Universidade da Maia

**Departamento de Ciências Sociais e do Comportamento**

Dissertação de mestrado

# A Rede Semântica do Discurso de Ódio na Internet

Jessica Sofia Neto Andradez, Nº 31553

Mestrado em Psicologia Clínica Forense – Intervenção com Agressores e Vítimas

Orientação:

Professor Doutor Tiago Bento da Silva Ferreira

Jessica Sofia Neto Andradez

31553

# A Rede Semântica do Discurso de Ódio na Internet

Dissertação de Mestrado em Psicologia Clínica Forense - Intervenção com Agressores e Vítimas

Trabalho realizado sob a orientação do Professor Doutor Tiago Bento da Silva Ferreira, Universidade da Maia

Outubro de 2024

**Agradecimentos**

Os meus agradecimentos vão para aqueles que estiveram presentes em todas as horas ao longo deste ano tão importante e marcante para mim.

Agradeço aos meus orientadores ao Dr. Tiago Bento da Silva Ferreira por estar sempre disponível e de maneira tão genuína em todo este processo de orientação.

Agradeço também ao Dr. Joaquim Silva Rocha, à minha família, ao meu irmão, a minha madrinha e aos meus avós pelo acompanhamento e encorajamento constante e por estarem sempre presentes. Mas o maior agradecimento familiar vai para a minha mãe. Todos os dias lhe estou grata pelas possibilidades que ela me deu e por todos os esforços e sacrifícios que fez ao longo da vida para eu conseguir alcançar todos os meus sonhos. Obrigada, Mãe, és o meu maior exemplo e a minha heroína.

Os meus agradecimentos vão também para Isabel Braga e para o meu namorado, Pedro Rangel pelo carinho, paciência e incentivo e para os meus melhores amigos: a Ana Teixeira, a Rita Albuquerque, o João Santos, o Gonçalo Varela, o António Couto, a Daniela Barbosa, a Juliana Santos, Raquel Silva, Lilibeth Barboza, Leonor Rio e o Jorge pela presença constante, por terem estado sempre ao meu lado durante este processo, sempre com palavras de encorajamento.

Outros agentes importantes neste processo foram os utentes que tive o privilégio de acompanhar, possibilitando-me a oportunidade de desenvolver as minhas competências e crescer tanto a nível pessoal como profissional.

O meu muito obrigado a todos aqueles que direta ou indiretamente fizeram parte do meu percurso académico.

**Resumo**

As redes sociais são veículos de difusão de opiniões, plataformas com perfis variáveis que servem também de local para disseminação de discursos de ódio. É fundamental que sejam criadas ferramentas para identificar, controlar, bloquear, filtrar e remover estas publicações de forma automática, na comunicação mediada por computador. O objetivo desta dissertação é caracterizar as propriedades psicolinguísticas do discurso de ódio através de um método de análise das redes semânticas, distinguindo-o de outros tipos de discurso. Foi reanalisado um corpus linguístico previamente anotado por Kennedy et al. (2022), cujo objetivo do seu estudo era medir o discurso de ódio por ser um problema de direitos humanos, através da utilização do Medindo o Discurso De Ódio Corpus (MHS) constituído por 50070 tweets publicados entre Março e Agosto de 2019. Com base nos bigramas, foram estimadas duas redes semânticas caracterizando os tweets classificados como contendo discurso de ódio e os tweets classificados como discurso de suporte. Estas redes foram comparadas com base nas suas propriedades topológicas: densidade, diâmetro, sortatividade, comprimento médio do discurso, transitividade e centralização. Concluiu-se que relativamente a sortatividade e comprimento do discurso médio, transitividade apresentam maiores valores no discurso de suporte, enquanto que propriedades como densidade, centralização e diâmetro no discurso de ódio. No discurso de suporte existe maior tendência de conexões com os semelhantes, menor facilidade de passagem de informação, grupo mais restrito com relações concisas, maduras e eficientes. No discurso de ódio, as redes sociais são mais densas e com maior fluxo de informação e de intensidade de rede o que espelha uma visão de conetividade, dependente de um elemento central e as relações não são tão eficientes. Com este estudo foi possível demonstrar a possibilidade de aprofundar esta temática e a necessidade de futuras pesquisas que ajudem na deteção do discurso de ódio nas redes sociais.

***Palavras-chave**:* violência, internet, *cyberbullying*, *cyberviolence*, discurso de ódio

## Abstract

Social media are vehicles for the dissemination of opinions, platforms with variable profiles that also serve as spaces for the spread of hate speech. It is essential to create tools to identify, control, block, filter, and automatically remove these posts in computer-mediated communication. The objective of this dissertation is to characterize the psycholinguistic properties of hate speech through a method of semantic network analysis, distinguishing it from other types of discourse. A previously annotated linguistic corpus by Kennedy et al. (2022) was reanalyzed, with the aim of measuring hate speech as a human rights issue, utilizing the Measuring Hate Speech Corpus (MHS) consisting of 50,070 tweets published between March and August 2019. Based on bigrams, two semantic networks were estimated to characterize tweets classified as containing hate speech and those classified as supportive discourse. These networks were compared based on their topological properties: density, diameter, assortativity, average discourse length, transitivity, and centralization. It was concluded that in terms of assortativity and average discourse length, transitivity showed higher values in supportive discourse, while properties like density, centralization, and diameter were greater in hate speech. Supportive discourse exhibited a higher tendency for connections with similar entities, less ease of information flow, and a more restricted group with concise, mature, and efficient relationships. In hate speech, social networks are denser and have greater information flow and network intensity, reflecting a view of connectivity dependent on a central element, with less efficient relationships. This study demonstrated the potential for deepening this theme and the need for future research to aid in the detection of hate speech on social media.

***Keywords:*** violence, internet, cyberbullying, cyberviolence, hate speech

# Índice

## A Rede Semântica do Discurso de Ódio na Internet

A violência caracteriza-se pela intenção de magoar o(s) outro(s), através de ameaças ou ações, com o objetivo de provocar danos físicos ou psicológicos, morte, perturbação do desenvolvimento ou privações (Krug et al., 2002).

Ao longo dos últimos anos, as redes sociais e a Internet têm sido usadas como ferramentas para potenciar comportamentos violentos, conhecidos como violência digital. Este fenómeno tem vindo a tornar-se mais frequente pelo que se tem vindo a fazer um maior esforço para entender o que é a violência digital e que tipo de atividades engloba (Peterson & Densley, 2017). A violência digital é definida por atos de violência que têm como finalidade lesar o outro através dos meios de comunicação. Estes atos podem ser insultos, difamações, *cyberbulling, cyberstalking* ou assédio (Šincek et al., 2017).

O discurso de ódio *(hate speech*) é um tipo de violência digital levado a cabo por uma pessoa ou grupo que acredita ser superior a outro e que tem como objetivo exercer violência e discriminação, sendo potencialmente causador de depressão, ansiedade e, em alguns casos mais graves, tentativas de suicídio da vítima (Deschamps & McNutt, 2016; Ioannou et al., 2018).

É importante que se detete precocemente o discurso de ódio nas redes sociais com o intuito de evitar problemas de saúde graves no indivíduo/comunidade, tanto físicos como psíquicos e sociais pela violação dos direitos humanos. O esforço dedicado a desenvolver ferramentas de deteção automática de discurso de ódio tem sobretudo vindo das ciências da computação e estatística através do desenvolvimento de métodos de modelação dos aspetos semânticos dos conteúdos colocados nas redes sociais (Esteve et al., 2018; Miró-Llinares et al., 2018; Sachdeva et al., 2022).

Uma rede semântica é uma estrutura de conhecimento que representa as relações entre os conceitos expressos num texto, a forma como se interconectam e qual o seu sentido. Da perspetiva das ciências da computação, as redes semânticas permitem tornar a informação

veiculada nas redes sociais pelos seus utilizadores legível e interpretável por meios computorizados permitindo assim o desenvolvimento de ferramentas tecnológicas que possibilitam a deteção precoce de discursos de ódio nas redes sociais e assim evitar as suas consequências (Fklih & Al-Turaif, 2023).

Diversos autores (Esteve et al., 2018; Miró-Llinares, 2018; Sachdeva et al., 2022; Watanabe et al., 2018) apontam para a necessidade de serem desenvolvidas técnicas mais eficientes para a correta deteção de discursos de ódio, que têm de ser adequadamente distinguidos de outras expressões discursivas que não constituem discurso de ódio. Assim, de modo a contribuir para este esforço, o objetivo desta dissertação é caraterizar as propriedades psicolinguísticas do discurso de ódio nas redes sociais através da análise das redes semânticas de conteúdos colocados nas redes sociais, contrastando-as com outros tipos de discursos.

Através desta caracterização, pretende-se que seja possível a compreensão dos mecanismos implicados neste tipo de violência e contribuir para o desenvolvimento de formas eficientes de o identificar.

## Violência Digital

A violência é definida como a utilização premeditada de poder ou força física com o objetivo de provocar dano físico ou psicológico ao próprio ou a terceiros (Krug et al., 2002). Assim definida, a violência é um fenómeno multifacetado quanto ao seu objeto, sendo possível distinguir vários tipos de violência como a violência doméstica, que consiste em violência que ocorre entre parceiros íntimos em contexto de coabitação, que causa danos físicos, psicológicos ou sexuais (Leal, 2021); a violência sexual, definida como atos ou tentativas de atos sexuais indesejados (Baigorria et al., 2017); a violência física, caracterizada como qualquer forma de agressão física que causa dano à integridade física da vítima; ou violência emocional, que inclui

qualquer ação que provoque sofrimento, aflição ou angústia ao indivíduo (Alves & Serrão, 2018).

O *bullying* é também uma forma de violência que se caracteriza por atos intencionais de instigar, agredir, perseguir, difamar, ridicularizar e/ou assediar alguém considerado indefeso, bem como divulgar publicamente factos privados de outra pessoa e infligir sofrimento emocional intencional à mesma (Figueiredo & Matos, 2017; Lee& Shin, 2017; Watts et al., 2017). Com a quantidade de equipamentos eletrónicos e a intensificação da presença das pessoas em ambientes digitais, os meios de comunicação digitais rapidamente se tornaram veículos importantes para os comportamentos de bullying (Figueiredo & Matos, 2018; Lee, & Shin, 2017; Watts et al., 2017), facilitando e intensificando modos de agressão relacional indireta, onde se é cruel para com os outros enviando, reencaminhado ou expondo *online* conteúdos privados com o intuito de prejudicar alguém (Figueiredo & Matos, 2018). O intuito de quem pratica *bullying* é causar algum tipo de lesão (física ou emocional) à vítima, o que pode ocorrer quando o agressor e a vítima estão em presença física um do outro, mas também e atualmente com muito maior incidência, através dos meios digitais, já que o *bullying* transcende a presença física e os agressores têm acesso permanente às suas vítimas (Watts et al., 2017). Com o aumento exponencial do uso das tecnologias, uma tendência preocupante em todo o mundo, é então o aumento dos comportamentos de violência digital.

Fabriz & Mendonça (2022) definem uma rede social como um serviço da Internet que permite que um indivíduo possa criar o seu próprio perfil para produzir conexões com uma lista de outros utilizadores. Estas ferramentas de comunicação, cada vez mais diversificadas e em número crescente, têm um papel muito importante na atual propagação e proliferação de comportamentos violentos. Através delas, os comportamentos violentos ganharam um novo significado, alcance e rapidez.

Neste sentido, a violência digital tornou-se uma questão importante, sobretudo na população juvenil de todo o mundo. O mesmo, por intermédio de redes sociais como Facebook, Instagram, Twitter e YouTube parece estar a ganhar popularidade devido à capacidade das massas de testemunhar e/ou participar nos ataques (Watts et al., 2017). Diversos estudos têm comprovado que há uma associação relevante entre a má utilização das redes socias e a saúde dos seus utilizadores. Yeu & Rich et al. (2023), numa recente revisão da literatura, revelaram existir uma correlação positiva entre o uso prolongado das redes sociais e sintomas elevados de depressão, ansiedade entre adolescentes e níveis mais baixos de satisfação com a vida. O uso prolongado das redes sociais está também associado a distúrbios de sono, incluindo menos duração de sono, maior latência do sono e despertares mais frequentes durante o sono. Existe também uma correlação positiva entre o envolvimento nas redes sociais e atos de *cyberbullying*. Isto tem sido confirmado por outras revisões (e.g., Montag et al., 2024; Weigle & Shafi et al., 2024) que têm sublinhado que números substanciais de crianças e adolescentes enfrentam problemas com a utilização das redes sociais. Problemas como distúrbios de atenção e de sono, depressão, medo de perder associado à dependência comportamental, ansiedade social, automutilação e suicídio são frequentes nos jovens que convivem mais de perto com as redes sociais; e que o uso problemático das redes sociais na infância e adolescência está também associado a níveis mais elevados de vitimização entre pares.

Os agressores parecem aproveitar o "efeito de desinibição *online*" ou "anonimato" para a prática de atos abusivos sem sofrerem qualquer tipo de consequências (Figueiredo & Matos, 2018; Lee, & Shin, 2017; Watts et al., 2017). A capacidade de permanecerem anónimos pode manter os perpetradores alheios às consequências dos seus atos, o que facilita a sua perpetuação.

Neste contexto, tem sido levado a cabo um esforço em distinguir diferentes formas de violência digital (e.g, Santos, 2015): *flaming*, que consiste no envio de mensagens insultuosas, vulgares e com raiva acerca de uma pessoa, por email ou SMS, para um grupo *online* ou para a

própria pessoa; *online harassment* ou assédio *online* caracteriza-se pelo envio constante de mensagens ofensivas a um indivíduo através de correio eletrónico ou mediante outro mecanismo de envio de mensagens de texto; *cyberstalking*, perseguição no ciberespaço, incide no assédio *online* e inclui ameaças de dano ou excessivamente intimidantes; denegrir e humilhar através do envio de declarações prejudiciais, simuladas, ou cruéis sobre uma pessoa para outras pessoas ou publicação desse material *online*; a dissimulação caracteriza-se por assumir a identidade de outra pessoa e enviar ou publicar material *online* para lesar a vítima; *outing* que consiste em enviar ou publicar *online* mensagens de texto ou de imagens que contêm informação sensível, privada ou embaraçosa, acerca de uma pessoa; a exclusão cruel de alguém de um grupo *online*.

Os meios utilizados para divulgação e propagação da violência digital são vários, podendo existir diferentes tipos de *bullying*: *bullying* através de mensagens de texto que são muito recorrentes no fenómeno da violência digital e podem ser enviadas por telemóveis ou por outro tipo de tecnologias que permita difundi-las (Cheminais, 2008; Katz, 2012; Santos, 2015, Tarshis, 2010); *bullying* por imagens/videoclip através de câmaras de telemóveis remetidas para inúmeras pessoas ou publicadas num espaço *online* de acesso público, com o objetivo de humilhar ou lesar alguém; *bullying* por chamada telefónica através de telemóvel, caraterizado pelas "chamadas silenciosas" que o agressor faz para a vítima, ou então através do envio de mensagens abusivas em que o ofensor oculta a sua identidade ou utiliza o telemóvel de outra pessoa. Para além disso, o ofensor pode ainda utilizar o telemóvel da vítima para ofender outras pessoas, pensando estas que o proprietário do telemóvel é o responsável em causa (Cheminais, 2008; Katz, 2012); *bullying* por correio eletrónico, um tipo de comunicação onde não existe necessidade de resposta imediata e é acessível através da internet, que passa pelo envio de mensagens que podem ser difundidas para uma ou mais pessoas, sendo estas recebidas em contas particulares de correio eletrónico; *bullying* em salas de *chat* são um tipo de comunicação

que é necessário resposta em simultâneo e geralmente existem para que os utilizadores possam interagir sobre alguma área de interesse em específica (Willard, 2007). Muitas vezes, os jovens que no mundo real dispõem de uma fraca rede de amigos recorrem a estas salas de forma a obter amizade ou algum tipo de intimidade com outras pessoas, podendo ser vítimas de mensagens embaraçosas ou ameaçadoras (Cheminais, 2008; Katz, 2012). Importa referir que, frequentemente, os utilizadores destas salas não apresentam a sua verdadeira identidade, falseando a idade, o género, a ocupação, o que pode ser útil para certos grupos de pessoas, nomeadamente os *cyberstalkers* e os predadores sexuais (Kowalski et al., 2012); *bullying* através de mensagens instantâneas, um recurso bastante popular não só entre os jovens, como também entre os adultos. As mensagens instantâneas permitem que indivíduos afastados geograficamente possam comunicar de forma instantânea, havendo programas que também incluem a possibilidade de utilização de microfones e de câmaras (Hinduja & Patchin, 2008). Neste âmbito, e de forma simples, o ofensor envia à vítima mensagens desagradáveis ou ameaçadoras em tempo real (Cheminais, 2008; Hinduja & Patchin, 2008; Katz, 2012); *bullying* através de *websites* sendo possível a criação de blogues difamatórios, *websites* que têm por objetivo humilhar alguém (Cheminais, 2008; Katz, 2012). Em relação aos blogues, estes podem ser caracterizados como uma espécie de diário pessoal interativo na Internet, em que a pessoa responsável pela página publica frequentemente conteúdo e solicita comentários daqueles que a visitam (Willard, 2007); *bullying* através de sites em redes sociais são plataformas na Internet que têm as seguintes características: a interação social entre duas ou mais pessoas; permitem a criação de perfis pessoais nos quais as pessoas podem divulgar informação; permitem a comunicação entre pessoas através de serviços de mensagens instantâneas ou de correio eletrónico e incluem funções de pesquisa para que o utilizador seja capaz de pesquisar outros utilizadores com os quais possa comunicar. Estes tipos de sites são propícios à prática de violência digital, não só porque permitem a publicação de comentários, fotografias e vídeos,

mas também porque é possível criar perfis ou contas falsas sobre alguém (Bauman, 2009). Segundo Kowalski et al. (2012) estes tipos de redes sociais podem ser utilizados como *burn pages*, (páginas na Internet) onde os jovens publicam boatos ou outro tipo de informação negativa sobre os seus colegas de escola e ainda a possibilidade de criar *social web sites*.

Relativamente à violência digital, refira-se ainda o envio de mensagens geralmente de conteúdo ameaçador, insultuoso, ou prejudicial para a vítima tendo como finalidade difundir falsos comentários, humilhar, ou ainda excluir alguém (Cheminais, 2008; Katz, 2012; Tarshis, 2010), afirma que, por norma, o ofensor utiliza um pseudónimo inventado ou então o correio eletrónico ou o nome de outra pessoa de forma a não ser detetado, preservando o seu anonimato (Cheminais, 2008; Katz, 2012). Uma das razões pela qual o correio eletrónico é uma das formas mais recorrentemente utilizadas na violência digital prende-se com o facto de o ofensor ser capaz de divulgar e propagar de uma só vez para centenas de pessoas, imagens ou outro tipo de informação sobre a vítima (Kowalski et al. 2012).

**Discurso de Ódio**

Como vimos, a violência digital pode adquirir diversas formas, seja por via da divulgação de conteúdos íntimos das vítimas, seja por via da utilização de um discurso agressivo que manipula a realidade das vítimas.

Neste contexto, o discurso de ódio surge sob a forma de expressões e mensagens que visam difamar, ofender, humilhar ou discriminar um indivíduo ou grupo com base em características como raça, religião, género, orientação sexual, nacionalidade, entre outros. Pode também surgir como incitação ao ódio, à discórdia e à promoção de ataques violentos entre grupos sociais (Fabriz & Mendonça et al., 2022). Trata-se de uma forma de comunicação agressiva e prejudicial que visa atingir uma pessoa ou grupo específico de pessoas que compartilham uma identidade comum (Rocha, 2021), quer seja por questões de raça, orientação

sexual, religião ou género. O discurso de ódio é entendido por Amores et al. (2021), como a existência de mensagens que implicam rejeição, humilhação, assédio, descrédito, estigmatização de indivíduos/comunidades baseados em atributos particulares.

O uso das redes sociais para a expressão do discurso de ódio permite que esta comunicação seja amplamente difundida, sob anonimato. Isto desinibe o agressor de demonstrar os seus preconceitos e ódio, sem enfrentar as consequências das suas ações. As vítimas podem ser lesadas diretamente, quando o discurso provoca dano de forma imediata, ou indiretamente, quando o dano é sofrido pela perpetuação da agressão por outros indivíduos ou comunidade, que adotam o discurso de ódio contra a vítima, sem ser necessária a ação direta do autor inicial (Chetty & Alathur, 2018). Este tipo de violência também pode ser visto como uma violência sobretudo simbólica, cujos efeitos se podem manter nesse âmbito ou passar à violência física (Martins, 2019). O discurso de ódio é construído e articulado através de cinco elementos fundamentais: infração de regras, indução de vergonha nas vítimas, indução de medo nas vítimas através de ameaças e intimidação, tentativa de desumanização da vítima e desinformação nas pessoas ou grupos a que estes pertencem (Blanco et al., 2022).

O discurso de ódio distingue-se dos modos convencionais de discurso de ódio pelo facto de poder ser anónimo (impossível identificar o agressor), ter um maior alcance, permanência e propagação (o discurso ódio alcança um número elevado de indivíduos, por um tempo indeterminado podendo chegar a qualquer lugar do mundo), permitir a invisibilidade (devido à distância que existe entre agressor e vítima este não tem noção dos danos causados na vítima), influenciar a comunidade (qualquer pessoa pode participar no discursos de ódio e atrair um público *online* que partilhe os mesmos pensamentos e ideais) e permitir a instantaneidade (a propagação do discurso de ódio acontece rapidamente) (Silva et al., 2021).

As caraterísticas do discurso de ódio são a persistência, a capacidade de pesquisa, a replicabilidade e a audiência invisível. A persistência refere-se às mensagens armazenadas e à

duração das mesmas no sistema. A capacidade de pesquisa corresponde à possibilidade de pesquisar e recuperar as mensagens a qualquer momento. A replicabilidade define-se pela possibilidade de reproduzir as mensagens por outros autores e de as difundir em diversas redes sociais. A audiência invisível, refere-se às centenas de indivíduos que participam no discurso de ódio, devido à rápida propagação e acessibilidade destas mensagens (Silva et al., 2021).

O discurso de ódio incentiva a atos de discriminação ou violência por motivos de raça, xenofobia, orientação sexual, entre outros, formas que propiciam climas de hostilidade que podem levar a atos discriminatórios ou até violentos de indivíduos/comunidades (Amores et al.,2021). O agressor visa atingir alguém, ou grupo específico de pessoas que compartilham uma identidade comum e as comunicações negam a condição de outros membros da sociedade como cidadãos livres e iguais (Rocha, 2021).

Silva et al. (2021) refere que ao tentar agredir uma pessoa no mundo digital, aqueles que compartilham das mesmas caraterísticas, ao entrarem em contato com este tipo de violência, compartilham desta mesma violação. Estes podem criar comunidades na internet para violentar, difamar de forma intencional o outro (Neves, 2022). O *bullying* e a violência digital, englobam conflitos nas relações, baseando-se nas intolerâncias às diferenças (Silva & Mascarenhas, 2010).

Pinheiro (2009) menciona que na violência digital é possível incluir certos indivíduos em certos estereótipos o que tem como consequência, a alteração de identidade social dessa pessoa, de forma negativa e que se irá refletir na vida social da vítima. Este tipo de discurso tem consequências potencialmente graves nos indivíduos e na comunidade. Segundo Oliveira (2008), o discurso de ódio afeta as pessoas, pois transmite emoções negativas que as desqualifica, refletindo-se nas suas vidas pessoais e podendo abranger a sua família, amigos, trabalho, vida social entre outros.

Constatou-se a existência de uma correlação entre a crescente utilização do discurso de ódio nas redes sociais e o aumento da criminalidade (Amores et al., 2021). Na violência digital, a situação intensifica-se uma vez que as testemunhas e os agressores são imensuráveis e são as redes sociais que possibilitam a transcendência das fronteiras do tempo, pois a violência demonstrada pode manter-se indefinidamente no espaço virtual e até no espaço pessoal e físico levando inevitavelmente a prejuízos na socialização/comunidade, pois as vítimas, como consequência, recorrem ao isolamento para se protegerem (Silva & Mascarenhas, 2010).

Ao efetuar a alteração da imagem da vítima através da violência digital, esta provocará a alteração do comportamento das vítimas e do seu bem-estar psicológico, o que levará ao afastamento à ausência de qualquer relacionamento social. A vítima, por sua vez, vai-se isolar afastando-se da família, parentes e amigos, podendo até fazê-las acreditar que são um incómodo para a sociedade (Silva & Mascarenhas, 2010).

O discurso de ódio e a violência digital levam a problemas psicológicos e ao sofrimento emocional, medo, stress intensificados, bem como o seu alcance a longo prazo pode implicar a mudança de comportamento e atitude com o intuito de defesa, levando a pessoa a limitar a sua capacidade de participar plenamente na sociedade. Pode-se concluir que o discurso de ódio e a violência digital, além de afetar a saúde e a dignidade de quem é vítima, é também uma ameaça à sua integração sociedade, sendo este um problema atual que não só envolve o indivíduo, mas também a família, grupo em grande escala a comunidade (Santos, 2015).

## Perfil dos Agressores

Não podemos falar de violência digital, sem dar ênfase aos agressores. Qualquer indivíduo constitui um potencial agressor no que à violência digital diz respeito e não existe um perfil definido e igual para todos, uma vez que as suas características são muito diversificadas (Neves, 2022). Normalmente, o agressor tem um papel de liderança num pequeno grupo de

amigos, embora seja rejeitado pela maioria dos companheiros. Ele gosta de dominar os outros, mostra dificuldade no cumprimento de normas e regras, mas tem normalmente boa autoestima, construída com base no domínio sobre os outros e no protagonismo que as condutas agressivas lhe proporcionam (Martins, 2019).

Os agressores, segundo Santos (2015), caracterizam-se pela exibição de comportamentos negativos e hostis, tendo como intuito provocar, humilhar ou excluir a vítima. Estes são, por norma, agressivos e socialmente dominantes, utilizando o seu poder para humilhar terceiros. Porém, a agressividade exibida por estes poderá ser a resposta de uma baixa autoestima, mas que resulta numa posição dominante devido à posição submissa da vítima.

Segundo Neves (2022), existem dois tipos de perfis dos indivíduos implicados na violência digital: os agressores pró-ativos (indivíduos que concretizam as suas ações para atingir uma determinada finalidade prejudicando terceiros) e os agressores reativos (indivíduos que quando sujeitos a uma provocação ou ameaça percecionada, seja real ou imaginada, executam uma ação). Em diversos estudos estes indivíduos evidenciaram altos níveis de stress, depressão e ansiedade (Šincek & Milié et al, 2017).

Segundo Seixas et al. (2016), os agressores apresentam impulsividade e baixa tolerância à frustração, extrema necessidade de dominar os outros, dificuldade em aceitar o cumprimento de normas e regras, maior uso de expressões e de comportamentos e atitudes agressivas e ou violentas, reduzida empatia perante as vítimas das agressões, reforçada pelo facto de não verem em tempo real o impacto das suas ações. Kowalski (2012) refere que estes apresentam menor desempenho escolar, notas mais baixas, problemas de concentração, depressão, ansiedade, consumo de tabaco, álcool e drogas, violência e delinquência nos jovens, ansiedade, ausência de empatia, comportamento agressivo e criminal e, por último, dependência da tecnologia.

Os agressores recorrem a várias estratégias como a persuasão, a criação de estereótipos, a substituição de nomes, o apelo à autoridade, a afirmação, a repetição e a seleção de factos

favoráveis ao seu ponto de vista, o que proporciona a criação de inimigos. Por outro lado, fazem uso da ausência de contraposição direta e imediata destas mensagens de ódio, recorrendo ao uso de técnicas de manipulação emocional, que propiciam o aumento da aceitação do seu discurso. Tanto quem inicia o discurso como quem o incentiva tem como objetivo intensificar a discriminação (Martins, 2019). Por último, com o estudo de Sahin & Ayaz-alkaya (2023) foi possível concluir que os adolescentes vítimas de *cyberbullying* experimentam consequências negativas como a depressão, uso de drogas ilegais, comportamento antissocial e envolvimento em comportamentos criminosos, que podem persistir na idade adulta.

**Perfil e Impacto nas Vítimas**

No que diz respeito ao perfil da vítima, não se consegue apresentar um perfil devidamente definido e homogéneo, podendo apresentar perfis totalmente díspares (Seixas et al., 2016). A vítima pode ser alguém bem-sucedido e devidamente integrado a nível profissional, mas que a dada altura, acabou por ser humilhado ou ameaçado por um agressor; alguém que sofreu mudança de interação dentro de um determinado grupo, levando a uma situação de exclusão de determinado elemento; indivíduos que por não se sentirem adaptados ao contexto onde estão integrados, têm maior predisposição para se sujeitarem a agressões, humilhação fazendo tudo o que está ao seu alcance para pertencer ou pelo menos, não ser afastados do grupo. Quando as vítimas são crianças e jovens revelam dificuldade em gerir as suas relações interpessoais no que diz respeito à assertividade e defesa dos seus direitos fundamentais, por isso apresentam uma certa dificuldade em criar e ou manter amizades possuindo uma rede de apoio emocional relativamente frágil ou inexistente em contexto escolar ou fora dele (Seixas et al.,2016).

As vítimas poderão ser caracterizadas como indivíduos sensíveis, respeitosos, honestos, criativos, com um grande sentido de desportivismo, um elevado nível de integridade e baixa

propensão à violência (Anderson, & Smith, 2007), porém são também indivíduos que poderão dispor de uma autoestima reduzida e graves problemas emocionais. Podem estar zangados, tristes, deprimidos, magoados, stressados, desamparados, sozinhos e confusos; experimentar sentimentos de depressão, baixa autoestima, desamparo, ansiedade social, concentração reduzida, alineação e ideação suicida (Santos, 2015). Segundo Seixas et al. (2016), a vítima pode apresentar níveis mais elevados de depressão, ansiedade, medo, baixa autoestima, sentimento de raiva, frustração, impotência, nervosismo, irritabilidade, insegurança, tristeza, perturbação sono, tentativa de suicídio, menor bem-estar psicológico, dificuldade de concentração, comprometimento do desempenho escolar, absentismo e abandono escolar (Seixas et al.,2016). Segundo o estudo de Salazar et al. (2024), refere-se que as vítimas de *cyberbullying* sofrem consequências negativas para a saúde mental, relativas à sua exposição, como o aumento de stress, ansiedade, depressão, bem como, efeitos na saúde física, como a perda de sono, cefaleias, alteração do apetite e dores de estômago. Ainda pode levar à redução da satisfação no trabalho, menor produtividade e maiores taxas de rotatividade. Iffland et al. (2023) referem que a vitimização por *cyberbullying* na adolescência por si só, aumenta o risco de desenvolvimento de uma gama de psicopatologias. O aumento significativo de vitimização por *cyberbullying* entre os 10 e os 17 anos está associado a problemas de internalização, como a depressão, baixa autoestima, solidão, ansiedade, ansiedade social e somatização. O suicídio emergiu como consequência mais preocupante da vitimização por *cyberbullying*, daí Sahin & Ayaz-Alkaya (2023) referirem que as vítimas podem apresentar problemas de saúde mental como depressão, baixa autoestima, ansiedade, raiva, ideação suicida, sintomas psicossomáticos, como insónia, cefaleias, problemas digestivos, tonturas, e ainda problemas sociais, como a solidão e sentimentos de impotência e apresentar ainda comportamentos de risco como fumar, abuso de substâncias e comportamentos suicidas.

De acordo com Santos (2015), a violência digital pode causar danos psicológicos variados desde a introversão, baixa autoestima, perturbação de pânico, insegurança, angústia, depressão, perturbações do sono, perturbações psicossomáticas, insucesso escolar advindo das dificuldades de concentração e do absentismo elevado, consumo excessivo de substâncias aditivas principalmente álcool, relutância em utilizar as novas tecnologias, ou, em situações extremas, o suicídio. As vítimas, segundo Martins (2007), apresentam-se socialmente isoladas, sem amigos, têm baixa autoestima, problemas de saúde física (sintomas psicossomáticos) e problemas de saúde mental (sintoma depressivo, ansiedade, insegurança, ainda podem apresentar medo dos agressores, vulnerabilidade, serem incapazes de se defenderem perante intimidação ou em alguns casos pertencer a famílias sobre protetoras). Segundo as autoras Silva & Mascarenhas (2010), existem dois tipos de vítimas: vítima típica que é caraterizada como alguém mais tímido, tranquilo, submisso e sensível, com baixa autoestima, inseguro, pouco sociável, fisicamente mais frágil do que seus agressores, apresenta poucos recursos para se defender das agressões e por último a depressão; a vítima provocativa apresenta depressão, baixa autoestima e ansiedade como a vítima típica, mas com a diferença de que este tipo de vítima apresenta medo em agir e pode apresentar hiperatividade, inquietação, dispersão e comportamentos agressivos.

O discurso de ódio afeta os indivíduos pela transmissão de emoções negativas que provocam a invasão das suas vidas e, eventualmente, nas suas famílias, trabalho, vida social, entre outros. Artigos como o de Figueiredo & Matos (2018) e Šincek et al. (2017) relatam que as vítimas deste tipo de violência apresentam inúmeros sentimentos negativos, entre os quais a irritabilidade, nervosismo, tristeza, mágoa, raiva, frustração e impotência, baixa autoestima, alterações do sono, medo, diminuição do desempenho escolar ou profissional, dificuldades de concentração, problemas a nível da saúde física, psicológica e emocional. Em alguns casos apresentam também distanciamento social, tentativas de suicídio, uso de substâncias proibidas

e stress pós-traumático. Podem ainda desenvolver ansiedade, depressão, automutilação, problemas como a agressividade e comportamento desviantes (Neves, 2022).

Podemos constatar que as vítimas do discurso de ódio pertencem a um grupo que normalmente são minorias. O efeito deste discurso não atinge somente as vítimas, mas o mundo onde esta se encontra inserida. A vítima pode vir a sofrer de insegurança evitando espaços públicos e viver em sociedade ou isolando-se ou até levar ao cálculo minucioso dos lugares a frequentar. Pode ter como consequência sentimentos de inferioridade que causam danos à saúde mental como depressão, vergonha, ansiedade, e até sentimentos de culpa. Pode ainda comprometer a autoestima da vítima (Harff et al., 2020). O desconhecimento de quem é o agressor, fatores como a inexistência de um lugar seguro, sensação de que estão a ser observadas, perseguidas, agredidas, leva a situação de stress da vítima, aumentando consideravelmente os níveis de pressão psicológica ao que estão sujeitas. Deste modo, depreende-se que o dano vivenciado na violência digital assume, essencialmente, uma natureza psicológica, física, emocional, social que perturba indiscutivelmente a saúde da vítima do abuso.

**Sociodemografia da Violência Digital**

A relação entre violência digital e género ainda apresenta conclusões bastante inconsistentes, ou seja, a prevalência do género relativamente à violência digital (quem é o agressor e quem é a vítima) não está determinada, não tendo os estudiosos da matéria chegado a um consenso, mas, na sua maioria, as investigações anteriores referem que os agressores são mais frequentemente do sexo masculino, e as vítimas do sexo feminino (Gonzalves-Cabrera et al., 2023; Gradinger et al., 2009; Mishna et al.., 2012; Sincek et al., 2017; Weigle & Shafi et al., 2024).

No que diz respeito à prevalência do género nos outros países, Li (2005) verificou que no Canadá cerca de 60% das vítimas de violência digital são do sexo feminino e mais de 50% dos agressores são do sexo masculino, voltando a verificar a mesma conclusão um ano depois, concluindo que o sexo masculino tem maior probabilidade da prática de violência digital em comparação com o sexo feminino (Li, 2006). Wang et al. (2009), à escala nacional dos Estados Unidos da América, verificou que os rapazes têm maior probabilidade de serem agressores e as raparigas vítimas. No estudo efetuado por Taiwan et al. (2010) em contexto asiático, conclui-se que o sexo masculino é mais vítima e agressor do que o sexo feminino. Calvete et al. (2010), no seu estudo realizado em Espanha, concluiu que o sexo masculino estaria mais frequentemente envolvido na prática da violência digital do que o sexo feminino. Walrave & Heirman (2011), com o seu estudo efetuado na Bélgica, obtive como resposta ao seu estudo que o sexo feminino tem maior probabilidade de ser vítima enquanto o sexo masculino tende a ser o agressor. Segundo Smith et al. (2006), na Inglaterra, depararam-se com o facto de o sexo feminino tem maior probabilidade de ser vitimizada.

Relativamente à relação entre a idade e a violência digital, constata-se que existe uma maior prática de violência digital no início do 9º ano de escolaridade, prolongando-se até ao secundário com jovens de idade entre os 13 e os 15 anos (Kowalski, 2012; Campell et al., 2013; Price e Dalgleish, 2010; Tokunaga, 2010; Jones et al., 2013; Williams & Guerra, 2007).

De acordo com os vários estudos, conclui-se que no período da adolescência existe uma maior propensão para o *bullying* entre pares e um número elevado de vítimas de *cyberbullying* e violência digital, o que também se deve ao facto de estes jovens apresentarem uma maior facilidade, autonomia e habilidade para o uso da tecnologia digital e maior para socializar (Sahin & Ayaz-Alkaya., 2023; Iffland et al., 2023; Seixas et al., 2016). Assim, e porque estes tipos de violência são praticados desde cedo e durante o tempo escolar, uma das consequências mais visíveis é o insucesso escolar, pois as vítimas apresentam níveis mais baixos de

rendimento escolar e faltas de assiduidade. Também se pode verificar que as fracas relações entre pares, falta de apoio social, clima negativo, a ansiedade dos jovens de estarem sozinhos ou sem amigos podem ser fatores de risco para os adolescentes passarem a ser agressores e a pôr em prática a violência entre pares ou a praticar violência digital (Price & Dalgleish, 2010; Sahin & Ayaz-Alkaya, 2023). O estudo de Salazar et al. (2024) concluiu também que o *cyberbullying* parte, na maior parte das vezes, de colegas, sendo que pessoas com traços de personalidade mais agradáveis estariam menos sujeitas a *cyberbullying*, ao contrário das pessoas com nível mais baixo de agradabilidade. As pessoas mais agradáveis, gentis, bem-humoradas, confiantes, cooperativas, generosas têm menor propensão a envolverem-se em conflitos interpessoais. Os que apresentam neuroticismo têm maior probabilidade de sofrer de *cyberbullying* e são mais propensos a experimentar eventos negativos na vida, explicado pelos seus estados emocionais negativos, como a ansiedade, medo, raiva, ansiedade social e depressão. Portanto, indivíduos com maior neuroticismo têm maior probabilidade de serem perpetuadores ou vítimas de *cyberbullying*. Relativamente à extroversão, consciência, e abertura este estudo demonstrou que não existia relação com o *cyberbullying.*

Por último, Seixas et al. (2016) evidenciou que existe uma associação da prática da violência digital sobre as chamadas hipotéticas minorias, sejam elas étnicas, raciais, religiosas ou de orientação sexual, necessidades especiais, sejam elas físicas, sensoriais, ou mentais, bem como pessoas pertencentes a grupos vulneráveis. O simples facto de se considerarem algumas pessoas diferentes, pode ser, só por si, um fator para converter o outro em potencial vítima. Para corroborar este estudo, Patchin et al. (2023) refere que são os adolescentes não heterossexuais os mais propensos à prática de *cyberbullying* e a pensamentos sérios de tentativa de suicídio.

## Características Linguísticas do Discurso de Ódio

Considerando a extensão e o impacto do discurso de ódio na Internet, são importantes os estudos que se desenvolveram e continuam a desenvolver para ajudar a detetar esse discurso e analisar as suas caraterísticas, sejam elas fonéticas, fonológicas, sintáticas, semânticas, pragmáticas, estilísticas ou outras. Estes estudos permitiram também mapear algumas das características linguísticas deste discurso de ódio.

Os marcadores que mais facilmente identificam o discurso de ódio são a linguagem informal (calão), expressões literárias alteradas que demonstram explicitamente a intenção do autor em causar danos à reputação do visado, menosprezar o assunto em discussão ou causar associações negativas à pessoa ou grupo de pessoas. Este tipo de linguagem recorre ainda à estereótipos e à anexação de rótulos, o humor amargo, os apelos diretos à discriminação e ao ódio, o incitamento dissimulado ou vago ao ódio e à violência, a repetição frequente de palavras, frases, sentenças ou pensamentos para enfatizá-los e a distorção de nomes próprios (Haladzhun et al., 2021).

Constata-se também que existe uma relação entre o discurso de ódio e os consequentes sentimentos negativos que daí advêm. De acordo com Schmidt & Wiegand (2017), o número de palavras negativas, conjunto, lista de insultos, declarações polares negativas, calúnias, ou até inclusão de imagens, vídeos e áudio depreciativos parecem também caracterizar o discurso de ódio. Pode-se recorrer a palavras inflamatórias, a comentários mais ou menos extensos, ao comprimento médio da palavra, ao número de pontuações, número de pontos, pontos de interrogações, aspas e pontuação repetida, número de letras maiúsculas, número de palavras de insulto, número de palavras desconhecidas ou com erros ortográficos (Nobata et al., 2016).

Segundo Fortuna &Nunes (2018), é importante detetar palavras consideradas insultos, palavrões, que tenham conotação negativa. Muitas vezes, as palavras de ódio são obscurecidas

com erros ortográficos intencionais, pela substituição de um único caracter, pelo uso repetitivo da mesma palavra no corpus, o uso de pronome pessoal na primeira e segunda pessoa, verbos na terceira pessoa do singular, adjetivos, determinantes, o uso de letras maiúsculas, pausas no texto, e outros, são também caraterísticas a ter em conta na deteção de discurso de ódio. Relativamente a polaridade negativa está em qualquer expressão enfatizada, o que só a semântica permite pelos diferentes sentidos que se podem dar à palavra ou expressão, bem como na importância que a pontuação pode aqui assumir pelo recurso internacional de um determinado número de pontos de exclamação, de pontos de interrogação, pontos finais, palavras em letra maiúscula, citações, interjeições, expressões de riso ou até de palavras e as vezes que esta é repetida, tudo ajuda na deteção do discurso de ódio (Watanabe et al., 2018).

Garcia-Diaz et al. (2020), quando refere as características semânticas do discurso de ódio, refere o uso de um discurso forte e ofensivo como palavrões, dando ênfase a características relacionadas com erros ortográficos e palavras mal escritas. Assim, é na articulação das diferentes classes gramaticais como os substantivas, preposições, verbos, entre outras, e o lugar que ocupam na frase que a morfossintaxe estará presente neste discurso de ódio. Os media têm também um papel fundamental na identificação da misoginia (sentimento de repulsa, adversão), pelo uso de hashtags (uso do símbolo "#" antes de uma palavra ou expressão) e do calão nas redes sociais, podendo também encontrar-se emojis negativos.

A pragmática é uma disciplina que estuda as relações existentes entre os signos e os sujeitos falantes, no sentido de descrever o uso que estes fazem da língua nas mais diversas situações de comunicação. Assim, qualquer palavra pode ser a escolhida do agressor para conseguir os seus intentos, porque ele sabe o contexto em que a está a utilizar e conhece os efeitos que ela vai produzir. Muitas vezes é através de elementos tão simples como o uso de sinónimos na linguagem figurada, o recurso a expressões relacionadas com animais ou ao sexo e risco, de negações como o "não", "nunca" ou "ninguém", que o agressor concretiza os seus

objetivos. Neste momento entra a psicolinguística que estuda a relação entre o uso da linguagem e o trabalho mental do emissor e do recetor da mensagem. O agressor estuda a vítima e conhece a linguagem que terá de utilizar para despertar nela sentimentos negativos.

### O Presente Estudo

Nesta dissertação serão reanalisados dados anteriormente recolhidos por Kennedy et al. (2022), cujo objetivo era medir o discurso de ódio, utilizando métodos estatísticos para mapear as suas caraterísticas, através da utilização do Medindo o Discurso De Ódio Corpus (MHS). Assim, serão caracterizados as propriedades psicolinguísticas do discurso de ódio e a construção de uma rede semântica através de um método de conteúdos das redes sociais - que têm o mérito de mapear dinamicamente as interações entre as palavras - permitindo aceder aos aspetos estruturais de utilização da língua, comparando-as com outros tipos de discursos.

## Método

### Participantes

Esta dissertação partiu, como já referido anteriormente, de dados recolhidos por Kennedy et al. (2022), cujo objetivo do seu estudo era medir o discurso de ódio por ser um problema de direitos humanos, através da utilização do Medindo o Discurso De Ódio Corpus (MHS). No estudo realizado por Kennedy et al., foram inicialmente recrutados 11.143 anotadores do *Amazon Mechanical Turk* que possuíam um total de 50.070 comentários. Para completar o instrumento de rotulagem utilizou-se o IP de cada anotador para garantir que estes pertenciam aos Estados Unidos (EUA). Foi examinada a qualidade dos rótulos de cada utilizador através de uma estatística média quadrática *Infit* que excluía os anotadores cuja estatística média fosse fora do intervalo 0,37-1,9. Os anotadores com média superior a um são os que têm mais imprevisibilidade ou ruído nas suas respostas; com valores de dois ou mais,

são vistos como os mais degradantes e os anotadores com valores inferiores a um pertencem aos que representam menor imprevisibilidade do que esperado, o que sugere que pode ter favorecido certas opções de resposta. Portanto, foram excluídos anotadores cuja estatística média quadrada *infit* fosse muito baixa, anotadores com parâmetros de gravidade extrema, anotadores com endereços IP fora dos EUA vinculados a serviços *proxy* conhecidos ou associados a mais de quatro tarefas de anotação; também foram excluídos avaliadores que não marcaram um número suficiente de grupos de identidade direcionados. A aplicação destes critérios de exclusão deixou 8472 anotadores, com 39565 comentários de acompanhamento. Neste estudo foram construídas redes semânticas para os *tweets* que foram classificados como discurso de ódio e discurso de suporte.

**Procedimento**

Na recolha de dados feita anteriormente por Kennedy et al. (2022) foram selecionadas três plataformas: *Youtube*, *Twitter* e *Reddit*, no período compreendido entre março e agosto de 2019 de onde foram recolhidos 50070 comentários (comentários escritos principalmente em inglês e que possuíssem entre 4 a 600 caracteres). Para a elaboração do corpus foi utilizada a API pública de cada plataforma na seguinte proporção: 40% dos comentários foram recolhidos no *Twitter* utilizando o API de *streaming* do Twitter que é uma amostra aleatória que facilita a recolha dos *tweets*; 20% dos comentários foram recolhidos no *Youtube*. A razão pela qual foi usada uma percentagem menor de comentários foi devido à dificuldade de recolha do conteúdo e pelo facto dos comentários serem pouco extensos e com linguagem menos complexa; os restantes 40% dos comentários foram recolhidos no *Reddit* de publicações realizadas em tempo real através do r/all da plataforma. No *Youtube*, procedeu-se à pesquisa de vídeos nas proximidades das trezentas cidades mais populosas dos EUA com comentários em inglês de autores americanos. Após a colheita dos comentários das plataformas selecionadas, realizou-se

o pré-processamento simples, removendo URLS, número de telefone e espaços em branco e acentos contíguos.

Para a construção da teorização, levou-se a cabo uma revisão manual dos comentários nas redes sociais selecionando um pequeno corpus de texto de referência para cada nível conceitual. Para isso foram selecionados dez comentários para servir de exemplo para cada um dos níveis teóricos, perfazendo setenta comentários (níveis teóricos: apoio, contradiscurso, neutro, hostilidade, desumanização, violência e genocídio). Conjuntamente com o desenvolvimento de construções e recorrendo a literatura existente, foi realizada manualmente a revisão de milhares de comentários do nosso corpus. Acrescentaram-se comentários do conjunto de referência para cada nível originados pelos diferentes grupos-alvo, tamanho do texto e estilos linguísticos. Posteriormente, foram selecionados os comentários que explicavam melhor os níveis de discurso de ódio. Foi desenvolvido um instrumento de rotulagem e anotação de dados que continha três componentes: i) um conjunto de dez itens de pesquisa autorizou o anotador a interrogar o comentário referente a diversas caraterísticas do discurso de ódio; ii) especificação de quaisquer grupos de identidade refletidos nos comentários; iii) pergunta sobre as informações demográficas do anotador. Este processo de anotação foi aprovado pelo conselho de revisão institucional da universidade da Califórnia, Berkeley. Foi respeitado o anonimato. Os anotadores recrutados do *Amazon Mechanical Turk* receberam vinte e seis comentários, dos quais seis eram do conjunto de referência. O tempo médio para a pesquisa foi de quarenta e nove minutos e os anotadores foram pagos para realizar a sua tarefa tendo a oportunidade para fornecer o *feedback* sobre o processo de rotulagem. Após a observação dos comentários, os anotadores propuseram classificações em escala de estilo *Likert* obtendo dez rótulos ordinais de pesquisa diferente (sentimento, respeito, insulto, humilhação, desumanização, violência, genocídio, ataque/defesa, status inferior/superior e uma classificação

binária do discurso de ódio). Deve ter-se em conta que quanto maior for a classificação, mais ódio representa.

Foi examinada a qualidade dos rótulos efetuados por cada anotador através de uma estatística média quadrática *infit*, um diagnóstico de ajuste do avaliador que é calculado durante a escala Rasch. A teoria de medição de Rasch fornece uma estrutura para a elaboração de uma escala de medição para um problema. Esta medição permite capturar vários recursos que influenciam a criação de um rótulo, incluindo o conteúdo da amostra de dados, a perspetiva do anotador e a tarefa em questão. O seu objetivo é medir um atributo não aparente nas redes sociais e através desta teoria torna-se possível transformar observações e anotações em variáveis que refletem uma escala subjacente. Os recursos são medidos numa escala contínua e esta teoria motiva, não apenas o desenvolvimento do conjunto de dados desagregados adequados para análise perspetivista, mas também, dados adequados para medição. Através desta teoria é possível transformar observações, anotações e variáveis que refletem uma escala subjacente. Esta teoria é capaz de avaliar múltiplas contribuições para os rótulos observados através da criação de uma escala de medição, juntamente com um modelo probabilístico multinível que capta contribuições separadas para as classificações nos seus parâmetros. Este modelo coloca os parâmetros ajustados numa escala comum e contínua.

Através desta análise, não podemos deixar de referir a importância dos componentes dos dados de rotulagem, pois pretende-se quantificar a intersubjetividade de cada conjunto. O modelo que incorpora o parâmetro da perspetiva do anotador como uma entrada auxiliar pode gerar previsões a nível de rótulo e pontuação condicionando a perspetiva do mesmo. O corpus inclui os alvos de grupo de identidade de cada comentário com oito grupos (raça/etnia, religião, nacionalidade ou situação de cidadania, sexo, orientação sexual, idade, deficiência e ideologia politica e quarenta e dois subgrupos de identidade-alvo). Relativamente aos dados demográficos, obtivemos seis grupos e quarenta subgrupos, o que facilitou a análise de

interação entre identidade do anotador e nível de comentário, ou seja, a perspetiva do anotador relacionada com a identidade.

**Análise de Dados**

Tendo o trabalho de Kennedy sugerido a realização de futuros conjuntos de dados de discurso de ódio que possam melhorar a construção e o instrumento da rotulagem do corpus MHS dando como sugestão revisões qualitativas, com medições mais precisas, surgiu este novo estudo que tem como objetivo caraterizar as propriedades psicolinguísticas do discurso de ódio, através do método da análise das redes semânticas de conteúdos colocados nas redes sociais, comparando-as com outros tipos de discursos o que permite mapear as interações entre palavras, conseguindo assim o acesso aos aspetos estruturais de utilização da língua.

Para isso foram pré-processados os *tweets* de modo a excluir símbolos, emojis, emoticions, URLS e as contrações foram expandidas (por exemplo, “ASAP” “as soon as possible”. Foram também removidas a pontuação, os números, e as *stop words.* As palavras restantes foram lematizadas. Foi construída uma rede semântica para os *tweets* classificados como discurso de ódio e outra para os *tweets* classificados como discurso de suporte.

Os *tweets* foram classificados com base na variável "*hate speech score*". Se este era maior do que 0.5, considerou-se o *tweet* discurso de ódio; se o “hate_speech_score” era menor do que um considerou-se discurso de suporte. Cada uma das redes foi depois construída com base nos bigramas. As propriedades das redes semânticas são: densidade, diâmetro, sortatividade, comprimento médio do discurso, transitividade e centralização. Densidade corresponde à medida do nível de arestas conectadas dentro de uma rede onde a possibilidade de valor e de retorno é entre o valor decimal de 1 e 0. Diâmetro é o mais longo de todos os comprimentos do caminho. Sortatividade quantifica a tendência de nós individuais se conectarem com outros nós semelhantes (homofilia). Comprimento médio do discurso

corresponde ao caminho mais curto entre os pares de nós, mostrando também o número de passos que leva para chegar de um nó da rede a outro. Transitividade determina a maior das possibilidades de uns nós estabelecerem relações com outros nós na rede. Centralização quantifica a extensão em que um grafo possui um ou mais nós que são significativamente mais centrais do que os outros (Boccaletti et al., 2006). A análise foi realizada no estudo, utilizando o R versão 4.4.0; para o pré-processamento foram utilizados os pacotes do R: udipipe, textclean, tidytext; para a construção da representação gráfica das redes foram utilizados os pacotes: igraph e ggraph.

## Resultados

A Figura 1 representa as redes semânticas do discurso de ódio e do discurso de suporte ou neutro. A Tabela 1 apresenta as características descritivas da topologia de cada uma destas redes.

**Figura 1**

*Redes semânticas do discurso de ódio e do discurso de suporte*

A. Discurso de ódio

B. Discurso de suporte

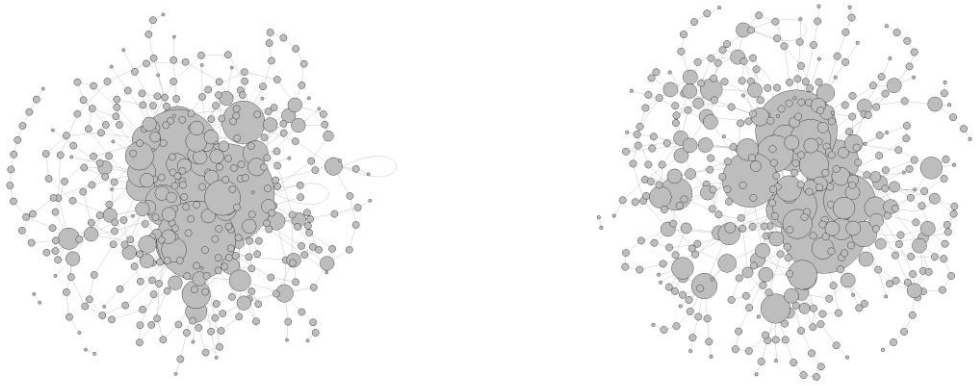

**Tabela 1**

*Características descritivas da topologia do discurso de ódio e discurso de suporte*

| | Sortatividade | Comprimento médio | Densidade | Transitividade | Centralização | Diâmetro |
|---|---|---|---|---|---|---|
| Discurso de ódio | 0.02026693 | 5.761905 | 0.007311723 | 0.02020202 | 0.06775530 | 21 |
| Discurso de suporte | 0.03965517 | 6.207992 | 0.006689777 | 0.02197802 | 0.05010035 | 18 |

Referente ao discurso de ódio, foi possível identificar, que relativamente à sortatividade o valor encontrado foi de 1.020; seguidamente o comprimento médio corresponde a 5.762, sendo que o valor de 0.007 reflete a densidade; a transitividade atingiu o valor de 0.020; na centralização 0.0678 e, por último, no diâmetro 21. No discurso de suporte, o valor da

sortatividade ronda os 0.039. Relativamente ao comprimento médio, o valor é de 6.208, o valor 0.007 é relativo à densidade na transitividade 0.022, seguida da centralização com 0.050 e, por último, com o diâmetro de 18.

Após a análise das propriedades semânticas verificou-se que relativamente à sortatividade, comprimento médio, bem como à transitividade apresentam-se maiores valores no discurso de suporte, o que leva a concluir que neste tipo de discurso existe maior tendência de conexões com os outros semelhantes. O comprimento médio sendo mais alto identifica menor facilidade de passagem de informação, no grupo todos se conhecem, tendo um grupo mais restrito com relações mais concisas, maduras e eficientes como indica o valor de densidade, sendo neste tipo de discurso correspondente a um valor mais baixo. Densidade, centralização e diâmetro têm um valor maior no discurso de ódio. No discurso de ódio existem redes sociais mais densas, com maior fluxo de informação e de intensidade de rede, que espelha uma visão geral de conetividade, mais central, portanto com maior grau de conexão e dependente de um elemento central que fomenta a interação e informa que as relações não são tão eficientes, concisas e maduras.

## Discussão

O cyber*bullying* é um problema da sociedade atual, muito potenciado pela quantidade de equipamentos eletrónicos e do número cada vez maior de pessoas que dominam os ambientes digitais, importantes veículos para os comportamentos de *bullying* (Figueiredo & Matos, 2018; Lee, & Shin, 2017; Watts et al., 2017). Este transcende a presença física e os agressores têm acesso permanente às suas vítimas (Watts et al.,2017). As redes sociais permitem que um indivíduo possa criar o seu próprio perfil, conectando-se a uma lista de outros utilizadores, podendo estas, ser um veículo de propagação e proliferação de comportamentos violentos como o discurso de ódio, que ganham um novo significado, alcance e rapidez (Fabriz & Mendonça,

2022). O discurso de ódio é entendido por Amores et al. (2021) como a existência de mensagens que implicam rejeição, humilhação, assédio, descrédito, estigmatização de indivíduos/comunidades baseados em atributos particulares. Silva et al. (2021) refere que ao tentar agredir uma pessoa no mundo digital, aqueles que compartilham das mesmas caraterísticas, ao entrarem em contato com este tipo de violência, compartilham desta mesma violação. Estes podem criar comunidades na Internet para violentar, difamar de forma intencional o outro (Neves, 2022) e afetar gravemente as pessoas, pois transmite emoções negativas que se refletem nas suas vidas pessoais, podendo abranger a sua família, amigos, trabalho, vida social, entre outros (Oliveira 2008).

Tendo como ponto de partida a abrangência desta problemática na saúde das pessoas, urge o enfoque na deteção do discurso de ódio na Internet, por isso é cada vez mais importante promover a literacia mediática e o desenvolvimento de mecanismos acessíveis de administração, registo e denúncia de situações de discriminação e discurso de incitamento à violência e ao ódio *online.* Embora se verifique uma elevada preocupação com este fenómeno de violência digital, as plataformas e aplicações informáticas para a deteção automática de discursos de ódio são limitadas, apesar destas redes abordarem o fenómeno do ódio *online* como uma prioridade nas suas políticas de conteúdo. Na prática, aplicam estratégias de deteção destes discursos, tendo como prioridade o poder de controlar, bloquear, filtrar e remover qualquer conteúdo agressivo que viole os seus termos. Atualmente, existe uma aposta em programas e aplicações que são desenvolvidos para ações de caráter preventivo, também.

Uma rede semântica é uma estrutura de conhecimento que representa as relações entre os conceitos expressos num texto, a forma como se interconectam e qual o seu sentido. Da perspetiva das ciências da computação, as redes semânticas permitem tornar a informação veiculada nas redes sociais pelos seus utilizadores legível e interpretável por meios computorizados permitindo assim o desenvolvimento de ferramentas tecnológicas que

possibilitam a deteção precoce de discursos de ódio nas redes sociais e assim evitar as suas consequências (Fklih & Al-Turaif, 2023).

O estudo tem como objetivo caracterizar as propriedades psicolinguísticas do discurso de ódio através de um método de análise das redes semânticas e distinguir discurso de ódio de outro tipo de discurso partindo de um estudo realizado por Kennedy que mede o discurso de odio utilizando métodos estatísticos para mapear as suas características através da utilização do (MHS) constituído por 50070 tweets publicados entre Março e Agosto de 2019.

Existe literatura que comprova as inúmeras caraterísticas deste discurso, sejam elas fonéticas, fonológicas, sintáticas, semânticas, pragmáticas, estilísticas ou outras, que podem e devem ser identificadas para evitar a propagação deste discurso. Os marcadores que mais facilmente identificam o discurso de ódio são a linguagem informal (calão), expressões literárias alteradas que demonstram explicitamente a intenção do autor em causar danos à reputação do visado, menosprezar o assunto em discussão ou causar associações negativas à pessoa ou grupo de pessoas. Este tipo de linguagem recorre ainda à estereótipos e à anexação de rótulos, o humor amargo, os apelos diretos à discriminação e ao ódio, o incitamento dissimulado ou vago ao ódio e à violência, a repetição frequente de palavras, frases, sentenças ou pensamentos para enfatizá-los e a distorção de nomes próprios (Haladzhun et al., 2021). De acordo com Schmidt & Wiegand (2017), o número de palavras negativas, conjunto, lista de insultos, declarações polares negativas, calúnias, ou até inclusão de imagens, vídeos e áudio depreciativos parecem também caracterizar o discurso de ódio. Garcia-Diaz et al. (2020), quando refere as características semânticas do discurso de ódio, refere o uso de um discurso forte e ofensivo, como palavrões, dando ênfase a características relacionadas com erros ortográficos e palavras mal escritas. Os media têm também um papel fundamental na identificação da misoginia (sentimento de repulsa, adversão), pelo uso de hashtags (uso do símbolo “#” antes de uma palavra ou expressão) e do calão nas redes sociais, podendo também

encontrar-se emojis negativos, entre outros. Diversos autores (Esteve et al., 2018; Miró-Llinares, 2018; Sachdeva et al., 2022; Watanabe et al., 2018) apontam para a necessidade de serem desenvolvidas técnicas mais eficientes para a correta deteção de discursos de ódio, que têm de ser adequadamente distinguidos de outras expressões discursivas que não constituem discurso de ódio.

Segundo sugestão do estudo anterior poder-se- á realizar um futuro conjunto de dados de discurso de ódio para melhorar a construção e o instrumento de rotulagem do discurso de corpus (MHS) dando como sugestão: revisões qualitativas da construção teorizada, respostas ordinais aos itens de pesquisa de maior subjetividade podem ser ajustadas, as perguntas demográficas do anotador podem ser aprimoradas, rodadas adicionais de anotação podem incluir mais enfase na explicação do anotador para as suas escolhas, que podem ajudar a melhorar a deteção do discurso odio nas redes sociais.

Assim a partir deste novo estudo que tem por base caraterizar as propriedades psicolinguísticas do discurso de ódio, houve lugar à construção de uma rede semântica através de um método de conteúdos das redes sociais - que têm o mérito de mapear dinamicamente as interações entre as palavras - permitindo aceder aos aspetos estruturais de utilização da língua, comparando-as com outros tipos de discursos. Para isso foram pré-processados os *tweets* de modo a excluir símbolos, emojis, emoticions, URLS e as contrações foram expandidas (por exemplo, “ASAP” “as soon as possible”, houve lugar a remoção da pontuação, dos números, e as *stop words.* As palavras restantes foram lematizadas. Foi construída uma rede semântica para os *tweets* classificados como discurso de ódio e outra para os *tweets* classificados como discurso de suporte e analisadas as propriedades das redes semânticas: densidade, diâmetro, sortatividade, comprimento médio do discurso, transitividade e centralização.

Após a análise das propriedades semânticas verificou-se que relativamente à sortatividade, comprimento médio, bem como à transitividade apresentam maiores valores no

discurso de suporte, o que leva a concluir que neste tipo de discurso existe maior tendência de conexões com os outros semelhantes. Densidade, centralização e diâmetro são propriedades com maior valor no discurso de ódio. No discurso de ódio existem redes sociais mais densas, com maior fluxo de informação e de intensidade de rede, que espelha uma visão geral de conetividade, mais central, portanto com maior grau de conexão e dependente de um elemento central que fomenta a interação. Podemos referir, com base neste estudo e através dos resultados apresentados, que estamos em sintonia com o autor, pois é possível, através de pesquisa complementar, obter resultados que, sem dúvida, ajudarão na deteção precoce do discurso de ódio nas redes sociais. Apesar de neste estudo ter comprovado a importância da continuidade deste tipo de trabalhos, sugerimos a continuidade de mais investigação nesta área, pois torna-se uma limitação a abrangência deste tema e as suas repercussões para o indivíduo/família/comunidade.

Em conclusão, a realidade atual sugere a necessidade de uma maior investigação no âmbito dos discursos de ódio, o seu alcance e o seu impacto social, bem como a sensibilização para esta temática. As redes sociais são simultaneamente extensão e complemento da vida física. Por isso, é necessário insistir na ideia de que o que acontece *online* tem impactos reais na vida de cada indivíduo.

## Referências

Anderson, T. &.Sturm.B.W. (2007). Cyberbullying from playground to computer. Young Adult Library Services, 24-27.

Antonio, R., Guerra,R., & Moleiro,C. (2020). *Cyberbullying em Portugal durante a pandemia do covid-19*. https://www.unicef.pt/global-pages/_/porfimaviolencia-nas-escolas/

Alves, C. S., & Serrão, C. (2018). Fatores de risco para a ocorrência de violência contra a pessoa idosa: revisão sistemática. *PAJAR - Pan-American Journal of Aging Research*, *6*(2), 58. https://doi.org/10.15448/2357-9641.2018.2.29964

Amores, J. J., Blanco-Herrero, D., Sánchez-Holgado, P., & Frías-Vázquez, M. (2021). Detecting ideological hatred on Twitter. Development and evaluation of a political ideology hate speech detector in tweets in Spanish. *Cuadernos.Info*, *49*, 98–124. https://doi.org/10.7764/cdi.49.27817

Anjos, J. C. V. (2022). "As garras do feminismo": discurso de ódio antifeminista no Facebook e o senso de urgência controlada. *Intercom: Revista Brasileira de Ciências Da Comunicação*, *45*. https://doi.org/10.1590/1809-58442022119pt

Aricak, T., Siyahhan, S., Uzunhasanoglu, A., Saribeyoglu, S., Ciplak, S., Yilmaz, N., & Memmedov, C. (2008). Cyberbullying among Turkish adolescents. CyberPsychology & Behavior, 11(3), 253-261.

Baigorria, J., Warmling, D., Neves, C. M., Delziovo, C. R., & Coelho, E.B.S. (2017). Prevalência e fatores associados da violência sexual contra a mulher: revisão sistemática. *Revista de Salud Pública*, *19*(6), 818–826. https://doi.org/10.15446/rsap.v19n6.65499

Bauman, S. (2009). Cyberbullying: A virtual menace. University of Arizon

Blanco-Alfonso, I., Rodríguez-Fernández, L., & Arce-García, S. (2022). Polarización y discurso de odio con sesgo de género asociado a la política: análisis de las interacciones en Twitter. *Revista de Comunicación*, *21*(2), 33–50. https://doi.org/10.26441/RC21.2-2022-A2

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. (2006). Complex networks: Structure and dynamics. Physics Reports, 424(4–5), 175–308. https://doi.org/10.1016/j.physrep.2005.10.009

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. In InterJournal: Vol. Complex Systems (p. 1695). https://igraph.org

Calvete, E., Orue, I., Estévez, A., Villardón, L., & Padilla, P. (2010). Cyberbullying in adolescents: Modalities and aggressors' profile. Computers in Human Behavior 26, 1128–1135.

Campbell, M. A., Slee, P. T., Spears, B., Butler, D., & Kift, S. (2013). Do cyberbullies suffer too? Cyberbullies' perceptions of the harm they cause to others and to their own mental health. School Psychology International, 34(6), 613-629.

Cheminais, R. (2008). *Every Child Matters: A Practical Guide for Teaching Assistants*. https://doi.org/10.4324/9780203462744.

Chen, W.-Y., & Corvo, K. (2011). *Community Violence*. Springer New York. https://doi.org/https://doi.org/10.1007/978-1-4419-1695-2_139

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118. https://doi.org/10.1016/j.avb.2018.05.003Choi, K.-S., & Lee, J. R. (2017). Theoretical analysis of cyber-interpersonal violence victimization and offending using cyber-routine activities theory. *Computers in Human Behavior*, *73*, 394–402. https://doi.org/10.1016/j.chb.2017.03.061

Deschamps, R., & McNutt, K. (2016). Cyberbullying: What's the problem? *Canadian Public Administration*, *59*(1), 45–71. https://doi.org/10.1111/capa.12159

Digiampietri, L. A. (n.d.). Análise de Redes Sociais. https://www.each.usp.br/digiampietri/AnaliseDeRedesSociais/03_AnaliseEstrutural.pdfEsteve, M., Miró, F., & Rabasa, A. (2018). Classification of tweets with a mixed method based on pragmatic content and meta-information. *International Journal of Design & Nature and Ecodynamics: A Transdisciplinary Journal Relating to Nature, Science and the Humanities*, 13(1), 60–70. https://doi.org/10.2495/dne-v13-n1-60-70Fabriz, D. C., & Mendonça, G. H. (2022). O papel das plataformas de redes sociais diante do dever de combater o discurso de ódio no Brasil. *Revista Da Faculdade de Direito UFPR*, *67*(1), 127. https://doi.org/10.5380/rfdufpr.v67i1.83904

Figueiredo, F., & Matos, A. (2018). *Agressão apoiada pelas tecnologias: o cyberbullying e o autocyberbullying*.

Fkih, F., & Al-Turaif, G. (2023). Threat modelling and detection using semantic network for improving social media safety. Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia, International Journal of Computer Network and Information Security, 15(1), 39–53. https://doi.org/10.5815/ijcnis.2023.01.04

Flach, R. M. D., & Deslandes, S. F. (2017). Abuso digital en relaciones afectivo-sexuales: Un análisis bibliográfico. In *Cadernos de Saude Publica* (Vol. 33, Issue 7). Fundacao Oswaldo Cruz. https://doi.org/10.1590/0102-311X00138516

Fortuna, P., & Nunes, S. (2019). A survey on automatic detection of hate speech in text. ACM Computing Surveys, 51(4), 1–30. https://doi.org/10.1145/3232676

García-Díaz, J. A., Jiménez-Zafra, S. M., García-Cumbreras, M. A., & Valencia-García, R. (2023).Evaluating feature combination strategies for hate-speech detection in Spanish

using linguistic features and transformers. Complex & Intelligent Systems, 9(3), 2893–2914. https://doi.org/10.1007/s40747-022-00693-xGonzález-Cabrera, J., Díaz-López, A., Caba-Machado, V., Ortega-Barón, J., Echezarraga, A., Fernández-González, L., & Machimbarrena, J. M. (2023). Epidemiology of peer cybervictimization and its relationship with health-related quality of life in adolescents: A prospective study. *Journal of Adolescence*, *95*(3), 468-478.

Gradinger, P., Strohmeier, D., & Spiel, C. (2009). Traditional bullying and cyberbullying: Identification of risk groups for adjustment problems. Journal of Psychology, 217(4), 205-213.

Haladzhuna ,Z.,Harmatiya , O.,Bidzilyab ,Y., Kunanetsa, N & Shunevycha, K., (2021). Hate Speech in Media Towards the Representatives of Roma Ethnic Community. Lviv Polytechnic Nationality University, State University 'Uzhhorod National University.'

Harff, G., & Duque, M. S. (2020). Discurso de ódio. *Revista Da Faculdade de Direito Da Universidade Federal de Uberlândia*, *48*(2), 264–295. https://doi.org/10.14393/RFADIR-v48n2a2020-54296

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. Deviant behavior, 29(2), 129-156.

Iffland, B., Bartsch, L. M., Kley, H., & Neuner, F. (2023). Growing relevance of reports of adolescent cyberbullying victimization among adult outpatients. *BMC Public Health*, *23*(1), 1503.

Ioannou, A., Blackburn, J., Stringhini, G., Cristofaro, E., Kourtellis, N., & Sirivianos, M. (2018). From risk factors to detection and intervention: a practical proposal for future work on cyberbullying. *Behaviour & Information Technology*, *37*(3), 258–266. https://doi.org/10.1080/0144929X.2018.1432688

Jones, L. M., Mitchell, K. J., & Finkelhor, D. (2013). Online harassment in context: Trends from three Youth Internet Safety Surveys (2000, 2005, 2010). Psychology of Violence, 3(1), 53-69.

Katz, A. (2012). *Cyberbullying and E-safety*: *What educators and other professionals need to know*. Jessica Kingsley

Klein, L. C. A., Gadelha, G. M. D. B., & Coura, A. C. (2022). Crise Democrática Brasileira e Disfuncionalidade e o Direito – À Liberdade de Expressão: Críticas ao Discurso de Ódio Sob o Viés da Teoria do Discurso Jurídico Habermasiano. *Revista Internacional Consinter de Direito*, 79–93. https://doi.org/10.19135/revista.consinter.00014.02

Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. Journal of adolescent health, 41(6), 22-30.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618

Kowalski, R. M., Morgan, C. A., & Limber, S. P. (2012). Traditional bullying as a potential warning sign of cyberbullying. *School Psychology International*, *33*(5), 505–519. https://doi.org/10.1177/0143034312445244

Krug, E. G., Mercy, J. A., Dahlberg, L. L., & Zwi, A. B. (2002). The world report on violence and health. *The Lancet*, *360*(9339), 1083–1088. https://doi.org/10.1016/S0140-6736(02)11133-0

Leal, E. V. S. (2021). *Violência Doméstica-Uma análise do discurso de juízes/as desembargadores/as nos acórdãos do Tribunal da Relação de Lisboa*.

Lee, C., & Shin, N. (2017). Prevalence of cyberbullying and predictors of cyberbullying perpetration among Korean adolescents. *Computers in Human Behavior*, *68*, 352–358. https://doi.org/10.1016/j.chb.2016.11.047

Li, Q. (2005). Cyberbullying in Schools: Nature and Extent of Canadian Adolescents' Experience.

Li, Q. (2006). Cyberbullying in schools a research of gender differences. School psychology international, 27(2), 157-170.

Martins, A. C. L. (2019). Hate speech in social networks and recognition of the other: The M. case. *Revista Direito GV*, *15*(1). https://doi.org/10.1590/2317-6172201905

Martins, M. J. D. (2007). Violência interpessoal e maus-tratos entre pares, em contexto escolar. *Revista Da Educação*, *XV*(2), 51–78.

Miró-Llinares, F., Moneva, A., & Esteve, M. (2018). Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments. *Crime Science, 7*(1), 15. https://doi.org/10.1186/s40163-018-0089-1

Mishna, F., Khoury-Kassabri, M., Gadalla, T., & Daciuk, J. (2012). Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims. Children and Youth Services Review, 34(1), 63-70.

Montag, C., Demetrovics, Z., Elhai, J.D., Grant, D., Koning, I., Rumpf, H.J., M Spada M, Throuvala, M., van den Eijnden, R. (2024). Problematic social media use in childhood and adolescence. *Addictive Behaviors*,153(107980), 107980. https://doi.org/10.1016/j.addbeh.2024.107980

Montero, A. I., Laforgue-Bullido, N., & Abril-Hervás, D. (2022). Hate speech: a systematic review of scientific production and educational considerations. *Revista Fuentes*, *2*(24), 222–233. https://doi.org/10.12795/revistafuentes.2022.20240

Moraes, P. A. , & Pamplona, D. A. (2019). O discurso de ódio como limitante da liberdade de expressão. *REVISTA QUAESTIO IURIS*, *12*(2). https://doi.org/10.12957/rqi.2019.37081

Neves, L. H. P. (2022). *Aplicação para deteção automática de discurso de ódio em lingua portuguesa recorrendo a aprendizagem computacional*. Instituto Politécnico de Leiria Escola Superior de Tecnologia e Gestão

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. Proceedings of the 25th International Conference on World Wide Web. (pp. 145–153). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883062

Oliveira, S. (2008). Ciberbullying: um fenómeno sem rosto. Educare.Pt. http://www.educare.pt/educare/Actualidade.Noticia (Acedido 21 de maio de 2024)

Patchin, J. W., Hinduja, S., & Meldrum, R. C. (2023). Digital self-harm and suicidality among adolescents. *Child and adolescent mental health*, *28*(1), 52-59.

Pedersen, T. L. (2024). ggraph: An implementation of grammar of graphics for graphs and networks (Version 2.2.1) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=ggraph

Peterson, J., & Densley, J. (2017). Cyber violence: What do we know and where do we go from here? *Aggression and Violent Behavior*, *34*, 193–200. https://doi.org/10.1016/j.avb.2017.01.012

Pinheiro, L. O. (2009). Cyberbullying em Portugal: uma perspectiva sociológica. Dissertação de Mestrado, Instituto de Ciências Sociais – Universidade do Minho, Portugal.

Price, M., & Dalgleish, J. (2010). Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people. Youth Studies Australia, 29(2), 51-59.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria URL https://www.R-project.org/.

Raigada, J. L. P., Solana, M. Y. M., & García, M. T. M. (2022). Una exploración del capital cognitivo ante discursos del odio por racismo. *Perspectivas de La Comunicación*, *15*(2), 59–98. https://doi.org/10.56754/0718-4867.1502.059

Rocha, G. M. (2021). O discurso de ódio e a violência nele contida. *Revista de Teorias Da Democracia e Direitos Políticos*, *7*(1), 147. https://doi.org/10.26668/IndexLawJournals/2525-9660/2021.v7i1.7856

Rollnert Liern, G. (2019). El discurso del odio: una lectura crítica de la regulación internacional. *Revista Española de Derecho Constitucional*, *115*, 81–109. https://doi.org/10.18042/cepc/redc.115.03

Rothenburg, W. C., & Stroppa, T. (2015). Liberdade de expressão e discurso do ódio: o conflito discursivo nas redes sociais freedom of expression and hate speech: the discursive conflict in social networks. *Anais Do 3º Congresso Internacional de Direito e Contemporaneidade*. http://www.ufsm.br/congressodireito/anais

Sachdeva, P. S., Barreto, R., Bacon, G., Sahn, A., Von Vacano, C., & Kennedy, C. J. (2022). *The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism*. https://huggingface.co/datasets/

Şahin, S. S., & Ayaz-Alkaya, S. (2023). Prevalence and predisposing factors of peer bullying and cyberbullying among adolescents: A cross-sectional study. *Children and Youth Services Review*, *155*, 107216.

Salazar, L. R., Weiss, A., Yarbrough, J. W., & Sell, K. M. (2024). Cyberbullying of university faculty: An examination of prevalence, coping, gender, and personality factors. *Computers in Human Behavior*, 108186.

Santos, B. M. de S., & de Farias, W. S. (2022). Who is afraid of the teacher? the hate speech addressed to teachers in virtual space. *Cadernos de Pesquisa*, *52*. https://doi.org/10.1590/198053149348

Santos, M. F. T. (2015). *Cyberbullying na adolescência: perfil psicológico de agressores, vítimas e observadores.* Universidade de Lisboa.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media.

Schneider, S. K., O'Donnell, L., Stueve, A., & Coulter, R. W. (2012). Cyberbullying, school bullying, and psychological distress: A regional census of high school students. American Journal of Public Health, 102(1), 171-177.

Seixas, S. Fernandes, L. & Morais, T. (2016). *CyberBullying – Um guia para pais e educadores* (platano, Ed.).

Seo, M. (2021). textclean: Text Cleaning Tools (Versão 0.9.3) [R package]. https://CRAN.R-project.org/package=textclean

Sierra, L. M. G. (2022). La discriminación vive en las calles. Análisis de vivencias rutinarias que configuran discriminación contra las mujeres. *Estudios Socio-Jurídicos*, *24*(1). https://doi.org/10.12804/revistas.urosario.edu.co/sociojuridicos/a.10009

Silge, J., & Robinson, D. (2022). tidytext: Text Mining using 'dplyr', 'ggplot2', and other tidy tools (Versão 0.3.4) [R package]. https://CRAN.R-project.org/package=tidytext

Silva, M.P., & Silva, L.S. (2021). Disseminação de discursos de ódio em comentários de

notícias: uma análise a partir de notícias sobre o universo lgbt em cibermeios sul-mato-grossenses no facebook. *Intercom, Revista Brasileira ciências*, 44 (2), 137 – 155. https://doi.org/10.1590/1809-5844202127

Silva, J. L., & Mascarenhas, Suely. A. N. (2010, July). Gestão do bullying e cyberbullying na universidade- desafios para a orientação educativa e convivência social e ética no ensino superior-estudo com estudantes da UFAM/brasil. *Revista AMAzônica*, , *3*, 46–55.

Silva, L. R. L., Botelho-Francisco, R. E., Oliveira, A.A.A., & Pontes, V. R. (2019). A gestão do discurso de ódio nas plataformas de redes sociais digitais: um comparativo entre Facebook, Twitter e Youtube. *Revista Ibero-Americana de Ciência Da Informação*, *12*(2), 470–492. https://doi.org/10.26512/rici.v12.n2.2019.22025

Smith, P. K., Mahdavi, J., Carvalho, M., & Tippett, N. (2006). An investigation into cyberbullying, its forms, awareness and impact, and the relationship between age and gender in cyberbullying.

Šincek, D., Duvnjak, I., & Milić, M. (2017). Psychological Outcomes of Cyber-Violence on Victims, Perpetrators and Perpetrators/Victims. *Hrvatska Revija Za Rehabilitacijska Istraživanja*, *53*(2), 98–110. https://doi.org/10.31299/hrri.53.2.8

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? Scandinavian journal of psychology, 49(2), 147-154.

Tarshis, T. P. (2010). *Living with Peer Pressure and Bullying*. Infobase Publishing

Teles, Á. R. (2019). *Deteção de haters nas redes sociais*. Universidade de Caxias do Sul.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior, 26, 277–287.

Walrave, M., & Heirman, W. (2011). Cyberbullying: Predicting victimization and perpetration. Children & Society, 25(1), 59-72.

Wang, J., Iannotti, R. J., & Nansel, T. R. (2009). School bullying among adolescents in the United States: Physical, verbal, relational, and cyber. Journal of Adolescent health, 45(4), 368-375.

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access*, *6*, 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394

Watts, L. K., Wagner, J., Velasquez, B., & Behrens, P. I. (2017). Cyberbullying in higher education: A literature review. *Computers in Human Behavior*, *69*, 268-274. https://doi.org/10.1016/j.chb.2016.12.038

Weigle, P. E., & Shafi, R. M. (2024). Social media and youth mental health. *Current psychiatry reports*, *26*(1), 1-8.Wijffels, J., BNOSAC, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic, Straka, M., & Straková, J. (2023). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit (Version 0.8.11) [R package]. CRAN. https://CRAN.Rproject.org/package=udpipe

Willard, N. (2007). The Authority and Responsibility of School Officials in Responding to Cyberbullying. *Journal of Adolescent Health*, 41, 564-565.

Williams, K. R., & Guerra, N. G. (2007). Prevalence and predictors of internet bullying. Journal of Adolescent Health, 41(6), 14-21.

Yue, Z., & Rich, M. (2023). Social Media and Adolescent Mental Health. *Current Pediatrics Reports*, *11*(4), 157-166.