UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas Josino de Paula Gonçalves

ANÁLISE DE *BULLYING* NO TRABALHO USANDO TEXTOS DE REDES SOCIAIS

Uberlândia, Brasil

2021

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Lucas Josino de Paula Gonçalves

ANÁLISE DE *BULLYING* NO TRABALHO USANDO TEXTOS DE REDES SOCIAIS

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Orientador: Profa. Dra. Elaine Ribeiro de Faria

Universidade Federal de Uberlândia – UFU

Faculdade de Ciência da Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2021

Lucas Josino de Paula Gonçalves

ANÁLISE DE *BULLYING* NO TRABALHO USANDO TEXTOS DE REDES SOCIAIS

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Profa. Dra. Elaine Ribeiro de Faria
Orientador

Prof. Dr. Paulo H. Ribeiro Gabriel
Examinador

Prof. Dr. Wendel A. Xavier de Melo
Examinador

Uberlândia, Brasil
2021

Dedico este trabalho a todos os empregados que enfrentam o assédio em seu ofício. Que a justiça e o respeito chegue até vocês!

Agradecimentos

Agradeço imensamente a minha orientadora, Profa. Dra. Elaine Ribeiro de Faria por todo o processo de orientação que me tornou uma pessoa muito melhor e principalmente por me manter motivado até mesmo nos momentos mais difíceis da construção deste trabalho.

A minha família, pelo imenso suporte e apoio em toda a minha graduação. O carinho e o amor deles deixou o desafio de estudar sozinho em outra cidade menor e me manteve cada vez mais confiante em ser uma pessoa melhor.

Aos meus amigos, Da Mandelaje e hoje da Francanalha, que estiveram comigo em todo o processo de desenvolvimento deste trabalho, discutindo, lendo e sugerindo pontos.

Ao Programa de Educação Tutorial do curso de Sistemas de Informação (PET-SI) por me auxiliar em todo o processo de graduação com apoio financeiro e em meus estudos e por proporcionar experiências que me capacitaram para as mais adversas situações.

A todos os docentes da Faculdade de Computação da UFU, que pela educação, motivaram minha vontade de conhecer mais e mudaram muito das minhas perspectivas e objetivos.

"É preciso força pra sonhar e perceber que a estrada vai além do que se vê."

—Los Hermanos

Resumo

O trabalho é uma das atividades cotidianas da vida de grande parte da população mundial. Mesmo com os inúmeros benefícios produzidos ao exercer uma função no mercado de trabalho, existem riscos de violência relativos a cargos e profissões. O *Bullying* é uma violência psicológica ou física repetitiva e existe em qualquer ambiente social e afeta diretamente a saúde mental do indivíduo. Quando no setor laboral, pode ser feita por funcionários com cargos mais altos ou em grupos, não efetivando tanto a prevenção das entidades contratantes, quanto a denúncia da vítima. As redes sociais, um ambiente virtual para comunicação, se tornam uma grande base de dados para os mais variados relatos, pois fornecem uma sensação maior de liberdade para com os usuários. Este trabalho busca captar episódios e fenômenos ligados ao *Bullying* no mercado de trabalho, por meio dessas redes sociais. Para isso foram utilizadas técnicas de descoberta do conhecimento e analisadas as possíveis relações do *Bullying* com as informações e relatos obtidos e classificados em tópicos através de algoritmos de modelagem de tópicos. Após a aplicação dos métodos foi possível observar uma relação dos *Tweets* com *Bullying* ligado a hierarquia do trabalho onde os cargos superiores são relatados como agressores, também foi possível observar situações de descontentamento com empresas, pois esses relatos supõem que essas organizações não se preocupam com medidas tratativas e tem conhecimento das agressões.

Palavras-chave: Modelagem de tópicos, *Latent Dirichlet Allocation*, Mineração de Textos, Assédio, *Bullying*, Trabalho, Descoberta do conhecimento.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O trabalho é parte essencial da vida de grande parte das pessoas. É para alguns, o pilar central de todo o seu cotidiano e tem sua importância não só em fonte de renda, mas em sua relação com a saúde mental (NAVARRO; PADILHA, 2007). A grande valia dessa atividade para a sociedade pode motivar a preocupação de entidades públicas como o Ministério Público, o qual constitui-se como defensor dos direitos gerais públicos (ARANTES, 1999).

Dentre as formas de atingir a quebra desses direitos dos trabalhadores, a agressão e o assédio são caminhos possíveis. A existência de agressão psicológica no ambiente de trabalho ganhou notoriedade nessa década, através de estudos, como o de (BARRETO; HELOANI, 2015), que fornece conclusões sobre a manifestação e prevenção desse fenômeno e salienta o aspecto de abuso como prejudicial ao funcionamento de empresas, por favorecer o ataque a saúde mental dos trabalhadores.

Algumas das preocupações existentes, quanto às agressões, podem ser veiculadas a ataques de *Bullying*, uma violência baseada em intencionalidade e repetição, podendo atingir nível psicológico, físico e social (TEIXEIRA; FERREIRA; BORGES, 2016). Um dos movimentos para o *Bullying* atingir inúmeros setores, como o próprio mercado de trabalho, foi a sua proliferação em redes de comunicação. O avanço das tecnologias de comunicação auxiliou no desenvolvimento de uma nova modalidade, denominada *Cyber-Bullying*, sendo um assédio que fomenta a interação agressiva através de dispositivos de comunicação (AMADO et al., 2016).

Das várias ferramentas de comunicação na Internet é notável que as redes sociais ocupam grande parte da porcentagem de ferramentas mais utilizadas por fazerem parte do cotidiano dos usuários. As redes sociais, geralmente fomentam o compartilhamento de informações da vida pessoal. Como ferramentas contidas no meio digital, fornecem acesso amplo e fácil como um canal de comunicação, amplificando a possibilidade de alcance e velocidade de interações (AMADO et al., 2016) o que pode viabilizar agressões.

Devido à quantidade de dados produzidos pelas redes sociais, diversos autores produziram trabalhos com o objetivo de identificar diferentes fenômenos através de opiniões de usuários nas redes. Dentre esses trabalhos, pode-se citar: Estudo da ocorrência de *cyberbullying* contra professores na rede social *Twitter* por meio de um algoritmo de classificação bayesiano (ALMEIDA, 2012). Detecção de *Bullying* em redes sociais e classificação do papel da vítima ou agressor (SILVA; SILVA; DIAS, 2018) e Detecção de *Bullying* Escolar em Redes sociais e suas implicações na Educação de Adolescentes (URTIGA; CASTRO, 2018).

Uma característica comum dos trabalhos citados é que eles aplicam um conjunto de técnicas para extração de conhecimento a partir dos dados obtidos das redes sociais. Dentre as técnicas usadas destacam-se o pré-processamento e limpeza dos dados, a conversão do texto num conjunto de atributos e os métodos de mineração de dados, que automaticamente classificam, agrupam ou sumarizam as postagens das redes sociais.

1.1 Justificativa

O *Bullying* é uma agressão intencional focada em rebaixar a vítima através de atos repetitivos. Essa violência frequente se torna algo sistemático no cotidiano da vítima (SCHREIBER et al., 2015), atingindo e descontrolando sua saúde e vida social (TEIXEIRA; FERREIRA; BORGES, 2016). O impacto causado por esse assédio pode ser ampliado para o nível digital. Essa vertente do *Bullying* chamada *CyberBullying* pode atingir a vítima com mais frequência, já que aumenta o alcance e a frequência dos ataques, pois permite que ocorram por redes sociais e outra ferramentas de comunicação afetando a vítima em qualquer momento e em qualquer lugar com o acesso rápido e fácil a *internet* (AMADO et al., 2016). O alcance conseguido com a *internet*, junto a outros fatores como a ausência de medidas de conscientização, auxilia para que o assédio virtual atinja com maior impacto outros patamares como, por exemplo, o mercado de trabalho.

O assédio nos ambientes laborais é uma temática recente que tem ganhado notoriedade nesta década através de estudos como o de (BARRETO; HELOANI, 2015). Essa violência é de efeito altamente nocivo para empresas, como mostrado pelo estudo de (FREITAS, 2007) e pode afetar o empregado gerando desordens na vida psíquica, social, profissional, familiar e afetiva. As agressões também atingem âmbito organizacional trazendo prejuízos como acidentes de trabalho, queda de produtividade, perda de equipamentos, etc. (FREITAS, 2007). Os impactos negativos no ofício motivam o desenvolvimento de formas de detecção de *Bullying* e a consequente tomada de medidas preventivas. Infelizmente, em alguns momentos as organizações podem se fragilizar ao não tratar a questão do assédio.

Essa violência é motivada geralmente pela intolerância e a carência de poder entre os funcionários, podendo ainda ser fortificada com a categorização na empresa. Níveis hierárquicos, terceirização e empregos com alto nível de rotina, que repetem atividades, geralmente são encarados como robotizados ou subcategorias e aumentam as probabilidades de assédio laboral, como o caso de profissões ligadas a *CallCenters* (FREITAS, 2007).

A existência do assédio laboral já preocupa organizações governamentais. Existem ações contra esse problema executadas pelo Ministério Público do Trabalho como a Campanha de prevenção e combate ao assédio moral no ambiente de trabalho (MINIS-

TéRIO PúBLICO FEDERAL, 2018). Essa campanha tem a proposta de conscientizar a população sobre o tema. Há um estudo de (ANABUKI, 2016), que demonstra que parte das ações tomadas pelo MPT durante o período de 2014 a 2015 em várias regiões do Brasil estavam ligadas a casos de assédio. Portanto, é evidente a existência de atuação de entidades públicas contra o assédio laboral, motivando a busca de novas formas de identificação de episódios.

A falta de dados sobre assédio, devido à falta de medidas empresariais (FREITAS, 2007) podem ser superados por meio da detecção dos episódios nas redes sociais, pois seu ambiente de liberdade permite as vítimas se expressarem de forma melhor, fato que se dá por conta da ausência de barreiras até então existentes nas empresas e no mundo físico (FREITAS, 2007).

As redes sociais são aplicações que fornecem espaços para a criação e compartilhamento de informação e do conhecimento (TOMAEL; ALCARA; CHIARA, 2005). Fazendo parte do âmbito de crescimento rápido da *internet*, essas plataformas de interação fornecem uma quantidade grande de dados sobre vários aspectos das pessoas (SILVA, 2013). Segundo o CGI (*Comitê Gestor da Internet*), 77% do produto utilizado por quem tem acesso à *internet* são as redes sociais, o que justifica o tamanho da quantidade de dados produzidos. O ambiente de liberdade de conteúdo e a ausência de grande parte dos limites impostos por situações reais, favorecem cenários de expressão, que geralmente são incomuns, serem naturais nas redes sociais. As comunicações agressivas e repetitivas também fazem parte das interações existentes, constituindo episódios de assédio (AMADO et al., 2016). Portanto, a usabilidade de tecnologias de comunicação, especificamente de redes sociais, pode fornecer relatos e situações presenciadas por vítimas, agressores e relacionados.

A detecção de fenômenos por redes sociais tem sido comprovada como útil, tanto pela grande base de dados provida pela exposição do usuário nessas ferramentas cotidianas como quanto a existência de trabalhos cuja funcionalidade é identificar e inferir vários movimentos. Recentes trabalhos tem o foco em detecção de episódios de violência como o próprio *Bullying*, o estudo de Silva, Silva e Dias (2018) detectou de forma automática o assédio em redes sociais. Já em (URTIGA; CASTRO, 2018) o autor busca identificar *Bullying* no âmbito escolar. O estudo de (ALMEIDA, 2012) discorre sobre a ocorrência de *CyberBullying* contra professores nas redes sociais. Dos inúmeros trabalhos que lidam com a detecção de fenômenos em redes sociais há a ausência de estudo perante o mercado de trabalho e o assédio envolvido nele. Há pouca literatura no que tange a *Bullying* no mercado de trabalho.

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho visa analisar, em seu formato geral, o*Bullying* no mercado de trabalho, expresso na rede social *Twitter* no período de 2015 a 2018. Foram selecionadas postagens que continham informações relativas ao assédio laboral e então submetidas a técnicas convencionais de extração de conhecimento que puderam caracterizar os principais assuntos contidos nesses textos.

1.2.2 Objetivos Específicos

- Construção de uma base de dados sobre *Bullying* no mercado de trabalho a partir de dados da rede social *Twitter*.
- Extração dos principais tópicos e assuntos dos textos de redes sociais sobre *Bullying* relacionado ao trabalho usando técnicas como modelagem de tópicos e nuvens de palavras.

1.3 Organização do Trabalho

O trabalho foi organizado da seguinte forma.

- **Capítulo 2 - Revisão Bibliográfica:** São abordados nesse capítulo, trabalhos da literatura para compreensão dos conceitos relacionados ao tema deste estudo. Há a apresentação do *Bullying* e sua relação com o meio digital, as redes sociais virtuais e sua relação com o cotidiano, as consequências dos assédios no ambiente de trabalho e uma visão geral da extração de conhecimento a partir de bases de dados. Por último este capítulo descreve os principais trabalhos relacionados.
- **Capítulo 3 - Trabalhos Relacionados:** São apresentados nesse capítulo os trabalhos utilizados como refêrencia para a criação e execução da metodologia deste estudo.
- **Capítulo 4 - Método para análise de Bullying a partir de dados de redes sociais:** Nesse capítulo são expostos os processos de coleta e análise dos dados das redes sociais com o fim de obter relações entre o trabalho e o *Bullying*.
- **Capítulo 5 - Resultados:** Os resultados conseguidos após a etapa de desenvolvimento são avaliados visando identificar a existência de *Bullying* relacionado ao trabalho onde foram obtidos resultados favoráveis ao assédio hierárquico, denúncias e sonegação de medidas tratativas pelas empresas.
- **Capítulo 6 - Conclusão:** são apresentadas as contribuições providas por este trabalho, bem como as principais conclusões obtidas que indicam uma relação de

agressão provida de cargos superiores, conhecimento e ignorância por parte das empresas sobre questões de assédios.

2 Revisão Bibliográfica

Este capítulo apresenta os conceitos teóricos necessários para compreensão e desenvolvimento deste trabalho, além de trabalhos relacionados encontrados na literatura.

A seção 2.1 apresenta os conceitos principais sobre *Bullying* e sua relação com o ambiente laboral. A seção 2.2 explica a importância e o funcionamento de redes sociais e sua posição no *Bullying*. A seção 2.3 explica o funcionamento e as técnicas usadas para coletar e analisar dados de redes sociais.

2.1 *Bullying*

Uma agressão que tem como característica atitudes de humilhação e intimidação pode ser tratada como *Bullying* (WEST et al., 2014). Também é possível defini-lo como a obtenção de poder através de ações agressivas (NETO, 2005) e, geralmente, praticadas de forma sistemática (PALÁCIOS; REGO, 2006). Sua prática é grande em âmbito escolar, cerca de 17% dos alunos de uma pesquisa feita no Rio de janeiro com 11 escolas se consideravam vítimas de *Bullying* (PALÁCIOS; REGO, 2006). Nas escolas, esse assédio é cometido em grupos e dividido nos papéis de agressor, vítima e espectador (SCHREIBER et al., 2015), sendo o agressor, o responsável por reivindicar poder através de atitudes nocivas para com a vítima que, pode ou não receber ajuda do espectador.

O avanço considerável das tecnologias de comunicação criou um cenário muito mais acessível para troca de informações e permitiu também que formas de agressão atingissem uma escala digital. O *CyberBullying* é a definição do *Bullying* realizado por ferramentas de comunicação e se distingue desse por tornar-se muito mais nocivo já que chega à vítima com um maior quadro de repetição e com maior continuidade por estar no cenário digital (SCHREIBER et al., 2015).

Com a *internet*, movimentos de agressão ganham uma cobertura muito maior, podendo atingir o mercado de trabalho. O *CyberBullying* se encaixa como um desses movimentos, já que é executado por tecnologias de comunicação. Há hoje uma frequência de utilização grande das tecnologias de comunicação no nosso cotidiano que auxiliam a proliferação de várias categorias de violências, pois essas ferramentas fornecem uma ausência gradativa de limites físicos e temporais (WEST et al., 2014). Essa ausência de limites influencia nos comportamentos menos cívicos e profissionais por distorcerem parte da compreensão dos direitos e limites existentes. Considerando o trabalho, pesquisas recentes comprovam a ocorrência de *Bullying* nesse ambiente. Em uma pesquisa feita a profissionais de enfermagem, 50% afirmaram ter sofrido agressões (TEIXEIRA; FER-

REIRA; BORGES, 2016) e também mencionaram sobre as dificuldades perante sua vida social e física criadas pelo *Bullying* laboral.

2.1.1 Assédio no Trabalho

Ainda que por ser um fenômeno hoje de grande preocupação para a sociedade, o assédio não está definido em consenso internacional fazendo com que cada nacionalidade adote termos diferentes para descrevê-lo. São exemplos de termos *Bullying*, *Mobbing* e assédio moral (ARAÚJO, 2009). Há a possibilidade de diferentes características entre os termos apresentadas por vários autores (MARCONDES; DIAS, 2011), mas, em geral, os termos expressam a existência de uma violência nociva ao individual, social e organizacional e que carece de prevenção e defesa (ARAÚJO, 2009).

O assédio no Brasil é um fenômeno antigo que perpetua na cultura há muito tempo. Os casos de agressão e situações de assédio no que diz a ofício se evidenciam desde o período colonial (HELOANI, 2004).

Os assédios laborais podem ser definidos como condutas de comportamentos abusivos, palavras, atos, gestos e outras formas de denegrir ou atacar partes físicas e psíquicas de uma pessoa impondo poder e superioridade (FREITAS, 2001). O assédio é advindo antes mesmo do conceito de trabalho como é conhecido hoje, já é algo explícito desde os primórdios da sociedade vigente como no período colonial, quando colonizados eram assediados fisicamente e de outras formas para confirmar a superioridade do colonizador (FREITAS, 2001).

Pode-se dizer que os motivos da proliferação do assédio se relacionam diretamente com a necessidade de soberania no ambiente de trabalho. Com a ausência de medidas contra o fenômeno, há uma relação de uma cultura de superioridade por gestores, que ocupam vagas altas na hierarquia empresarial e desmotivam empregados com a possibilidade de ameaçar seu cargo (FREITAS, 2007).

O assédio pode ocorrer de várias formas e níveis no ambiente laboral podendo ser classificado em duas vertentes, assédio vertical e horizontal. O assédio vertical é de fato o mais característico e consiste basicamente no domínio de chefia ou nível superior como agressor, buscando diminuir, excluir e motivar a demissão de funcionário de menor nível (NASCIMENTO, 2004), sendo esse ataque, em grande parte das vezes, motivado para escapar de vertentes de leis trabalhistas como pagamentos de rescisão. Já o assédio horizontal tem maior hipótese de acontecer em situações de promoção por ser praticado, em geral, por parte dos empregados de mesmo nível ou próximo na hierarquia da empresa. Em definição geral, é o assédio praticado por empregados próximos da vítima, ofendendo e denegrindo sua imagem através de chacota e boatos negativos estragando o funcionamento do ambiente laboral e a saúde psicológica da vítima (NASCIMENTO, 2004). Outra

classificação é o *Mobbing* combinado e ascendente (MOTHÉ, 2006). O primeiro se dá pela união do chefe e colegas de trabalho a praticarem um assédio conjunto, já o outro é tido como o inverso, ocorrendo com a união dos empregados para praticarem assédio contra o chefe motivando a perda de cargo superior conforme demonstrado no estudo de Mothé (2006).

2.1.2 Consequências do Assédio

O assédio em seu papel de agressão pode atingir de várias maneiras a vítima e causar os mais variados sintomas degradantes a sua saúde ou mesmo a morte (NASCIMENTO, 2004). Portanto, qualquer agressão oferece riscos nocivos aos envolvidos. Dessa forma, é possível dizer que no ambiente laboral, o assédio pode funcionar como uma ferramenta prejudicial aos empregados, criando desestabilizações culminando em um desempenho baixo, desatenção e ao fracasso. Esses comportamentos em várias das situações podem levar a prejuízos, principalmente de cunho financeiro para as organizações (FREITAS, 2007). Os prejuízos financeiros se dão quando, por exemplo, alguma ferramenta ou máquina da empresa é operada de forma errônea, necessitando reparo. A crescente desmotivação no ambiente de trabalho faz com que o empregado tenha baixo desempenho aumentando as hipóteses de estragar equipamentos obrigando a empresa a dispender de custo de reposição. Ainda em cenários mais tardios, com a necessidade de demissão, a organização carecerá de pagamento de direitos previstos na constituição (FREITAS, 2007).

Os prejuízos para a vítima, mesmo com a existência de danos financeiros, como arcar com custos de erros, estão mais ligados a saúde mental, já que o assédio é guiado através de ataques que, em sua maioria, atingem o oprimido causando terror psicológico (NASCIMENTO, 2004). As dificuldades encontradas em primeira instância pela vítima são o despreparo de gestores de projetos, e recursos humanos que não enxergam os cenários passíveis de agressão e acobertam os episódios, deixando o assédio atingir com frequência e em grande escala os empregados (NASCIMENTO, 2004). Como há ausência de medidas para contornar as agressões, logo o trabalhador ficará exposto a ações por parte dos agressores, que veiculam gestos, palavras, e atitudes que o atingem afetando sua personalidade e saúde mental (MOTHÉ, 2006). Após os ataques, as consequências surgem de formas variadas, em primeiro momento o assediado sente estresse, os primeiros sintomas de depressão, ansiedade, etc. (MOTHÉ, 2006). Com a repetição dos ataques ocorre uma maior distorção da personalidade da vítima, podendo causar consequências mais graves, como a sensação de nulidade, a incompetência fruto da dificuldade de concentração, afastamento e perda do emprego, que motivam a outros ciclos de problemas psicológicos, como tentativas de suicídio, alcoolismo e uso de drogas (FREITAS, 2007).

2.2 Redes sociais

O ser humano tem em seu cotidiano o convívio em sociedade, que motiva o relacionamento em grupos e desenvolve um conjunto de conexões entre pessoas. Esse conceito tem funcionamento tanto presencial quanto virtual (TOMAEL; ALCARA; CHIARA, 2005). Na visão virtual, as redes sociais funcionam utilizando dois elementos, os chamados "atores"que podem ser usuários, páginas, etc. e as conexões existentes entre os atores, permitindo a interação entre vários deles criando conteúdos (RECUERO, 2009). As redes sociais hoje atingem uma grande parte da população mundial por sua acessibilidade com os usuários permitindo a conexão entre grupos com objetivos específicos e contatos próximos (TOMAEL; ALCARA; CHIARA, 2005).

Conforme dados de balanço da rede social *Facebook*, em 2018 foram computados 1,49 Bilhões de usuários diários ativos (FACEBOOK, 2018c), indicando que as redes sociais fazem parte de grande parcela de presença no cotidiano mundial. Por mais que elas sigam um determinado padrão para comunicação e relação entre os usuários, cada aplicação tem um foco específico. Algumas das principais redes sociais serão abordadas a seguir explicando seu funcionamento geral.

- O *Facebook* é uma rede social de funcionamento global, que permite postagens de texto, fotos e vídeos, criação de páginas de conteúdo e grupos além de eventos. Permite a interação entre pares através de mensagens, comentários em publicações e chamadas de vídeo e áudio (FACEBOOK, 2018a).
- *Instagram* é uma rede social com a proposta de veicular em suas postagens fotos e vídeos, sendo dessa forma uma rede mais ligada a mídias. Também provê troca de mensagens e comentários nas publicações (INSTAGRAM, 2018).
- *Twitter* é uma rede social com a proposta de veicular publicações e notícias. Sua forma de publicação é chamada *tweet* e se limita a 280 caracteres permitindo também na postagem anexos de fotos e vídeos. A rede social permite também a troca de mensagens e comentários, porém os comentários, diferentes das outras redes sociais, funcionam como outro *tweet* (TWITTER, 2018). O *Twitter* vem sendo amplamente utilizado em pesquisas acadêmicas por limitar o número de caracteres oferecendo dados mais sucintos, por gerir todo seu conteúdo baseado em *tweets* e principalmente por fornecer acesso aos dados por consultas com chaves de busca em sua API, facilitando a extração (TWITTER, 2019a).

2.2.1 Bullying nas redes sociais

A *internet*, além de estreitar laços, desprende de características que assimilam aspectos não muito usuais no nosso dia-a-dia físico, sua possibilidade de construção de conteúdos permitem a disseminação de informações e comportamentos contraditórios ao

nosso cotidiano (CASTELLS, 2003). As redes sociais presentes na *internet* herdam essa característica e se tornam ferramentas para a execução de atividades diferentes ao nosso cotidiano como é o caso do *CyberBullying*. Medidas como a Central de Prevenção ao *Bullying*, criada pelo *Facebook* com a finalidade de inibir assédios em redes sociais, fortalecem o conceito da existência e da importância de conscientização nas redes sobre esse movimento (FACEBOOK, 2018b).

A dificuldade de prevenção dos ataques e a frequência deles se tornam muito altas quando feitos por redes sociais. Os ataques nas redes permitem situações de anonimato, sendo que o atacante pode se esconder em perfis falsos (AZEVEDO; MIRANDA; SOUZA, 2012). Há também fenômenos exclusivos, que podem auxiliar o processo nas redes sociais, pois a agressão pode ficar armazenada podendo ser reutilizada ou revivida por outros usuários. Pode ser também destinada a grupos de pessoas relacionados a vítima por terem uma conexão na plataforma facilitando seu alcance (TAVARES, 2012).

2.3 Análise de dados textuais

Com a acessibilidade da *internet*, houve um engrandecimento da criação e compartilhamento de dados textuais (GONÇALVES, 2002). Esses dados podem ter significados particulares, podendo inferir fenômenos por processos de Descoberta de Conhecimento em Bases de Dados (DANTAS et al., 2008). O processo de Descoberta de Conhecimento em Bases de Dados é um conjunto de etapas para inferir conhecimento sobre um determinado conjunto de dados (DANTAS et al., 2008).

A análise de dados utilizando os princípios de Descoberta de Conhecimento em Bancos de Dados permitem a identificação de fenômenos, fatores e opiniões em redes sociais (FREITAS et al., 2008). Com o grande número de usuários em redes sociais, como exemplo, os milhões de usuários ativos presentes no *Facebook* (FACEBOOK, 2018c), o número de dados criados nessas plataformas motivaram vários trabalhos da literatura como (SILVA; SILVA; DIAS, 2018) e (PORTO et al., 2012). Esses trabalhos, aplicando o processo de Descoberta de Conhecimento em Bases de Dados, envolvem um conjunto de etapas específicas, análogas em grande parte da literatura, envolvendo a análise de dados textuais.

2.3.1 Coleta

A etapa de coleta das informações analisadas pode ser tratada como fundamento do projeto. Primeiramente, é realizada a escolha da ferramenta de comunicação da qual os dados serão coletados. Com a escolha da ferramenta se faz necessária a pesquisa sobre as formas de extração possíveis.

Uma grande parte das redes sociais detém de uma interface para gerenciar os conteúdos criados pelos usuários, chamada API *(Application Programming Interface)*. Com essas implementações, é possível fazer requisições que levam termos de busca e devolvem conteúdos criados por usuários (FRANÇA et al., 2014). Uma API pode facilitar a extração, porém possuir algumas limitações. Um exemplo de limitação é a restrição no número de requisições. Quem faz a coleta de forma gratuita fica limitado a coletar postagens de até 7 dias no *Twitter* (MARCONDES, 2017).

2.3.2 Pré-processamento

A etapa de pré-processamento acontece após a coleta da base de dados e seu funcionamento pode ser tratado como o ato de padronização de todos os dados coletados. Geralmente, os conteúdos publicados na *internet* podem deter de vários elementos que são sem significado para os processos de classificação.

O uso de abreviações é rico na *internet* por ser um fenômeno relativo à integração da linguagem com a tecnologia e da carência de velocidade na comunicação que se tornou mais rápida (KOMESU, 2006). Abreviações advindas da *internet* são desafios frequentes, que motivam o pré-processamento, pois inviabilizam processos como o de classificação de texto pela ausência de significado direto e por não poder ser interpretada pelos classificadores, tendo uma obtenção de resultados muitas vezes simplista (BATISTA et al., 2003). Outra dificuldade frequente nos conteúdos extraídos é o uso de outros produtos na construção de conteúdo nas redes sociais como áudios, vídeos e imagens, que não proporcionam predição por sua natureza, impossível de interpretar em algumas metodologias como a análise por textos.

Na etapa de pré-processamento podem ser feitos, geralmente, além dos citados, os processos de:

- Substituições em casos de texto com abreviações para a palavra abreviada, substituições em utilização de figuras e caracteres para representação de sentimentos.
- Remoção de Duplicatas em casos que a frequência não é analisada.
- Remoção de itens sem caracteres úteis para classificação como números e caracteres especiais.
- Remoção de *spams*.
- Remoção de registros com tamanho pequeno, como, por exemplo, postagens com uma palavra.
- Remoção de alguns conectivos e palavras frequentes na língua sem um significado para a classificação
- Transcrição do texto em sua forma radical, como, por exemplo, Pezão e Pezinho transcritos para Pé. Esse processo se chama lematização.

O pré-processamento pode ser feito de forma manual, mas com grandes volumes geralmente encontrados nas coletas pode ser custoso fazer a padronização (MATSUBARA; MARTINS; MONARD, 2003), demandando o uso de ferramentas e construção de algoritmos.

2.3.3 Mineiração de Dados

A etapa de mineração de dados é uma das partes mais importantes do processo de Descoberta de Conhecimento em Bases de dados, pois nessa etapa são utilizadas técnicas que possibilitam distinguir conjuntos de fenômenos a partir das bases (CÔRTES; PORCARO; LIFSCHITZ, 2016). Quaisquer técnicas que possibilitem extrair conhecimento de um volume de dados podem ser definidas como técnicas de mineração de dados (QUONIAM et al., 2001).

Na etapa de mineração de dados, há duas análises existentes. A análise de prognóstico contém etapas de classificação, predição e estimação, podendo envolver o uso das mais variadas técnicas para predizer categorias, valores e comportamento dos dados analisados (CÔRTES; PORCARO; LIFSCHITZ, 2016). Já a análise descritiva visa descobrir eventos específicos nos dados através de reconhecimento de um conjunto de padrões já existentes nos dados, mas não visíveis (CÔRTES; PORCARO; LIFSCHITZ, 2016). Algumas das principais técnicas de descobrimento utilizadas na mineração de dados serão descritas a seguir:

A sumarização é uma das técnicas mais importantes dentre as existentes (DANTAS et al., 2008). Seu funcionamento consiste definir um grupo de características gerais no conjunto de dados (DANTAS et al., 2008), com isso é possível definir características ou uma descrição do conjunto de dados (SANTOS et al., 2009).

O agrupamento ou *clustering* é uma técnica que divide os dados em vários conjuntos utilizando como parâmetro para definir cada conjunto, atributos semelhantes entre os dados, resultando em vários grupos, onde um grupo tem características semelhantes internamente e seja distinto o máximo possível dos outros grupos (CÔRTES; PORCARO; LIFSCHITZ, 2016).

A modelagem probabilística de tópicos é um assunto de estudo na área de aprendizado de máquina, seu objetivo é, através de técnicas que analisam as relações de dados textuais em um grande conjunto, extrair valores temáticos que o definam (BIANCHINI, 2018). O retorno, com a aplicação da técnica, são um grupo de termos que definem os tópicos tratados no conjunto.

A classificação é uma técnica de análise prognóstica (CÔRTES; PORCARO; LIFSCHITZ, 2016), muito usual em vários trabalhos da literatura como (SILVA; SILVA; DIAS, 2018). Consiste em definir se um determinado registro de um conjunto de dados poderia

ser rotulado em uma das classes pré-definidas ou não definidas, onde nesse caso os algoritmos escolhidos treinam um modelo de classificação que pode deduzir uma categoria para esse determinado registro (SILVA; SILVA; DIAS, 2018).

2.3.4 Validação

O processo de validação é a etapa de definição da qualidade dos resultados. É nessa etapa que os resultados da análise dos dados são verificados e devem ser expressos de forma sucinta para permitir a avaliação (BARION; LAGO, 2015). Há vários métodos e ferramentas para expressar os resultados.

Em modelagem de tópicos o fator de validação de maior importância é a capacidade humana de interpretar um conjunto de tópicos (FALEIROS; LOPES, 2020). Para mensurar a boa interpretação, existem várias métricas propostas na literatura tais como, PMI que mede a associatividade entre duas palavras e Cv que mede a coerência das palavras vigentes em um tópico (BIANCHINI, 2018).

A utilização de recursos gráficos também pode servir como ferramenta de análise dos resultados principalmente quando os avaliadores podem estar em contextos diferentes como Inteligência de Negócios (CÔRTES; PORCARO; LIFSCHITZ, 2016). São possíveis várias aplicações de gráficos e tabelas, conforme apresentados no trabalho de (QUONIAM et al., 2001), que tornam o processo de validação mais sucinto.

3 Trabalhos relacionados

A ascensão das redes sociais motivou o desenvolvimento de estudos para a área de computação. A quantidade de dados disponíveis, bem como o crescimento de trabalhos da literatura sobre processamento de dados, motivaram a criação de inúmeras ferramentas e elaboração de vários estudos (OLIVEIRA et al., 2006), em várias linhas de pesquisa, como a mineração de dados em textos. Nessa seção são apresentados os trabalhos relacionados com os temas abordados. A Tabela 1 resume os trabalhos relacionados que serão apresentados a seguir.

O trabalho de (SILVA; SILVA; DIAS, 2018) utiliza redes sociais para fazer a análise da existência de *Bullying* em postagens. O estudo tem como foco a construção de um corpus para análise de postagens na rede social *Twitter*. As postagens foram coletados pela API da rede social usando sinônimos de *Bullying* como palavras de busca. Em seguida, foram removidos *spams* e textos com menos de 120 caracteres. Após a etapa de normalização da base de dados foram utilizados 2.000 *tweets* para o processo de classificação manual, onde avaliadores definiam a presença ou ausência de *Bullying* no texto. Na etapa de mineração automática dos textos foram utilizadas as técnicas de classificação SVM, Naive Bayes e Regressão Logística, sendo que todos conseguiram alcançar acurácia maior que 60%. Há também uma funcionalidade de classificação do papel de indivíduo dentro de uma situação de *Bullying*, que segue a mesma metodologia aplicada na primeira etapa.

No estudo de (XU; ZHU; BELLMORE, 2012), há a utilização de técnicas de detecção de sentimento nas postagens que contenham traços de *Bullying*. A base de dados foi obtida pela API da rede social *Twitter*. Os dados coletados passaram por uma etapa de pré-processamento, removendo *stopwords* e pontuação, *tweets* repetidos, os termos usados na busca da API e em seguida foram convertidos em bigramas e unigramas. Os dados então são classificados usando o algoritmo de SVM. Como base para o aprendizado supervisionado do SVM foram utilizadas definições de *Bullying* trazidas do *Wikipédia* e definições trazidas da busca no *Twitter*. O trabalho classificou os *tweets* em 7 classes de sentimentos, utilizando na construção de sua implementação sinônimos dessas classes para englobar maior amplitude na detecção. Como resultado foi possível identificar frequência dos sentimentos medo, tristeza, raiva e alívio como mais frequentes.

O estudo de (ALMEIDA, 2012) também é relacionado com a descoberta de violência virtual em redes sociais. Nesse trabalho, o autor detectou na rede social *Twitter* a existência de *Bullying* contra professores por coleta das postagens dos usuários, pré-processamento e utilização de algoritmos de aprendizado de máquina. O processo de coleta foi feito utilizando a API, buscando pelos termos meu professor e minha professora

durante uma semana, feita 6 horas em cada dia, resultando em 6900 registros coletados. Após a conclusão da extração, foi feito o processo de organização dos *tweets*, onde foram removidos termos desnecessários, os termos da busca, referências ao usuário, links e outros itens desnecessários. Ao final, os dados passam pelo classificador Bayesiano. O classificador foi treinado com um conjunto de 200 *tweets* dos 300 classificados de forma manual, definindo a existência de *Bullying* por avaliadores. A classificação teve como resultado a rotulação em 4 classes com porcentagem geral de acerto de 80%.

É encontrado no trabalho de (URTIGA; CASTRO, 2018) a detecção de *Cyber-Bullying* escolar utilizando conceitos de descoberta de conhecimento e mineração de dados em redes sociais. Foram extraídos dados do *Twitter* por meio do ambiente computacional R com um pacote especifico que trabalha com a rede social e permite extrações diretas da API pelo *Software*. Para o processo de descoberta, o autor utiliza das técnicas de *clustering* para agrupar os dados e a sumarização para extrair a essência das postagens agrupadas. Também foi aplicada uma entrevista onde alunos de uma escola de ensino médio relataram experiências na escola e fora dela. Com as bases coletadas no *Twitter* junto as entrevistas aplicadas, o autor obteve como resultado que 57% das meninas e 43% dos meninos já sofreram alguma agressão na vida escolar. E concluiu haver a existência de *Bullying* tanto no âmbito escolar como nas redes sociais.

No trabalho de (NUNES; FREITAS; PARAISO, 2009) o foco da detecção é encontrar assédio moral em *e-mails*. Para detectar foram levantadas duas bases, uma com definições e sinônimos do assédio moral em fóruns, dicionários e outros lugares, tendo como princípio as leis que enquadram o assédio moral e a outra base com relatos de pessoas assediadas, os quais foram extraídos de redes sociais. Houve então uma etapa de refinamento dos dados extraídos na primeira base agrupando as palavras em classes gramáticas. Para assimilar o conteúdo do *e-mail* à agressão, foi criado pelo autor um algoritmo que converte todos os termos para uma representação em *N-gram* e com essa representação são comparados cada vetor de cada um dos termos entre às duas bases propostas em busca de uma similaridade, se encontrada, é então definido um possível episódio de assédio moral. Como resultado no experimento de busca de similaridade dos *e-mails* foram detectados em 72% a ocorrência de assédio. A segunda classificação utilizou 33 palavras para serem tratadas como caracterizadoras de assédio moral onde os resultados obtiveram 90,91% de acerto.

3.1 Considerações Finais

As redes sociais ocupam hoje grande fatia da comunicação e estão presentes na vida de boa parte da população brasileira. O volume de dados criado dentro dessas tecnologias bem como o ambiente propício para exposição dos usuários permitem que seja possível a

Tabela 1 – Trabalhos Relacionados

Trabalho	***Objetivo***	***Coleta***	***Pré-processamento***	***Técnicas de mineração de dados***
(SILVA; SILVA; DIAS, 2018)	Detecção de *Bullying* em redes sociais e classificação do papel da vítima ou agressor.	Fonte: Twitter. Ferramenta: API do Twitter. Período: 3 meses. Palavras-chave: *Bullying*, seus sinônimos e variações.	Remoção de textos contendo menos de 120 caracteres, Remoção de textos repetidos. Conversão em *bag-of-words* de unigramas e bigramas	Classificadores: Naive-Bayes, Regressão Lógistica e SVM.
(XU; ZHU; BELLMORE, 2012)	Desenvolvimento de metodologia de aprendizado rápida para análise de sentimento em postagens contendo traços de *Bullying*	Fonte: Twitter. Ferramenta: API do Twitter. Palavras-chave: *bully*, *bullied*, *Bullying*.	Remoção de *tweets* repetidos. Remoção de *stopwords*, pontuação e termos de busca nos *tweets*. Conversão em bigramas e unigramas. Remoção dos termos com frequência menor que 5.	Classificador: SVM treinado por meio das definições encontradas no *Wikipédia* e no Twitter sobre *Bullying*.

extração de fenômenos como o *Bullying*.

A descoberta de conhecimento nas bases de dados criadas por redes sociais podem fornecer inúmeras características a serem analisadas e motivam trabalhos em todo o mundo com fins de detecção e classificação de fenômenos. Alguns com foco em detecção de *Bullying* como (SILVA; SILVA; DIAS, 2018) e (XU; ZHU; BELLMORE, 2012), por exemplo.

Neste estudo serão coletados *tweets* visando analisar casos de *Bullying*, mas com o diferencial da relação com o mercado de trabalho motivando a detecção desse assédio em empresas e no setor laboral. Para o desenvolvimento do estudo serão coletados os *tweets*, aplicadas medidas de limpeza e padronização nos dados, processos de análise com mineração dos dados e validação dos resultados da mineração.

Trabalho	***Objetivo***	***Coleta***	***Pré-processamento***	***Técnicas de mineração de dados***
(ALMEIDA, 2012)	Detecção de *Bullying* contra professores no *Twitter*.	Fonte: Twitter. Ferramenta: API do Twitter. Periodo: uma semana. Palavras-chave: *meu professor* e *minha professora*.	Remoção dos termos da busca, referências ao usuário, links e outros itens desnecessários.	Classificador: Naive Bayes.
(URTIGA; CASTRO, 2018)	Detecção de *Bullying* no âmbito escolar.	Fonte: *Twitter* Palavras-chave: Baseadas em nuvem de palavras gerada na rede social sobre *Bullying*	Remoção dos termos desnecessários, os termos da busca, referências ao usuário, links e outros itens desnecessários.	Classificador: Naive-Bayes.
(URTIGA; CASTRO, 2018)	Detecção de *Bullying* em âmbito escolar.	Fonte: Twitter. Ferramenta: API do Twitter. Palavras-chave: são algumas das palavras pertencentes a nuvem de palavras sobre *Bullying* gerada na rede social. Foram feitas também entrevistas com alunos de uma escola pública	Remoção de alguns *tweets* que não tem relação com *Bullying*.	Agrupamento e sumarização para identificar fenômenos que se relacionam com as entrevistas.
(NUNES; FREITAS; PARAISO, 2009)	Detecção de Assédio moral em *emails*	Fontes: Redes sociais, fóruns, dicionários. Palavras-chave: busca feita em locais que lidam com assédio moral, não há palavras-chave.	Agrupamento das palavras em classes gramaticais e sinônimos. Atribuição de significado a cada uma das palavras.	Classificador: Criação de algoritmo próprio.

4 Método para análise de *Bullying* a partir de dados de redes sociais

Neste capítulo serão demonstradas as ações utilizadas para o desenvolvimento do trabalho com o objetivo de analisar o *Bullying* relacionado ao trabalho em redes sociais.

A seção 4.1 explica o processo de coleta dos dados utilizado neste trabalho. A seção 4.2 descreve o funcionamento de cada estratégia para limpeza dos dados utilizada no pré-processamento. Na seção 4.3 são descritos os processos utilizados para a mineração dos dados. A seção 4.4 apresenta as métricas escolhidas para quantificar a coerência entre os tópicos para fins de validação.

4.1 Coleta dos Dados

Foi escolhida a rede social *Twitter* por ser muito popular permitindo extrair um grande montante de dados e com isso aumentar a probabilidade de uma seleção de *postagens* mais próximas do intuito da análise.

O *Twitter* permite que postagens públicas dos últimos 7 dias possam ser extraídas de forma gratuita por um conjunto de palavras-chave e atributos (TWITTER, 2019b). De modo a obter uma base de dados com um número satisfatório de instâncias, foram coletados os *tweets* entre os anos de 2015 e 2018. Assim, houve a necessidade da utilização de uma biblioteca com um conjunto de funções para extração dos dados chamada *GetOldTweets*.

Figura 1 – Exemplo de código para busca de *tweets*.

```
tweetCriteria = got.manager.TweetCriteria().setUsername("barackobama")\
                                           .setTopTweets(True)\
                                           .setMaxTweets(10)
tweet = got.manager.TweetManager.getTweets(tweetCriteria)[0]
```

Fonte: Autor.

A biblioteca foi implementada em várias linguagens, sendo que neste trabalho foi utilizada sua implementação em *Python*[1]. Com a biblioteca é possível fazer buscas na rede social sem a restrição de datas, além de possibilitar a definiçõa de um conjunto específico

[1] https://github.com/Jefferson-Henrique/GetOldTweets-python

de critérios que podem ajudar a filtrar conteúdos específicos (HENRIQUE, 2019). Alguns exemplos das funções:

- *setSince*: define que deverão ser retornados apenas *tweets* criados a partir da data.
- *setUntil*: define que deverão ser retornados apenas *tweets* criados antes dessa data.
- *setQuerySearch*: define quais serão os termos que deverão estar presentes nos *tweets* retornados.
- *setNear*: define a localização que os *tweets* buscados deverão estar.

Preenchidos os critérios para fazer a busca, a ferramenta retorna um objeto para cada *tweet* encontrado. Esse objeto detém de um conjunto de atributos relativos ao *tweet* como: o texto do *tweet*, usuário, data de criação, número de *retweets*, entre outros.

Neste trabalho foram utilizadas como chaves de busca alguns termos relativos a assédio, *Bullying* e trabalho. Foram criadas dez chaves de busca com flexões dos termos pesquisados, parte deles baseados em (SILVA; SILVA; DIAS, 2018). Essas buscas usaram o conjunto de datas definidos acima, portanto, em cada uma das dez consultas foram utilizados os anos entre 2015 e 2018.

Foi criado um algoritmo em *Python* que utiliza *GetOldTweets* com os termos de busca e critérios já apresentados. Ao fim da busca o algoritmo gera um arquivo com a extensão *.txt* contendo todos os *tweets*. Os outros atributos foram ignorados nas etapas seguintes.

4.2 Pré-processamento dos dados

Após a coleta foram empregadas técnicas para filtrar e organizar a estrutura dos dados coletados no *Twitter*. Nesta etapa, o objetivo é remover termos irrelevantes e adequar os dados para a etapa de análise textual. As etapas de pré-processamento utilizadas neste trabalho serão descritas abaixo.

4.2.1 Tratamento de caracteres e palavras

A primeira etapa do pré-processamento é a adaptação e remoção de parte do conjunto de palavras de cada *tweet*. As técnicas utilizadas são:

- Converter *tweets* em minúsculos: transforma os textos todos em minúsculos. Tem como fim padronizar todos os *tweets*.
- Remoção de *links*, citações e *hashtags*: são removidos todos os *link*s para outros *sites*, perfis e *fotos*. Também são removidas citações de outros usuários e *hashtags*, pois neste trabalho não há a necessidade desses elementos na análise.

- Conversão de abreviações: o uso de abreviações é frequente na *internet*, também conhecidas como *internetês* (KOMESU, 2006). Quando abreviadas, perdem seu valor semântico em vários algoritmos de análise de dados. Essa etapa consiste em converter a abreviação em sua palavra normal utilizando dicionário próprio construído pelo autor.
- Remoção de caracteres não-alfabéticos: são removidos todos os caracteres não existentes no alfabeto brasileiro, por serem irrelevantes ao trabalho.
- Remoção de *tweets* pequenos: são removidos *tweets* com menos de 5 caracteres, pois não há relevância em *tweets* com uma palavra.

4.2.2 *Tokens*

Com a remoção de grande parte de dados irrelevantes, os *tweets* são convertidos em um conjunto de termos, cada termo recebe o nome de *token*. Essa técnica é muito importante na etapa de pré-processamento, pois transforma os textos coletados na estrutura necessária para o processo de análise dos dados. Abaixo estão as etapas aplicadas em cada conjunto de *tokens*.

- Remoção de *tokens*: são removidos *tokens* com menos que 3 caracteres e *tokens* que são um conjunto de termos repetidos representando risadas, por exemplo.
- Remoção de *stopwords*: são removidos *tokens* que são considerados *stopwords*. Neste trabalho a remoção foi baseada em um conjunto de *stopwords* encontrado em fóruns e em bibliotecas disponíveis para *Python*.
- Remoção de conjuntos de *tokens* repetidos: após os processos anteriores, o conjunto de *tokens* que se repetir no corpus é removido, por exemplo, *"fazer trabalho sobre assédio com pessoas sofrendo"* e *"fazer trabalho sobre assédio com pessoas sofrendo rt"*, após as etapas anteriores resultariam no mesmo conjunto de tokens e seriam removidos.

4.2.3 Lematização

O processo de lematização consiste em converter todas as flexões de verbos em seu infinitivo e converter substantivos e adjetivos em sua forma singular masculina (LUCCA; NUNES, 2002). Esse processo padroniza as palavras em uma única flexão favorecendo melhores resultados nos algoritmos de análise dos dados.

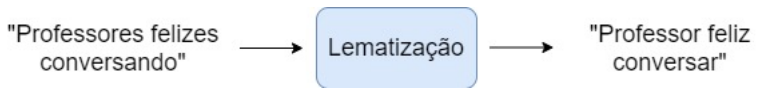
Neste trabalho, para o processo de lematização, foi utilizada uma implementação em *Python* de um conjunto de funções de processamento de linguagem natural chamada *NLPyPort*. Essa implementação contém um conjunto de melhorias em relação a biblioteca *NLTK* e funções para pré-processamento focadas inteiramente no idioma português (FERREIRA; OLIVEIRA; RODRIGUES, 2019).

Para o processo de lematização, é primeiramente executada a função de *Tag* da *NLPyPort*. O objetivo da função é mapear a palavra conforme as classes gramáticas. Com essa informação é possível executar o fluxo da função de lematização descrito em (FERREIRA; OLIVEIRA; RODRIGUES, 2019), que consiste em:

1. Módulo 1: pesquisa o termo em um dicionário já com sua representação lematizada.
2. Módulo 2: caso não seja encontrado um léxico correspondente, executa um conjunto de regras no termo, tendo como base a classe gramatical recebida da função *tag* e a cada aplicação, utiliza-se o módulo 1 para pesquisar o léxico.
3. Executa a regra de normalização para advérbios.
4. Executa a regra de normalização para léxicos no plural.
5. Executa a regra de normalização para léxicos em superlativo.
6. Executa a regra de normalização para léxicos em aumentativo.
7. Executa a regra de normalização para léxicos em diminutivo.
8. Executa a regra de normalização para léxicos no gênero feminino.
9. Executa a regra de normalização para verbos.
10. Caso o léxico não seja encontrado em nenhuma das várias regras aplicadas presume-se que ele já esteja lematizado.

A Figura 2 exemplifica uma frase e o resultado em um processo de lematização.

Figura 2 – Exemplo de frase lematizada.

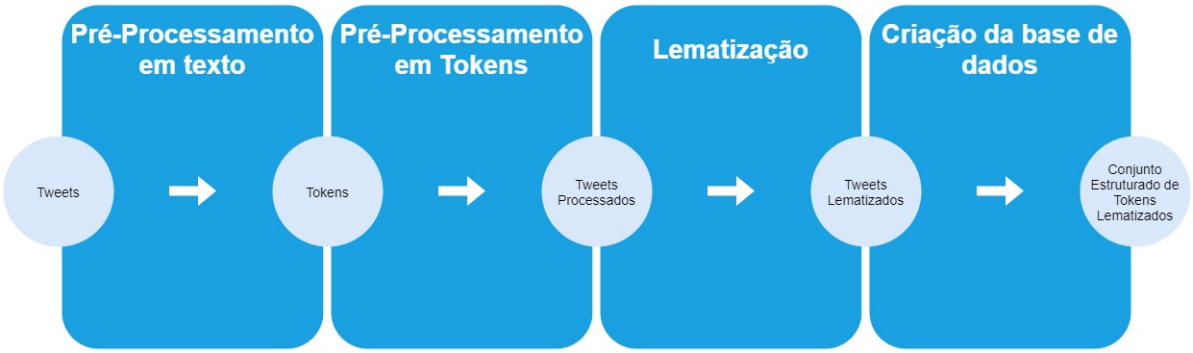


Fonte: Autor.

4.2.4 Criação da base de dados

Após a lematização, é construída uma função que, seleciona o conjunto retornado pela *NLPyPort*, converte para estruturas aceitas pelos algoritmos usados na etapa de análise de dados como matrizes e vetores, sendo melhor descritas na seção sobre modelagem de tópicos, em seguida o conjunto de dados é salvo em um arquivo para ser utilizado na etapa seguinte. O fluxo do pré-processamento é representado pela Figura 3

Figura 3 – Etapas do pré-processamento.



Fonte: Autor.

4.3 Mineração de Dados

Concluído o pré-processamento dos dados, foram executadas as técnicas de mineração de dados com o intuito de identificar temas relativos ao *Bullying* no mercado de trabalho. Nesta seção serão descritas as técnicas de análise textual empregadas no trabalho.

4.3.1 Modelagem de Tópicos

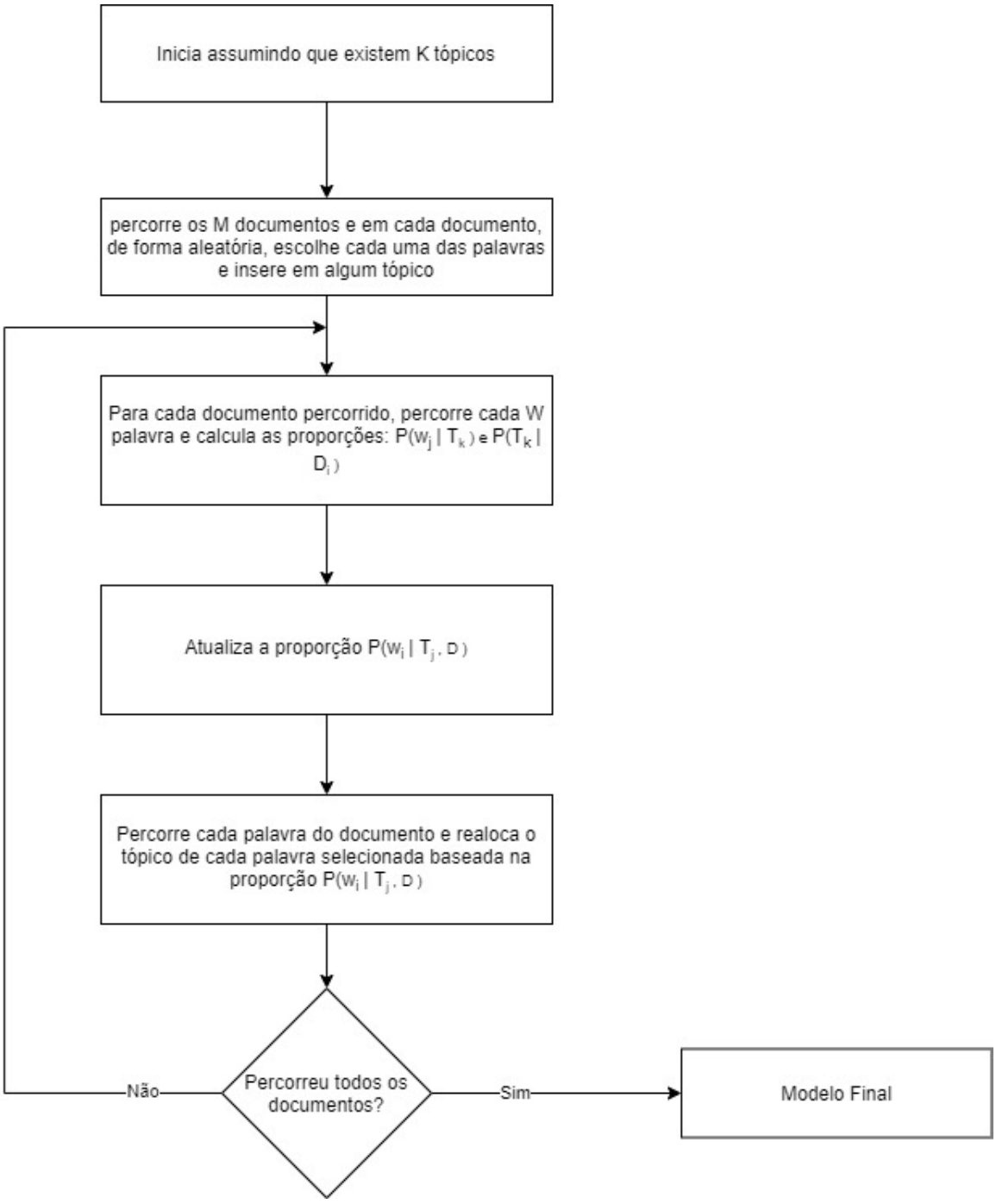
Com o intuito de extrair um conjunto de tópicos que representem os assuntos no corpus de *tweets* foi escolhido como algoritmo de modelagem de tópicos o *LDA* (*Latent Dirichlet Allocation*).

Esse algoritmo produz tópicos baseados na relação entre os documentos de uma coleção (BIANCHINI, 2018). Seu diferencial consiste em produzir tópicos latentes (desconhecidos) utilizando inferência estática (SANTOS et al., 2015).

O algoritmo parte da suposição que cada documento é um conjunto de tópicos e cada tópico é um conjunto de palavras desses documentos (SANTOS et al., 2015), A partir dessa relação é possível produzir valores probabilísticos entre as palavras dos documentos e sua relação com cada um dos possíveis tópicos. O processo de execução descrito em (GREATLEARNING, 2020) é apresentado na figura 4 e consiste em:

1. Assumir a existência de K tópicos
2. Percorrer M documentos e, de forma aleatória, selecionar cada uma das palavras do documento e alocar em um dos K tópicos
3. Percorre cada documento e cada palavra W desse documento, calcula a proporção de palavras nesse documento que são pertencentes a um tópico K exceto a palavra

Figura 4 – Diagrama do fluxo de execução do *LDA* traduzido de (GREATLEARNING, 2020).



Fonte: Autor.

W. Se a proporção for alta, a palavra W tem grande probabilidade de pertencer aquele tópico. Essa proporção é dada por P(W_j | T_k) onde W_j representa a palavra percorrida e T_k o tópico percorrido. Também calcula a proporção de quantos documentos com a palavra W estão em um tópico K, ou seja, quantos documentos estão em um Tópico K por causa daquela palavra. Essa proporção é dada por P(T_k | D_i) sendo D_i o i-ésimo documento.

4. Atualiza a proporção P(w_i | T_j , D) dada por P(w_j | T_k) x P(T_k | D_i). Essa proporção representa o quanto uma palavra é próxima do tópico, tanto pela sua frequência em outros documentos pertencentes ao tópico, quanto ao seu documento ter uma relação de proximidade. Se, por exemplo, a palavra fosse muito próxima do tópico por estar presente em vários outros documentos, mas seu documento não tivesse proximidade, ela teria uma baixa proporção de presença no tópico.

5. Executa a regra de normalização para léxicos em aumentativo.

6. Caso o número de documentos existentes tenha sida percorrido, apresenta o modelo, caso não tenha sido, inicia o passo 3 novamente.

O *LDA* é implementado em várias bibliotecas em *Python*. A biblioteca escolhida para este trabalho foi a *Gensim*, uma biblioteca com várias ferramentas disponíveis para análise de conjunto de dados (ŘEHŮŘEK; SOJKA, 2010).

Foi utilizada neste trabalho a implementação do *LDA* contida na biblioteca *Gensim*. Além da implementação também são utilizadas algumas funções antes do processo de modelagem que convertem o conjunto de dados em duas estruturas esperadas na execução do algoritmo.

São utilizadas duas funções, a *corpora.Dictionary*, essa função gera uma estrutura que mantém cada *token* e seu identificador. Essa estrutura é chamada *Dictionary* (ŘEHŮŘEK; SOJKA, 2010). Um exemplo é mostrado na figura 5.

Figura 5 – Exemplo de criação de um *Dictionary*.

```
dictionary = corpora.Dictionary(tweetsTokenizados)
```

Fonte: Autor.

A outra função utilizada para estruturar o conjunto de *tweets* pré-processados é uma função da própria *Dictionary* criada anteriormente. A função *doc2bow* transforma o conjunto de *tokens*, já presentes na *Dictionary*, em uma *Bag-Of-Words*, essa estrutura contém o identificador do *token* e sua frequência em todo o corpus. Um exemplo é mostrado na figura 6.

Figura 6 – Exemplo de uso da função *doc2bow*.

```
corpus = [dictionary.doc2bow(textos) for textos in tweetsTokenizados]
```

Fonte: Autor.

Com as duas estruturas definidas é possível executar o *LDA*. Para a execução foram informados os atributos:

- *Corpus*: esse atributo espera uma *Bag-Of-Words*, portanto, é informada a estrutura provida pela função *doc2bow*.
- *id2word*: nesse atributo é informado o conjunto que relaciona a identificação com o *token*, é informado o *Dictionary* provido pela função da *Gensim*.
- *num_topics*: o número de tópicos que a função deve extrair.
- *passes*: o número de repetições que a modelagem deve fazer no corpus.

Após a execução da função é gerado um conjunto de informações com os tópicos. Com esse conjunto é possível extrair os tópicos modelados pelo algoritmo. Um exemplo de execução é mostrado na figura 7.

Figura 7 – Exemplo de uso da implementação *LDA*.

```
ldamodel = gensim.models.ldamodel.LdaModel(corpus, num_topics, id2word, passes)
topics = ldamodel.print_topics(num_topics=10,num_words=3)
```

Fonte: Autor.

O retorno é o conjunto com os tópicos ordenados pelo nível de pontuação. Como exemplo, A figura 8 mostra um tópico onde 0 representa sua posição no *array* de tópicos gerados, entre aspas o termo em questão e o valor ao lado esquerdo de cada termo é a probabilidade utilizada para definir aquele termo naquele tópico.

Figura 8 – Exemplo de um tópico obtido.

```
(0, 0.069*"homem"+0.016*"perder"+0.014*"caso"+0.012*"sofrer"+0.009*"denunciar"+0.008*"empresa")
```

Fonte: Autor.

4.3.2 Nuvem de Palavras

As nuvens de palavras ou nuvens de termos foram uma das abordagens selecionadas para este trabalho, especialmente, por tornar muito sucinta a análise de grandes conjuntos de termos graficamente.

Para a implementação foi utilizado uma biblioteca de construção de nuvem de palavras chamada *WordCloud*. Essa biblioteca é implementada em *Python* e é amplamente utilizada na literatura.

O primeiro elemento para ser utilizado na construção das nuvens é o conjunto de dados que a função analisará. É selecionado como entrada então um arquivo *.txt* com os dados totalmente pré-processados nas etapas anteriores.

Figura 9 – Exemplo da captura de dados.

```
dataset = open("lematizados.txt",encoding='utf-8').read()
```

Fonte: Autor.

Nessa etapa é informado também um conjunto de parâmetros ligados ao número de termos e questões gráficas da nuvem, geralmente relacionados a coloração e forma.

Figura 10 – Exemplo da inserção da coleta de dados.

```
wc = WordCloud(background_color = "white",max_words = 50)
```

Fonte: Autor.

Após especificar os parâmetros, é executada a função para gerar a nuvem de palavras e exportar ela em uma imagem onde é especificado o nome.

Figura 11 – Exemplo de execucação da nuvem de palavra.

```
wc.generate(dataset)
wc.to_file("wordCloud.png")
```

Fonte: Autor.

Com isso foi gerada uma nuvem de palavras com todos os termos de todos os documentos após o processo de lematização.

4.4 Validação

O processo de validação dentro de várias áreas incluindo a modelagem de tópicos, pode ser tanto qualitativo, que mensura se o resultado é bom ou não, quanto quantitativo, que enumera os resultados (BIANCHINI, 2018).

O *LDA*, algoritmo escolhido neste trabalho, por ser não-supervisionado e probabilístico, dificulta a criação de métricas qualitativas e quantitativas (KAPADIA, 2019), que podem ser facilmente extraídas em outros algoritmos de mineração. Por ser uma metodologia muito presente na literatura entre trabalhos sobre modelagem de tópicos, foi escolhido para esse trabalho validar os resultados baseados em coerência.

A coerência de tópicos é o valor de quão um conjunto de termos dentro de um tópico estão relacionados (KAPADIA, 2019). Algumas métricas utilizam a representação de janelas de palavras, uma janela de palavras é uma estrutura que seleciona um conjunto de palavras do tópico. Como, por exemplo, um conjunto de tópicos $T = w_1, w_2, w_3$ pode conter como janelas $W = w_1, w_2$, $W = w_2, w_3$, etc.

As métricas escolhidas para validação neste trabalho são:

- *NPMI*: É obtida a partir da probabilidade de duas palavras dentro de um tópico estarem em uma mesma janela de palavras. Essa métrica é baseada em cada par de palavras em um tópico (PUERARI, 2019).
- *CV*: É uma variação de *NPMI* que utiliza a *sliding window*. Essa técnica funciona selecionando uma janela de palavras de tamanho N e desliza sobre outro conjunto baseado em N, por exemplo, a janela W = w2, w3, w4 após o processo resultaria em W = w3, w4, w5 (PUERARI, 2019).
- *UMass*: É gerada a partir da contagem de co-ocorrência de dois termos em todos os documentos (PUERARI, 2019).
- *UCI*: É gerada pela mesma técnica aplicada em *CV*, porém sem a normalização entre -1 e 1 (BIANCHINI, 2018).

As métricas foram geradas no site *Palmetto*. Esse sistema utiliza como referência para calcular a coerência entre termos de um tópico a ocorrência entre esses mesmos termos em artigos no wikipedia (BIANCHINI, 2018).

5 Resultados

Neste capítulo são apresentados os resultados obtidos das etapas desenvolvidas anteriormente. A seção 5.1 mostra os resultados da coleta dos dados A seção 5.2 discorre sobre o desempenho das técnicas utilizadas para pré-processamento. Na seção 5.3 são analisadas a execução das técnicas nuvem de palavras e modelagem de tópicos. A seção 5.4 faz considerações sobre os valores obtidos nas métricas de coerência.

5.1 Coleta de Tweets

O conjunto de *tweets* coletados sem nenhum pré-processamento durante os quatro anos propostos resultaram em um total de 29.738. A média foi de mais de 7000 *tweets* por ano, sendo que o periodo de 2018 destaca-se com maior número de coleta, 11.318.

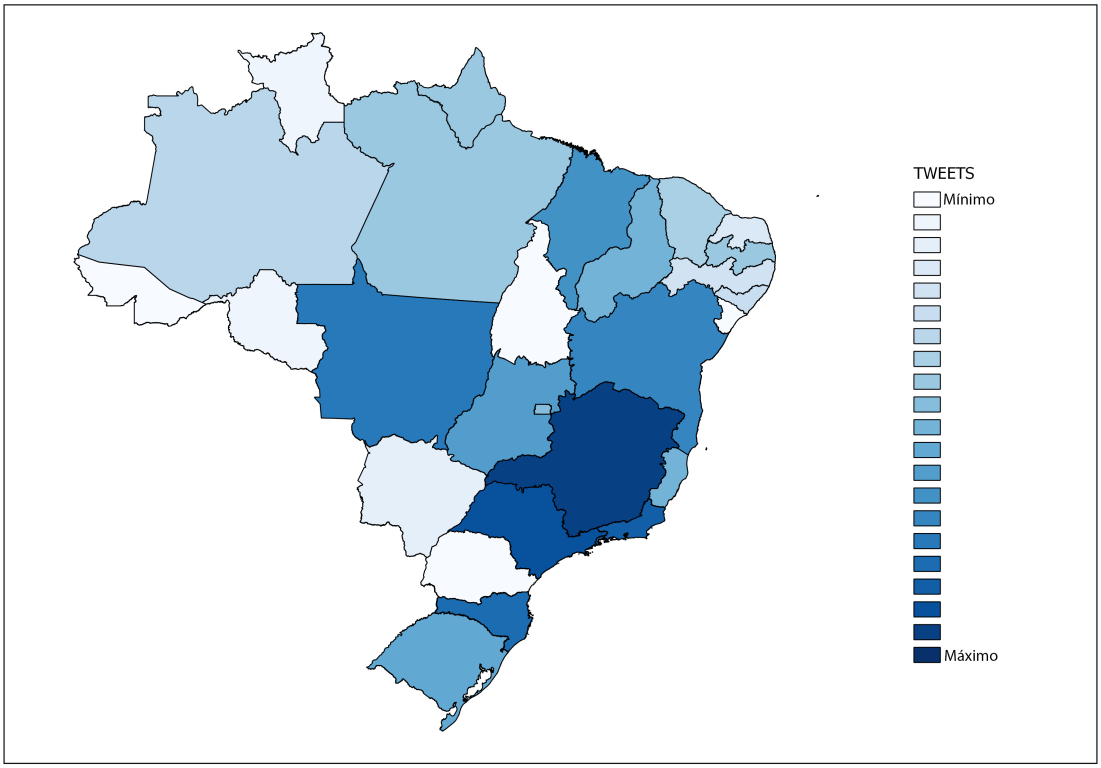
É possível observar, conforme a tabela 2, que alguns termos pesquisados são muito efetivos na busca, enquanto outra parcela apresenta a quase inexistência de *tweets* com os conjuntos pesquisados na rede social. Esse fator tem relação com os vários termos utilizados para definir o *Bullying* no trabalho, já que com as frequentes campanhas de conscientização do assédio no trabalho, ambos os termos destacam-se e são de maior conhecimento da população. Ambos os termos, "assédio", *"Bullying"* e "trabalho", obtiveram uma grande porcentagem sobre os outros termos. Cerca de 92,42% de todo o conjunto coletado em 2015, por exemplo.

Tabela 2 – Tweets coletados por ano

Termos	***2015***	***2016***	***2017***	***2018***
"bully trabalho"	32	30	58	75
"Bullying trabalho"	1466	1274	1107	1886
"assedio trabalho"	2808	1953	4242	6342
"bully emprego"	3	1	4	4
"Bullying emprego"	90	38	76	162
"bullyng emprego"	94	42	85	170
"assedio emprego"	159	168	372	667
"assedio callcenter"	2	2	5	3
"Bullying callcenter"	1	3	6	8
"assedio telemarketing"	83	36	96	39
"Bullying telemarketing"	13	3	4	8

A veiculação de ataques e depoimentos em redes sociais de situações de *Bullying* relacionados ao trabalho estão diretamente ligados ao acesso à *internet*, consequentemente, nos números da coleta. No Brasil o acesso à *internet* tem proporções diferentes conforme a região (CETIC, 2018) e isso pode afetar a coleta diretamente. A falta de conhecimento sobre *Bullying* no trabalho pode impactar a coleta, já que se não compreendidas as características dessa violência, não há denúncias nem relatos nas redes sociais. Esse impacto é visto no número de casos registrado em Pernambuco de 855 pelo TRT (TV GLOBO, 2019) contra mais 16,9 mil em São Paulo, estado onde o Ministério Público do Trabalho já investiu em fortes campanhas de conscientização (MINISTÉRIO PÚBLICO PúBLICO DO TRABALHO EM SãO PAULO, 2015).

Figura 12 – Mapa de calor do valor das coletas por estado.



Fonte: Autor.

Conforme mostrado na figura 12, a região sudeste do país, que detém também do maior acesso à *internet*, tem um número maior exposto de *tweets* coletados (CETIC, 2018). A região nordestina do país detém de um baixo número de conjunto de dados que pode ter uma relação com o baixo número de brasileiros conectados, sendo a região com o menor frequência de computadores com acesso à *internet* no país (CETIC, 2018).

Ainda há regiões com baixo número de coleta, mas com bom acesso à *internet*. Isso pode se dar pela questão que nessas regiões os trabalhadores associam relatar episódios

com represálias e a possibilidade do desemprego, como o Paraná (G1 PARANá, 2019). É importante destacar que a violência no trabalho é presente há um longo período na sociedade, já os estudos sobre o *Bullying* organizacional em âmbito de saúde e trabalho são um tema pesquisado a poucos anos (BARRETO, 2005).

Uma possível queda em 2016 do número de dados coletados pode ser veiculada ao medo do desemprego, (G1 ECONOMIA, 2017), bem como a diminuição do trabalho formal com carteira assinada, que vem perdendo força desde o segundo semestre de 2016 (AGÊNCIA BRASIL, 2017). Um trabalho não formal anula certos comportamentos, já que o trabalhador pode não estar inserindo dentro de um núcleo organizacional, diminuindo consideravelmente a possibilidade de abusos e violência por parte de outros trabalhadores de nível hierárquico maior.

5.2 Pré-processamento dos dados

Foi escolhido, a priori, um conjunto pequeno de técnicas muito utilizadas na literatura. Porém, conforme a qualidade dos resultados não atingiam os resultados esperados, foram acrescidas novas funções para filtrar e eliminar texto sem valor semântico para o trabalho. Com isso, um grande número de funções foram utilizadas, sendo que esta etapa foi agrupada em 3 conjuntos de técnicas.

5.2.1 Tratamento de caracteres e palavras

Essa é a primeira etapa, nela foram removidos os termos comuns de um *tweet* como *links*, *hashtags*, etc. Após esse processo há uma conversão das abreviações pelos seus termos representativos, além da remoção de caracteres não alfabéticos.

Durante essa etapa, um número massivo de *tweets* em inglês foi retornado, provavelmente pelos termos utilizados na busca serem muito utilizados em países onde o idioma é falado, assim também é aplicada uma função de detecção de idioma que elimina *tweets* em inglês. Todos os *tweets* que restaram desse processo que contém menos de cinco caracteres também foram removidos. A tabela 3 mostra os *tweets* afetados por esta etapa.

5.2.2 Tratamento de *Stopwords*

Na segunda etapa, a primeira função aplicada é a conversão em *tokens* dos termos presentes em cada *tweet*, que uma função que afeta todos os *tweets*. A cada conversão de *tweet*, cada *token* passa por um conjunto de funções que poderão implicar em uma possível remoção na frase. Todas as funções dessa etapa servem como verificações que validam se há relevância no *token* para os fins deste trabalho. A tabela 4 mostra o conjunto de *tweets* afetados nesta etapa.

Tabela 3 – Tweets Alterados na primeira etapa

Função	***Número de Tweets Afetados***
Remoção de *links*	13160
Remoção de *hashtags*	105
Remoção de citação de usuários	2500
Conversão de abreviações	3569
Remoção de caracateres não-alfabéticos	23751
Remoção de tweets com tamanho menor de 5	1026

O número afetado de *tweets* na segunda etapa é motivado em grande parte pela frequência de *stopwords* presentes em cada conjunto de *tokens*. A função de remoção de *Stopwords* que afetou 13079 *tweets* identifica cada termo como *Stopword* utilizando uma base de referência. Neste trabalho foram utilizadas 3 bases de dados, das quais 2 foram construídas pelo próprio autor e a terceira é uma base implementada pela biblioteca já utilizada em outras partes do código, a *NLTK*. Infelizmente a base da biblioteca contava apenas com 204 *Stopwords*, por isso houve a necessidade de construção das outras bases. Às bases juntas somam 514 *Stopwords*.

5.2.3 Lematização

Como o processo de lematização consiste em converter um *token* para uma forma comum entre suas variações, um grande número de *tokens* poderão repetir, portanto, como se trata de uma conversão, nessa etapa não há nenhuma remoção dos conjuntos de *tokens*.

O número de *tweets* após a segunda etapa foi de 16543, dos quais 15190 passaram pelo processo de lematização. Esse grande número já era esperado visto que no português, um verbo pode vir a ter inúmeras conjugações motivadas por tempo verbal, pronome e outros elementos presentes em uma frase. Foi necessário lematizar algumas palavras diretamente no código por problemas de implementação na biblioteca, que retornavam flexões inexistentes no português.

Tabela 4 – *Tweets* Alterados na segunda etapa

Função	***Número de Tweets Afetados***
Remoção de risadas	110
Remoção de *Stopwords*	19989
Remoção do *tweets* com menos de 3 tokens	8608

Após o processo de lematização foram constatadas algumas anomalias na coleta. Devido ao fato da palavra *Bullying* ser frequente em campanhas escolares, foi decidido remover todos os *tweets* que continham a palavra "*escola*". O tamanho do corpus após essa remoção foi de 14369.

5.3 Mineração de Dados

Na etapa de mineração, o objetivo de busca de padrões foi executada utilizado LDA e para apoiar o entendimento e a avaliação do pré-processamento, foram geradas nuvens de palavras.

5.3.1 Nuvem de Palavras

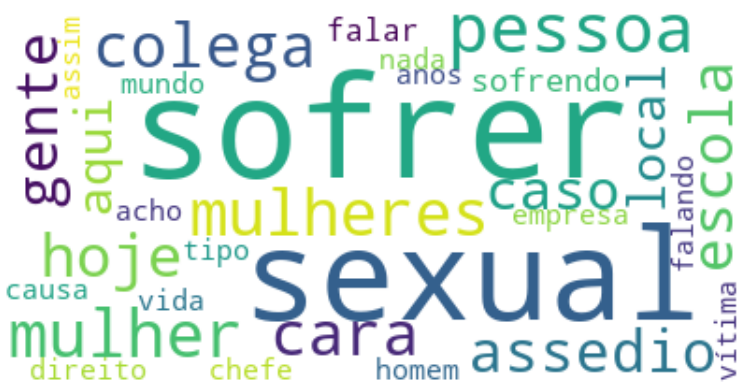
Em busca de sintetizar o entendimento geral da coleta e avaliar a qualidade do pré-processamento, foram criadas nuvens de palavras durante cada etapa de limpeza. O tamanho de cada palavra nas figuras descreve a frequência, portanto, a relevância de cada termo sobre todo o conjunto coletado.

Figura 13 – Nuvem palavras dos textos coletados

Fonte: Autor.

A figura 13 mostra os termos coletados sem qualquer filtragem e correção. Houve uma frequência muito grande de termos como: *"pic"*, *"bit"* e *"https"*. Essas palavras estão diretamente ligadas ao compartilhamento de *links* e a imagens anexadas ao *tweets*. Também há uma presença grande de conectivos que não são de relevância para o algoritmo de modelagem de tópicos. Por mais que já esperada uma maior frequência de palavras insignificantes para a mineração ainda é possível observar algumas duplas de palavras como *"assedio moral"*, que tratam do assunto esperado pela coleta e *"local de"*, *"ambiente de"* que referenciam a relação entre assédio laboral e ambientes de trabalhos.

Figura 14 – Nuvem palavras dos textos pré-processados sem lematização



Fonte: Autor.

A figura 14 representa o conjunto de termos após as técnicas de pré-processamento excluindo a lematização. A nuvem de palavras dessa etapa retornou valores de muita significância no assunto abordado como os termos *"vítima"*, *"sofrer"*, *"empresa"* e *"chefe"*. São palavras muito frequentes no assédio laboral e presume o fenômeno já abordado em (FREITAS, 2007) sobre a relação do assédio em violência com o nível hierárquico. Também se observa os termos *"colega"* e *"local"* que podem descrever a existência do assédio no local de trabalho bem como possíveis papéis de um colega em um episódio. Por fim destacam-se as palavras *"mulher"*, *"mulheres"*, *"sofrendo"* e *"sofrer"* que motivaram a utilização da técnica de lematização. A palavra *"escola"* também motivou a remoção de termos durante a etapa de lematização, já comentados acima.

A figura 15 foi gerada a partir de todo o conjunto de termos após todas as etapas de pré-processamento. Ela representa todos os dados que serão consumidos pelo algoritmo de mineração com todas as etapas executadas.

Figura 15 – Nuvem palavras dos textos pré-processados sem lematização

Fonte: Autor.

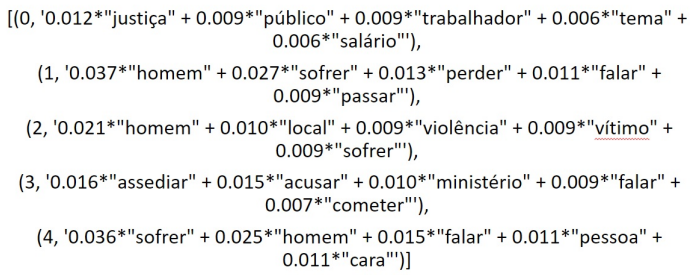
É possível notar que não há flexões de um mesmo termo evidentes na etapa anterior onde não havia acontecido ainda o processo de lematização. A palavra *"amigo"* não vista antes nas outras nuvens fortalece o indício de possíveis papéis em um episódio de assédio. Houve um aumento nas palavras *"homem"* e *"cara"* que podem indicar uma relação sexista em casos de assédio no local de trabalho. Com essa nuvem foi possível confirmar que o trabalho do algoritmo de lematização foi bastante efetivo.

5.3.2 Modelagem de tópicos

O processo de implementação para a modelagem dos tópicos foi feito utilizando um conjunto de algoritmos disponibilizados em *Python* tendo como base para a geração o *LDA* implementado pela biblioteca *Gensim*.

A primeira abordagem de testes teve como escolha das variáveis o número fixo de 5 tópicos e testou algumas variações do atributo *passes* nos valores 50, 75, 100, 150, 175 e 200. Esse atributo diz ao algoritmo quantas vezes ele deve manter a repetição pelo corpus (GENSIM, 2021). Não foram obtidos resultados satisfatórios com esse conjunto de atributos então foi aplicada uma abordagem comum em vários trabalhos da literatura. A figura 16 mostra um exemplo de tópicos gerados com apenas a variação em *passes*. Os resultados obtidos com essa abordagem foram muito genéricos e não-interpretáveis.

Figura 16 – Exemplo de tópico obtido com a abordagem de variações de *passes*



Fonte: Autor.

Ao passo que a variação de *passes* não foi efetiva, foi escolhido variar o número de tópicos no algoritmo. Para definir os números testáveis foi feita uma seleção baseada no valor de coerência *Cv* de cada conjunto de tópicos indo de 4 a 28 tópicos. Esta abordagem utilizando valor de coerência não representa garantia, visto que o algoritmo utilizado se encaixa como não-supervisionado, portanto, as métricas não são tão precisas conforme abordado em (CASTRO, 2020) e (KAPADIA, 2019). Para gerar as métricas foi utilizada a função de cálculo dos valores de coerência presente na biblioteca *Gensim* utilizada na etapa de validação do número de tópicos. Na tabela 5 são mostrados os valores coletados.

Foram coletados os conjuntos de números iguais a 26, 10, 22, 16 e 6. Foram retornados os valores dos 5 primeiros tópicos.

Tabela 5 – Valores de coerência obtidos com a variação dos números de tópicos

Número de tópicos	***Cv***
2	0.3626
4	0.3685
6	0.3726
8	0.3604
10	0.3841
12	0.3691
14	0.3627
16	0.3774
18	0.3419
20	0.3663
22	0.3794
24	0.3517
26	0.3901
28	0.3439

Houve uma proximidade muito grande entre os números de tópicos escolhidos. Dentre os cinco tópicos de melhor valor três deles detém de métricas praticamente iguais. Também ficou claro que o aumento do número de tópicos não resultariam em melhores valores de coerência já que o tópico dez teve a segunda melhor pontuação.por isso e pela legibilidade fornecida entre os tópicos de menor conjunto ficaram escolhidos o conjunto com dez e com seis tópicos.

O conjunto de tópicos obtidos com o valor 10 é mostrado na tabela 6. Esse conjunto obteve bons resultados e foi possível identificar os quatro melhores tópicos. As palavras *"sofrer"*, *"abuso"* e *"empresa"* indicam um possível ambiente de trabalho nocivo, já os termos *"perder"*, *"caso"* e *"denúncia"* podem indicar a possibilidade de fracasso em combater situações de *Bullying* laboral. O segundo grupo contém os termos *"sofrer"*, *"passar"*, *"cargo"* e *"macho"* que indicam possivelmente assédio hierárquico veiculado por homens. Os termos *falar*, *"pessoa"*, *"conseguir"* e *"denunciar"* podem estar ligados a situações positivas de denúncia. O termo *"conseguir"* aparenta indicar certa dificuldade em fazer alguma denúncia, isso reforça o trabalho de (NASCIMENTO, 2004) onde é dito que fatores internos como o despreparo da equipe de recursos humanos e uma cultura interna ligada ao assédio podem dificultar medidas corretivas bem como relatos.

O terceiro tópico propõe com os termos *"acusar"*, *"caso"* e *"operador"* relatos sobre acusações, já os termos *"agente"* e *"condição"* são relacionados ao artigo 216A da lei No

Figura 17 – Exemplo de *tweet* sobre abuso de superiores

Fonte: Autor.

10.224 muito citado em *tweets* coletados que diz: "Constranger alguém com o intuito de obter vantagem ou favorecimento sexual, prevalecendo-se o agente da sua condição de superior hierárquico", esse conjunto de termos indica a possibilidade de acusações contra pessoas que ocupam cargos superiores. A figura 17 mostra um dos *tweets* que demostram o desrespeito de superiores.

Tabela 6 – Conjunto de tópicos com K = 10

Assunto	***Tópicos***
Intenção de denúncia mas com perspectiva do chefe ganhar	0.056*"homem"+0.026*"sofrer"+0.023*"cara"+ 0.021*"perder"+0.016*"querer"+0.012*"ganhar"+ 0.011*"chefe"+0.011*"ficar"+0.010*"falar"+ 0.010*"passar"
Denúncia de Assédio hierarquico contra homem	0.072*"homem"+0.042*"sofrer"+0.017*"perder"+ 0.013*"falar"+0.012*"passar"+0.011*"cargo"+ 0.009*"macho"+0.009*"pessoa"+0.009*"conseguir"+ 0.008*"denunciar"
Acusação de operadores contra agentes de condição superior de sua empresa	0.048*"homem"+0.019*"acusar"+0.017*"operador"+ 0.013*"brasil"+0.012*"caso"+0.012*"saber"+ 0.012*"questão"+0.012*"agente"+0.011*"condição"+ 0.011*"empresa"
Comparação do assédio entre as profissões de atendente e ator	0.025*"sofrer"+0.017*"homem"+0.015*"atendente"+ 0.014*"respeito"+0.012*"saber"+0.012*"ficar"+ 0.011*"ator"+0.011*"pressão"+0.011*"falta"+ 0.010*"ninguém"
Desconhecido	0.019*"situaçõe"+0.015*"aceitar"+0.014*"colega"+ 0.012*"servidor"+0.012*"tipo"+0.012*"machismo"+ 0.012*"banco"+0.011*"chamar"+0.011*"parte"+ 0.010*"sofrer"

É possível notar no quarto conjunto os termos *"sofrer"*, *"atendente"*, *"ator"*, *"falta"*, *"pressão"* e *"ninguém"* que indicam possivelmente uma comparação entre o assédio sofrido por atendentes enquanto o ator que é denunciado por cometer o assédio não sofre pressão de ninguém.

O termo *"atendente"* bem como um ator já apareceram em conjuntos de tópicos diferentes já citados, nos tópicos anteriores onde o termo *"atendente"* também é dito que há um desrespeito com esses profissionais bem como o tópico sobre o ator em questão indica um caso de grande repercussão no período coletado, ambos os conjuntos possivelmente justificam uma comparação feita nesse período.

O último tópico apresentado na tabela 6 aparenta dizer sobre possíveis situações de machismo sofridas por servidores públicos e a aceitação desse assédio no setor de trabalho. Porém, o grupo não conseguiu apresentar um conjunto de termos fortes o suficiente para indicar exatamente o assunto abordado por isso foi tratado como desconhecido.

Na tabela 7 são mostrados os conjuntos gerados com 6 tópicos. É possível destacar no primeiro grupo as palavras *"denúncia"*, *"sofrer"* e *"abuso"* que presumem relatos e denúncias no ambiente de trabalho. Os termos *"perder"* e *"caso"* podem indicar uma acusação falha, muito sustentada pelos termos *"empresa"* e *"saber"*, onde a empresa tendo ciência dos atos e não aplicando medidas corretivas tende a sustentar o ambiente nocivo e dificultar as denúncias (FREITAS, 2007).

No segundo grupo os termos *"sofrer"*, *"local"* indicam possivelmente *Bullying* no trabalho bem como os termos *"ministério"*, *"querer"*, *"falar"* e *"empresa"* indicam um interesse em denunciar a empresa ao ministério público.

No terceiro tópico o conjunto de termos *"falar"*, *"assediar"* e *"cara"* indicam o possível papel de agressor e uma situação de *Bullying* conforme dito em (SCHREIBER et al., 2015). Os termos *"alguém"* e *"acusar"* indicam denúncias que tiveram como resultado o rebaixamento de cargo do agressor conforme expresso nos termos *"perder"* e *"cargo"*.

Os termos *"operador"*, *"sofrer"*, *"oferecer"* e *"risco"* presentes no quarto tópico possivelmente indicam uma situação frequente de assédio contra operadores. O conjunto de termos *"querer"*, *"processar"* e *"empresa"* possivelmente dizem respeito a vontade de denunciar a empresa por expor os operadores a esse ambiantes nocivos.

O último grupo de termos possivelmente indica a presença de situações de assédio veiculas as jornalistas com os termos *"passar"*, *"sofrer"*, bem como a presença de *"função"*, *"trabalho"* e *"receber"* que podem indicar o assédio direcionado a quem exerce essa função.

Os resultados obtidos mostrados foram se tornando mais legíveis conforme a diminuição do número de tópicos utilizados no algoritmo. Os conjuntos baseados em 10 e em 6 tópicos careceram de pouca ou nenhuma consulta a base coletada para terem seus assuntos definidos e foram considerados os mais legíveis pelo autor.

Tabela 7 – Conjunto de tópicos com K = 6

Assunto	***Tópicos***
Perder caso de denúncia	0.069*"homem"+0.016*"perder"+0.014*"caso"+ 0.012*"sofrer"+0.011*"denúncia"+0.009*"denunciar"+ 0.008*"empresa"+0.007*"abuso"+0.007*"falar"+ 0.006*"saber"
Intenção de denúncia da empresa no ministério	0.022*"sofrer"+0.011*"local"+0.011*"perder"+ 0.009*"pessoa"+0.008*"ministério"+0.008*"querer"+ 0.007*"falar"+0.007*"feminista"+0.006*"cara"+ 0.006*"empresa"
Perda de cargo por acusação de assédio	0.028*"falar"+0.016*"assediar"+0.014*"gente"+ 0.013*"trabalho"+0.012*"cara"+0.010*"sofrer"+ 0.009*"perder"+0.009*"cargo"+0.009*"alguém"+ 0.008*"acusar"
Intenção de operador processar empresa	0.011*"operador"+0.010*"risco"+0.010*"processar"+ 0.010*"querer"+0.010*"sofrer"+0.007*"oferecer"+ 0.007*"público"+0.006*"continuar"+0.006*"causa"+ 0.006*"empresa"
Jornalistas sofrem assédio enquanto trabalham	0.012*"crime"+0.011*"jornalista"+0.010*"passar"+ 0.008*"funcionário"+0.008*"achar"+0.007*"sofrer"+ 0.007*"função"+0.007*"trabalho"+0.007*"receber"+ 0.007*"pessoa"

5.4 Validação

Para o processo de validação todos os tópicos de cada conjunto foram lançados na plataforma *Palmetto*. O resultado das métricas de coerência de cada conjunto de tópicos serviram de apoio para validações e considerações a respeito do resultado da modelagem.

A métrica de *Cv* apresentada é a gerada no processo de seleção dos tópicos já que com ela comportamentos inesperados são possíveis na ferramenta *Palmetto* conforme mostrado em (GITHUB, 2017).

Tabela 8 – Métricas de coerências *Cv* e *UCI* de 10 e 6 tópicos

Número de tópicos	***CV***	***UCI***
10	0,3841	-0,88538695
6	0.3726	-1,23994011

Conforme representados na tabela 8 os dois tópicos são próximos em seu valor *Cv* não possibilitando conclusões a respeito da métrica. Já em *UCI* onde não há normalização dos valores (BIANCHINI, 2018) é possível observar uma melhor colocação no conjunto com 10 tópicos.

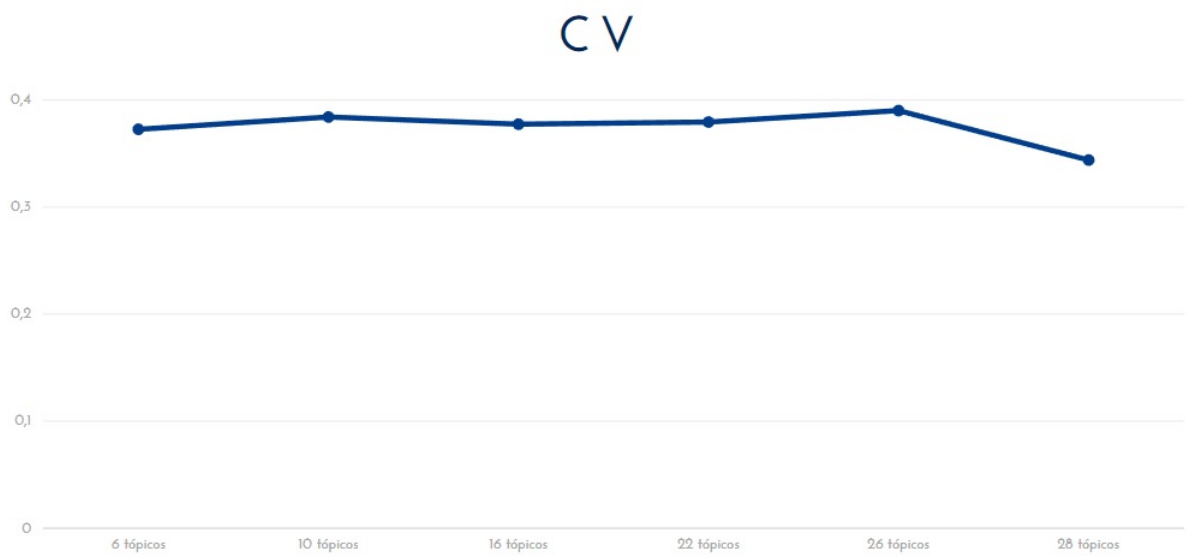
Tabela 9 – Métricas de coerências de *NPMI* e *UMass* de 10 e 6 tópicos

Número de tópicos	***NPMI***	***UMass***
10	-0,03204322	-12,20831433
6	-0,04410811	-11,62244757

A tabela 9 representa os dois conjuntos selecionados com as métricas remanescentes. Nota-se um destaque no conjunto com seis tópicos obtendo melhor métrica em *NPMI*. Para o conjunto de dez tópicos houve melhor métrica em *UMass*, destaca-se ainda as métricas *Cv* e *UMass* que não resultaram em uma grande diferença do entre os dois conjuntos de tópicos apresentados.

Pelas métricas apresentadas o número de tópicos com melhor coesão seria o gerado com dez, porém o conjunto não atingiu majoritariamente bom desempenho em todas as métricas, diferente de algumas obras na literatura como (BIANCHINI, 2018).

Figura 18 – Gráfico dos dados coletados de *Cv*



Fonte: Autor.

A figura 18 mostra o gráfico que representa os melhores valores da métrica *Cv* já apresentados na tabela 5 e o último número selecionado. Há alguns aumentos no conjunto com dez e vinte seis itens, mas seguidos de quedas e ao final do intervalo avaliado uma queda maior com o valor de vinte oito tópicos. Com isso é possível notar na figura que o crescimento do número de tópicos não impactou em melhorias nas medidas que avaliam coesão.

6 Conclusão

Foram apresentados neste trabalho características, definições e consequências sobre violência relacionada ao mercado de trabalho. O processo de desenvolvimento foi focado em utilizar conteúdos expressos em redes sociais tendo como base o *Twitter* onde há aspectos interessantes para coleta e análise.

O objetivo deste trabalho foi coletar e analisar os *tweets* sobre violência relacionada com o trabalho. Para a execução foram utilizadas as técnicas de mineração de textos: coleta de *Tweets*, pré-processamento, análise utilizando modelagem de tópicos, nuvem de palavras e validação. LDA foi o algoritmo utilizado para a modelagem de tópicos.

Mesmo não sabendo qual o papel exato em situações de violência e trabalho, é possível concluir que nos relatos do *Twitter*, homens tem uma relação de grande impacto em casos de *Bullying* laboral. Os termos *"homem"* e *"cara"* se destacaram na nuvem de palavras por terem uma grande frequência nos *tweets* coletados bem como a presença em 6 tópicos dos dois conjuntos apresentados neste trabalho.

É possível concluir que os usuários do *Twitter* enxergam uma relação de proximidade entre assédio e hierarquia nas empresas. Ficou evidente pela presença dos termos *"chefe"* e *"cargo"* em parte dos tópicos e na nuvem de palavras.

Conclui-se que há uma certa desconfiança por parte dos usuários da rede social utilizada neste trabalho quanto as empresas em casos de assédio. Dos tópicos gerados que mencionaram *"empresa"* também seguiam em conjunto com o termo *"saber"*, essas palavras juntas podem indicar um possível descomprometimento da empresa com casos de *Bullying*.

6.1 Trabalhos futuros

Neste trabalho foram coletados os dados da rede social *Twitter* buscando por *tweets* relacionados a violência e trabalho. A utilização de uma base de dados baseada em casos denunciados que passaram por entidades públicas como registro de denúncias no ministério público ou processos que correram em tribunais de justiça poderiam fornecer dados mais esclarecedores como as profissões e papéis envolvidos.

A análise principal neste trabalho se dá pelo algoritmo LDA de modelagem de tópicos, a utilização de outros métodos de mineração, se baseadas em outras bases de coleta como mencionado, surtiriam um efeito minucioso na análise podendo agrupar papéis, profissões e setores que detém de maior envolvimento em casos de assédio.

Outra abordagem poderia trabalhar com outros algoritmos de modelagem de tópicos para fazer análise comparativa entre as métricas de coerência geradas por cada um e por fim obter os tópicos com o algoritmo de melhor desempenho, portanto, com maior valor de coesão.

O tema e as conclusões aqui apresentados poderiam ser acrescidos de uma análise classificatória dos sentimentos expostos. A relação dos sentimentos perante o tema abordado podem expor um conjunto de informações não cobertas neste trabalho quanto a visão dos envolvidos nos relatos.

Outra abordagem utilizando as mesmas técnicas aplicadas neste trabalho, mas se concentrado em extrair também as respostas dos *tweets* extraídos bem como as *hashtags* e os *emojis* poderiam trazer um maior conjunto de informações relacionadas aos dados coletados.

Referências

AGÊNCIA BRASIL. *Trabalho informal puxou aumento da taxa de ocupação, diz Ipea*. 2017. Disponível em: <http://agenciabrasil.ebc.com.br/economia/noticia/2019-03/trabalho-informal-puxou-aumento-da-taxa-de-ocupacao-diz-ipea>. Acesso em: 14 jan 2020. Citado na página 40.

ALMEIDA, R. J. de A. Estudo da ocorrência de cyberbullying contra professores na rede social twitter por meio de um algoritmo de classificação bayesiano. *Texto Livre: Linguagem e Tecnologia*, v. 5, n. 1, p. 77–83, 2012. Citado 4 vezes nas páginas 11, 13, 24 e 27.

AMADO, J.; MATOS, A.; PESSOA, T.; JAGER, T. Cyberbullying: Um desafio à investigação e à formação. revistas.rcaap.pt, p. 23 – 29, 06 2016. Disponível em: <https://revistas.rcaap.pt/interaccoes/article/view/409/363>. Citado 3 vezes nas páginas 11, 12 e 13.

ANABUKI, L. N. D. C. AssÉdio moral organizacional como dano ao meio ambiente do trabalho e a atuaÇÃo do mpt: Um retrato empÍrico de um encontro necessÁrio. IDP, 2016. Disponível em: <http://dspace.idp.edu.br:8080/xmlui/handle/123456789/2023?show=full>. Citado na página 13.

ARANTES, R. B. Direito e política: o ministério público e a defesa dos direitos coletivos. *Revista brasileira de ciências sociais*, SciELO Brasil, v. 14, n. 39, p. 83–102, 1999. Citado na página 11.

ARAÚJO, M. S. G. de. *Preditores individuais e organizacionais de bullying no local de trabalho*. Tese (Doutorado) — Universidade do Minho (Portugal), 2009. Citado na página 17.

AZEVEDO, J. C.; MIRANDA, F. A. de; SOUZA, C. H. M. de. Reflexões a cerca das estruturas psíquicas e a prática do Ciberbullying no contexto da escola. Intercom: Revista Brasileira de Ciências da Comunicação, 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1809-58442012000200013&lng=en&nrm=iso&tlng=pt>. Citado na página 20.

BARION, E. C. N.; LAGO, D. Mineração de textos. *Revista de Ciências Exatas e Tecnologia*, v. 3, n. 3, p. 123–140, 2015. Citado na página 23.

BARRETO, M.; HELOANI, R. Violência, saúde e trabalho: a intolerância e o assédio moral nas relações laborais. *Serviço Social & Sociedade*, SciELO Brasil, n. 123, p. 544–561, 2015. Citado 2 vezes nas páginas 11 e 12.

BARRETO, M. M. S. *Assédio moral: a violência sutil-análise epidemiológica e psicossocial no trabalho no Brasil*. Tese (Doutorado) — Pontifícia Universidade Católica de São Paulo, 2005. Citado na página 40.

BATISTA, G. E. d. A. P. et al. *Pré-processamento de dados em aprendizado de máquina supervisionado*. Tese (Doutorado) — Universidade de São Paulo, 2003. Citado na página 21.

BIANCHINI, L. *Análise exploratória dos tópicos no Stack Overflow usando LDA (Latent Dirichlet Allocation)*. 2018. Monografia (Bacharel em Ciência da Computação), Universidade Federal da Fronteira Sul, Brazil. Citado 6 vezes nas páginas 22, 23, 32, 37, 48 e 49.

CASTELLS, M. *A Galáxia Internet: reflexões sobre a Internet, negócios e a sociedade*. [S.l.]: Editora Zahar, 2003. Citado na página 20.

CASTRO, B. Y. S. de. *Como modelar tópicos através de Latent Dirichlet Allocation (LDA) através da biblioteca Gensim*. 2020. Disponível em: <https://medium.com/somos-tera/como-modelar-t%C3%B3picos-atrav%C3%A9s-de-latent-dirichlet-allocation-lda-atrav%C3%A9s-da-biblioteca-gensim-1fa17357ad4b>. Acesso em: 14 jan 2021. Citado na página 44.

CETIC. *CETIC*. 2018. Disponível em: <https://www.cetic.br/tics/domicilios/2018/domicilios/A4/>. Acesso em: 14 jan 2020. Citado na página 39.

CÔRTES, S. d. C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados–funcionalidades, técnicas e abordagens. 2002. *Acesso em 2020*, v. 26, 2016. Citado 2 vezes nas páginas 22 e 23.

DANTAS, E. R. G.; ALMEIDA, J. C.; JÚNIOR, P.; LIMA, D. S. de; PESSOA-UNIPÊ, J. O uso da descoberta de conhecimento em base de dados para apoiar a tomada de decisões. *V Simpósio de Excelência em Gestão e Tecnologia-SEGeT*, v. 1, p. 50–60, 2008. Citado 2 vezes nas páginas 20 e 22.

FACEBOOK. *Central de Ajuda*. 2018. Disponível em: <https://www.facebook.com/help/?ref=pf>. Acesso em: 15 nov 2018. Citado na página 19.

_____. *Central de Prevenção ao Bullying*. 2018. Disponível em: <https://www.facebook.com/safety/bullying>. Acesso em: 12 jan 2019. Citado na página 20.

_____. *View all Press Releases Facebook Reports Third Quarter 2018 Results*. 2018. Disponível em: <https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-Third-Quarter-2018-Results/default.aspx>. Acesso em: 17 nov 2018. Citado 2 vezes nas páginas 19 e 20.

FALEIROS, T. d. P.; LOPES, A. d. A. Modelos probabilisticos de topicos: desvendando o latent dirichlet allocation. 2016. *Instituto de Ciencias Matemáticas e de Computação (ICMC/USP), Sao Carlos–SP*, 2020. Citado na página 23.

FERREIRA, J.; OLIVEIRA, H. G.; RODRIGUES, R. Improving nltk for processing portuguese. In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FUER INFORMATIK. *8th Symposium on Languages, Applications and Technologies (SLATE 2019)*. [S.l.], 2019. Citado 2 vezes nas páginas 30 e 31.

FRANÇA, T. C.; FARIA, F.; MICELI, C.; RANGEL, F.; OLIVEIRA, J. Big social data: Princípios sobre coleta, tratamento e análise de dados sociais. *Anais do SBBD (Porto Alegre). SBC*, p. 1–40, 2014. Citado na página 21.

FREITAS, C.; NEDEL, L. P.; GALANTE, R.; LAMB, L. C.; SPRITZER, A. S.; FUJII, S.; OLIVEIRA, J. P. M. de; ARAÚJO, R. M.; MORO, M. M. Extração de conhecimento e análise visual de redes sociais. *SEMISH-Seminário Integrado de Software e Hardware, Belém do Pará, Brasil, SBC*, p. 106–120, 2008. Citado na página 20.

FREITAS, M. E. d. Assédio moral e assédio sexual: faces do poder perverso nas organizações. *Revista de administração de Empresas*, SciELO Brasil, v. 41, n. 2, p. 8–19, 2001. Citado na página 17.

FREITAS, M. E. D. Quem paga a conta do assédio moral no trabalho? RAE-eletrônica, 2007. Disponível em: <http://www.redalyc.org/html/2051/205114655011>. Citado 6 vezes nas páginas 12, 13, 17, 18, 43 e 47.

G1 ECONOMIA. *Desemprego fica em 12% no 4º trimestre de 2016 e atinge 12,3 milhões*. 2017. Disponível em: <https://g1.globo.com/economia/noticia/desemprego-fica-em-12-no-4-trimestre-de-2016.ghtml>. Acesso em: 14 jan 2020. Citado na página 40.

G1 PARANá. *TRT do Paraná é o 4º com maior número de ações por assédio moral, diz CNJ*. 2019. Disponível em: <https://g1.globo.com/pr/parana/>. Acesso em: 14 jan 2020. Citado na página 40.

GENSIM. *LdaModel*. 2021. Disponível em: <https://radimrehurek.com/gensim/models/ldamodel.html>. Acesso em: 14 jun 2021. Citado na página 44.

GITHUB. *Coleta de Denúncias*. 2017. Disponível em: <https://github.com/dice-group/Palmetto/issues/12>. Acesso em: 22 may 2021. Citado na página 48.

GONÇALVES, L. S. M. *Categorizaçao em text mining*. Tese (Doutorado) — Universidade de São Paulo, 2002. Citado na página 20.

GREATLEARNING. *Understanding Latent Dirichlet Allocation (LDA)*. 2020. Disponível em: <https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation>. Acesso em: 12 nov 2020. Citado 3 vezes nas páginas 7, 32 e 33.

HELOANI, R. Assédio moral: um ensaio sobre a expropriação da dignidade no trabalho. *RAE-eletrônica*, SciELO Brasil, v. 3, n. 1, 2004. Citado na página 17.

HENRIQUE, J. *GetOldTweets*. 2019. Disponível em: <https://github.com/Jefferson-Henrique/GetOldTweets-python>. Citado na página 29.

INSTAGRAM. *Central de Ajuda*. 2018. Disponível em: <https://help.instagram.com/>. Acesso em: 15 nov 2018. Citado na página 19.

KAPADIA, S. *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. 2019. Disponível em: <https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0>. Citado 2 vezes nas páginas 37 e 44.

KOMESU, F. Visões da língua (gem) em comentários sobre internetês não é língua portuguesa. *Filologia e Linguística Portuguesa*, Universidade de São Paulo, n. 8, p. 425–437, 2006. Citado 2 vezes nas páginas 21 e 30.

LUCCA, J. D.; NUNES, M. d. G. V. *Lematização versus Stemming*. 2002. Disponível em: <http://www.nilc.icmc.usp.br/nilc/download/lematizacao_versus_steming.pdf>. Citado na página 30.

MARCONDES, A. L. N.; DIAS, R. Características do bullying como um tipo de assédio moral nas organizações. *Revista Pensamento Contemporâneo em Administração*, Universidade Federal Fluminense, v. 5, n. 1, p. 80–87, 2011. Citado na página 17.

MARCONDES, R. R. *Análise dos sentimentos expressos na rede social Twitter em relação aos filmes indicados ao Oscar 2017*. 2017. Monografia (Bacharel em Sistemas de Informação), Universidade Federal de Uberlândia, Brazil. Citado na página 21.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. *Techinical Report*, v. 209, p. 4, 2003. Citado na página 22.

MINISTÉRIO PÚBLICO PúBLICO DO TRABALHO EM SãO PAULO. *MPT em São Paulo lança campanha contra assédio moral*. 2015. Disponível em: <http://www.prt2.mpt.mp.br/252-mpt-em-sao-paulo-lanca-campanha-contra-assedio-moral>. Acesso em: 14 jan 2020. Citado na página 39.

MINISTéRIO PúBLICO FEDERAL. *Dia de Combate ao Assédio Moral: MPF promove campanha para prevenir e enfrentar a prática*. 2018. Disponível em: <http://www.mpf.mp.br/pgr/noticias-pgr>. Acesso em: 17 nov 2018. Citado na página 13.

MOTHÉ, C. B. O assédio moral nas relações de trabalho. *Revista de Direito Trabalhista*, v. 12, n. 3, p. 12–3, 2006. Citado na página 18.

NASCIMENTO, S. A. M. O assédio moral no ambiente do trabalho. Revista LTR, 2004. Citado 3 vezes nas páginas 17, 18 e 45.

NAVARRO, V. L.; PADILHA, V. Dilemas do trabalho no capitalismo contemporâneo. *Psicologia & Sociedade*, SciELO Brasil, v. 19, n. spe, p. 14–20, 2007. Citado na página 11.

NETO, A. A. L. Bullying - comportamento agressivo entre estudantes. Jornal de Pediatria, 2005. Disponível em: <http://www.scielo.br/pdf/jped/v81n5s0/v81n5Sa06.pdf>. Citado na página 16.

NUNES, A. V.; FREITAS, C. O.; PARAISO, E. C. Detecção de assédio moral em e-mails. In: *I Student Workshop on Information and Human Language Technology, São Carlos. POA: SBC*. [S.l.: s.n.], 2009. v. 1, p. 01–05. Citado 2 vezes nas páginas 25 e 27.

OLIVEIRA, A. B. de; MATHEUS, R. F.; PARREIRAS, F. S.; PARREIRAS, T. S. et al. Análise de redes sociais como metodologia de apoio para a discussão da interdisciplinaridade na ciência da informação. *Ciência da Informação*, v. 35, n. 1, 2006. Citado na página 24.

PALÁCIOS, M.; REGO, S. Bullying: mais uma epidemia invisível? Revista Brasileira de Educação Médica, 2006. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-55022006000100001>. Citado na página 16.

PORTO, R. M. A. B.; BAX, M. P.; FERREIRA, L. G. da F.; SILVA, G. C. Análise de sentimento sobre veÍculos em redes sociais. XIII Encontro Nacional de Pesquisa em Ciência da Informação, 2012. Disponível em: <http://enancib.ibict.br/index.php/enancib/xiiienancib/paper/viewFile/3863/2986>. Citado na página 20.

PUERARI, I. *Análise exploratória sobre registros eletrônicos de saúde do setor de unidade de terapia intensiva utilizando modelagem de tópicos*. 2019. Monografia (Bacharel em Ciência da Computação), Universidade Federal da Fronteira Sul, Brazil. Citado na página 37.

QUONIAM, L.; TARAPANOFF, K.; JÚNIOR, R. d. A.; ALVARES, L. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o brasil. *Ciência da informação*, SciELO Brasil, v. 30, n. 2, p. 20–28, 2001. Citado 2 vezes nas páginas 22 e 23.

RECUERO, R. *Redes sociais na internet*. [S.l.]: Editora Meridional, 2009. Citado na página 19.

ŘEHŮŘEK, R.; SOJKA, P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010. p. 45–50. <http://is.muni.cz/publication/884893/en>. Citado na página 34.

SANTOS, R. et al. Conceitos de mineração de dados na web. *XV Simpósio Brasileiro de Sistemas Multimídia e Web, VI Simpósio Brasileiro de Sistemas Colaborativos–Anais, MM Teixeira, CAC Teixeira, FAM Trinta, e P. PM Farias, Eds*, p. 81–124, 2009. Citado na página 22.

SANTOS, R. E.; SOUZA, E. P.; CORREIA-NETO, J. S.; MAGALHÃES, C. V.; VILAR, G. Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. *Revista de Sistemas e Computação-RSC*, v. 4, n. 2, 2015. Citado na página 32.

SCHREIBER; CASTRO, F. C. de; ANTUNES; CRISTINAR, M. Cyberbullying: do virtual ao psicológico. scielopepsic, p. 109 – 125, 01 2015. Disponível em: <https://revistas.rcaap.pt/interaccoes/article/view/409/363>. Citado 3 vezes nas páginas 12, 16 e 47.

SILVA, G. L. A. da. *Text Mining, um estudo a partir da rede social Twitter*. 2013. Monografia (Bacharel em estatística), URCAMP (Universidade federal do Rio Grande do Sul), Brazil. Citado na página 13.

SILVA, G. M. e; SILVA, N. F. F. da; DIAS, M. de S. Detecção de bullying: Como identificar automaticamente essa prática em redes sociais? Revista de Sistemas de Informação da FSMA, n. 21, p. 11 – 19, 06 2018. Disponível em: <http://www.fsma.edu.br/si/edicao21/FSMA_SI_2018_1_Principal_05.pdf>. Citado 8 vezes nas páginas 11, 13, 20, 22, 23, 24, 26 e 29.

TAVARES, H. Cyberbulling na adolescência. Nascer e Crescer, 2012. Disponível em: <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0872-07542012000300016>. Citado na página 20.

TEIXEIRA, A.; FERREIRA, T.; BORGES, E. Bullying no trabalho: percepção e impacto na saúde mental e vida pessoal dos enfermeiros. scielopt, p. 23 – 29, 06 2016. Disponível em: <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S1647-21602016000100004&nrm=iso>. Citado 3 vezes nas páginas 11, 12 e 17.

TOMAEL, M. I.; ALCARA, A. R.; CHIARA, I. G. D. Das redes sociais à inovação. scielo, v. 34, p. 93 – 104, 08 2005. ISSN 0100-1965. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652005000200010&nrm=iso>. Citado 2 vezes nas páginas 13 e 19.

TV GLOBO. *Assédio moral no trabalho: TRT-PE recebeu 855 denúncias em 2018; entenda o que é.* 2019. Disponível em: <https://g1.globo.com/pe/pernambuco/noticia/2019/05/02/trt-pe-registra-media-de-dois-processos-de-assedio-moral-no-trabalho-por-dia-em-2018.ghtml>. Acesso em: 14 jan 2020. Citado na página 39.

TWITTER. *Central de Ajuda*. 2018. Disponível em: <https://help.twitter.com/pt>. Acesso em: 15 nov 2018. Citado na página 19.

_____. *Overview Twitter API*. 2019. Disponível em: <https://developer.twitter.com/en/docs/tweets/search/overview>. Citado na página 19.

_____. *Twitter Products*. 2019. Disponível em: <https://developer.twitter.com/en/products/twitter-api>. Citado na página 28.

URTIGA, T.; CASTRO, T. Detecção de bullying escolar em redes sociais e suas implicações na educação de adolescentes. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1693. Citado 4 vezes nas páginas 11, 13, 25 e 27.

WEST, B.; FOSTER, M.; LEVIN, A.; EDMISON, J.; ROBIBERO, D. Cyberbullying at work: In search of effective guidance. Laws, 2014. Disponível em: <https://www.mdpi.com/2075-471X/3/3/598>. Citado na página 16.

XU, J.-M.; ZHU, X.; BELLMORE, A. Fast learning for sentiment analysis on bullying. In: ACM. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*. [S.l.], 2012. p. 10. Citado 2 vezes nas páginas 24 e 26.