Feminismo e Redes Sociais Online: uma Análise de Tweets sobre o Dia Internacional da Mulher

Geandreson de S. Costa[1], **Danielle C. C. Couto**[1],
Antonio F. L. Jacob Junior[2], **Fábio M. F. Lobato**[2,3]

[1]Universidade Federal do Pará (UFPA), Belém, Pará, Brasil

[2]Universidade Estadual do Maranhão, São Luís, Maranhão, Brasil

[3]Universidade Federal do Oeste do Para (UFOPA), Santarém, Pará, Brasil

danifc@ufpa.br, fabio.lobato@ufopa.edu.br

Abstract. *Social networks are playing an increasingly important role in supporting the discourses and agendas of the current feminist movement. Aiming to identify which themes are addressed by the feminist agenda worldwide and which polarities are present in these manifestations, this paper analyzes data collected from Twitter related to International Women's Day. For this, topic modeling and sentiment analysis were applied. The data used was collected in real-time during the days before and after the 8th of March in the years 2020 and 2021. The results showed that the topics found vary from one year to the next, but all are confluent with the movement. Also, some topics are always addressed, and the polarity towards these manifestations tends to be mostly neutral.*

Resumo. *As redes sociais estão desempenhando um papel cada vez mais importante no suporte a discursos e agendas do movimento feminista atual. Visando identificar quais as temáticas abordadas pela agenda feminista ao redor do mundo e quais polaridades estão presentes nessas manifestações, este trabalho analisa dados coletados do Twitter relacionados ao Dia Internacional da Mulher. Para isso, foram aplicadas modelagem de tópicos e análise de sentimento. Os dados utilizados foram coletados em tempo real durante os dias anteriores e posteriores ao 8 de março nos anos de 2020 e 2021. Os resultados mostraram que as temáticas encontradas variam de um ano para o outro, mas todos estão confluentes com o movimento. E ainda, existem tópicos que sempre são abordados e que a polaridade em relação a essas manifestações tende a ser de maioria neutra.*

1. Introdução

Estima-se que, atualmente, existam cerca de 4,2 bilhões de contas ativas em Redes Sociais no mundo [We Are Social 2021]. Consequentemente, estas têm ganhado cada vez mais protagonismo, uma vez que se apresentam como o principal meio de interações sociais, no qual seus usuários podem comunicar-se entre si, manifestar seus desejos, pensamentos, opiniões e debater sobre eles [Tajudeen et al. 2018].

As interações entre atores de uma rede social produzem um grande volume de dados, chamada de Conteúdo Gerado pelo Usuário (*User-Generated Content* - UGC)

[Russell 2013]. Estes dados representam matéria-prima de grande valor para diversas áreas, da academia à indústria. Por meio de *insights* obtidos a partir das análises desses dados é possível, por exemplo, adaptar um produto, serviço ou estratégia de negócio, de modo a melhor se adequar aos anseios dos consumidores [Lobato et al. 2016]. Além disso, esses dados podem permitir a analise mais profunda da natureza de um movimento político-social, suas formas de comunicação, conteúdos abordados, *etc.* [Mundt et al. 2018].

Nos últimos anos, um dos movimentos sociais mais proeminentes nos ambientes digitais é o movimento feminista. Os feminismos, diversos em suas demandas e objetivos, caracterizam-se principalmente por terem em comum o fato de buscar equidade de gênero e vêm ganhando espaço para as suas agendas nos ambientes virtuais [Locke et al. 2018]. Com isso, percebe-se uma maior exposição a termos pertencentes a estes movimentos, conjugado a uma busca em compreender seus significados. Por exemplo, em 2017 o dicionário Merrian-Webster elegeu o feminismo como a palavra do ano [Merriam-Webster 2017]. Dados obtidos por meio do *Google Trends*[1] mostram que nos últimos cinco anos a busca pelo termo feminismo vem se popularizando. Ao analisar esses dados, observa-se que as buscas por este termo coincidem com o mês de março, quando anualmente é comemorado o Dia Internacional da Mulher, conhecido também pelo seu acrônimo em inglês IWD, de *International Women's Day*.

Redes sociais como o Twitter costumam ser sensíveis às dinâmicas de popularidade de determinados assuntos. Por exemplo, no IWD apresenta um grande fluxo de conteúdos de apoio ou discussão das causas do movimento feminista, ou ainda, tentativas de desmerecê-los, reproduzindo ações misóginas e machistas contra ao qual o movimento tanto luta [Fuchs and Schäfer 2019]. Consequentemente, um grande volume de UGC é gerado. Diante desse contexto, o presente trabalho tem o objetivo de apresentar uma análise das manifestações do movimentos feministas no Twitter ao redor do mundo por ocasião IWD. Para isso, buscou-se identificar quais as principais temáticas abordadas pelas agendas feministas e quais os sentimentos relacionados à essas manifestações. Para tal, foram implementadas algumas técnicas de Processamento de Linguagem Natural (PLN) como Modelagem de Tópicos e Análise de Sentimentos (AS). O conjunto de dados utilizado foi construído a partir da coleta de *tweets* em tempo real nos anos de 2020 e 2021.

A motivação para o trabalho surgiu do interesse em contribuir com pesquisas sobre as temáticas feministas, apresentando-as mais uma análise do tema sob a ótica netnográfica. Embora haja aumento recente no número de trabalhos que abordam e analisam a temática sob essa ótica, a maioria ainda é composta por análises em que aspectos qualitativos constituem no cerne das investigações em detrimento de abordagens quanti-quali [Stauffer and O'Brien 2018].

O artigo está organizado da seguinte maneira: na Seção 2, é feita uma discussão dos trabalhos correlatos ao presente estudo. Na Seção 3 é apresentada a metodologia que guiou o trabalho, seguido da Seção 4, onde são apresentados os resultados das análises implementadas. Por fim, as conclusões, aspectos éticos e sugestões de trabalhos futuros são dadas na Seção 5.

[1]Disponíveis em `https://bit.ly/3a9AlTf` e `https://bit.ly/3wLpEy5`. Acesso em 10 dez. 2021

2. Trabalhos Relacionados

A atenção aos movimentos feministas nas últimas décadas vem promovendo uma maior interação entre ciências sociais, ciências da informação e computação [Leung et al. 2019]. Este fenômeno é motivado, além do grande volume e disponibilidade do UGC, pela criação, desenvolvimento e evolução de novos métodos de análise de mídias sociais [Reyes-Menendez et al. 2020]. Nesse sentido, [Deriu and Iezzi 2020] chamam a atenção para esse novo cenário de multidisciplinaridade orientado a dados. Nele, pesquisas de vários campos poderiam ser expandidas analisando dados textuais usando técnicas de mineração de texto e PLN.

Análises feitas a partir de termos e *hashtags* do Twitter são as mais comuns, pois possibilitam a coleta e criação de base de dados de melhor qualidade [Hino and Fahey 2019, Lima Jr. et al. 2020]. O trabalho de [Puente et al. 2021] examinou demonstrações no Twitter, que continham as *hashtags* mais comumente utilizadas, como *#8M* e *#NiUnaMenos*, vindas do movimento feminista espanhol durante o 08 de março de 2017. [Goel and Sharma 2020] analisaram *tweets* que continham a termo *#MeToo*, os quais foram inspirados nos relatos publicados pelo jornal *The New York Times* sobre assédios e abusos sexuais em *Hollywood* [2]. [Lommel et al. 2019] explorou o conceito de identidade coletiva de ativistas feministas no Twitter durante os protestos de janeiro de 2017 nos Estados Unidos, contra o então recém-eleito presidente americano Donald Trump. [Rodrigues et al. 2019] analisaram comentários de notícias sobre tentativa de feminicídio da paisagista Elaine Caparroz, ocorrido no Rio de Janeiro em 2019. Os resultados evidenciaram que os comentários convergem para a culpabilização da vítima.

Em [Rodriguez et al. 2020] os autores analisaram o impacto social da performance "um estuprador em seu caminho" - *"un violador em tu caminho*", em espanhol - proposto pela coletivo feminista chileno *La Tesis*. As performances ocorreram primeiramente em cidades chilenas e, logo depois, foram replicadas em várias cidades do mundo. A repercussão foi analisada por meio de *tweets* que mencionaram o termo *#LaTesis*. [Yagui et al. 2017] objetivaram identificar quais as opiniões e sentimentos que estavam relacionados ao movimento em relação ao "Bela, recatada e do lar", expressão popularizada por meio de uma matéria de uma revista de circulação nacional [3]. [Dilai and Levchenko 2018] buscaram compreender qual a percepção que os ucranianos tinham a respeito de feminismo, utilizando *tweets* como fonte de dados.

Alguns trabalhos não ficam somente nos ambientes online e buscam avaliar o quanto o contexto influencia o discurso. Por exemplo, [Scarborough and Helmuth 2021] investigaram qual a relação entre os contextos *online* e *offline* dos discursos do movimento feminista, avaliando se características locais e culturais pode prever o apoio ou não a este movimento social no Twitter. Já em [Scarborough 2018], o autor se propõe a encontrar uma estimativa que represente a opinião pública a respeito do feminismo, calculando a nível regional, estadual e municipal; e examinar se essa estimativa representa as atitudes de diversas populações, diferentes em raça, gênero e classe social.

[2]Disponível em `https://nyti.ms/38vTPAN`. Acesso em 04 abr. 2022

[3]Disponível em `https://bit.ly/3r73QdZ`. Acesso em 22 mar. 2022

3. Materiais e Métodos

Por ser amplamente utilizada em projetos de mineração e análise de dados, optou-se pelo processo *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [Schröer et al. 2021]. O CRISP-DM é composto por seis etapas, a saber: i) entendimento do negócio; ii) entendimento dos dados; iii) preparação dos dados; iv) modelagem; v) avaliação; e vi) entrega. Trata-se de um processo incremental e cíclico, sendo possível retornar a etapas anteriores sempre que necessário. A seguir é apresentado como as etapas do CRISP-DM foram conduzidas no presente estudo.

3.1. Entendimento do negócio

Nesta etapa realizou-se uma breve revisão de literatura sobre alguns temas como redes sociais, movimentos sociais, ativismo digital e feminismo. Isto auxiliou no conhecimento e assimilação de vários conceitos que permeiam essas temáticas. Em seguida, foi definido que seriam usados os dados da rede social Twitter, coletados em tempo real, para a construção de uma base com os dados a serem analisados. O Twitter dispõe de uma *Application Programming Interface* (API) para a coleta dos dados, de fácil manuseio, contrastando com outras redes sociais como o Facebook e o Instagram.

Definiu-se, também, que seriam desenvolvidos *scripts* na linguagem Python, pelo vasto conjunto de ferramentas para tarefas de mineração de dados textuais oferecidos nesta linguagem [Piatetsky 2019]. A base de dados foi criada no banco de dados não relacional MongoDB, pois este baseia-se em documentos na estrutura chave-valor, não sendo necessária modelagens e normalizações preliminares. Sendo assim, os dados são inseridos na estrutura que são coletados, no caso, os resultados retornados pela API no formato *JavaScript Object Notation* (JSON).

3.2. Entendimento e Preparação dos dados

A coleta dos dados começou sempre no dia anterior ao dia internacional da mulher e se estendeu até um dia após, ou seja, entre os dias 07 e 09 de março, conforme mostrado mais detalhadamente no Tabela 1. O intuito era coletar uma vasta quantidade de *tweets* oriundos de vários países, em diferentes fuso horários, publicados antes, durante e após a comemoração e que tivessem relação com a data.

Tabela 1. Datas e horários das coletas.

Ano	Início	Fim
2020	07/03 21h00 UTC -3	09/03 17h00 UTC -3
2021	07/03 19h00 UTC -3	09/03 19h00 UTC -3

Este processo foi realizado usando redundância com mais de três servidores, visando garantir que a extração dos *tweets* superassem percalços que poderiam ocorrer, como instabilidade na conexão; problemas com a API do Twitter (expiração dos *tokens* de acesso); ou com a infraestrutura dos computadores utilizados. As *queries* utilizadas para a busca dos *tweets* foram, primeiramente, baseadas em *hashtags* e termos mapeados a partir de um levantamento feito numa base de dados pré-existente. Essa base continha pouco mais de 400 mil *tweets* e foi coletada no IWD de 2019 pelo coordenador deste estudo, servindo como protótipo para a metodologia aqui apresentada.

De forma sucinta, a definição dos termos de busca ocorreu da seguinte forma. Inicialmente, o grupo havia mapeado as principais *hashtags* utilizadas no IWD por meio de levantamento bibliográfico e consulta a estudiosas na área. Uma lista não exaustiva das *hashtags* mapeadas inclui: *#8M*, *#8M<ANO>*, *#IWD*, *#feminismo*, *#8deMarco*, *etc*. Importante destacar que alguns termos, como o *#8M*, agregam postagens de diversos idiomas. Com esta lista definida, deu-se inicio ao processo de coleta. Em tempo real, também, era realizado o mapeamento de co-ocorrência de *hashtags* (e.g.: *"#IWD"* e *"#MeeToo"*) e, também, da consulta às tendências disponibilizadas no sítio web *trends24* [4], o qual dispunha da identificação de *trending topics* em vários países. As tendências do ano sob escrutínio relacionadas ao IWD eram então incluídas na lista de busca.

Após o mapeamento inicial, a busca contava com 22 *queries* em inglês, espanhol e português. Ao fim da última atualização, a lista possuía 49 itens, contando com idiomas como alemão, francês, russo, turco, entre outros. A lista completa das *queries* para o ano de 2020 é mostrada nas Tabela 2. Para o ano de 2021, foram utilizados os mesmos termos de busca utilizados no ano anterior, atualizando suas grafias, como por exemplo, *#8M2021* e *#IWD2021*. Vale notar que o passo de atualização usando tendências, também, foi realizado neste ano.

Tabela 2. *Queries* utilizadas na busca e coleta dos *tweets* para o ano de 2020.

Queries iniciais	8M, 8M2020, 8deMarco, DiaInternacionalDaMulher, DiaDaMulher, MulheresNaHistória, MaisJuntasQueNunca, ElaMeInspira, HuelgaFeminista2020, DiaDeLaMujer, DiaInternacionalDeLaMujer, FelizDiaDeLaMujer, 8marzo, InternationalWomensDay, WomensDay, HappyWomensDay2020, IWD, IWD2020, IWDay2020, Dia Internacional da Mulher, International Womens Day, Dia Internacional de la Mujer, EachforEqual
Queries adicionadas	8martdunyakadinlargunu, sororidade, sorority, allaboutwomen, Happy IWD, يوم_المراه_العالمي, MarcheFéministe, TuNesPasCoupable, weltfrauentag2020, JourneeDesDroitsDesFemmes, marchefeministe, 8marts, Weltfrauentag, 8Marzo, InternationellaKvinnodagen, 8March, Marcha8M, 8марта, марта, С 8 Марта, #международныйженскийдень, Международным, kvinnedagen, internationalevrouwendag, 8mars, GenerationEquality, Παγκόσμια ημερα της γυναίκας.

Com o intuito de se avaliar o impacto/engajamento com os *tweets*, foi realizado um processo de atualização dos mesmos. O *script* consistia na consulta do *tweet* pelo seu identificador usando a API e na atualização das informações do mesmo (*e.g., curtidas, compartilhamento, comentários e edições*. Considerando aspectos éticos, caso o *tweet* tenha sido deletado pelo usuário, o mesmo era removido da base.

Em relação ao pré-processamento, considerando as especificidades idiomáticas, o primeiro passo foi dividir a base por idioma. Foram selecionados para o presente estudo os três idiomas mais prevalentes, a saber: inglês, espanhol e português. Isto foi feito visando

[4]Disponível em `https://trends24.in/` Acesso em 10 dez. 2021

um melhor acompanhamento dos experimentos e, também, considerando a quantidade de técnicas já desenvolvidas ou adaptadas para analisá-los [Pereira 2021].

O pré-processamento do conteúdo textual foi realizado baseando-se nos trabalho de [Cirqueira et al. 2018, Duong and Nguyen-Thi 2021], utilizando a biblioteca *Pandas* [5]. As etapas de pré-processamento aplicadas aos dados foram: remoção de links, remoção de menções à usuários, remoção de *hashtags*, remoção de quebras de linhas, remoção de sinais de pontuação, remoção de caracteres numéricos, remoção de espaços duplos, conversão para caracteres em minúsculo, remoção de caracteres repetidos em sequência, remoção de *stopwords*, remoção de palavras de tamanho menor de 3 caracteres, remoção de acentuação e remoção de caracteres *non-ASCII*. Ao final desta fase, foram criados arquivos no formato *comma separeted values* (CSV), utilizados como entrada para a modelagem dos dados.

3.3. Modelagem dos Dados

A primeira análise realizada nessa etapa foi a Modelagem de Tópicos. Essa é uma técnica de mineração de textos que objetiva a descoberta de estruturas temáticas subjacentes, dado um conjunto de documentos, sem que para isso ocorra alguma espécie de treinamento prévio do algoritmo. Uma modelagem de tópicos consiste em se encontrar os k tópicos - um padrão recorrente de co-ocorrência de palavras mais proeminentes nos documentos. Cada tópico é representado por uma lista ranqueada de termos fortemente correlacionados, onde cada documento pode estar associado a um ou mais tópicos [Belford et al. 2018].

A *Non-Negative Matrix Factorization* (NMF) é uma abordagem não-supervisionada para a redução de dimensionalidade de matrizes não negativas. Dada uma matriz $A_{n\times m}$, esta é decomposta em duas outras matrizes, $W_{n\times k}$ e $H_{k\times m}$, onde $A \cong WH$. Nessa abordagem, as k colunas de W podem ser interpretadas como sendo os tópicos, composto pelos principais n termos ponderados, enquanto que a matriz H provê conhecimento da relação dos m documentos (colunas) com os k tópicos (linhas). Em contextos em que os textos possuem um tamanho diminuto, frequentemente existente na comunicação via redes sociais, a NMF consegue produzir tópicos mais coerentes em comparação a outras abordagens [Nugroho et al. 2020]. A aplicação da NMF ocorreu nas seguintes etapas, propostas por [Greene et al. 2014] e adaptadas para o contexto deste trabalho:

1. Criação da matriz de entrada usando o método *Term Frequency - Inverse Document Frequency* (TF-IDF);
2. Definição de um intervalo de diferentes valores para k;
3. Para cada valor de k, aplicar a NMF na matriz de entrada;
4. Avaliação e validação dos cenários obtidos.

Outra tarefa realizada nesse estágio foi a Análise de Sentimentos. Esta técnica objetiva a classificação de dados textuais de acordo com polaridades, como por exemplo, positivo, negativo ou neutro [Medhat et al. 2014]. Com a recente popularidade do campo nos últimos anos, vários algoritmos, métodos, abordagens e aplicações com essa finalidade vêm sendo desenvolvidos [Araújo et al. 2020].

[5] Disponível em `https://pandas.pydata.org/`. Acesso em 04 abr. 2022

Para realizar essa análise no presente trabalho utilizou-se o iFeel 2.0, sistema que classifica a polaridade dos textos implementando 17 métodos de AS [Araújo et al. 2016]. Este sistema utiliza a escala 1, 0 e -1 para a classificação em positivo, neutro e negativo, respectivamente. O iFeel 2.0, também, possui suporte multi-idiomático por meio do tradutor Yandex. Apesar de robusto, o iFeel limita o processamento à arquivos com 1.000 linhas de entrada. Visando contornar tal restrição, a base de dados foi subdividida em vários arquivos CSV de 1.000 linhas e a automatização do processamento foi feito usando a biblioteca *os* do Python. Por fim, os dados de saída foram reagrupados para se computar as estatísticas para toda a base.

3.4. Avaliação e Entrega

As últimas duas etapas do CRISP-DM consistem na avaliação dos resultados para posterior entrega. A avaliação do presente estudo foi realizada de forma similar à [Rodrigues et al. 2022] e consistiu na: i) validação e anotação dos tópicos pelos autores, buscando o consenso, abordagem baseada na Metodologia Delphi; ii) validação da análise de sentimentos por meio de *ground truth* - onde *tweets* foram aleatoriamente selecionados e inspecionados qualitativamente pelos autores. A entrega consiste no presente documento, apresentações e, também, na construção de relatório técnico a ser disponibilizado para a comunidade científica.

4. Resultados

Foram obtidos um total de 9.647.404 e 11.193.522 *tweets* para os anos de 2020 e 2021, respectivamente. Conforme sinalizado na subseção 3.2, as coletas ocorreram em mais de um local e, dessa forma, foram utilizadas três bases na coleta realizada em 2020 e duas bases para o ano de 2021. Conforme esperado, sub-bases de cada ano possuíam *tweets* repetidos, por isso, na etapa de união das bases foi realizada a exclusão de registros repetidos e atualização dos dados. Nesse caso, a base de 2020 ficou composta de 5.558.171 *tweets*, enquanto a base de 2021 continha 5.360.203. A Tabela 3 mostra a quantidade de *tweets* de cada base de coleta, bem como o tamanho de cada uma.

Tabela 3. Quantidade de *tweets* e tamanho das bases.

Bases	Quantidade	Tamanho (GB)
2020		
Base 1	1.878.069	12,4
Base 2	3.703.971	28,4
Base 3	4.065.364	30,0
Base geral	5.558.171	37,1
2021		
Base 1	5.631.905	36,8
Base 2	5.561.617	36.4
Base geral	5.360.203	39,1

Dando prosseguimento, foram criadas as sub-bases para os idiomas inglês, espanhol e português. A Tabela 4 mostra a quantidade de *tweets* presentes em cada uma dessas sub-bases.

Na modelagem de tópicos a avaliação qualitativa levou à escolha de um número baixo de tópicos, variando entre 5 e 10 tópicos, conforme mostra a Tabela 5. Isso ocorre

Tabela 4. Quantidade de *tweets* presentes nas sub-bases.

	2020	2021
Inglês	1.991.792	2.459.599
Espanhol	1.898.572	1.441.175
Português	274.636	124.444

em grande parte devido ao contexto em que os dados foram obtidos, isto é, uma data temática em que os assuntos abordados tendem a ser mais específicos. Convém pontuar que a anotação dos tópicos foi feita em Português, independente do idioma, tal como disposto na Tabela 5.

Tabela 5. Rótulos dos tópicos anotados para cada idioma/ano.

Idioma	Quant.	Descritores
2020		
Inglês	5	Sororidade, Celebração da mulher, Características da mulher, Equidade, Política
Espanhol	10	Comemoração, Memória, Marco histórico, Desejos e anseios,Lutas diárias, Manifestações, Protestos, Violência de gênero, Sororidade, Papéis familiares
Português	5	Marco histórico, Homenagens cotidianas, Sarcasmo, Sororidade, Prestar tributo
2021		
Inglês	10	Empoderamento, Manifestações em redes sociais, Ativismo interseccional, Celebração, Inspiração, Música, Homenagem à profissionais de saúde, Mães solo, Desejos de mudança, Criação/educação
Espanhol	5	Denuncia, Resistência, Tributo, Celebração, Violência
Português	5	Felicitações, Política, Marco histórico, Campanhas, Lutas diárias

No ano de 2020, tivemos os seguintes cenários: no português, foram escolhidos 5 tópicos. Estes fazem referência a temas como acontecimentos históricos, homenagens e apoio mútuo entre as mulheres. Interessante destacar o tópico "Marco histórico" - que faz alusão ao episódio que deu origem a comemoração do Dia Internacional da Mulher[6]. No espanhol, foram escolhidos para esse idioma 10 tópicos. Ao analisar a Tabela, pode-se observar que os tópicos "Marco histórico" e "Sororidade", também, se fazem presente, assim como no português. Os tópicos "Manifestações" e "Protestos", apesar de estarem nomeados diferentemente, podem ser considerados sinônimos. Além desses, há tópicos relacionados a questões sempre presentes nas discussões feministas, como "Violência de gênero" e "Papéis familiares". Já no inglês, o cenário escolhido foi de 5 tópicos e, novamente, o tópico "Sororidade" se fez presente, com destaques para os tópicos "Política" e "Equidade" que apareceram pela primeira vez em todos os idiomas.

Já para o ano de 2021, os cenários foram os seguintes: no português, o cenário que melhor descrevia os tópicos mais relevantes na rede foi de 5 tópicos. O tópico "Marco histórico" mais uma vez se fez presente, mostrando que as manifestações sempre relembram esse episódio como conscientização de sua luta histórica. O tópico "Política" apareceu novamente, mostrando que há uma forte tendência a politização, oposição e descontentamento com os governos. No espanhol, o cenário escolhido foi de 5 tópicos,

[6]Disponível em `http://glo.bo/3JrN6oo`. Acesso em 08 mar. 2022

com destaque para “Violência”, que pode ser considerado semelhante ao “Violência de gênero” do ano anterior, apesar de o primeiro ter-se mostrado mais relacionado à inercia dos governos em resolver casos de feminicídio. Por fim, no inglês, a combinação escolhida foi de 10 tópicos, representando um aumento em comparação com o ano anterior. Consequentemente, há o surgimento de novos tópicos. Por exemplo, “Homenagem à profissionais de saúde” faz referência a milhares de mulheres que trabalhavam em hospitais na pandemia do Covid-19.

Resumidamente, foram registrados 17 tópicos diferentes para o ano de 2020, com o termo “Marco histórico” comum ao português e espanhol e “Sororidade”se fazendo presente nos três idiomas. Para 2021, ao todo foram contabilizados 19 tópicos diferentes e dessa vez não houve nenhum tópico comum aos três idiomas, somente o tópico “Celebração” esteve presente no espanhol e no inglês. Comparando os tópicos em relação aos anos, registram-se 33 tópicos diferentes - embora alguns possam ser similares entre eles - e os tópicos “Celebração”, “Lutas diárias”, “Marco histórico”, “Política” e “Sororidade” são comuns entre os dois anos da análise. Isso demonstra que, apesar de haver uma variabilidade nos temas tratados ao redor do mundo, existem tópicos que são centrais e caros a uma agenda global dos movimentos feministas.

Na análise de sentimento, a polaridade neutra é a predominante em todos os idiomas, semelhante ao trabalho de [Dilai and Levchenko 2018]. Para o ano de 2020, essa polaridade representa 62,73% do total para o inglês, porém, para o português e o espanhol, há uma quantidade bem elevada dessas categorias: quase 98% para ambos os idiomas. Este cenário, também, se repetiu para a análise feita com os dados de 2021: cerca de 62% de polaridade neutra para os *tweets* em inglês e cerca de 94% e 97% para o português e espanhol, respectivamente. A Tabela 6 mostra detalhadamente os valores percentuais para as outras polaridades dos idiomas nos dois cenários analisados.

Tabela 6. Percentuais de polaridades da análise de sentimentos.

	Neutro	Positivo	Negativo	Indefinido
Português	97,85%	1,22%	0,63%	0,30%
Espanhol	97,51%	1,26%	0,75%	0,48%
Inglês	62,73%	27,70%	5,93%	3,64%

(a) 2020

	Neutro	Positivo	Negativo	Indefinido
Português	94,22%	3,40%	1,01%	1,36%
Espanhol	97,02%	1,69%	0,81%	0,49%
Inglês	60,76%	29,67%	5,75%	3,82%

(b) 2021

5. Considerações Finais

Neste trabalho foi realizada uma análise de dados referentes ao Dia Internacional da Mulher nos anos de 2020 e 2021, utilizando (*tweets*) em inglês, português e espanhol. Estes dados foram coletados em tempo real por meio da API do Twitter, nos dias 07, 08 e 09 de março dos respectivos anos, utilizando *hashtags* relacionadas ao IWD. O processo adotado foi o CRISP-DM por ser um método bem consolidado na literatura. Na etapa de análise foram conduzidas a modelagem de tópicos e análise de sentimentos.

Os resultados da modelagem de tópicos mostraram que, por se tratar de uma data temática, uma quantidade menor - entre 5 e 10 tópicos - era mais adequada para representar os dados utilizados. Os tópicos identificados evidenciaram a agenda feminista para os anos e idiomas analisados, incluindo temas como empoderamento, sororidade, equidade, justiça, direitos e protestos. Embora a análise tenha sido feita em três idiomas diferentes, muitos dos tópicos eram semelhantes ou compartilhados, evidenciando assim uma espécie de agenda global comum.

Na análise de sentimentos, os resultados foram bem satisfatórios para o inglês nos dois anos de análise, com as classificações bem balanceadas. Contudo, no português e no espanhol, a grande quantidade de dados com polaridade neutra - alguns com mais de 95% do total - nos indicou que o iFeel não desempenhou bem para esses idiomas. Este fato foi corroborado pelo *ground truth* conduzido, o que evidencia a carência de métodos mais acurados para análise de sentimentos voltados para o português e o espanhol.

Além da modelagem de tópicos e AS, foram implementadas a análise exploratória de dados, análise de sentimento para cada tópico identificado e a divisão dos sentimentos por idioma para os tópicos em comum. Porém, devido ao caráter diminuto deste documento não puderam ser apresentadas, mas encontram-se disponíveis no repositório do *Github* do projeto, disponível em `https://github.com/fabiolobato/8m_brasnam2022`.

Como trabalhos futuros, pretende-se conduzir o estudo sob a ótica netnográfica, continuar a coleta de dados com o objetivo de construir uma base histórica, implementar e comparar novos métodos de modelagem de tópicos focados em textos curtos, aplicar métodos de análise de sentimentos desenvolvidos especificamente para português e espanhol e o desenvolvimento de uma plataforma para auxiliar na visualização do conhecimento extraído dos dados.

Agradecimentos

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - DT - 308334/2020; pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) - PRONEM-FAPESPA/CNPq nº 045/2021. Agradecemos também aos revisores(as) pelas sugestões que muito auxiliaram na melhora do trabalho.

Referências

Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512.

Araújo, M. L. D., Diniz, J. P., Bastos, L., Soares, E., Júnior, M., Ferreira, M., Ribeiro, F., and Benevenuto, F. (2016). ifeel 2.0: A multilingual benchmarking system for sentence-level sentiment analysis. In *Tenth International AAAI Conference on Web and Social Media*.

Belford, M., Mac Namee, B., and Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications*, 91:159–169.

Cirqueira, D., Pinheiro, M. F., Jacob, A., Lobato, F., and Santana, Á. (2018). A literature review in preprocessing for sentiment analysis for brazilian portuguese social media. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE.

Deriu, F. and Iezzi, D. (2020). Text analytics in gender studies. introduction.

Dilai, M. and Levchenko, O. (2018). Discourses surrounding feminism in ukraine: A sentiment analysis of twitter data. In *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, volume 2, pages 47–50. IEEE.

Duong, H.-T. and Nguyen-Thi, T.-A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, 8(1):1–16.

Fuchs, T. and Schäfer, F. (2019). Normalizing misogyny: hate speech and verbal abuse of female politicians on japanese twitter. In *Japan forum*, pages 1–27. Taylor & Francis.

Goel, R. and Sharma, R. (2020). Understanding the metoo movement through the lens of the twitter. In *International Conference on Social Informatics*, pages 67–80. Springer.

Greene, D., O'Callaghan, D., and Cunningham, P. (2014). How many topics? stability analysis for topic models. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 498–513. Springer.

Hino, A. and Fahey, R. A. (2019). Representing the twittersphere: Archiving a representative sample of twitter data under resource constraints. *International journal of information management*, 48:175–184.

Leung, L., Miedema, S., Warner, X., Homan, S., and Fulu, E. (2019). Making feminism count: integrating feminist research principles in large-scale quantitative research on violence against women and girls. *Gender & Development*, 27(3):427–447.

Lima Jr., E. G. S., Sousa, G. N., Jacob Jr., A. F. L., and Lobato, F. M. F. (2020). Ferramentas para análise de mídias sociais: Um levantamento sistemático. In *Computer on The Beach 2020*, pages 389–396. UNIVALE.

Lobato, F., Pinheiro, M., Jacob, A., Reinhold, O., and Santana, Á. (2016). Social crm: Biggest challenges to make it work in the real world. In *International Conference on Business Information Systems*, pages 221–232. Springer.

Locke, A., Lawthom, R., and Lyons, A. (2018). Social media platforms as complex and contradictory spaces for feminisms: Visibility, opportunity, power, resistance and activism.

Lommel, L. S., Schreier, M., and Fruchtmann, J. (2019). We strike, therefore we are? a twitter analysis of feminist identity in the context of# daywithoutawoman. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 20:2. DEU.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Merriam-Webster (2017). Merriam-webster's 2017 words of the year: Feminism.

Mundt, M., Ross, K., and Burnett, C. M. (2018). Scaling social movements through social media: The case of black lives matter. *Social Media+ Society*, 4(4).

Nugroho, R., Paris, C., Nepal, S., Yang, J., and Zhao, W. (2020). A survey of recent methods on deriving topics from twitter: algorithm to evaluation. *Knowledge and Information Systems*, 62(7):2485–2519.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Piatetsky, G. (2019). Python leads the 11 top data science, machine learning platforms: Trends and analysis.

Puente, S. N., Maceiras, S. D., and Romero, D. F. (2021). Twitter activism and ethical witnessing: Possibilities and challenges of feminist politics against gender-based violence. *Social Science Computer Review*, 39(2):295–311.

Reyes-Menendez, A., Saura, J. R., and Filipe, F. (2020). Marketing challenges in the# metoo era: Gaining business insights using an exploratory sentiment analysis. *Heliyon*.

Rodrigues, L., da Silva Junior, J., and Lobato, F. (2019). A culpa é dela! É isso o que dizem nos comentários das notícias sobre a tentativa de feminicídio de elaine caparroz. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Rodrigues, L., Prado, A., and Lobato, F. M. F. (2022). Pandemia de covid-19 no brasil: uma análise sobre notícias e comentários de usuários. *Culturas Midiáticas*, 16:26.

Rodriguez, S., Allende-Cid, H., Gonzalez, C., Alfaro, R., Elortegui, C., Palma, W., and Santander, P. (2020). Analyzing# lastesis feminist movement in twitter using topic models. In *International Conference on Human-Computer Interaction*. Springer.

Russell, M. A. (2013). *Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. "O'Reilly Media, Inc.".

Scarborough, W. J. (2018). Feminist twitter and gender attitudes: Opportunities and limitations to using twitter in the study of public opinion. *Socius*, 4:2378023118780760.

Scarborough, W. J. and Helmuth, A. S. (2021). How cultural environments shape online sentiment toward social movements: Place character and support for feminism. In *Sociological Forum*. Wiley Online Library.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

Stauffer, K. E. and O'Brien, D. Z. (2018). Quantitative methods and feminist political science. In *Oxford Research Encyclopedia of Politics*. Oxford University Press.

Tajudeen, F. P., Jaafar, N. I., and Ainin, S. (2018). Understanding the impact of social media usage among organizations. *Information & Management*, 55(3):308–321.

We Are Social (2021). Digital 2021 global overview report.

Yagui, M., Maia, L. F., Ugulino, W., Vivacqua, A., and Oliveira, J. (2017). "bela, recatada e do lar": Base de dados e aspectos do movimento social ocorrido na rede social online twitter. In *Anais do XIV Simpósio Brasileiro de Sistemas Colaborativos*. SBC.