conta do Google, e a partir disso é permitido que ele faça publicações. Para a visualização de publicações prévias ou dos *trending topics*, não é necessário que o usuário esteja logado.

5.2.2 Aplicação do servidor

Para o lado do servidor, foi utilizada a função de *API Routes* do *framework NextJS*, sendo a parte responsável pela conexão com o banco de dados para buscar as publicações e *trending topics* para preencher a interface do usuário (lado do cliente), além de buscar os dados do usuário no login ou cadastrar um novo usuário. As publicações ficam salvas no banco de dados *serverless* FaunaDB, contendo apenas três tabelas relacionais: tabela de usuários, tabela de postagens e tabela de *trending topics*.

Por se tratar do mesmo *framework* para os lados do cliente e servidor, ambas as funcionalidades estão na mesma aplicação, logo, ela pode ser hospedada em um só ambiente. A aplicação de demonstração está hospedada na plataforma *Vercel*, disponibilizada na internet para que qualquer pessoa acesse, enquanto que o banco de dados fica hospedado na plataforma do próprio FaunaDB.

Para a geração de *trending topics*, foi utilizado uma implementação do TF-IDF em Javascript, retirada de um repositório [3] no GitHub. A etapa de processamento das publicações e geração de *trending topics* ocorre no lado do servidor, que é onde foi utilizado o módulo de TF-IDF citado.

[3] `https://github.com/techfort/mimir`

Capítulo 6

Conclusão

Este capítulo tem como objetivo apresentar as conclusões deste trabalho e listar possíveis trabalhos futuros.

6.1 Objetivos Alcançados

As redes sociais transformaram a forma com que as pessoas interagem, e por conta disso se mostraram um meio eficaz de criar e estabelecer conexões entre seus usuários. Apesar de ainda não estarem tão estabelecidas quanto às redes centralizadas, as RSDs também trazem vantagens valiosas e benefícios quando aplicadas no contexto acadêmico. A adoção de ambientes virtuais voltados à aprendizagem pelas instituições de ensino pode ser um indicativo da integração do ensino formal e informal, que foram definidos em anteriormente em [12], onde também são discutidas estratégias de estímulo para incorporação de aspectos relativos à aprendizagem informal na graduação dos estudantes.

Voltando aos objetivos deste trabalho, entende-se que os mesmos foram contemplados de maneira satisfatória. Com a proposta de expansão das funcionalidades do CICFriend, a partir de uma implementação de uma ferramenta de geração de *trending topics*, foi realizada a contribuição para a criação de um meio de integração do ambiente de ensino formal com o informal do departamento CIC, onde os alunos e professores poderão, a partir dos *trending topics*, incorporar os assuntos em tendência e os assuntos que estão sendo mais comentados pelos usuários aos estudos em sala de aula. Essa integração também será realizada através da conexão entre os tópicos em tendência e os fóruns das disciplinas.

Remetendo aos outros objetivos também citados na seção 1.3, foram realizados os protótipos de tela do *add-on* de *trending topics*, a fim de demonstrar as possibilidades de funcionalidades existentes a partir dele, e apresentar a ideia inicial da ferramenta. Além

disso, foi realizado um protótipo funcional não integrado ao CICFriend, que realiza a geração de *trending topics* a partir de postagens.

Apesar de não ser o foco deste trabalho, e possuir teor motivacional e justificativo, foram apresentados estudos sobre os vários benefícios do uso de redes sociais no ambiente educacional, além de também ter sido mostrado os diferenciais que as redes sociais descentralizadas apresentam em relação às centralizadas, mais especificamente as funcionalidades do Friendica, como a arquitetura descentralizada, sem autoridade central ou propriedade sobre os dados dos usuários, privacidade e interoperabilidade. Entrando no contexto dos *trending topics*, foram apresentados diversos estudos do uso dessa ferramenta, e apesar de possuir aplicações nas áreas *marketing*, finanças, jornalismo, entre outros, o levantamento de trabalhos relacionados apresentou uma lacuna na literatura em relação à exploração desta ferramenta no campo educacional.

6.2 Trabalhos Futuros

Devido ao fato de ter sido teorizado, existem possibilidades de continuações viáveis e úteis para trabalhos futuros. Primeiramente, uma possível direção é a implementação completa do *add-on* de *trending topics* e implantação em versão beta integrado ao CICFriend, a fim de avaliar a usabilidade e o impacto dessa ferramenta na interação entre os usuários e na disseminação de informações relevantes para a comunidade acadêmica. Outra contribuição seria a criação de um ambiente para testes beta. Ainda, uma pesquisa de levantamento inicial de percepção de usuário pode ser feita com a comunidade para avaliação de sugestões e melhorias em relação à feature. Além disso, é importante considerar aspectos como a privacidade e a segurança dos dados dos usuários, bem como a possibilidade de vieses ou manipulações nos temas em destaque.

Tendo em vista os estudos apresentados neste trabalho, que relatam o uso das redes sociais em diversas áreas e para diferentes funcionalidades, existem também formas de conectar publicações de RSOs ao contexto universitário, a fim de expandir a barreira do conhecimento e tornar o processo de ensino-aprendizagem mais holístico. Internamente ao projeto, um possível avanço para ser realizado é a aprimoração da integração dos tópicos dos *trending topics* com os fóruns das disciplinas. A proposta de usar palavras-chave das ementas e dos documentos das matérias está aberta para automações, com o intuito de facilitar e acelerar esse processo de integração dos termos com os fóruns das disciplinas. Essa integração não está restrita à semântica de palavras-chave, podendo ser personalizada a partir de outros mapeamentos entre os ambientes formais (disciplinas) e informais (*trending topics*).

Por fim, seria interessante investigar a possibilidade de personalização da lista de *trending topics*, com base nas áreas de interesse de cada usuário, a fim de aumentar ainda mais a relevância e o valor dessa funcionalidade para a comunidade acadêmica. Ao permitir que os usuários possam selecionar e acompanhar temas específicos em suas áreas de interesse, a plataforma pode proporcionar uma melhor experiência, visando maximizar o valor e a utilidade da rede social para o CIC.

Referências

[1] Capobianco, Ligia: *Comunicação e literacia digital na internet: estudo etnográfico e análise exploratória de dados do programa de inclusão digital acessasp - ponline.* Escola de Comunicações e Artes, Universidade de São Paulo, São Paulo., 2010. 2

[2] Maia, B. R., Dias P. C.: *Ansiedade, depressão e estresse em estudantes universitários: o impacto da covid-19.* Estudos de Psicologia (Campinas), 2020. http://dx.doi.org/10.1590/1982-0275202037e200067. 2

[3] Sommerville, I.: *Engenharia de software.* Pearson Prentice Hall, 9ª edição, 2011, ISBN 9788579361081. https://books.google.com.br/books?id=H4u5ygAACAAJ. 2

[4] Zhang, Xiaoming, Xiaoming Chen, Yan Chen, Senzhang Wang, Zhoujun Li e Jiali Xia: *Event detection and popularity prediction in microblogging.* Neurocomputing, 149:1469–1480, 2015, ISSN 0925-2312. https://www.sciencedirect.com/science/article/pii/S0925231214010893. 3

[5] Economic, Quantifying the e Cultural Biases of Social Media through Trending Topics.: *Early identification of personalized trending topics in microblogging.* 2015. https://doi.org/10.1609/icwsm.v11i1.14943. 3, 14

[6] Kwak, Haewoon, Changhyun Lee, Hosung Park e Sue Moon: *What is twitter, a social network or a news media?* Association for Computing Machinery, 2010, ISBN 9781605587998. https://doi.org/10.1145/1772690.1772751. 3

[7] Johan Bollen, Huina Mao, Xiao Jun Zeng: *Twitter mood predicts the stock market.* Journal of Computational Science, 2011. 3

[8] Nóbrega, Germana e Fernando Cruz: *Rumo a um ecossistema educacional apoiado por computador e socialização em rede descentralizada.* 2021. https://sol.sbc.org.br/index.php/sbsc_estendido/article/view/16033. 3, 10, 17

[9] Silva Oliveira, Jéssica da: *Rede social descentralizada em contexto acadêmico: caracterização e potencialidades.* 2021. https://bdm.unb.br/handle/10483/28307. 3, 4, 10, 17, 60

[10] Davi Martins Torres, Gabriel de Oliveira Estevam: *Implantação da rede social descentralizada cicfriend e levantamento inicial de percepção de usuária(o) discente.* https://bdm.unb.br/bitstream/10483/31175/1/2022_DaviTorres_GabrielEstevam_tcc.pdf. 3, 17

[11] Nóbrega, Germana, Gabriel T. da Silva Thiago V. R. Silva: *Um projeto estruturante para orientações de tcc em cursos de computação: que oportunidades para ihc?* 2022. https://sol.sbc.org.br/index.php/weihc/article/view/22854. 3

[12] Egler, Pedro Henrique Pires: *Agregando a aprendizagem informal à formal badges digitais para participação discente nas conversas em rede social descentralizada sobre disciplinas.* https://bdm.unb.br/bitstream/10483/31237/1/2021_PedroHenriquePiresEgler_tcc.pdf. 4, 17, 22, 79

[13] Leila Christina Dias, Rogério Leandro Lima da Silveira: *Redes, sociedades e territórios.* 3ª edição, 2021. http://hdl.handle.net/11624/3125. 7, 8

[14] Paul, Thomas, Sonja Buchegger e Thorsten Strufe: *Decentralized Social Networking Services.* 2011, ISBN 978-88-470-1817-4. http://10.1007/978-88-470-1818-1_14. 7

[15] Jiang, Le e Xinglin Zhang: *Bcosn: A blockchain-based decentralized online social network.* IEEE Transactions on Computational Social Systems, 2019. 7

[16] Rejeb, Abderahman, Karim Rejeb, Alireza Abdollahi e Horst Treiblmaier: *The big picture on instagram research: Insights from a bibliometric analysis.* Telematics and Informatics, 73:101876, 2022, ISSN 0736-5853. https://www.sciencedirect.com/science/article/pii/S0736585322001095. 8

[17] Daroda, Raquel Ferreira: *As novas tecnologias e o espaço público da cidade contemporânea.* Tese de Mestrado, Universidade Federal Do Rio Grande Do Sul, 2012. http://hdl.handle.net/10183/67063. 8

[18] Simon, K: *DataReportal (2023), Digital 2023 Global Overview Report.* 2023. https://datareportal.com/reports/digital-2023-global-overview-report. 8

[19] Amaral, Inês: *Redes Sociais na Internet: Sociabilidades Emergentes.* dezembro 2016, ISBN 9789896543525. 8

[20] Chawinga, Winner: *Taking social media to a university classroom: teaching and learning using twitter and blogs.* International Journal of Educational Technology in Higher Education, 14, dezembro 2017. 9

[21] Almulla, Mohammed Abdullatif: *Social media use for educational purposes: Systematic literature review in higher education of middle east countries (mec).* International Journal of Advanced Trends in Computer Science and Engineering, 2020. 9, 10

[22] Greenhow, C. e S. Galvin: *Teaching with social media: evidence-based strategies for making remote higher education less remote.* Information and Learning Sciences, 2020. 9

[23] Manca, Stefania: *Snapping, pinning, liking or texting: Investigating social media in higher education beyond facebook.* The Internet and Higher Education, 44:100707, 2020, ISSN 1096-7516. https://www.sciencedirect.com/science/article/pii/S1096751619304257. 9

[24] Davis III, C.H.F., Deil Amen R. Rios Aguilar C. amp; González Canché: *M.s.social media and higher education: A literature review and research directions.* 2020. https://www.academia.edu/1220569/Social_Media_in_Higher_Education_A_Literature_Review_and_Research_Directions. 9

[25] Olubunmi, Funmilola e Olugboyega Salami: *Use of social media for knowledge sharing among students.* Asian Journal of Information Science and Technology, 8, agosto 2018. 9

[26] Ben Dyson, Yanhua Shen, Wen Xiong: *How cooperative learning is conceptualized and implemented in chinese physical education: A systematic review of literature.* 2020. https://orcid.org/0000-0002-7165-6598. 9

[27] Chelly, Magda e Hana Mataillet: *Social media and the impact on education: Social media and home education.* Em *2012 International Conference on E-Learning and E-Technologies in Education (ICEEE)*, 2012. 10

[28] Luciana Oliveira, Alvaro Figueira: *Analysing relevant interactions by bridging facebook and moodle.* IATED, 2016, ISBN 978-84-608-5617-7. https://doi.org/10.21125/inted.2016.0971. 10

[29] MEC: *Diretrizes curriculares nacionais para os cursos de graduação na área da computação.* 2016. http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=52101-rces005-16-pdf&category_slug=novembro-2016-pdf&Itemid=30192. 10, 11

[30] Calsavara, Alcides, Ana Paula Gonçalves Serra, Zampirolli, Francisco de Assis, Leandro Silva Galvão de Carvalho, Miguel Jonathan e Ronaldo Celso Messias. Correia: *Método baseado nos referenciais de formação da sbc para reestruturação de descritivos de disciplinas de ciência da computação em conformidade com as dcn de 2016.* Workshop sobre Educação em Computação (WEI), 2018. 11

[31] Zorzo, A. F., Nunes D. Matos E. Steinmacher I. Leite J. Araujo R. M. Correia R. Martins S: *Referenciais de Formação para os Cursos de Graduação em Computação.* Sociedade Brasileira de Computação (SBC), 2017, ISBN 978-85-7669-424-3. https://sol.sbc.org.br/livros/index.php/sbc/catalog/book/63. 11, 12

[32] Liang Wu, Xia Hu, Huan Liu: *Early identification of personalized trending topics in microblogging.* Proceedings of the International AAAI Conference on Web and Social Media, 2017. 12, 13, 14

[33] Annamoradnejad, Issa e Jafar Habibi: *A comprehensive analysis of twitter trending topics.* Em *2019 5th International Conference on Web Research (ICWR)*, páginas 22–27, 2019. 12

[34] Hachaj, Tomasz e Marek R. Ogiela: *Clustering of trending topics in microblogging posts: A graph-based approach.* Future Generation Computer Systems, 67:297–304, 2017, ISSN 0167-739X. https://www.sciencedirect.com/science/article/pii/S0167739X16300863. 12

[35] Hanna, K., D. Swerdloff e C. Welliver: *407 google search trends for topics in men's health.* The Journal of Sexual Medicine, 17(1, Supplement 1):S118, 2020, ISSN 1743-6095. https://www.sciencedirect.com/science/article/pii/S1743609519317709, 20th Annual Fall Scientific Meeting of SMSNA. 12

[36] Miao, Zhongchen, Kai Chen, Yi Fang, Jianhua He, Yi Zhou, Wenjun Zhang e Hongyuan Zha: *Cost-effective online trending topic detection and popularity prediction in microblogging.* ACM Trans. Inf. Syst., 35(3), dec 2016, ISSN 1046-8188. https://doi.org/10.1145/3001833. 13

[37] Hamadeh, Moutaz e Sherief Abdallah: *Discover Trending Topics of Interest to Governments.* 2018, ISBN 978-3-319-64860-6. https://10.1007/978-3-319-64861-3_34. 14

[38] Yang, Tian e Yilang Peng: *The importance of trending topics in the gatekeeping of social media news engagement: A natural experiment on weibo.* Communication Research, 49:009365022093372, junho 2020. 14

[39] Paul, Thomas, Antonino Famulari e Thorsten Strufe: *A survey on decentralized online social networks.* Comput. Netw., 75(PA):437–452, dec 2014, ISSN 1389-1286. https://doi.org/10.1016/j.comnet.2014.10.005. 15, 16

[40] Baran, Paul: *On distributed communications: I. introduction to distributed communications networks.* 1964. 15

[41] Myers West, S.: *Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms.* 2018. https://doi.org/10.1177/1461444818773059. 15

[42] Maathuis, Clara e Iddo Kerkhof: *Social media manipulation awareness through deep learning based disinformation generation.* International Conference on Cyber Warfare and Security, 18:227–236, fevereiro 2023. 16

[43] Santos FALCONI, Talita Gouvea de Oliveira SOBREIRO Clarissa Manzano dos: *Liberdade de expressão na era das notícias falsas e manipuladas de conteúdo político-eleitoral.* 2018. http://intertemas.toledoprudente.edu.br/index.php/ETIC/article/view/7014. 16

[44] Falch, Morten, Anders Henten, Reza Tadayoni e Iwona Windekilde: *Business models in social networking.* janeiro 2009. 16

[45] Guha, Saikat, Kevin D. Tang e Paul Francis: *Noyb: privacy in online social networks.* Em *Workshop on Online Social Networks*, 2008. 16

[46] Nagulendra, Sayooran e Julita Vassileva: *Minimizing social data overload through interest-based stream filtering in a p2p social network.* Em *2013 International Conference on Social Computing*, páginas 878–881, 2013. 16

[47] Bamman, David, Brendan O'Connor e Noah Smith: *Censorship and deletion practices in chinese social media.* First Monday, 2012. 16

[48] Lobato, Fábio MF, Gleyce C de Sousa e Antonio FL Jacob Jr: *Brasnam em perspectiva: uma análise da sua trajetória até os 10 anos de existência.* Em *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, páginas 217–228. SBC, 2021. 18, 19

[49] Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta Harshit Surana: *Practical Natural Language Processing - A Comprehensive Guide to Building Real-World NLP Systems.* junho 2020, ISBN 978-1-492-05405-4. https://www.oreilly.com/library/view/practical-natural-language/9781492054047/. 28, 31, 32, 34, 36, 37, 39, 42

[50] Kouloumpis, Efthymios, Theresa Wilson e Johanna Moore: *Twitter sentiment analysis: The good the bad and the omg!* janeiro 2011. 29

[51] Braun, Peter, Alfredo Cuzzocrea, Carson K. Leung, Adam G.M. Pazdor e Kimberly Tran: *Knowledge discovery from social graph data.* Procedia Computer Science, 96:682–691, 2016, ISSN 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050916320610, Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 20th International Conference KES-2016. 29

[52] Mikolov, Tomas, Kai Chen, Greg Corrado e Jeffrey Dean: *Efficient estimation of word representations in vector space*, 2013. 29, 44

[53] Géron, Aurélien: *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* setembro 2019, ISBN 978-1-492-03264-9. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/. 32, 35

[54] Patil, R.S., Kolhe S.R.: *Supervised classifiers with tf-idf features for sentiment analysis of marathi tweets.* Social Network Analysis and Mining, 12, 2022. https://link.springer.com/article/10.1007/s13278-022-00877-w. 39

[55] Google: *Word2vec pre-trained model.* https://code.google.com/archive/p/word2vec/, acesso em 24/03/2023. 43, 72

[56] Facebook: *fasttext.* https://fasttext.cc, acesso em 24/03/2023. 44

[57] Lidwell, William: *Universal Principles of Design, Revised and Updated.* ISBN 978-1592535873. https://universalprinciplesofdesign.com/books. 44, 47

[58] Loranger, Hoa: *Ux without user research is not ux.* https://www.nngroup.com/articles/ux-without-user-research/, acesso em 05/02/2022. 44

[59] Murphy, Christopher: *Comprehensive guide to ux design: Ux laws – part 2.* https://xd.adobe.com/ideas/guides/comprehensive-guide-ux-design-ux-laws-part-2/, acesso em 05/02/2022. 45

[60] Liu, Wanyu, Julien Gori, Olivier Rioul, Michel Beaudouin-Lafon e Yves Guiard: *How relevant is hick's law for hci?* Em *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, página 1–11, New York, NY, USA, 2020. Association for Computing Machinery, ISBN 9781450367080. https://doi.org/10.1145/3313831.3376878. 45, 46

[61] Nielsen, Jakob: *College students on the web*. https://www.nngroup.com/articles/college-students-on-the-web/, acesso em 05/02/2022. 45, 46

[62] Salazar, Kim: *The anatomy of a list entry*. https://www.nngroup.com/articles/list-entries/, acesso em 05/02/2022. 46

[63] Budiu, Raluca: *Information foraging: A theory of how people navigate on the web*. https://www.nngroup.com/articles/information-foraging/, acesso em 05/02/2022. 46

[64] Nielsen, Jakob: *Progressive disclosure*. https://www.nngroup.com/articles/progressive-disclosure/, acesso em 05/02/2022. 47

[65] Nielsen, Jakob: *Horizontal attention leans left (early research)*. https://www.nngroup.com/articles/horizontal-attention-original-research/, acesso em 05/02/2022. 48

Apêndice A

Apêndice

A.1 Instalação da aplicação para testes locais

Para a execução da aplicação em um ambiente local, é necessário instalar algumas bibliotecas, no caso, o *Node.js* que foi mencionado anteriormente. A instalação do *Node.js* é vital para a execuação da aplicação. De forma adicional, deve ser instalado o gerenciador de pacotes NPM[1], para a instalação das bibliotecas internas da aplicação.

A.1.1 Instalando o NVM

Para facilitar o processo de instalação do *Node.js* e do NPM, existe o NVM (*Node Version Manager*, do inglês, "gerenciador de versões do Node")[2]. O NVM permite controlar as versões instaladas do *Node.js* na máquina, permitindo buscar atualizações, mudar entre versões, e até remover versões antigas do *Node.js*. Além disso, a instalação do NVM também instala o NPM.

O tutorial para a instalação se encontra no repositório da biblioteca publicado no *GitHub*[3]. Para instalar a biblioteca, é necessário instalar um *script* e em seguida adicioná-lo às variáveis de ambiente do sistema.

O comando abaixo baixa o *script* de instalação:

```
1    curl -o- https://raw.githubusercontent.com/nvm-sh/nvm/v0.39.3/
   install.sh | bash
```

Após isso, o seguinte trecho deve ser adicionado às variáveis de ambiente do sistema:

```
1    export NVM_DIR="$([ -z "${XDG_CONFIG_HOME-}" ] && printf %s "${HOME
   }/.nvm" || printf %s "${XDG_CONFIG_HOME}/nvm")" [ -s "$NVM_DIR/nvm.sh
   " ] && \. "$NVM_DIR/nvm.sh" # This loads nvm
```

[1] https://www.npmjs.com/
[2] https://github.com/nvm-sh/nvm
[3] https://github.com/nvm-sh/nvm

Para conferir se a instalação procedeu de forma correta, deve executar o comando abaixo para verificar a versão instalada do *Node.js*. Se a instalação ocorreu sem problemas, o número da versão instalada será exibido.

```
1    node -v # v18.12.1
```

A.1.2 Baixando o projeto

O projeto está hospedado no *GitHub*[4]. Para baixar o projeto, é necessário ter a biblioteca `git`[5] instalada na máquina. Então, execute o seguinte comando para baixar o repositório da aplicação:

```
1    git clone https://github.com/italomarcos1/friendica-demo.git
```

Após baixar o repositório, entre na pasta, abra o terminal e execute o seguinte comando para instalar as bibliotecas do projeto:

```
1    npm install
```

Para executar o projeto localmente na máquina, por fim, execute o comando:

```
1    npm run dev
```

No terminal irá aparecer um endereço como `http://localhost:3000`, que deve ser utilizado no navegador para acessar a aplicação.

[4] `https://github.com/italomarcos1/friendica-demo`
[5] `https://git-scm.com/book/en/v2/Getting-Started-Installing-Git`