UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS CABRAL CARNEIRO DA CUNHA

FAKEWHATSAPP.BR: DETECÇÃO DE DESINFORMAÇÃO E DESINFORMADORES EM GRUPOS PÚBLICOS DO WHATSAPP EM PT-BR

FORTALEZA

2021

LUCAS CABRAL CARNEIRO DA CUNHA

FAKEWHATSAPP.BR: DETECÇÃO DE DESINFORMAÇÃO E DESINFORMADORES EM GRUPOS PÚBLICOS DO WHATSAPP EM PT-BR

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Orientador: Prof. Dr. José Maria da Silva Monteiro Filho

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C978f Cunha, Lucas Cabral Carneiro da.
FakeWhatsApp.BR: detecção de desinformação e desinformadores em grupos públicos do WhatsApp em PT-BR / Lucas Cabral Carneiro da Cunha. – 2021.
130 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2021.
Orientação: Prof. Dr. José Maria da Silva Monteiro Filho.

1. Desinformação. 2. Desinformadores. 3. WhatsApp. 4. Aprendizado de Máquina. 5. Processamento de Linguagem Natural. I. Título.

CDD 005

LUCAS CABRAL CARNEIRO DA CUNHA

FAKEWHATSAPP.BR: DETECÇÃO DE DESINFORMAÇÃO E DESINFORMADORES EM GRUPOS PÚBLICOS DO WHATSAPP EM PT-BR

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Ciência da Computação

Aprovada em: 22 de Dezembro de 2021

BANCA EXAMINADORA

Prof. Dr. José Maria da Silva Monteiro Filho (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Ms. José Wellington Franco da Silva
Universidade Federal do Ceará (UFC)

Prof. Dr. César Lincoln Cavalcante Mattos
Universidade Federal do Ceará (UFC)

Profa. Dra. Jonice de Oliveira Sampaio
Universidade Federal do Rio de Janeiro (UFRJ)

Para Mia, Maya e Lua.

AGRADECIMENTOS

Agradeço à minha mãe, Marilene Cabral, pela infinita paciência, amor e carinho que me dedicou durante toda minha vida. Sem você eu não teria chegado até aqui.

À minha esposa, Lia Aguiar, pelo amor, companheirismo, incentivo e apoio constante que foram essenciais para minha vida e para a realização desse trabalho. Sou muito feliz por dividir a vida com você. Obrigado por tornar esse mestrado possível e pela revisão desta dissertação.

Ao Prof. Ms. José Wellington Franco da Silva pela amizade e orientação durante todo esse processo e pelos ensinamentos na área de processamento de linguagem natural. Nossas conversas durante o café da tarde no Diniz deixaram saudades.

Ao Prof. Dr. José Maria da Silva Monteiro Filho pela orientação deste trabalho e inspiração na luta por uma sociedade mais justa.

Aos amigos do grupo ARiDa, pela camaradagem, solicitude e troca de conhecimentos que ocorreram no laboratório durante o período anterior a pandemia. Aos colegas Artur Franco, Cristiano Melo, Thiago Vennuto, Gustavo Moraes, Ivandro Claudino, José "Barão" Augusto, Arlino Magalhães e demais.

Ao Prof. Dr. César Lincoln, por todo os ensinamentos na área de aprendizado de máquina e pela presteza em contribuir com esse trabalho. Suas aulas aumentaram ainda mais o meu interesse pro aprendizado de máquina e me motivaram a aprender cada vez mais.

À Profa. Dra. Jonice Sampaio pela prestatividade em contribuir de forma enriquecedora para esse trabalho.

À amiga Thays Lavor, pela amizade, instigação e por inadvertidamente ter sido responsável por eventos que culminaram na escolha do tema desta pesquisa.

Aos amigos do Lead: Eudênia Magalhães, Daniel Coutinho, Eduardo Montesuma, Adson Damasceno, João Victor Sampaio, Laura Esteche, e demais, cuja amizade, troca de experiência e incentivo foram essenciais para conciliar o trabalho com a escrita desta dissertação.

Ao Prof. Dr. Tobias Rafael Fernandes Neto, coordenador do Laboratório de Sistemas Motrizes (LAMOTRIZ) onde este *template* foi desenvolvido.

Ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), por financiar parcialmente essa pesquisa através da bolsa de mestrado.

“Não acredite em nada que você lê na rede. Exceto isto. Bem, incluindo isso, eu suponho.”

(Douglas Adams)

RESUMO

Nos últimos anos, a propagação em larga escala através das redes sociais de informações falsas, enganosas ou distorcidas tornou-se um grave problema social. A disseminação de desinformação tem prejudicado organizações e indivíduos e impactando negativamente processos democráticos, economia, saúde e segurança pública. Assim, o estudo e desenvolvimento de métodos para detecção automática de desinformação, bem como de usuários maliciosos que espalham desinformação em larga escala, ganhou a atenção da academia e da indústria. No Brasil e em diversos outros países, o aplicativo móvel de mensagens WhatsApp é um dos meios onde mais circula desinformação. Porém, encontram-se ainda poucos trabalhos na literatura que abordem a detecção de desinformação nesse cenário específico. Nessa dissertação, propomos a construção e disponibilização do FakeWhatsApp.Br: um conjunto de dados de mensagens obtidas de grupos públicos de WhatsApp, contendo informações de propagação (social e temporal), onde as mensagens compartilhadas mais de uma vez foram rotuladas como contendo ou não desinformação. A partir desse recurso, realizamos uma série de experimentos de classificação utilizando diferentes técnicas de aprendizado de máquina para detectar mensagens que contenham desinformação e detectar desinformadores. Foram comparados e discutidos métodos de classificação baseados em processamento de linguagem natural e em atributos de usuários, analisando as vantagens e limitações de cada abordagem e identificando as particularidades e desafios destes problemas. Os resultados obtidos neste trabalho fornecem contribuições iniciais para o estudo destas questões e realizam apontamentos para pesquisas futuras no contexto de desinformação no WhatsApp.

Palavras-chave: desinformação; desinformadores; aprendizado de máquina; WhatsApp; processamento de linguagem natural.

ABSTRACT

In recent years, the large-scale propagation through social midia of false, misleading, or distorted information, i.e. disinformation, has become a serious social problem, harming organizations and individuals and negatively impacting democratic processes, economy, health and public safety. Thus, the study and development of methods for automatic detection of misinformation, as well as the detection of malicious users that spread misinformation, gained the attention of academia and industry. In Brazil and in several other countries, the mobile messaging application WhatsApp is one of the midia in which misinformation circulates the most. However, there are still few works in the literature that address the detection of misinformation in this specific scenario. In this dissertation, we propose the construction of FakeWhatsApp.Br: a dataset of messages obtained from public WhatsApp groups, containing propagation information (social and temporal), where messages shared more than once were labeled as containing or not misinformation. From this resource, we carry out a series of classification experiments using different machine learning techniques to detect messages with misinformation and misinformation spreaders. Classification methods based on natural language processing and user attributes were compared and discussed, analyzing the advantages and limitations of each approach and identifying the particularities and challenges of these problems. The results obtained in this work provide initial contributions to the study of these problems and point to future research in the context of misinformation on WhatsApp.

Keywords: misinformation; misinformers; machine learning; WhatsApp; natural language processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

TF-IDF	*Term Frequency-Inverse Document Frequency* / Frequência do Termo - Inverso da Frequência dos Documentos
NLP	*Natural Language Processing* / Processamento de Linguagem Natural
TIC	Tecnologias da Informação e Comunicação
SVM	*Support Vector Machine* / Máquina de Vetores de Suporte
BoW	*Bag of Words* / Saco de Palavras
LIWC	*Linguistic Inquiry and Word Count* / Pesquisa Linguística e Contagem de Palavras
POS	*Part of Speech* / Partes do Discurso
SPO	Sujeito, Predicado e Objeto
KG	*Knowledge Graph* / Grafo de Conhecimento
MLP	*Multilayer Perceptron* / Perceptron Multicamada
AUC	*Area Under the ROC Curve* / área sobre a curva ROC
CSV	*Comma-separated values* / valores separados por vírgula
LGPD	Lei Geral de Proteção de Dados Pessoais
KDE	*Kernel Density Estimation* / estimativa de densidade kernel
ROC	*Receiver Operator Characteristic Curve* / Curva Característica de Operação do Receptor
KNN	*K-Nearest Neighbors* / K-vizinhos mais próximos
PCA	*Principal Component Analysis* / Análise dos Componentes Principais
MDI	*Mean Decrease in Impurity* / decaimento médio de impureza
GNN	*Graph Neural Networks* / Redes Neurais para Grafos

SUMÁRIO

1 INTRODUÇÃO

1.1 Motivação

Nos últimos anos, a ascensão das redes sociais virtuais alterou significativamente o modo como produzimos, compartilhamos e consumimos informação. De acordo com Shu *et al.* (2019), através de tais plataformas é mais rápido e barato para usuários obter acesso a notícias, quando comparado a mídias tradicionais como jornais e televisão. Além disso, as redes sociais, como são mais comumente chamadas, permitem que usuários compartilhem, comentem e debatam notícias com outros usuários, participando de forma ativa da transmissão de informação e gerando engajamento. Em geral, redes sociais tem fácil acesso, baixo custo e permitem a disseminação de conteúdo em grande volume e velocidade.

Contudo, ao mesmo tempo que redes sociais permitem o amplo acesso à informação de qualidade, o seu ambiente descentralizado e desregulado também é um ambiente fértil para a disseminação em massa de informações falsas ou enganosas, a chamada desinformação (VOSOUGHI *et al.*, 2018; GUO *et al.*, 2019a; SU *et al.*, 2020). De modo geral, qualquer usuário nas redes sociais pode criar e compartilhar conteúdo sem compromisso com a veracidade ou com as consequências dessa publicação (SHU *et al.*, 2017).

Além disso, dada a facilidade de criação de contas de usuário e relativo grau de anonimidade, há uma forte presença de usuários maliciosos que, intencionalmente e repetidamente, espalham desinformação de forma desproporcional. Referenciados neste trabalho como desinformadores, esses usuários muitas vezes agem em grupos coordenados ou recebendo pagamento por estes serviços. Orlov e Litvak (2018) refere-se a esses usuários como "propagandistas". Há também um grande número de contas controladas por softwares, os *bots* sociais. Por atuarem de forma automatizada, *bots* possuem a capacidade de disseminar conteúdo enganoso em velocidade e volume muito maior que um usuário humano poderia alcançar (ABDIN, 2019; WANG *et al.*, 2018).

Essa combinação de fatores permite que a desinformação seja propagada em larga escala, enganando grande quantidade de usuários em um curto espaço de tempo e prejudicando indivíduos, organizações e a sociedade de forma geral. Shu *et al.* (2018) afirma que a prevalência da desinformação possui o potencial de mudar a forma como o público responde à verdade e quebrar a confiabilidade em fontes legítimas de informação.

O conceito de desinformação é amplamente definido por Su *et al.* (2020) como

informação deturpada, seja esta forjada, enganosa, falsa ou distorcida. Essa definição abrangente engloba diversos conceitos específicos encontrados na literatura e que podem se sobrepor, tais como *Fake News* (LAZER *et al.*, 2018), rumores (SHU *et al.*, 2017), enganação (MAALEJ, 2001) e outros. Em particular, o termo *Fake News*, embora descreva um tipo específico de desinformação, escrita de modo a imitar o estilo de uma notícia jornalística, se tornou muito presente na cultura popular e muitas vezes é utilizado como sinônimo de desinformação de forma indiscriminada.

Desinformação é usualmente criada com intenções maliciosas para manipular a opinião pública, prejudicar indivíduos, organizações ou grupos sociais e obter ganhos econômicos ou políticos (INTERVOZES, 2019). De acordo com Lazer *et al.* (2018), a desinformação se espalha de forma mais rápida, profunda e ampla nas redes sociais do que a informação legítima. Devido ao alto volume de informações a que estamos expostos ao utilizar redes sociais, os humanos têm uma capacidade limitada de distinguir informações verdadeiras de informações falsas (VOSOUGHI *et al.*, 2018; QIU *et al.*, 2017). Segundo Guo *et al.* (2019a), a disseminação da desinformação é um problema social de nível global, causando danos a democracia, justiça, economia, saúde e segurança públicas. Alguns casos famosos podem ser exemplificados:

- A disseminação coordenada e automatizada de desinformação desempenhou um papel considerável nas eleições presidenciais do Estados Unidos em 2016, favorecendo o então candidato Donald Trump (ALLCOTT; GENTZKOW, 2017). Segundo Machado *et al.* (2018), uma situação similar ocorreu nas eleições presidenciais do Brasil em 2018.
- Em 2020, associada à pandemia do novo Coronavírus (SARS-CoV-2, o COVID-19) surgiu também uma infodemia, como foi definido pela Organização Mundial da Saúde (ORGANIZATION *et al.*, 2021): o excesso de informação, incluindo informações falsas ou enganosas em ambientes digitais e físicos durante um surto de doença. Desinformação sobre COVID-19 causou um impacto significativo na saúde pública, promovendo o descrédito da ciência e das instituições globais de saúde pública, enfraquecendo a adesão da população à vacinas e aos cuidados de prevenção e promovendo tratamentos ineficazes (GALLOTTI *et al.*, 2020; GALHARDI *et al.*, 2020). A desorientação causada pela desinformação elevou os níveis de infecção, levando a perda de muitas vidas.
- Desde 2012 ocorreram na Índia diversos casos de linchamentos por multidões ocasionados por informações falsas, resultando em ferimentos, morte e trauma de indivíduos falsamente acusados de crimes como rapto de crianças ou tráfico de órgãos humanos (BANAJI *et al.*,

2019). Um caso em particular ocorrido em 2019 resultou na morte de 40 pessoas.

Nesse contexto, a detecção automática de desinformação através de métodos computacionais tem atraído atenção da academia e da indústria. A tarefa de detecção de desinformação é definida de forma ampla por Su *et al.* (2020) como a tarefa de avaliar a veracidade, credibilidade ou autenticidade de alegações em um fragmento de informação. Algoritmos de detecção de desinformação podem ser utilizados para identificar precocemente conteúdos enganosos tão rápido quanto estes surgem nas redes sociais. A detecção automática pode alertar jornalistas, entidades legais ou as próprias plataformas para que seja realizada uma checagem de fatos e uma possível mitigação dos danos, seja bloqueando a disseminação do conteúdo enganoso ou alertando usuários da falsidade do mesmo.

De acordo com Shu *et al.* (2019), o ecossistema de informação em redes sociais envolve três dimensões: a dimensão do conteúdo ("**o quê**"), a dimensão social ("**quem?**") e a dimensão temporal ("**quando?**"). A dimensão do conteúdo descreve a relação entre as informações propriamente ditas, sejam na forma de texto, imagens, vídeos, etc. A dimensão social envolve a credibilidade dos publicadores, usuários que divulgam, consomem ou interagem com a informação. A dimensão temporal descreve o comportamento da informação na rede ao longo do tempo. Todas as dimensões podem ser exploradas e combinadas para caracterizar e detectar desinformação em diferentes contextos. Em particular, na dimensão do conteúdo, métodos que combinam técnicas de Aprendizagem de Máquina com *Natural Language Processing* / Processamento de Linguagem Natural (*NLP*) tem se destacado na literatura e alcançado bons resultados. De acordo com Shu *et al.* (2017), essa abordagem baseia-se na hipótese de que textos escritos com a intenção de enganar possuem padrões linguísticos que os distinguem de textos não-enganosos. Além disso, esses padrões podem ser aprendidos e discriminados por modelos de Aprendizagem de Máquina a partir de uma amostra razoável de dados. Técnicas baseadas em *NLP* são importantes pois grande parte da desinformação é criada e compartilhada na forma de texto.

Outra tarefa intimamente relacionada com detecção de desinformação em redes sociais, e por vezes explorada de forma auxiliar nesta, é a identificação de usuários maliciosos. Chamados neste trabalho de **desinformadores**, esses usuários atuam de forma consistente e intencional disseminando desinformação em larga escala, comumente violando os padrões de comunidade das plataformas onde atuam. Embora usuários regulares possam ocasionalmente compartilhar desinformação, os desinformadores diferem-se pelo volume desproporcional de

disseminação.

Desinformadores podem ser humanos engajados ou *bots*. É notório também a existência dos chamados *cyborgs*, que alternam entre atividade humana e automática no controle da conta, tornando sua atividade mais difícil de ser detectada (CHU *et al.*, 2010). Identificar esses usuários de forma automática é uma ação importante para mitigar a disseminação de desinformação. Através de sua identificação, as plataformas das redes sociais onde atuam podem bloquear sua atividade, o que pode ser mais eficaz que bloquear a desinformação em si. Essa tarefa também pode ser executada através de técnicas de Aprendizado de Máquina, uma vez que hajam dados representativos dos usuários.

Detecção de desinformação e de desinformadores são tarefas fundamentais para a criação de métodos que efetivamente realizem a mitigação da disseminação de desinformação. Contudo, é notório que o problema da disseminação de desinformação é extremamente complexo, cuja compreensão e solução é intrinsecamente interdisciplinar, permeando as ciências sociais, a psicologia, o jornalismo, dentre outros campos de estudo. Nesse sentido, este trabalho busca oferecer uma contribuição na área da Computação, mas reconhecendo suas limitações.

1.2 Principais Desafios

Para criação e validação de métodos eficientes de detecção de desinformação ou de desinformadores em redes sociais é necessário a obtenção de dados representativos do problema alvo. Na detecção de desinformação textual, esses dados correspondem aos textos reais que são compartilhados nas plataformas, com um rótulo associado indicando a veracidade do texto. Esse rótulo é o que deseja-se que modelos de Aprendizado de Máquina aprendam a predizer.

Já no caso de detecção de desinformadores, os dados devem representar o comportamento do usuário na plataforma em questão. Os tipos de dados podem variar de acordo com a plataforma. Embora a obtenção dos dados possa ser feita de forma automática ou semiautomática utilizando APIs ou *web-crawlers*, a atribuição de rótulos aos textos (também chamada de anotação) é um processo que costuma envolver uma grande quantidade de trabalho manual. De acordo com Rubin *et al.* (2015), a anotação dos textos é feita por especialistas que realizam checagem de fatos nas alegações presentes nos textos, que é um processo demorado e custoso.

Na literatura encontra-se uma grande quantidade de trabalhos que treinam e validam métodos com textos coletados de plataformas como Facebook[1] (POTTHAST *et al.*, 2017;

[1] https://www.facebook.com/

SANTIA; WILLIAMS, 2018; MITRA; GILBERT, 2015; GRANIK; MESYURA, 2017) e Twitter[2] (ZUBIAGA *et al.*, 2016; ZERVOPOULOS *et al.*, 2020). Esses conjuntos de dados são disponibilizados publicamente para fins de pesquisa e desenvolvimento e avaliação de novos métodos de detecção de desinformação. Facebook e Twitter tem em comum o fato de possuírem conteúdo público que pode ser acessado pela *Web*, facilitando a coleta automatizada de dados. É importante salientar que a maioria desses conjuntos de dados textuais estão na língua inglesa, havendo uma grande lacuna de dados em português.

1.2.1 Desinformação no contexto do WhatsApp

Embora redes sociais como Twitter e Facebook possuam um papel importante na comunicação de muitas pessoas, em países como Brasil, Índia, México e outros, um dos principais meios de circulação de desinformação é o aplicativo móvel de mensagens WhatsApp[3]. Um dos propósitos do WhatsApp é permitir que usuários enviem livremente mensagens privadas uns aos outros através de seus *smartphones*. As mensagens podem ser compostas de texto, imagens, vídeos ou áudios.

Apesar de ser principalmente utilizado para conversas individuais, o WhatsApp possui o recurso de grupos de conversa, onde até 256 usuários podem participar e encaminhar mensagens. Esses grupos podem ser públicos, onde qualquer usuário pode participar apenas clicando em um link. Grupos públicos podem ser grandes canais de comunicação, permitindo que usuários debatam temas específicos com outros usuários, por vezes desconhecidos, compartilhando e recebendo conteúdo em larga escala. No Brasil, o WhatsApp é o aplicativo móvel mais utilizado (INSIGHT), 2020), com mais de 136 milhões de usuários, e uma ferramenta de comunização já enraizada na cultura popular. Não à toa, é um meio bastante utilizado para a disseminação de desinformação. Alguns dados relevantes podem ser destacados:

- Um relatório da agência Reuters (NEWMAN *et al.*, 2020) de 2020 indicou que no Brasil cerca das 35% de notícias enganosas são compartilhadas através do WhatsApp;
- Estudando grupos públicos de WhatsApp, Resende *et al.* (2019) indicou que cerca de 40,7% de mensagens falsas continuaram sendo compartilhadas mesmo após serem checadas como falsas;
- O estudo de Galhardi *et al.* (2020) apontou que 10,5% das notícias falsas acerca do

[2] https://twitter.com/
[3] https://www.whatsapp.com/

COVID-19 foram publicadas no Instagram, 15,8% no Facebook e 73,7% circularam via WhatsApp ;

– Uma pesquisa feita pela Câmara dos Deputados e o Senado do Brasil apontou que o WhatsApp é a principal fonte de informação dos brasileiros, utilizado por 79% dos entrevistados. Outros meios de comunicação mencionados foram canais de televisão (50%), vídeos no YouTube (49%), Facebook (44%), sites de notícias (38%), Instagram (30%), rádio (22%), jornais impressos (8%) e Twitter (7%) (AGÊNCIA BRASIL, 2019).

Essa questão é bem conhecida pela empresa desenvolvedora do WhatsApp, que afirma buscar medidas para reduzir cada vez mais a disseminação de desinformação, como mencionado por Exame (2020). Essas medidas ocorrem desde 2018 e incluem limitar a quantidade de membros em grupos, limitar o encaminhamento de mensagens e sinalizar mensagens que são encaminhadas com alta frequência. De acordo com Época Negócios (2018), a empresa também afirma que remove contas que praticam *spam*, identificadas através do uso de inteligência artificial. Contudo não é deixado claro quais métodos seriam utilizados para a identificação dessas contas.

1.2.2 Desafios no contexto do WhatsApp

Apesar do cenário de propagação de desinformação no WhatsApp, encontram-se na literatura científica poucos trabalhos que desenvolvam técnicas de detecção de desinformação para conteúdo coletado diretamente do WhatsApp. Em parte, isso deve-se à sua natureza privada, o que torna a coleta de dados desafiadora do ponto de vista técnico. Trabalhos como o de Garimella e Tyson (2018) e Resende *et al.* (2018) propõem sistemas que permitem o monitoramento de grupos públicos, extraindo diversas informações relevantes através de análises exploratórias. Porém, não encontramos na literatura trabalhos que realizem a detecção de desinformação em dados em português extraídos diretamente do WhatsApp. Tampouco temos conhecimento de conjuntos de dados com essa especificidade disponibilizados publicamente para pesquisa. Esses dados são necessários para o desenvolvimento de modelos com boa performance no contexto do WhatsApp. Quando trata-se de abordagens baseadas em *NLP*, a performance de um modelo é altamente dependente dos padrões linguísticos, tópicos e do vocabulário presente nos dados utilizados para treiná-lo.

O WhatsApp possui particularidades relevantes em relação as já mencionadas redes sociais. Enquanto o Facebook é uma rede de compartilhamento público de fotos, atualizações e notícias gerais com membros conectados com você, e o Twitter é um microblog onde membros interagem com seguidores através de mensagens curtas, o WhatsApp possui uma natureza conversação interpessoal. Segundo Waterloo *et al.* (2018), esse fator torna único o conteúdo compartilhado através desta plataforma e a forma como os seus usuários se expressam e interagem entre si. Os autores ainda pontuam como particularidades as mensagens curtas, o uso recorrente de ícones de emoção, os *emojis*, e o estilo informal de escrita.

Rosenfeld *et al.* (2018) pontua que a forma de registro no WhatsApp é feita exclusivamente através de um número de telefone celular, e o *smartphone* é a principal interface para enviar e receber mensagens. Muitas operadoras de celular oferecem planos de dados que permitem uso ilimitado do WhatsApp, sendo portanto mais acessível para uma parcela da população com acesso limitado à internet e computadores, o que corresponde a uma parcela significativa da população no Brasil e em outros países em desenvolvimento . Esses fatos contribuem para que o WhatsApp tenha uma base de usuários maior e mais diversa.

Dadas essas diferenças, é razoável assumir que um modelo de detecção de desinformação textual treinado com dados coletados do Twitter ou do Facebook teria uma performance baixa quando utilizado para classificar mensagens do WhatsApp. Além disso, em problemas de detecção de desinformação utilizando *NLP* e Aprendizado de Máquina, a escolha do modelo, etapas de pré-processamento e extração de atributos possui grande influência no desempenho da solução. No caso de textos coletados do WhatsApp, desconhece-se quais as melhores combinações de técnicas.

Destaca-se ainda que informações sobre o comportamento dos usuários no WhatsApp, ou seja, a dimensão social, também poderiam ser utilizadas para realizar a detecção da desinformação e também de desinformadores. Também não encontra-se na literatura trabalhos que abordem a detecção de usuários maliciosos no contexto do WhatsApp. Essa tarefa também merece atenção, pois os dados que descrevem o comportamento dos usuários são inerentemente diferentes de outras redes sociais, considerando as particularidades já mencionadas.

1.2.3 Possíveis estratégias para mitigar o problema da desinformação no WhatsApp

Dada a importância do WhatsApp como meio de comunicação e o sua má utilização como canal de desinformação, torna-se necessária a coleta de dados reais do WhatsApp para estudo, desenvolvimento e avaliação de métodos de detecção de desinformação e desinformadores nesse contexto. Destaca-se ainda que a detecção da desinformação é um primeiro passo fundamental para criação de métodos de mitigação. Possíveis aplicações para esses métodos seriam:

- Um *bot* ou serviço *Web* para que usuários consultem a probabilidade de uma mensagem recebida ser enganosa.
- Em sistemas de monitoramento de grupos públicos de WhatsApp, como os já mencionados em Garimella e Tyson (2018) e Resende *et al.* (2018), um modelo de detecção de desinformação poderia detectar as mensagens enganosas tão rápido quanto estas surgissem, permitindo que pesquisadores analisassem o volume, velocidade de disseminação e outros dados que aumentassem a compreensão da propagação de desinformação nessas redes.
- Ainda considerando o cenário anterior, jornalistas que monitorassem grupos poderiam ser alertados em tempo real sobre desinformação, realizar checagem de fatos e espalhar a contra-informação verdadeira para mitigar o espalhamento da desinformação na rede.
- Um modelo capaz de identificar os usuários maliciosos, que de forma repetida e intencional espalham desinformação, poderia ser utilizado pela plataforma para bloquear a atividade desses usuários. Ou ainda, considerando sistemas de monitoramento de grupos públicos, o modelo poderia alertar especialistas humanos que, após confirmação da detecção, poderiam denunciar ações danosas desses usuários à plataforma ou mesmo às autoridades legais, reduzindo a falta de responsabilização existente no ambiente de WhatsApp.

1.3 Questões de pesquisa e contribuições esperadas

Dado o cenário exposto, é possível identificar duas importantes lacunas de pesquisa:

- A ausência de conjuntos de dados representativos do contexto de desinformação no WhatsApp em português disponibilizados publicamente para pesquisa.
- A ausência de estudos de métodos de detecção de desinformação e desinformadores no contexto do WhatsApp.

Com base nos problemas observados, neste trabalho de dissertação propomos um estudo avaliativo de métodos de detecção de desinformação e de desinformadores baseados em Aprendizado de Máquina em dados coletados do WhatsApp em português brasileiro. Para esse fim, buscamos responder as seguintes questões de pesquisa:

Q1. O quão desafiadora é a detecção de desinformação textual no WhatsApp utilizando técnicas de *NLP* e Aprendizado de Máquina Supervisionado?

Q2. Quais atributos podem ser extraídos para descrever o comportamento dos usuários no contexto do WhatsApp e como podem ser explorados para auxiliar a detecção de desinformação?

Q3. Quais combinações de métodos de pré-processamento, extração de atributos e algoritmos de classificação podem ser adequadamente explorados para a tarefa de detecção de desinformação no WhatsApp?

Q4. Quais as limitações das melhores abordagens de detecção de desinformação avaliadas para responder a Q3?

Q5. Que atributos e métodos podem ser adequadamente explorados para detecção de usuários desinformadores no contexto do WhatsApp?

Para responder essas questões, este trabalho traz as seguintes contribuições:

1. A criação e disponibilização de um conjunto de dados de larga-escala, parcialmente rotulado e anonimizado, de mensagens coletadas diretamente de grupos públicos do WhatsApp em português, o chamado FakeWhatsApp.BR[4].Esse conjunto de dados também contém informações que permitem que a detecção de desinformação seja explorada na dimensão do conteúdo, social e temporal. Até onde sabemos, esse é o primeiro conjunto de dados público desse tipo e permitirá que outros pesquisadores estudem novas soluções para esse problema.
2. Um procedimento de rotulação de textos entre textos que contém desinformação e textos que não contém desinformação, baseado em checagem de fatos, e uma estratégia de expansão automática de dados rotulados baseada em similaridade de texto.
3. A condução de uma série de experimentos de classificação de texto, combinando diferentes métodos de extração de atributos e de algoritmos de classificação para discriminar os textos com ou sem desinformação. Essa série de experimentos estabelece uma base de performance para esse problema e indica quais abordagens dentre as testadas são mais

[4] https://github.com/cabrau/FakeWhatsApp.Br

adequadas, provendo informações sobre esse problema ainda não-explorado.

4. A proposta de uma série de atributos da dimensão social para descrever o comportamento de usuários no contexto do WhatsApp e a avaliação do uso desses desses atributos no problema de detecção de desinformação.
5. A proposta de uma definição de desinformadores no contexto do WhatsApp e a análise de métodos de detecção com base nos atributos dos usuários.

1.4 Contribuições científicas

Com base nos experimentos e resultados obtidos durante o desenvolvimento desta dissertação foram publicados os seguintes artigos científicos:

Cabral, L., Monteiro, J., Franco da Silva, J., Mattos, C., Mourão, P. "FakeWhastApp.BR: NLP and Machine Learning Techniques for Misinformation Detection in Brazilian Portuguese WhatsApp Messages". Trabalho publicado em Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1, p. 63-74.

Claudino de Sá, I., Monteiro, J., Franco da Silva, J., Medeiros, L., Mourão, P., **Cabral, L.**, "Digital Lighthouse: A Platform for Monitoring Public Groups in WhatsApp". Trabalho publicado em Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021) - Volume 1, p. 297-304.

Martins, A. D. F., **Cabral, L.**, Mourão, P. J. C., Monteiro, J. M., Machado, J. (2021, June). "Detection of Misinformation About COVID-19 in Brazilian Portuguese WhatsApp Messages". Trabalho publicado em International Conference on Applications of Natural Language to Information Systems (NLDB 2021) - p. 199-206. Springer, Cham.

Martins, A. D. F., **Cabral, L.**, Mourao, P. J. C., de Sá, I. C., Monteiro, J. M., Machado, J. (2021, October). COVID19. BR: A Dataset of Misinformation about COVID-19 in Brazilian Portuguese WhatsApp Messages. Trabalho publicado em Anais do III Dataset Showcase Workshop (pp. 138-147). SBC.

1.5 Organização da dissertação

Esta dissertação está organizada em sete capítulos, descritos brevemente a seguir:

– **Capítulo 1. Introdução**

  Este capítulo apresenta o problema de detecção de desinformação, o seu contexto social e desafios técnicos e científicos envolvidos, trazendo ainda uma visão geral da solução proposta pelo trabalho.

– **Capítulo 2. Fundamentação Teórica**

  Nesse capítulo são apresentados os conceitos fundamentais para a compreensão deste trabalho, incluindo definições de desinformação e termos correlatos, a dinâmica da desinformação nas redes sociais, definição formal das tarefas de detecção de desinformação e desinformadores e algumas das abordagens existentes nesse campo.

– **Capítulo 3. Trabalhos Relacionados**

  Este capítulo traz uma revisão bibliográfica comparativa de abordagens e resultados alcançados na literatura acerca de problemas similares aos enfrentados neste trabalho.

– **Capítulo 4. O Conjunto de Dados FakeWhatsApp.BR**

  Neste capítulo é descrito o procedimento de criação do conjunto de dados FakeWhatsApp.BR e uma análise exploratória desse conjunto de dados. O capítulo termina com uma discussão sobre as limitações e vieses contidos nos dados.

– **Capítulo 5. Detecção de Desinformação**

  Neste capítulo são descritos os experimentos de detecção de desinformação propostos. O capítulo termina com uma análise crítica dos resultados obtidos com esses métodos e suas limitações, destacando as as vantagens e desvantagens das abordagens experimentadas.

– **Capítulo 6. Detecção de Desinformadores**

  Neste capítulo é feita uma proposta de definição de desinformadores, a análise e comparação de um método de detecção de desinformadores baseado em detecção de *outliers* e um método baseada em aprendizado supervisionado.

– **Capítulo 7. Conclusão**

  O capítulo final conclui a dissertação, apontando as contribuições científicas, descobertas a partir dos experimentos, limitações das abordagens experimentadas, possíveis ameaças à validade, oportunidades de melhoria e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos fundamentais abordados neste trabalho. Inicialmente, discorremos sobre a definição de desinformação e conceitos relacionados, além das características de desinformação nas redes sociais. Em seguida, são discutidas as definições de detecção de desinformação, de desinformadores e as técnicas relacionadas a estes problemas.

2.1 Desinformação

A criação e transmissão de informações falsas e propositalmente enganosas é um fenômeno social antigo e que não surgiu com a Internet. De acordo com INTERVOZES (2019) e Posetti e Matthews (2018), isso já ocorria inicialmente através de conversas interpessoais e, posteriormente, através dos veículos de comunicação tradicionais, como jornais, rádio e televisão, que em diversos casos divulgaram informações desvirtuadas afim de manipular a opinião pública e favorecer seus interesses.

Entretanto, com o advento e popularização da Internet, em particular das redes sociais digitais, esse fenômeno ganhou uma nova magnitude. Segundo Vosoughi *et al.* (2018), com a facilidade de acesso e baixo custo, cada vez mais os usuários tem preferido consumir informação *online* ao invés dos meios tradicionais. Se antes a produção e divulgação de informação se centralizava nas mídias tradicionais, com as redes sociais, cada usuário é um produtor e divulgador em potencial.

De acordo com Conroy *et al.* (2015), se, por um lado, a descentralização do ecossistema de informação trouxe a possibilidade de produtores independentes aumentarem o seu alcance e que mais pessoas tenham acesso a informação de qualidade, por outro, trouxe a possibilidade que fontes sem credibilidade disseminassem informações enganosas com quase pouca ou nenhuma responsabilização sobre as consequências desse conteúdo. As redes sociais permitem a propagação desses conteúdos em alta velocidade e com alcance global quase imediato, fenômeno popularmente conhecido como “viralização”.

O problema da desinformação ainda é agravado pelo direcionamento segmentado de publicações, associado à coleta não-autorizada de dados pessoais dos usuários. Este fato tornou-se notório após o escândalo da ação da empresa *Cambridge Analytica* na eleição norte-americana de 2016, conforme demonstrado por Posetti e Matthews (2018). Segundo Guo *et al.* (2019a), as consequências da disseminação em massa de informação falsa, ou desinformação,

tem trazido sérias consequências negativas para a nossa sociedade como um todo, uma vez que ataca a credibilidade do ecossistema de informação, impactando diretamente na democracia, na segurança, economia, educação e saúde pública.

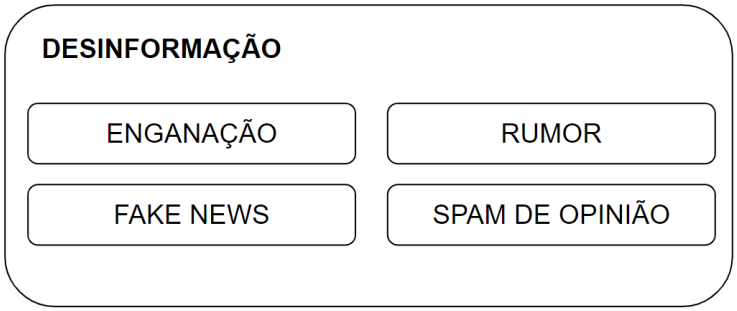
O fenômeno da disseminação em massa de desinformação é largamente estudado pela ótica das ciências humanas, especialmente pelos campos da Comunicação Social (INTERVOZES, 2019; POSETTI; MATTHEWS, 2018), Sociologia (MIHAILIDIS; VIOTTY, 2017; BRATU *et al.*, 2020), Psicologia (SUNDAR, 2016) e Economia (KSHETRI; VOAS, 2017). Nos últimos anos esse problema tem atraído cada vez mais o interesse de pesquisadores da área da Computação. Uma vez que é um problema potencializado pelas Tecnologias da Informação e Comunicação (TIC), é adequado que sejam buscadas soluções dentro desse campo para compreender, prevenir ou mitigar os danos sociais causados pela desinformação.

A seguir, discutimos alguns dos principais conceitos relacionados a desinformação conforme encontrado na literatura da Computação.

2.1.1 Conceitos relacionados

Desinformação é um conceito cuja definição varia bastante na literatura de acordo com diferentes autores. Adotaremos a definição abrangente de Su *et al.* (2020), que define desinformação como informação deturpada de forma geral, seja falsa, enganosa, forjada, imprecisa, descontextualizada ou distorcida. Ainda de acordo com Su *et al.* (2020), desinformação é usualmente criada com intenções maliciosas para atingir certos propósitos. Essa definição ampla engloba outros conceitos similares, mas que possuem suas especificidades, como enganação, *fake news*, rumores e *spam* de opinião, conforme ilustrado na Figura 1. A seguir, descrevemos as principais características desses conceitos já mencionados.

Figura 1 – Conceitos relacionados à desinformação.



Fonte: o autor.

Enganação (originalmente em inglês, *deception*) é definida por Zuckerman *et al.* (1981) como uma declaração intencionalmente enganosa, ou seja, que o enganador considera que é falsa, mas deseja convencer o receptor da mensagem de que ela é verdadeira. De acordo com An (2015), como a intencionalidade é parte essencial do conceito, a transmissão de uma informação falsa na qual o transmissor acredita na sua veracidade, seja por ter sido convencido, por um engano honesto ou por uma recordação imprecisa, não é considerado enganação. Vale destacar que nem sempre é possível conhecer a intenção do criador de uma informação.

Fake news, segundo Lazer *et al.* (2018), são caracterizadas por tratar-se de conteúdo comprovadamente falso que mimetiza o estilo de escrita de artigos jornalísticos reais, embora sem passar por um processo editorial ou de controle de credibilidade. Esse termo tornou-se fortemente presente na cultura popular, especialmente após os eventos das eleições norte-americanas de 2016 e o *Brexit*, onde houve um amplo debate na mídia sobre a influência das *fake news* em ambos os casos. *Fake news* foi considerada a "palavra do ano 2017" pelo influente *Collins Dictionary* (QUANDT *et al.*, 2019). De fato, esse é o termo relacionado mais encontrado na literatura e é comumente definido por alguns autores de forma intercambiável com o conceito geral de desinformação definido por Su *et al.* (2020), desconsiderando a especificidade de texto jornalístico.

Rumor é definido por Vosoughi *et al.* (2017) como um relato ou afirmação circulante cuja veracidade ainda não foi ou não pode ser verificada, muitas vezes gerando uma situação de ambiguidade. O autor define que um rumor pode acabar de três maneiras: pode ser identificado como factual, não-factual ou permanecer não resolvido. Diferente das *fake news*, que usualmente referem-se a eventos públicos que eventualmente são verificados como verdadeiros ou falsos, rumores podem ser de longo prazo, entrando nessa categoria as teorias da conspiração. Pelas definições apresentadas, pode-se também argumentar que uma *fake news* é inicialmente um rumor, até ser verificada como falsa.

***Spam* de opinião**, também chamado de *spam* de revisão, de acordo com Jindal e Liu (2008), são revisões forjadas que podem ser auto-promoção ou falsas declarações sobre os produtos revisados, para deliberadamente confundir consumidores para comprar ou evitar um produto.

Destaca-se ainda que, de acordo com o conceito adotado por Fallis (2014), há uma distinção para os termos *disinformation* e *misinformation*, onde o primeiro é utilizado para se referir a informações falsas divulgadas com o propósito de enganar, enquanto o último refere-se

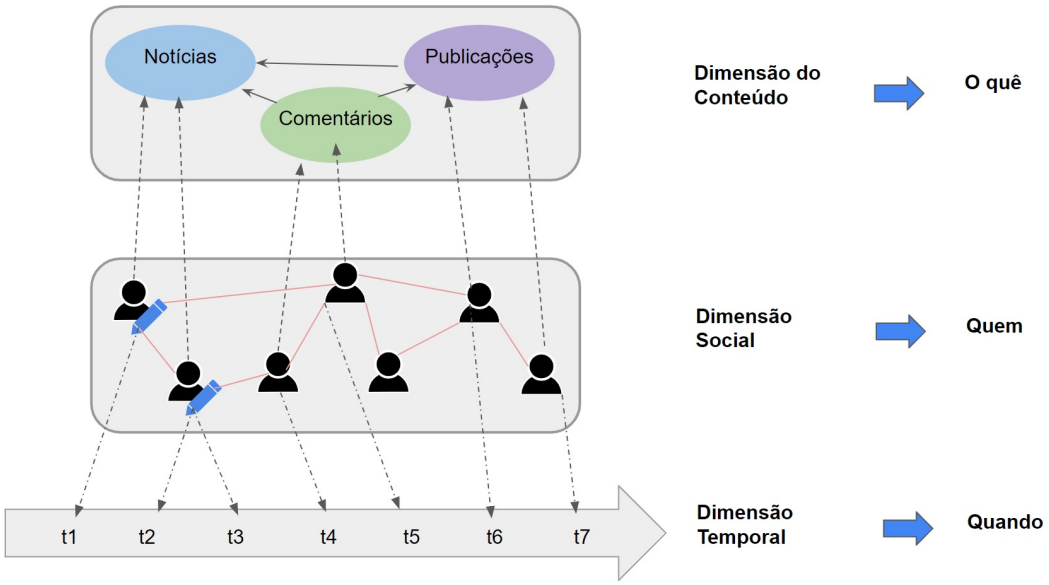
a informações falsas divulgadas por falta de conhecimento da informação verdadeira, ou seja, sem intencionalidade de enganar. Nota-se que em português ambos os termos são traduzidos como “desinformação”.

Observa-se do apresentado que existem possíveis interseções entre os conceitos, que também podem variar de acordo com a literatura consultada. Neste trabalho, iremos adotar a definição mais abrangente de desinformação, que engloba todos os conceitos apresentados. A seguir, são discutidas algumas das características da desinformação em redes sociais.

2.1.2 Desinformação em redes sociais

Segundo Shu *et al.* (2019), o ecossistema de disseminação de informação nas redes sociais envolve três dimensões, conforme ilustrado na Figura 2. A dimensão do conteúdo (*“o quê”*) descreve o conteúdo propriamente dito das publicações em redes sociais, incluindo mensagens e comentários acerca dessas publicações, que podem ser textos ou outras formas de mídia como imagens, áudios ou vídeos. A dimensão social (*“quem”*) descreve os comportamentos e relações entre os usuários que interagem com os conteúdos, sejam eles publicadores, consumidores ou divulgadores. Por fim, a dimensão temporal (*“quando”*) descreve a evolução das publicações e eventos na rede social ao longo do tempo.

Figura 2 – Dimensões do ecossistema de informação em redes sociais.



Fonte: adaptado de Shu *et al.* (2019)

Além dessas dimensões, a dinâmica das redes sociais possui características relevantes para o estudo da desinformação. Discutiremos algumas dessas características a seguir.

2.1.2.1 Câmera de eco

O processo de buscar e consumir informação em redes sociais possui uma maior autonomia quando comparado com o consumo de informação através das mídias tradicionais. Segundo Barberá *et al.* (2015), usuários de redes sociais tendem a se conectar com usuários de ideias semelhantes e, portanto, recebem informação que promove a sua narrativa preferida. Isso pode levar ao aumento da polarização de usuários, resultando no chamado efeito de *câmera de eco*.

Nas câmeras de eco, usuários compartilham e consomem a mesma informação, criando comunidades polarizadas e segmentadas. Segundo Paul e Matthews (2016), a câmera de eco facilita o processo pelo qual pessoas consomem e acreditam em desinformação pelos seguintes fatores psicológicos:

- **Credibilidade social**: é mais provável que as pessoas percebam uma fonte como confiável se outras a percebem como tal, especialmente quando não há informação suficiente para avaliar a veracidade daquela fonte.
- **Heurística de frequência**: usuários podem naturalmente favorecer informação que eles recebem mais frequentemente, mesmo que seja falsa.

2.1.2.2 Filtros bolha

Segundo Pariser (2011), um *filtro bolha* é um isolamento intelectual que ocorre quando redes sociais utilizam algoritmos para personalizar as informações recebidas por usuários. Os algoritmos fazem suposições sobre as preferências dos usuários baseado nos seus dados pessoais, tais como histórico de navegação, comportamento de cliques, histórico de buscas e localização. Assim, é mais provável que a rede social apresente a um usuário informação que reforce a suas atividades anteriores. Um filtro bolha pode reduzir a visibilidade de pontos de vista contraditórios, amplificando os desafios psicológicos individuais para identificar notícias falsas. Esses desafios incluem:

- **Realismo ingênuo**: de acordo com Ross *et al.* (1996), pessoas tendem a acreditar que as suas percepções da realidade são as únicas acuradas, enquanto outras percepções discordantes são consideradas desinformadas, irracionais ou tendenciosas.

– **Viés de confirmação**: de acordo com Nickerson (1998), pessoas preferem receber informações que confirmem suas visões pré-existentes.

2.1.2.3 Usuários individuais

Segundo Shu *et al.* (2019), no processo de disseminação de desinformação em redes sociais, os usuários individuais exercem papéis diferentes, classificando eles nos seguintes grupos: *persuasores*, que espalham desinformação com suas opiniões pessoais para influenciar outros a acreditarem neles; *usuários crédulos*, que são persuadidos a acreditarem em desinformação; e *clarificadores*, que apresentam pontos-de-vista céticos para esclarecer desinformação. De acordo com Vicario *et al.* (2016), a enxurrada de desinformação é disseminada não apenas por persuasores influentes mas também por uma massa crítica de indivíduos influenciados, ou seja, usuários crédulos.

Rotter (1980) define credulidade como um conceito diferente de confiança. Enquanto indivíduos com alta confiança são indivíduos que assumem que outras pessoas são confiáveis a menos que provem o contrário, indivíduos crédulos são insensíveis à percepção de características que revelam ausência de credibilidade. Shu *et al.* (2019) afirma que reduzir o recebimento de desinformação para usuários crédulos é fundamental para mitigar os efeitos da desinformação. Os esclarecedores podem espalhar opiniões opostas contra notícias falsas e evitar pontos de vista unilaterais. Por fim, os esclarecedores também podem espalhar notícias verdadeiras que podem imunizar os usuários crédulos contra uma desinformação e engajar esses usuários a propagar e espalhar notícias verdadeiras para outros usuários.

2.1.2.4 Contas maliciosas

Existem usuários de redes sociais que atuam com intenções maliciosas, desvirtuando o uso da rede. Em muitos casos, existem contas de usuários que sequer são controladas inteiramente por seres humanos. Contas maliciosas que podem amplificar a disseminação de desinformação incluem *bots sociais*, *trolls*, *propagandistas* e *cyborgs*.

De acordo com Shu *et al.* (2019), ***bots*** (uma contração do termo em inglês, *robots*) são contas de redes sociais controladas por um algoritmo que publica conteúdo e interage com outros usuários de forma automática. *Bots* podem ser entidades maliciosas designadas especificamente para manipular e disseminar desinformação em redes sociais em larga escala.

Já os ***trolls***, segundo Buckels *et al.* (2014), são humanos reais que visam prejudicar

comunidades *online*, enganar, provocar e ofender outros usuários para gerar uma resposta emocional. De acordo com Cheng *et al.* (2017), o comportamento de *trolls*, a trollagem (um aportuguesamento da expressão inglesa *trolling*), é altamente afetado pelo humor das pessoas e pelo contexto das discussões *online*, o que permite a fácil disseminação de desinformação entre comunidades outrora "normais".

Propagandistas ("propagandists") são definidos por Orlov e Litvak (2018) como um grupo de pessoas que intencionalmente espalham desinformação ou declarações tendenciosas, tipicamente agindo de forma coordenada e recebendo pagamento por essa tarefa. Vale notar que, de acordo com essa definição, propagandistas podem ser responsáveis por contas de *bots*, utilizando-as como ferramenta para os seus propósitos.

Por fim, ***cyborgs*** são deinidos por Chu *et al.* (2010) como contas cujo controle é feito ora por um ser humano, ora por um algoritmo. Contas *cyborgs* são usualmente registradas por um humano para disfarçar a atividade automatizada realizada nas redes sociais. A facilidade de alternar entre ação humana e automatizada permite a esses usuários oportunidades únicas de disseminar desinformação.

Neste trabalho, adotamos uma definição ampla de desinformadores para nos referir aos diversos tipos de usuários maliciosos que disseminam desinformação continuamente.

2.2 Detecção de desinformação e desinformadores

Na seção anterior, nós introduzimos uma caracterização conceitual de desinformação nas redes sociais. Baseada nessa caracterização, exploraremos a definição dos problemas de detecção de desinformação e desinformadores e as principais abordagens existentes.

2.2.1 Definição formal de detecção de desinformação

A definição formal de detecção de desinformação pode variar na literatura de acordo com a abordagem dos autores. Nesta dissertação adotaremos uma definição baseada na dada por Guo *et al.* (2019a), acrescentando ainda a informação da rede social, conforme feito por Freire e Goldschmidt (2019b). Essa definição modela a tarefa de detecção de desinformação como um problema de classificação binária. Nessa definição, a seguinte notação é utilizada:

– Seja uma uma informação (notícia, alegação, declaração, etc.) s, disseminada em uma rede social N. A informação s contém um conjunto $P = \{p_1, p_2, ..., p_n\}$ de n publicações

relacionadas e um conjunto $U = \{u_1, u_2, ..., u_m\}$ de m usuários relevantes. Cada p_i consiste de uma série de atributos que representam a publicação, incluindo texto, imagem, número de comentários, etc. Ou seja, representa a dimensão do conteúdo. Cada u_i consiste de uma série de atributos descrevendo o usuário, como número de publicações, número de conexões com outros usuários em N, etc. Ou seja, representa a dimensão social.

– Seja $E = \{e_1, e_2, ..., e_{n \times m}\}$ os engajamentos entre m usuários e n publicações. Cada e_i é definido como $e_i = \{p_i, u_j, a, t\}$, representando que um usuário u_j interagiu com a publicação p_i através da ação a (publicar, compartilhar, comentar, etc) no tempo t. A partir do conjunto E é possível representar a dimensão temporal.

Seja a informação $s = \{P, U, E\}$, em uma rede social N, com seu conjunto de publicações P, conjunto de usuários U e conjunto de engajamentos E. Seja a função $D(s) \mapsto [0, 1]$ um score de desinformação atribuído à s e ϕ um limiar de decisão. A detecção de desinformação é a tarefa de aprender uma função de decisão $F(s, \phi)$ satisfazendo:

$$F(s, \phi) = \begin{cases} 1 \text{ (contém desinformação), se} & D(s) \geqslant \phi \\ 0 \text{ (não contém desinformação), se} & D(s) < \phi \end{cases}$$

No contexto de aprendizado supervisionado, tanto $D(s)$ como ϕ são aprendidos a partir de dados. Observa-se que essa definição é ampla o suficiente para considerar diferentes abordagens de detecção de desinformação, quer explorem atributos da dimensão do conteúdo, da dimensão social ou uma combinação de ambas.

2.2.1.1 Detecção de desinformação por aprendizado supervisionado

Os métodos de classificação mais comuns utilizados na literatura para detecção de desinformação são algoritmos de aprendizado supervisionado. De acordo com Bishop (2006), o objetivo de algoritmos de aprendizado supervisionado em um problema de classificação é aprender uma superfície de decisão que mapeia um espaço de atributos de entrada para um espaço de saída de rótulos de classe. Neste problema em particular, se o vetor de atributos de entrada representa uma desinformação ou não.

Ainda segundo Bishop (2006), a "aprendizagem" do algoritmo geralmente refere-se à otimização de um conjunto de parâmetros, de modo a reduzir o erro de classificação em um conjunto de dados de treinamento, onde o "erro" é dado por uma função-custo. Diversos algoritmos foram utilizados com sucesso para detecção de desinformação em diversos cenários

e considerando diferentes conjuntos de atributos, desde modelos estatísticos como Regressão Logística, *Support Vector Machine* / Máquina de Vetores de Suporte (*SVM*) *Naïve Bayes*, à modelos de aprendizado profundo.

2.2.1.2 Outras formulações de detecção de desinformação

Encontram-se na literatura outras formulações para essa tarefa que divergem da apresentada. Nos casos em que a informação é parcialmente verdadeira e parcialmente falsa, a classificação binária não é completa o suficiente. Uma solução para esse problema é formular a detecção de desinformação como um problema de classificação multi-classe, adicionando novas classes com maior granularidade para s.

Por exemplo, Wang (2017), define as classes-alvo como *pants-fire* (uma expressão em inglês para uma mentira descarada), *false*, *barely-true*, *half-true*, *mostly-true* e *true*. Detecção de desinformação pode também ser formulada como uma tarefa de regressão. Nakashole e Mitchell (2014) formula a tarefa como a predição de um *score* numérico de veracidade.

Contudo, vale notar que o desenvolvimento e a avaliação de soluções baseadas nessas definições dependem da obtenção de conjuntos de dados rotulados com múltiplas classes ou com um *score* de veracidade. Uma vez que a criação desses rótulos é uma tarefa mais desafiadora que rótulos binários, essas abordagens são menos frequentes na literatura, sendo a classificação binária prevalecente.

Destaca-se ainda que existem definições que não são baseadas em aprendizado supervisionado, uma vez que a disponibilidade de dados rotulados ainda é baixa e a rotulação de dados é custosa. Métodos semi-supervisionados e não-supervisionados são propostos formulando detecção de desinformação como um problema de clusterização (RUBIN; VASHCHILKO, 2012; GUACHO *et al.*, 2018; YANG *et al.*, 2019).

A seguir, realizamos uma definição análoga para a tarefa de detecção de desinformadores.

2.2.2 Definição formal de detecção de desinformadores

A detecção de desinformadores é uma tarefa fortemente relacionada com o problema de detecção de desinformação. Mas, ao invés de classificar uma informação, busca-se classificar usuários. De maneira geral, o objetivo dessa tarefa é identificar usuários maliciosos, responsáveis por propagação de desinformação em larga escala, sejam *bots*, *trolls*, *cyborgs*, propagandistas,

dentre outros subtipos. A detecção de desinformadores é uma abordagem importante para combater a disseminação da desinformação, uma vez que permite identificar usuários que muitas vezes são as fontes ou os maiores propagadores de desinformação. Com essa identificação seria possível, por exemplo, bloquear o fluxo de desinformação desses usuários para outros usuários conectados com estes, mitigando a sua disseminação.

Assim como no caso da detecção de desinformação, a formulação do problema de detecção de desinformadores pode variar na literatura, de acordo com o problema específico abordado pelos autores, como detecção de *bots* ou de propagandistas. Propomos uma definição abrangente semelhante à Definição 2.2.1, formulada a seguir:

Seja um usuário $u = \{U_u, E_u\}$ de uma rede social N, que possui associado a ele um conjunto de $U_u = \{u_1, u_2, ..., u_m\}$ de outros m usuários com os quais ele possui uma conexão, um conjunto de engajamentos $E_u = \{e_1^u, e_2^u, ..., e_n^u\}$, onde cada $e_i^u = \{p_i, a, t\}$ representa um engajamento de u com a publicação p_i, pela ação a, no tempo t. Seja a função $Q(s) \mapsto [0,1]$ um score de desinformador atribuído à u e τ um limiar de decisão. A detecção da desinformadores é a tarefa de aprender uma função de predição $G(u,\tau) \mapsto [0,1]$, satisfazendo:

$$G(u,\tau) = \begin{cases} 1 \text{ (é desinformador), se} & Q(s) \geqslant \tau \\ 0 \text{ (não é desinformador), se} & Q(s) < \tau \end{cases}$$

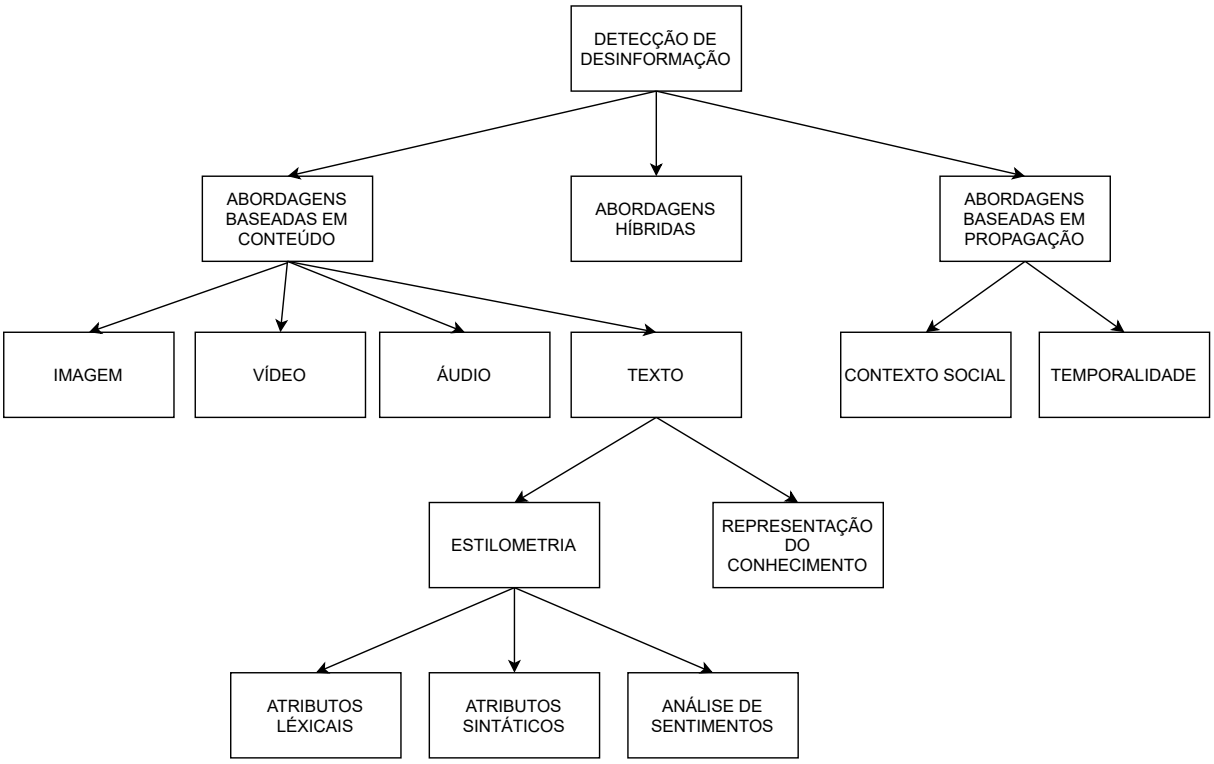
Uma definição específica para categorizar um usuário como desinformador pode variar de acordo com a rede social analisada ou o comportamento específico que se deseja detectar. Entretanto, deve ser considerado que o usuário publica ou compartilha desinformação com frequência ou proporção incomum em comparação com os outros usuários dessa rede social (ORLOV; LITVAK, 2018; CHU *et al.*, 2010; ZHANG; HARA, 2020). Ou seja, um desinformador publica uma alta quantidade de publicações enganosas, ou a maior parte de suas publicações contém informações falsas.

Não trata-se, portanto, de um usuário crédulo, que possui atividades regulares na rede social e eventualmente publica desinformação, mas sim usuários engajados na atividade de disseminar desinformação, de forma anormal em relação a usuários regulares. Vale ressaltar que, a depender da rede social, esse comportamento muitas vezes infringe as políticas de comunidade da mesma.

2.2.3 Abordagens de detecção de desinformação

De acordo com a Definição 2.2.1, as abordagens de detecção de desinformação existentes podem ser categorizadas em três grandes grupos, de acordo com o tipo de atributos utilizados para classificação: **baseadas em conteúdo**, **baseadas em propagação** e **híbridas**. A Figura 3 ilustra a classificação entre as abordagens, que serão discutidas a seguir. Em particular, neste trabalho são exploradas abordagens baseadas em conteúdo, em contexto social e abordagens híbridas.

Figura 3 – Categorias de abordagens de detecção de desinformação.



Fonte: o autor.

2.2.3.1 Abordagens baseadas em conteúdo

Conforme afirmado por Guo *et al.* (2019a), abordagens baseadas em conteúdo extraem atributos da mídia associada a uma publicação, podendo ser textos, imagens, áudio ou vídeos. Ou seja, dados relativos ao conjunto P da Definição 2.2.1. Este trabalho foca na utilização de técnicas detecção de desinformação na forma de texto utilizando técnicas de *NLP*, também chamadas de abordagens baseadas em estilo, ou estilometria (ZHOU; ZAFARANI, 2018; POTTHAST *et al.*, 2017).

De acordo com Guo *et al.* (2019a), a desinformação é fabricada para enganar o público e atrair a atenção das pessoas, então seu conteúdo textual geralmente tem padrões díspares quando comparada com informações verdadeiras. Como o nome indica, esse grupo de abordagens busca identificar esses padrões através de atributos do estilo de escrita, tais como padrões de frequências de palavras e correlação estatística entre textos, atributos lexicais, sintáticos, atributos de tópicos, atributos linguísticos, atributos de sentimentos e outros, conforme afirmado por Rubin e Lukoianova (2015).

Atributos lexicais incluem atributos do nível de caracteres e de nível de palavras. Segundo Su *et al.* (2020), dentre os atributos lexicais, a forma mais simples de representar textos é através da representação *Bag of Words* / Saco de Palavras (*BoW*), que considera cada palavra como uma unidade única e igualmente significativa. Na representação *BoW*, cada texto (ou documento) é representado pela presença ou frequência de "n-gramas" (sequências contígua de n signos linguísticos, comumente palavras), desconsiderando ordem e gramática. A Tabela 1 exemplifica a representação *BoW* para um pequeno conjunto de documentos.

Ainda segundo Su *et al.* (2020), o conjunto de atributos de nível de palavras mais frequentemente adotado é a representação *TF-IDF* de um documento. Na representação *TF-IDF* cada texto é representado pela frequência dos n-gramas ponderados pela ocorrência desses n-gramas em todos os documentos de um *corpus* (conjunto de documentos). Os valores *TF-IDF* buscam modelar a importância de termos em documentos de acordo com a sua frequência, ponderados pela quantidade de documentos do mesmo *corpus* que possuem esse termo. Se uma palavra ocorre com muita frequência em um documento e não ocorre em outros documentos do corpus, essa palavra deve ser importante nesse documento. Por outro lado, se um termo ocorre em todos os documentos, o valor de *TF-IDF* deste é penalizado. O cálculo do *TF-IDF* para cada termo é formalizado na Equação 2.1:

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{2.1}$$

onde t é um termo, d é um documento, D é o corpus com N documentos. $tf(t,d)$ é a frequência do termo t no documento d e $idf(t,D)$ é dado pela Equação 2.2:

$$IDF(t,D) = log\frac{|D|}{1+\{d \in D : t \in d\}} \tag{2.2}$$

Tabela 1 – Exemplo da representação *Bag of Words*

Documento	o	gato	sentou	no	chapéu	com
o gato sentou	1	1	1	0	0	0
o gato sentou no chapéu	1	1	1	1	1	0
o gato com chapéu	1	1	0	0	1	1

Tanto com o *BoW* como com o *TF-IDF*, um *corpus* é representado por uma matriz termo-documento, onde cada linha representa um documento e cada coluna representa um valor para um termo naquele documento. Como os documentos não costumam possuir a maior parte dos termos presentes no vocabulário do corpus, a matriz termo-documento costuma ser extremamente esparsa. Percebe-se que vetores *BoW* e *TF-IDF* podem ser significativamente grandes, a depender do tamanho do vocabulário dos documentos.

Outro conjunto de atributos lexicais que ganhou destaque na literatura pelo bom desempenho em problemas de classificação de texto são modelos de representação de vetores de palavras (*word embeddings*) como *Word2Vec* (MIKOLOV *et al.*, 2013). *Word embeddings* são modelos de linguagem aprendidas por redes neurais que codificam palavras em um espaço vetorial denso, onde palavras que possuem semântica semelhante, ou seja, que ocorrem frequentemente no mesmo contexto, são codificadas em vetores espacialmente próximos. Conceitualmente, modelos de *word embeddings* realizam o mapeamento matemático de um espaço com muitas dimensões por palavra para um espaço vetorial contínuo com uma dimensão muito inferior.

Ainda dentro dos atributos lexicais, pode ser citado o *Linguistic Inquiry and Word Count* / Pesquisa Linguística e Contagem de Palavras (*LIWC*) (POTTHAST *et al.*, 2017), que categoriza palavras em categorias derivadas da gramática e psicologia. É importante considerar que, apesar de a literatura mostrar que métodos baseados em atributos lexicais obtém um bom desempenho na tarefa de detecção da desinformação nas condições testadas, os padrões de atributos lexicais costumam ser altamente sensíveis ao ciclo de informação específico no qual o *corpus* de treinamento foi construído, conforme mostrado nos trabalhos de Schuster *et al.* (2020) e Zhou *et al.* (2019). Ou seja, os padrões de frequências de termos nos documentos que identificam uma desinformação podem mudar de acordo com os tópicos e assuntos discutidos em um determinado momento temporal.

Outros grupos de abordagens baseadas em estilo incluem **atributos sintáticos**, formados por atributos gramaticais e estruturais, como frequência da pontuação ou de *Part of Speech* / Partes do Discurso (*POS*) (QAZVINIAN *et al.*, 2011); e **análise de sentimentos**, que busca modelar a polaridade de emoções presentes em um texto, com a premissa que

desinformação tipicamente contém emoções intensas que buscam engajar o público (GUO *et al.*, 2019b).

Ainda considerando a desinformação em textos, existe um grupo de abordagens que separam-se das abordagens baseadas em estilo, que são as **baseadas em conhecimento**. Segundo Pan *et al.* (2018), nessas abordagens a informação é representada pelo conhecimento sistematicamente extraído na forma de um conjunto de triplas factuais contendo Sujeito, Predicado e Objeto (SPO), extraídos do texto. Por exemplo, as triplas (***LuísRobertoBarroso, Ser, Magistrado***), (***LuísRobertoBarroso, NascidoEm, Brasil***), (***LuísRobertoBarroso, Ser, MinistroDoSupremoTribunalFederal***) e (***LuísRobertoBarroso, Ser, MinistroDoTribunalSuperiorEleitoral***) representam conhecimento extraído da sentença *"Luís Roberto Barroso é um magistrado brasileiro, atualmente ministro do Supremo Tribunal Federal e do Tribunal Superior Eleitoral"*.

Um conjunto de triplas forma o chamado *Knowledge Graph* / Grafo de Conhecimento (*KG*), que pode ser construído de forma manual, automática ou semiautomática. Em problemas de detecção de desinformação, o *KG* é entendido como a *"verdade terrestre"* e o processo de detecção consiste em avaliar a veracidade destas informações, verificando a compatibilidade do conhecimento extraído do conteúdo em comparação com as representações do conhecimento no *KG* (PAN *et al.*, 2018; SHI; WENINGER, 2015; LIN *et al.*, 2018; TCHECHMEDJIEV *et al.*, 2019). Segundo Shi e Weninger (2015) e Lin *et al.* (2018), embora seja uma tarefa que pode ser entendida como detecção de desinformação, esse tipo de abordagem também é referenciada como checagem de fatos automática.

2.2.3.2 Limitações de abordagens baseadas em conteúdo

Abordagens baseadas em conteúdo são as mais frequentes na literatura, devido a praticidade de desenvolver modelos de predição com alta acurácia, uma vez que dados rotulados estejam disponíveis. Modelos de detecção de desinformação baseados em estilo possuem ainda a capacidade de classificar uma nova informação assim que ela surge, pois não necessitam de nenhum dado adicional além do próprio conteúdo da informação, permitindo a detecção precoce. Entretanto, elas também possuem limitações relevantes, como a já citada limitação temporal, que é relativa a própria mutabilidade da língua, exigindo que esses modelos estejam em contínuo aprimoramento para evitar a queda de desempenho.

Além disso, foi demonstrado por Zhou *et al.* (2019) que métodos desse tipo são vulneráveis a ataques adversariais. Esses ataques levam modelos a obter baixo desempenho

apenas com pequenas modificações no texto que alteram a sua semântica mas sem modificar o estilo de escrita. Isso indica que um usuário malicioso poderia criar textos propositalmente para enganar esses modelos. Mais ainda, segundo Mohseni *et al.* (2019) muitas dessas abordagens possuem uma baixa interpretabilidade, pois os modelos de alto desempenho são usualmente "caixas-preta". Ou seja, o usuário final não possui um entendimento do porquê o modelo afirma que uma dada informação é verdadeira ou falsa. A transparência nesse processo é fundamental em muitas aplicações, principalmente quando lida-se com a questão legal envolvida na comunicação em redes sociais.

2.2.3.3 Abordagens baseadas em propagação

Segundo Su *et al.* (2020), abordagens baseadas em propagação exploram contextos sociais e engajamentos de usuários com uma informação ao longo do tempo. Pode-se considerar que abordagens baseadas em propagação exploram padrões em atributos que representam as dimensões social e temporal, ou seja, atributos relacionados aos conjuntos U e E da Definição 2.2.1. Esses atributos podem ser separados em dois grupos: **contexto social** e **temporalidade**.

Abordagens baseadas em contexto social exploram atributos sobre os usuários que interagiram com a informação, seja publicando, curtindo ou compartilhando. Os atributos que podem ser extraídos de um usuário u dependem fortemente da rede social N em questão. No Twitter, por exemplo, alguns atributos explícitos são: se a conta do usuário é ou não verificada pela plataforma, o número de dias passados desde o registro da conta, o número de publicações, o número de ações favoritadas, o número de seguidores, o número de contas seguidas (SHU *et al.*, 2018), razão entre seguidores/seguidos, se a conta possui descrição, se a conta possui uma imagem de perfil, localização da conta (MA *et al.*, 2015), dentre outras.

Outros possíveis atributos de usuário, extraídos de forma indireta, são a modelagem de um *score* de credibilidade dos usuários (LI *et al.*, 2019; ZHANG; HARA, 2020), métricas de redes complexas, como centralidade do nó (ORLOV; LITVAK, 2018), dentre outras. Uma abordagem que alcançou bons resultados são representações latentes de *embedding* de usuários, aprendida a partir da matriz de adjacência entre usuários conectados, de modo que usuários similares são representados por vetores espacialmente próximos (SHU *et al.*, 2019; HAMDI *et al.*, 2020).

Abordagens baseadas em temporalidade exploram atributos relacionados a padrões de difusão da informação em uma dada rede social ao longo do tempo. Por exemplo, Jin

et al. (2013) utiliza modelos epidemiológicos para descrever o processo de disseminação de rumores no Twitter e propõe um modelo epidêmico aprimorado para detecção de desinformação. Já Ma *et al.* (2015) caracteriza padrões temporais de atributos de contexto social ao longo da disseminação de uma informação na rede, combinando as dimensões social e temporal.

De acordo com Wu e Liu (2018), abordagens baseadas em propagação são relevantes por não sofrerem as limitações já discutidas de abordagens baseadas em conteúdo, como a variabilidade temporal dos padrões linguísticos. Por exemplo, um modelo treinado com dados de conteúdo de um determinado período e contexto, como desinformações sobre política, pode ter um baixo desempenho quando utilizado em outro contexto, como em textos sobre saúde. Abordagens baseadas em propagação são mais robustas nesse aspecto, pois espera-se que os padrões de propagação não se alterem de acordo com a temática dos tópicos discutidos.

Porém, segundo Ruchansky *et al.* (2017), dados de propagação nem sempre são possíveis de serem obtidos na prática e suas características são muito dependentes de redes sociais específicas, sendo por isso menos comuns na literatura do que abordagens baseadas em conteúdo. Além disso, selecionar os atributos de propagação relevantes para identificar desinformação não é uma tarefa trivial. Por fim, conforme relatado por Qian *et al.* (2018), algumas abordagens podem necessitar que uma informação tenha um certo nível de propagação para fazer uma predição acurada, o que aumenta a latência e, consequentemente, reduz a detecção precoce.

2.2.3.4 Abordagens híbridas

O terceiro grande grupo de abordagens são as abordagens híbridas. De acordo com Ruchansky *et al.* (2017), abordagens híbridas combinam de forma complementar atributos de conteúdo e de propagação, combinando seus pontos fortes e mitigando suas fraquezas. Em particular, busca-se obter o alto desempenho de abordagens baseadas em conteúdo e a robustez de abordagens baseadas em propagação, reduzindo a dependência linguística do corpus de treinamento e a necessidade de uma grande quantidade de propagações de uma informação. Intuitivamente, esse grupo de abordagens aproxima-se do processo realizado por um ser humano para verificar a veracidade de uma informação, analisando tanto o conteúdo, como a credibilidade das pessoas que o publicaram.

Dentre as muitas combinações possíveis, podemos citar, por exemplo, o trabalho de Vedova *et al.* (2018), que combina a informação textual de uma publicação com os atributos

explícitos de usuários interagindo com elas, combinando as dimensões social e do conteúdo. Já Li *et al.* (2019) utiliza uma abordagem baseada em aprendizado profundo com mecanismo de atenção para extrair atributos para representar o texto, da propagação da notícia e a credibilidade dos usuários que compartilharam, combinando assim as dimensões do conteúdo, social e temporal. No trabalho de Ruchansky *et al.* (2017) também é utilizada uma abordagem baseada em aprendizado profundo para extrair atributos do texto, da fonte da notícia e da reposta dos usuários a esta notícia.

Abordagens híbridas estão no estado da arte em detecção de desinformação. Porém, tem como custo a necessidade de uma maior quantidade de informações, o que nem sempre é fácil ou possível de se obter. Além disso, assim como as abordagens baseadas em propagação, não é trivial identificar a correta combinação de atributos de propagação que permitem uma boa classificação.

2.3 Conclusão

Nesse capítulo foram discutidos os principais conceitos relacionados a desinformação e a sua dinâmica nas redes sociais. Em seguida, foram feitas as definições formais das tarefas de detecção de desinformação e de desinformadores, além de uma organização conceitual das diferentes abordagens de detecção de desinformação. No capítulo seguinte serão apresentados trabalhos do estado da arte relacionados aos problemas aqui abordados, evidenciando a lacuna de pesquisa existente onde esta dissertação faz suas contribuições.

3 TRABALHOS RELACIONADOS

Este capítulo tem por finalidade apresentar e discutir o estado da arte em relação à detecção automática de desinformação textual e desinformadores em redes sociais no contexto da língua portuguesa, incluindo os conjuntos de dados disponíveis, as abordagens propostas, métodos existentes e ferramentas comumente utilizadas. Em seguida, serão discutidas as abordagens e os métodos desenvolvidos especificamente para o contexto do WhatsApp. Por fim, será realizada uma análise comparativa entre os trabalhos apresentados e apontadas as lacunas onde esta dissertação busca realizar suas contribuições. Salienta-se que esta revisão da literatura foi do tipo narrativa, não seguindo critérios sistemáticos para a busca e seleção dos artigos analisados.

3.1 Detecção de desinformação na língua portuguesa

Apesar da grande quantidade de trabalhos investigando o problema de detecção de desinformação, ainda são poucos os que apresentam soluções para a língua portuguesa. Neste contexto, Monteiro *et al.* (2018b) apresenta o primeiro e maior corpus de *Fake News* em português, chamado **Fake.Br**. Esse corpus foi construído de forma semi-automática, coletando notícias falsas da *Web* através de *web scrapping*, validando e buscando a notícia real relacionada. Dessa forma, foi gerado uma quantidade igual de notícias verdadeiras e falsas para treinamento de modelos preditivos. Ao total, o conjunto de dados possui 7,200 notícias. Além da construção do conjunto de dados, os autores realizaram diversos experimentos de classificação, usando classificadores como *Naıve-Bayes*, *Random Forest*, e *Multilayer Perceptron* / Perceptron Multicamada (*MLP*) e diferentes atributos como Bag of Words, TF-IDF e classes gramaticais para detectar *Fake News*. É preciso destacar que os objetos de classificação são todos notícias, coletadas de páginas da *Web*. Portanto, trata-se do problema mais específico de detecção de *Fake News*, em contraste com detecção de desinformação, que é um problema mais abrangente.

Outros trabalhos ainda realizaram experimentos posteriores no Fake.Br, buscando estender os resultados do trabalho original. Silva *et al.* (2020b) propôs responder as seguintes questões de pesquisa: quais são os melhores métodos atuais para detecção automática de notícias falsas? Qual é o melhor conjunto de atributos para classificação de notícias falsas? Qual é o impacto das diferentes estratégias de classificação (por exemplo, *ensemble* e *stacking*) para detecção de notícias falsas? Este trabalho alcançou uma performance de F1 *score* de 0,97, utilizando uma combinação de atributos linguísticos com *BoW* e uma abordagem de *stacking*.

No trabalho de Ruiz (2020), os autores obtém bons resultados classificando o Fake.Br com uma abordagem baseada em aprendizado profundo e *word embeddings*. Neste trabalho é utilizado uma *Hierarchical Attention Network* (HAN) como método de classificação e o método GloVe para representação dos *word embeddings*, alcançando um F1 *score* de 0,97.

Outro trabalho que destaca-se por apresentar um conjunto de dados para estudo de desinformação em português é o de Moreno e Bressan (2019). Os autores criaram e disponibilizaram publicamente o FACTCK.BR, um conjunto de dados rotulados e atualizável de alegações para a tarefa de checagem automática de fatos. Cada alegação foi analisada individualmente e contém a URL do artigo original, a data de publicação, uma resenha da alegação e do texto, além de um *score* de veracidade em uma escala linear, ao invés do mais frequente rótulo binário de verdadeiro ou falso.

No estudo desenvolvido por Charles e Sampaio (2018) é apresentado o FakePedia, um conjunto de dados colaborativo de rumores no Brasil, disponibilizando também um *plug-in* para o navegador *Web* Google Chrome que permite a verificação rápida de conteúdo selecionado pelo usuário baseada em recuperação da informação por similaridade, bem como uma API para que outras aplicações consultem a FakePedia. A FakePedia apresenta ao todo 3825 instâncias falsas e 1033 verdadeiras. O trabalho de Moraes *et al.* (2019) também utiliza o FakePedia, onde realiza experimentos de classificação utilizando atributos variados da estrutura gramatical, como *POS*, quantidade e percentual de caracteres maiúsculos e pontos de exclamação, além de análise de sentimentos, obtendo uma acurácia de 0,94 com o classificador *SVM*.

No trabalho de Cordeiro e Pinheiro (2019), é apresentado o conjunto de dados FakeTweet.Br, com *tweets* coletados de forma semi-automática e rotulados manualmente entre verdadeiro e falso. Ao todo, esse conjunto de dados é formado por um conjunto de treino e um de teste pré-definidos, contendo 194 *tweets* falsos e 85 *tweets* verdadeiros no conjunto de treinamento, e 12 *tweets* falsos e 8 *tweets* verdadeiros no conjunto de teste. Neste trabalho, foram também conduzidos uma série de experimentos de classificação com uma grande variedade de algoritmos clássicos de aprendizado de máquina e com atributos *BoW* e *TF-IDF*, obtendo o melhor desempenho de F1 *score* de 0,73 com o classificador SGD. Vale ressaltar que este é o primeiro conjunto de dados rotulado de desinformação extraída do Twitter em língua portuguesa. Porém, contém somente informação a nível de conteúdo, não disponibilizando os dados de propagação.

Por fim, no trabalho de Silva *et al.* (2020a) é proposto o método FakeNewsSetGen,

um processo abrangente para criação de conjuntos de dados de *Fake News* contendo informações de propagação. Como estudo de caso, os autores utilizaram este processo para criar o FakeNews-Set, até então o primeiro conjunto de dados de desinformação na língua portuguesa com dados de propagação, coletado do Twitter. Ao todo, esse conjunto de dados contém 300 instâncias falsas e 300 instâncias verdadeiras. Os autores avaliam métodos de detecção baseados em propagação para esse conjunto de dados e alcançam um F1 *score* de 0,959, utilizando o método *Implict Crowd Signals* (ICS) proposta por Freire e Goldschmidt (2019a)

3.2 Detecção de desinformadores na língua portuguesa

A detecção de desinformadores em redes sociais ainda é um problema menos abordado no contexto social da língua portuguesa quando comparado com detecção de desinformação. A maior parte dos trabalhos encontrados aborda o problema de detecção de *bots*. O trabalho de Lêu *et al.* (2019) apresenta um estudo de caso sobre a ação de *bots* nas discussões no Twitter acerca das eleições presidenciais brasileiras de 2018. O conjunto de dados foi coletado utilizando a API do Twitter, contendo informações de 635.957 usuários, rotulando um subconjunto manualmente. Ao total são extraídos 22 atributos para representar os usuários, incluindo a idade da conta em dias, se o usuário utiliza a foto padrão do Twitter em seu perfil e se alterou ou não o perfil padrão inicial da conta, o número de seguidores, de *tweets*, se é verificado, se a localização é dada, dentre outros. É utilizada árvores de decisão e regressão linear para classificar os usuários. O melhor resultado apresentado possui uma *Area Under the ROC Curve* / área sobre a curva ROC (AUC) de 0,819.

O trabalho de Santos *et al.* (2021) estende os resultados apresentados no conjunto de dados criado por Lêu *et al.* (2019), realizando experimentos com os algoritmos Regressão Logística, Árvore Aleatória, Random Forests, *Naive Bayes* e MLP, além de utilizar a técnica de super amostragem SMOTE, uma vez que as classes são desbalanceadas, havendo mais usuários normais do que *bots*. O melhor resultado alcançou AUC de 0,81 com *Naive Bayes*, sem o uso de super amostragem.

No trabalho de Braz e Goldschmidt (2018) são empregadas Redes Neurais Convolucionais (CNN) para extração de atributos diretamente dos textos publicados por usuários em um conjunto de dados multilíngua do Twitter criado por Lee *et al.* (2011). Nessa abordagem, a CNN identifica mensagens suspeitas que são usadas para classificar usuários junto com seus atributos comportamentais: número de usuários que a conta segue, número de *tweets* postados pela conta,

razão da quantidade de usuários seguidos por quantidade de seguidores e número de seguidores. O melhor resultado apresentado possui acurácia de 0,921.

No trabalho de Leite *et al.* (2020) é proposto um conjunto de regras para descrever e classificar *bots* no Twitter. As regras são baseadas no comportamento dos usuários, e utilizam como dados de entrada a quantidade de *tweets* favoritados, o índice de *tweets* respondidos, quantidade de *tweets* favoritados e média de *retweets*. Através de uma árvore de decisão, os usuários podem ser classificados por essas regras. No experimento apresentado, foi obtido uma AUC e uma acurácia de 0,97 no conjunto de dado coletado por Cresci *et al.* (2017).

Por fim, no trabalho de Benevenuto *et al.* (2008) é abordado o problema de detecção de usuários maliciosos (*spammers*) da plataforma de vídeos YouTube. Os autores coletaram um conjunto de dados de usuários do YouTube na forma de um grafo de vídeo-respostas. Os usuários são representados por três grupos de atributos: **atributos de usuários**, composto por número de vídeos adicionados no YouTube, número de amigos, número de vídeos assistidos, número de vídeos adicionados como favoritos, número de vídeos resposta enviados e recebidos, número de inscrições, número de inscritos e o número máximo de vídeos adicionados em um dia; **atributos dos vídeos**, composto pelas médias de duração, número de exibições, avaliações, comentários, favoritos, menções honrosas e elos externos nos vídeos postados; e **atributos de redes sociais**, que são os atributos de grafo - coeficiente de clusterização, UserRank, *betweenness*, reciprocidade e assortatividade. Utilizando esses atributos, foi obtido um F1 *score* de 0.81 na tarefa de detecção desses usuários.

3.3 Desinformação no WhatsApp

Existem relativamente poucos estudos sobre desinformação no contexto do WhatsApp. Um dos motivos é que a coleta de dados é mais desafiadora. Algumas redes sociais, como Twitter[1] e Facebook[2], disponibilizam APIs públicas que permitem a coleta de dados. Outras, como YouTube ou Instagram, por serem acessadas publicamente por navegadores *Web*, podem ter seus dados coletados por *web-crawlers*, softwares que realizam a navegação e extração automática de informações de páginas HTML. Porém, devido a sua natureza de aplicativo de mensagens privado, nenhuma dessas abordagens é diretamente possível no WhatsApp, sendo necessário recorrer a outras alternativas.

[1] https://developer.twitter.com/en/docs/twitter-api
[2] https://developers.facebook.com/docs/

Nesse sentido, destaca-se o trabalho seminal de Garimella e Tyson (2018), que propôs uma metodologia para coletar e analisar mensagens de grupos públicos de WhatsApp. Neste trabalho, os autores criam uma conta no WhatsApp e, através dessa conta, entram em diversos grupos públicos encontrados por meio de links na *Web*. Uma vez nos grupos, os autores conseguem extrair as mensagens utilizando a extensão *Web* do WhatsApp [3] e realizando a raspagem de dados. Dessa forma, os autores construíram um conjunto de dados realizando um *crawling* de 178 grupos públicos, contendo cerca de 45 mil usuários e 454 mil mensagens de diferentes países e linguagens, tais como Índia, Paquistão, Rússia, Brasil e Colômbia.

Na detecção de desinformação destaca-se o trabalho de Gaglani *et al.* (2020), onde os autores propõem uma estratégia para detecção de desinformação baseada em aprendizado não-supervisionado. Neste trabalho, foram coletadas 1000 mensagens de 10 grupos públicos, do contexto indiano. A melhor acurácia apresentada por este trabalho foi de 0,78. No trabalho de Indumathi e Gitanjali (2020), é proposto um sistema para bloquear desinformação no WhatsApp baseado em palavras-chave, número de caracteres e *crowdsourcing* - através de avaliações de usuários.

No contexto brasileiro, no trabalho de Resende *et al.* (2018) os autores apresentam um sistema similar ao proposto por Garimella e Tyson (2018) para coletar, analisar e visualizar grupos públicos no WhatsApp. Além de descrever sua metodologia, os autores também fazem uma breve caracterização das 169.154 mensagens compartilhadas por 6.314 usuários em 127 grupos públicos com temática de discussão política, a fim de auxiliar jornalistas e pesquisadores a compreender a repercussão dos acontecimentos relacionados às eleições brasileiras de 2018.

No estudo apresentado em Machado *et al.* (2019), os autores coletaram e analisaram 298.892 mensagens de WhatsApp, de 130 grupos públicos, no período que antecedeu os dois turnos das eleições presidenciais brasileiras de 2018. Além disso, eles examinaram uma amostra de 200 vídeos e imagens, extraídos dessas mensagens do WhatsApp, e desenvolveram uma nova tipologia para classificar esse conteúdo de mídia.

Em Resende *et al.* (2019), os autores analisaram diferentes aspectos das mensagens do WhatsApp de grupos públicos de orientação política. As mensagens foram coletadas durante dois grandes eventos políticos no Brasil: a greve nacional de caminhoneiros e a campanha das eleições nacionais, ambas em 2018. Os autores analisaram os tipos de conteúdos compartilhados dentro de tais grupos, bem como as estruturas de rede que emergem das interações do usuário.

[3] https://faq.whatsapp.com/web/download-and-installation/about-whatsapp-web-and-desktop/

Além disso, identificaram a presença de desinformação entre as imagens compartilhadas por meio de rótulos fornecidos por jornalistas e por um procedimento automático baseado em buscas no Google. Tanto em Resende *et al.* (2019), como em Machado *et al.* (2019) e Resende *et al.* (2018), os dados coletados não foram rotulados, nenhum conjunto de dados foi disponibilizado publicamente e nenhuma solução para o problema de detecção de desinformação ou desinformadores foi apresentada.

No caso da detecção de desinformação em português, citamos o trabalho de Faustini e Covões (2019), que explora o uso da técnica de *One-Class Classification* (OCC) em alguns conjuntos de dados, incluindo o Fake.Br, um conjunto de dados extraído do Twitter e um conjunto de 177 mensagens rotuladas provenientes do WhatsApp, das quais 12 são verdadeiras e o restante são falsas. No conjunto de dados do WhatsApp, o melhor desempenho obtido foi um F1 *score* de 0,65. Os dados utilizados no trabalho de Faustini e Covões (2019) foram disponibilizados publicamente. Porém, destaca-se que no caso do conjunto de dados proveniente do WhatsApp, as mensagens foram coletadas de forma indireta, através de um site de checagem de fatos. Assim, somente o conteúdo dessas mensagens foi coletado, impossibilitando análises temporais ou sociais.

Por fim, com exceção dos trabalhos publicados a partir dos resultados desta dissertação (CABRAL *et al.*, 2021; SÁ *et al.*, 2021; MARTINS *et al.*, 2021), não encontramos na literatura nenhum trabalho que realize a detecção de desinformadores no contexto do WhatsApp, deixando uma lacuna importante a ser preenchida. Em Cabral *et al.* (2021) foram realizados uma série de experimentos de detecção de desinformação com os dados de WhatsApp propostos nesta dissertação. Em Martins *et al.* (2021) foram realizados experimentos similares, mas com dados coletados em 2020, acerca da pandemia do Covid-19. Já em Sá *et al.* (2021) foi proposto a plataforma Farol Digital, que busca viabilizar sistematicamente a coleta, armazenamento e análise de dados de grupos públicos de WhatsApp.

3.4 Análise comparativa

Nesta seção, realizamos uma análise comparativa dos trabalhos relacionados citados em termos de conjuntos de dados, incluindo o conjunto de dados proposto nesta dissertação como uma de suas principais contribuições. A Tabela 2 apresenta de forma comparativa os principais conjuntos de dados de tarefas correlatas a detecção de desinformação e desinformadores na língua portuguesa, informando o contexto de onde foi coletado, a tarefa original a qual ele

Tabela 2 – Resumo de conjuntos de dados relacionados a detecção de desinformação ou desinformadores na língua portuguesa.

Trabalho	Conjunto de dados	Contexto original	Tarefa	Conteúdo	Propagação	Instâncias positivas	Instâncias negativas
(MONTEIRO *et al.*, 2018b)	Fake.Br	Websites	Detecção de Fake News	Sim	Não	3600	3600
(MORENO; BRESSAN, 2019)	FACTCK.BR	Websites	Checagem automática de fatos	Sim	Sim	2000	2000
(CHARLES; SAMPAIO, 2018)	FakePedia	Websites	Detecção de Fake News	Sim	Não	3825	1033
(CORDEIRO; PINHEIRO, 2019)	FakeTweet.Br	Twitter	Detecção de Fake News	Sim	Não	206	93
(SILVA *et al.*, 2020a)	FakeNewsSet	Twitter	Detecção de Fake News	Sim	Sim	300	300
(Faustini; Covões, 2019)	-	WhatsApp	Detecção de Fake News	Sim	Não	165	12
(BENEVENUTO *et al.*, 2008)	-	YouTube	Detecção de usuários maliciosos	Sim	Sim	157	641
(LÊU *et al.*, 2019)	-	Twitter	Detecção de *bots*	Sim	Sim	65	577
Este trabalho	FakeWhatsApp.Br	WhatsApp	Detecção de desinformação e desinformadores	Sim	Sim	3718	4089

Fonte: o autor.

Tabela 3 – Trabalhos que analisam dados do WhatsApp no contexto do Brasil.

Trabalho/Conjunto de dados	# Mensagens	# Grupos	# Usuários	Rotulado	Público
(RESENDE *et al.*, 2018)	169.154	127	6.314	Não	Não
(MACHADO *et al.*, 2019)	298.892	130	Não informado	Não	Não
(RESENDE *et al.*, 2019) / Greve dos caminhoneiros	95.424	141	5.272	Não	Não
(RESENDE *et al.*, 2019) / Eleições brasileiras de 2018	591.162	136	18.725	Não	Não
Este trabalho	282.601	59	5.364	Parcialmente	Sim

Fonte: o autor.

foi designado, se possui informação de conteúdo e de propagação (social e/ou temporal) e a quantidade de instâncias positivas (notícias falsas ou usuários desinformadores) e negativas (notícias verdadeiras ou usuários regulares).

Na Tabela 3 são comparados os estudos que realizam análise de dados em grupos públicos de WhatsApp, contendo informação sobre a quantidade de grupos, usuários e mensagens analisadas em cada, se possui rótulos e se é disponibilizado publicamente.

Através das Tabelas 2 e 3, observa-se que há uma lacuna de conjuntos de dados rotulados para estudos de detecção de desinformação e desinformadores no contexto do WhatsApp em português, contendo dados de propagação. Nesta dissertação, propomos a criação e disponibilização de tal conjunto, que será detalhado no Capítulo 4

3.5 Conclusão

Neste capítulo foi apresentado e discutido trabalhos do estado da arte que se correlacionam com os desafios de pesquisa enfrentados nesta dissertação. Foi também realizada uma análise comparativa que pontua a contribuição inovadora do conjunto de dados criado neste trabalho. No capítulo seguinte será discutido com detalhes o processo de criação e uma análise exploratória deste conjunto de dados.

4 O CONJUNTO DE DADOS FAKEWHATSAPP.BR

Existe uma lacuna de trabalhos que desenvolvam métodos de detecção de desinformação e desinformadores voltadas para o contexto do aplicativo WhatsApp. Visando preencher essa lacuna, nessa dissertação propomos a criação e disponibilização de um conjunto de dados representativo desse contexto, além de uma avaliação experimental dessas tarefas.

Neste capítulo serão descritas todas as etapas envolvidas na criação do conjunto de dados proposto nesse trabalho, chamado de FakeWhatsApp.Br, em referência aos trabalhos de Monteiro *et al.* (2018a) e Cordeiro e Pinheiro (2019). Será também apresentada uma análise exploratória desses dados. Por fim, o capítulo encerra-se com uma discussão sobre as limitações e vieses presentes nos dados. Os dados e análises apresentados nesta dissertação podem ser encontrados em nosso repositório *online*[1]. Vale destacar ainda que os métodos apresentados nesse capítulo são extensíveis para outros aplicativos de mensagem de funcionamento similar ao WhatsApp, como o Telegram[2], por exemplo.

4.1 Coleta dos dados brutos

Os dados utilizados nessa pesquisa foram originalmente coletados no trabalho de Mourão (2020), durante a campanha das eleições presidenciais de 2018 no Brasil. Conforme mencionado no Capítulo 3, o WhatsApp não disponibiliza uma API pública para coleta de dados, exigindo que métodos alternativos sejam empregados.

Para esta coleta, o autor adotou a estratégia de criar uma conta de WhatsApp e utilizá-la para adentrar em grupos públicos de campanha política. Os grupos foram encontrados através de buscas na *Web* pela palavra-chave "chat.whatsapp.com/" que corresponde a links de grupos públicos de WhatsApp. Também foram encontrados links em grupos abertos na rede social Facebook. Por fim, após adentrar nesses grupos, novos links foram encontrados dentro dos próprios grupos. Ao total, a conta criada pelo autor entrou em 59 grupos que atendiam aos critérios pré-estabelecidos.

A conta permaneceu nesses grupos, sem realizar interações, entre julho e novembro de 2018, correspondendo ao período de campanha eleitoral brasileira. Após esse período, as mensagens foram extraídas utilizando um recurso do WhatsApp que permite salvar todas as mensagens na forma de arquivos de texto plano. As mensagens são armazenadas em linhas de

[1] https://github.com/cabrau/FakeWhatsApp.Br
[2] https://www.telegram.org/

texto, onde cada nova mensagem começa com a data/hora, o número do celular do autor da mensagem, e o texto da mensagem. Ao final do processo, o conjunto de dados brutos extraídos totalizaram 59 arquivos de texto, cada um correspondendo às mensagens de um determinado grupo. São esses dados em forma de arquivos de texto que foram inicialmente trabalhados durante esta pesquisa.

Devido ao método de extração, só obtivemos acesso aos textos das mensagens. O conteúdo de arquivos de mídia, abundantes no contexto do WhatsApp, como imagens, vídeos e áudios, não puderam ser acessados. Quando um arquivo de mídia é enviado, a mensagem armazenada por esse método de extração é o texto "<*Arquivo de mídia oculto*>". Dessa forma, nos dados brutos é registrado que um arquivo de mídia foi enviado, mas não pode-se determinar qual era o conteúdo desse arquivo ou mesmo o seu tipo (imagem, vídeo ou áudio). A Figura 4 ilustra o formato dos dados brutos na forma de texto plano. Observa-se a grande quantidade de mensagens com arquivos de mídia. Esse padrão repete-se em muitos outros grupos, por se tratarem de grupos de campanha política. Uma análise detalhada desses padrões será conduzida na Seção 4.4.

Figura 4 – Exemplo dos dados brutos na forma de arquivo de texto, com os número de celular dos usuários omitidos.

```
09/08/18 13:49 - +55 [redacted]: https://youtu.be/DfiOP8FLv84
09/08/18 13:49 - +55 [redacted]: https://noticias.uol.com.br/politica/eleicoes/2018/noticias/agencia-estado/2018/08/08/em-15-estados-pt-se-
alia-a-partidos-que-apoiaram-impeachment.htm
09/08/18 13:49 - +55 [redacted]: https://youtu.be/TUtOifMCRew
09/08/18 13:49 - +55 [redacted]: https://youtu.be/bG3qqYIW8ZY
09/08/18 14:19 - +55 [redacted]: - Repelido pelos corruptos do Centrão e abandonado pelos comunistas do PT, Ciro Gomes vota 17 - Bolsonaro
09/08/18 14:21 - +55 [redacted]: <Arquivo de mídia oculto>
09/08/18 15:48 - +55 [redacted]: https://youtu.be/pZ2jKB9qCMs
09/08/18 15:48 - +55 [redacted]: https://youtu.be/cBBmq8wEDmE
09/08/18 16:25 - +55 [redacted]: 3 homens são enforcados em praça publica em plena luz do dia https://www.xn---noticias-93a.com/2018/08/3-
homens-sao-enforcados-em-praca.html?m=1#.W2xKOcvSTbc.whatsapp
09/08/18 16:28 - +55 [redacted]: ☎
09/08/18 16:36 - +55 [redacted]: <Arquivo de mídia oculto>
09/08/18 16:54 - +55 [redacted]: Boa tarde! Venha passear em Angra dos Reis e fique no melhor condomínio. Clube com piscina, cachoeira,
trilha, praia, gente bonita e simpática . Aproveite o feriado de 7 de setembro.
09/08/18 16:53 - +55 [redacted]: <Arquivo de mídia oculto>
09/08/18 16:54 - +55 [redacted]: Ligue 24.998707427 e aluguel uma belíssima casa conosco. Venha! Venha! Curtir!
09/08/18 17:20 - +55 [redacted]: https://youtu.be/0fvDUOb1Izo
09/08/18 18:11 - +55 [redacted]: Não vou revelar meu voto, apenas darei pista...
09/08/18 18:11 - +55 [redacted]: <Arquivo de mídia oculto>
09/08/18 21:06 - +55 [redacted]: <Arquivo de mídia oculto>
09/08/18 21:07 - +55 [redacted]: Kkkkkkkk
09/08/18 21:08 - +55 [redacted]: Debate nesse momento dos candidatos a presidência
09/08/18 21:08 - +55 [redacted]: Band
09/08/18 21:08 - +55 [redacted]: <Arquivo de mídia oculto>
09/08/18 21:22 - +55 [redacted]: debate agora NA band Bolsonaro falou da Vila Histórica
09/08/18 22:21 - +55 [redacted]: <Arquivo de mídia oculto>
```

Fonte: o autor.

4.2 Normalização dos dados

A partir dos dados brutos em forma de texto plano, foi realizado um processamento para estruturar as informações em uma tabela, onde cada linha representa uma mensagem e cada coluna representa uma variável dessa mensagem. Para isso, foi feito o uso de expressões

regulares para reconhecer o padrão de data, hora, código telefônico de país (DDI), de estado (DDD, caso o número seja do Brasil), número do celular do usuário e texto da mensagem. Nessa extração foram descartadas mensagens geradas pelo próprio WhatsApp, como por exemplo, mensagens informando que um usuário saiu ou entrou no grupo, ou que a imagem ou nome do grupo foi alterado. Assim, para cada arquivo de texto que representa as mensagens de um grupo, foram aplicadas expressões regulares e obtidas essas variáveis de cada mensagem, além do grupo do qual foram obtidas. Esses dados foram então armazenados em um único arquivo *Comma-separated values* / valores separados por vírgula (*CSV*). A partir dessa tabela, seguiram-se outras etapas de processamento, descritas na lista a seguir:

- **Anonimização**: levando em consideração a privacidade dos usuários, foi realizada uma anonimização dos telefones utilizando uma função *hash* que transforma cada número de telefone em um identificador anônimo e único. Também foram criados pseudônimos para os grupos. Por fim, os números de telefone que porventura apareçam nos textos, são substituídos pelo valor simbólico "[TELEFONE]", através do uso de expressões regulares. Como os dados são extraídos de grupos públicos e não são coletados nem armazenados dados sensíveis dos usuários, esse estudo não fere a política de privacidade do WhatsApp[3] nem a Lei Geral de Proteção de Dados Pessoais (LGPD)[4]
- **Arquivos de Mídia**: Foi criada uma variável booleana que identifica se a mensagem é um arquivo de mídia ou texto, verificando se o texto da mensagem é igual a "<*Arquivo de mídia oculto*>".
- **Compartilhamentos**: Buscando identificar as mensagens compartilhadas ou encaminhadas, foram contadas quantas vezes cada texto aparece de forma idêntica no conjunto de dados. No entanto, para evitar que mensagens não relevantes como cumprimentos e outras mensagens curtas ("bom dia", "ok", "kkk", "sim") tivessem uma contagem alta, nós contamos somente as mensagens repetidas com 5 ou mais palavras. Chamamos essa contagem de "compartilhamentos" e as mensagens com mais de um compartilhamento são chamadas nesse contexto de mensagens "virais", criando uma variável booleana para indicar essa característica. Ou seja, são mensagens criadas propositalmente para que fossem recebidas por muitos usuários. Não foram contados os compartilhamentos das mensagens de mídia, uma vez que não temos como identificar o conteúdo de cada mensagem deste tipo.

Após essa normalização, temos um conjunto de dados estruturado, anonimizado e

[3] https://www.whatsapp.com/legal/privacy-policy/?lang=pt_br
[4] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm

não rotulado de mensagens de WhatsApp, contendo um total de 282.601 mensagens, enviadas por 5.364 usuários, de todos os estados do Brasil. A Figura 5 ilustra uma amostra do conjunto de dados nessa etapa. Vale ressaltar que uma vez que temos a data/hora e uma identificação de qual usuário enviou cada mensagem, podemos analisar as dimensões temporal e social. Ou seja, esse conjunto de dados possui um recorte de dados de propagação. Na Seção a seguir será apresentado o procedimento para rotular as mensagens.

Figura 5 – Amostra dos dados estruturados antes da rotulação.

id	timestamp	ddi	ddd	country	state	group	midia	shares	viral	text
174833288054493582	14/10/18 11:10	55	21	BRASIL	Rio de Janeiro	2018_32	1	1	0	<Arquivo de mídia oculto>
6870457668414668434	29/08/18 16:29	55	98	BRASIL	Maranhão	2018_59	0	1	0	Tenho pra mim que esse incêndio no Museum não ...
9153336480030841649	21/09/18 16:43	55	85	BRASIL	Ceará	2018_19	1	1	0	<Arquivo de mídia oculto>
5045562863966605913	04/09/18 18:19	55	53	BRASIL	Rio Grande do Sul	2018_59	0	1	0	Que chegue até quem pode determinar, pelo meno...
5884926763501609001	25/09/18 09:11	55	99	BRASIL	Maranhão	2018_45	1	1	0	<Arquivo de mídia oculto>
1586038897185087233	17/10/18 19:44	55	63	BRASIL	Tocantins	2018_42	0	2	1	Médico nordestino manda recado emocionante e v...
-9077122156671723433	09/09/18 22:07	+55	81	BRASIL	Pernambuco	2018_59	0	1	0	pensando nisto tuo
-8559750381028809722	16/08/18 12:51	55	11	BRASIL	São Paulo	2018_33	1	1	0	<Arquivo de mídia oculto>

Fonte: o autor.

4.3 Rotulação

Construir conjuntos de dados rotulados em larga escala é um dos maiores desafios para a detecção automática de desinformação. Esse processo envolve analisar cuidadosamente textos e realizar uma apuração das alegações contidas nele, que é desafiador e custoso. A tarefa de checagem de fatos é normalmente realizada manualmente por jornalistas ou outros especialistas treinados (RUBIN *et al.*, 2015). Neste trabalho, a rotulação foi feita manualmente pelo autor.

Como mencionado anteriormente, modelamos o problema de detecção de desinformação no WhatsApp como um problema de classificação binária, onde mensagens contendo desinformação formam a classe positiva (rótulo 1) e mensagens sem desinformação formam a classe negativa (rótulo 0). Para rotular nosso conjunto de dados com estas classes, selecionamos apenas o subconjunto das mensagens virais únicas, totalizando 5.284 mensagens. Esse recorte

se deve a dois fatores. Primeiro, conforme mencionado, o processo de rotulação é demorado, inviabilizando a rotulação de todas as mensagens. Segundo, de acordo com Vosoughi *et al.* (2018), a desinformação se espalha de forma mais rápida, profunda e ampla nas redes sociais do que a informação verdadeira, tendo, portanto, um caráter viral. Argumentamos que, dessa forma, evitamos rotular textos provenientes de diálogos no corpus, focando na detecção de mensagens criadas para serem disseminadas em larga escala.

O texto de cada mensagem foi individualmente usado como *query* em buscadores na *Web* a fim de obter mais informações sobre sua veracidade. A partir dessa busca, em muitos casos pode-se encontrar referências ao conteúdo da mensagem em sites de notícias ou checagem de fatos. Outras vezes, encontra-se a mensagem sendo reproduzida em outras redes sociais, como Twitter, Facebook e Youtube ou em portais de notícias.

Observamos empiricamente que o rotulação de dados em mensagens de WhatsApp é especialmente desafiadora e custosa. Diferentemente de conjuntos de dados de notícias, que são alegações factuais, mensagens de WhastApp contém uma variedade mais ampla de mensagens, incluindo rumores, textos humorísticos, sátiras, propaganda, textos de opinião, discurso de ódio, *trollagem*, dentro outros. Nem sempre é possível chegar a uma resposta objetiva, pois podem existir alegações que não podem ser facilmente comprovadas ou refutadas. Nesse cenário complexo, a rotulação envolve um certo nível de subjetividade do anotador. Para tornar o processo o mais rigoroso e objetivo, seguimos um protocolo de rotulação descrito a seguir, com exemplos de cada caso:

1. Se o texto contém alegações verificadamente falsas, nós rotulamos como desinformação. Para esse propósito, utilizamos buscas na *Web* e fazemos uso de plataformas brasileiras de checagem de fatos, como *Agência Lupa*[5] e *Boatos.org*[6].

   Exemplo: "Bolsa Ditadura se transformou em indústria: VC sabia que 20mil anistiados, entre eles, Chico Buarque, Gilberto Gil, Caetano Veloso, Marieta Severo, Taiguara, Lula, Zé Dirceu, Fernando Henrique Cardoso, recebem o Bolsa Ditadura mensalmente e são isentos de pagar Imposto de Renda? Sendo que dos 20 mil, 10 mil recebem indenizações mensais acima do teto constitucional(R$ 33.763,00) Essa esquerda maldita tira dos cofres públicos mensalmente a bagatela de R$ 365.000.000,00 (Trezentos e sessenta e cinco milhões) pagos por nós, otários!"[7].

[5] http://piaui.folha.uol.com.br/lupa/

[6] http://www.boatos.org/

[7] https://www.aosfatos.org/noticias/nao-e-verdade-que-governo-paga-bolsa-ditadura-20-mil-anistiados-politicos/

2. Se o texto contiver alegações que não podem ser comprovadas e que são imprecisas, tendenciosas, alarmistas e/ou prejudiciais a grupos ou indivíduos, rotulamos como desinformação.

   Exemplo: "O golpe da esquerda é o seguinte: a viagem do Ciro Gomes à Europa foi proposital! Um teatro p/ colocar a seguinte narrativa em prática: ele sai de cena, ou seja, teoricamente não está apoiando Haddad, de repente ele volta (e de fato, de acordo c/ o Estadão ele está chegando hoje) qdo não terá mais nenhuma pesquisa a ser divulgada, para q não se "comprove", se de fato aconteceria, a escalada q Haddad "terá" de milhões de votos em 2 dias. (...)"

3. Por decisão do Tribunal Superior Eleitoral, as pesquisas eleitorais informais, que não atendem aos requisitos formais e rigores científicos, foram proibidas nas eleições de 2018, evitando assim a desinformação do público. Assim, rotulamos mensagens contendo tais enquetes como desinformação.

   Exemplo: "Vota aí e repassa!!! Vamos ver se o ibope está certo? https://pt.surveymonkey.com/r/W85R38F"

4. Algumas das mensagens são textos curtos originalmente acompanhados de conteúdo de mídia (imagem, áudio ou vídeo) que não estão acessíveis. Nesses casos, procuramos na *Web* o conteúdo midiático referido e, se encontrarmos a mídia, atribuímos um rótulo seguindo os critérios anteriores.

   Exemplo: "Antes de decidir seu voto ouça o que diz o padre Marcelo Rossi". [8]

5. Se o conteúdo original da mídia não puder ser encontrado, procuramos indicações do Item 2 no próprio texto para julgar se contém desinformação.

   Exemplo: "Olha o que os partidos de esquerda defendem E se votarmos viraremos isso".

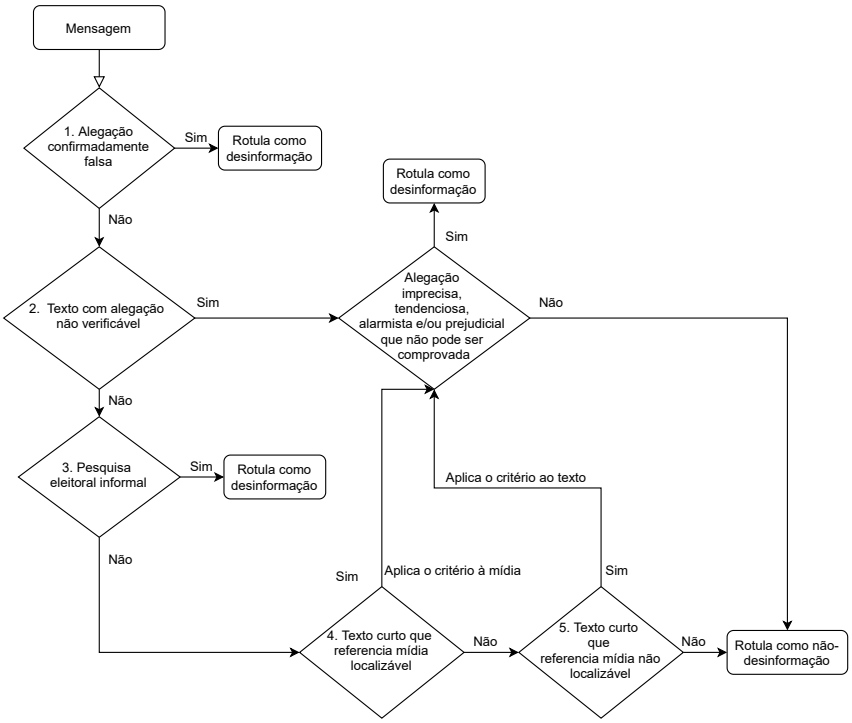
6. Se nenhuma das indicações anteriores for encontrada no texto, rotulamos que não contém desinformação. Levamos em consideração quando o texto é uma opinião em vez de uma afirmação ou é humorístico, atribuindo um rótulo de não desinformação em ambos os casos.

   Exemplo: "Relaxando no sofá, barriguinha plusize, 9mm na cintura, sem coldre, no pelo, com saque cruzado, relogio Cassio modelo 1985 no punho e xingando comunistas no insta... Esse é meu Presidente!".

O processo completo é ilustrado no formato de fluxograma pela Figura 6.

[8] https://www.boatos.org/religiao/padre-marcelo-rossi-grava-audio-brasil-bolsonaro-comunismo.html

Figura 6 – Fluxograma do protocolo de rotulação das mensagens.



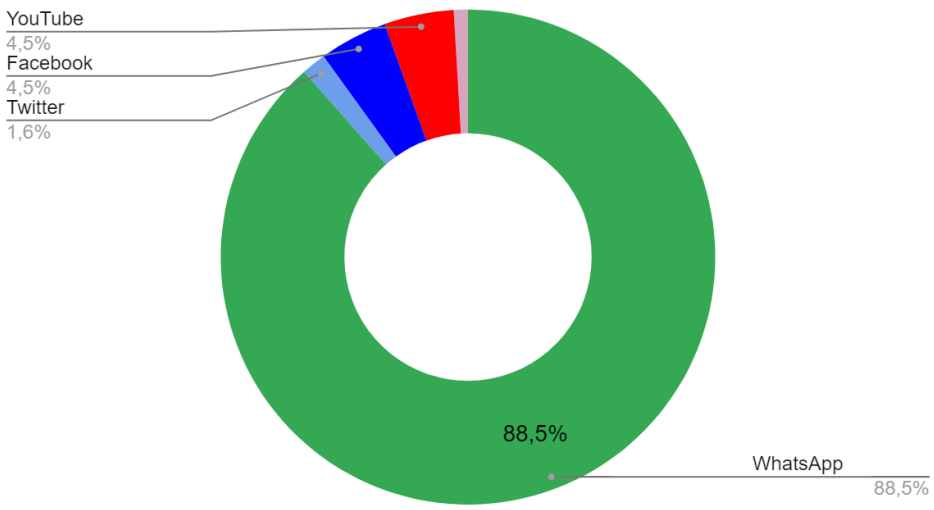
Fonte: o autor.

4.3.1 Mensagens em outras redes sociais

Durante o processo rotulação manual, onde o texto de cada mensagem foi buscado na *Web*, observou-se que algumas mensagens foram reproduzidas em outras redes sociais, como YouTube, Twitter e Facebook. Do total de 5.284 mensagens, 610 (11,5%) foram encontradas em outras redes sociais. Dessas, 85 foram encontradas no Twitter, 236 no Facebook, 240 no YouTube e 49 em páginas diversas na *Web*, como blogs e portais de notícias. A Figura 7 ilustra essas proporções.

A grande maioria das mensagens eram exclusivas do ambiente do WhatsApp, o que reforça nossa motivação inicial de que as particularidade do WhatsApp fazem com que o conteúdo que circula nele seja único em relação ao conteúdo que circula em outras redes. Nas mensagens exclusivas de WhatsApp, percebeu-se o uso de formatação específica dessa plataforma, como o uso de asteriscos para tornar o texto negrito, e a utilização intensa de *emojis*, em quantidade e variedade, quando comparado com mensagens de outras redes.

Figura 7 – Proporção de mensagens virais encontradas em outras redes sociais.



Fonte: o autor.

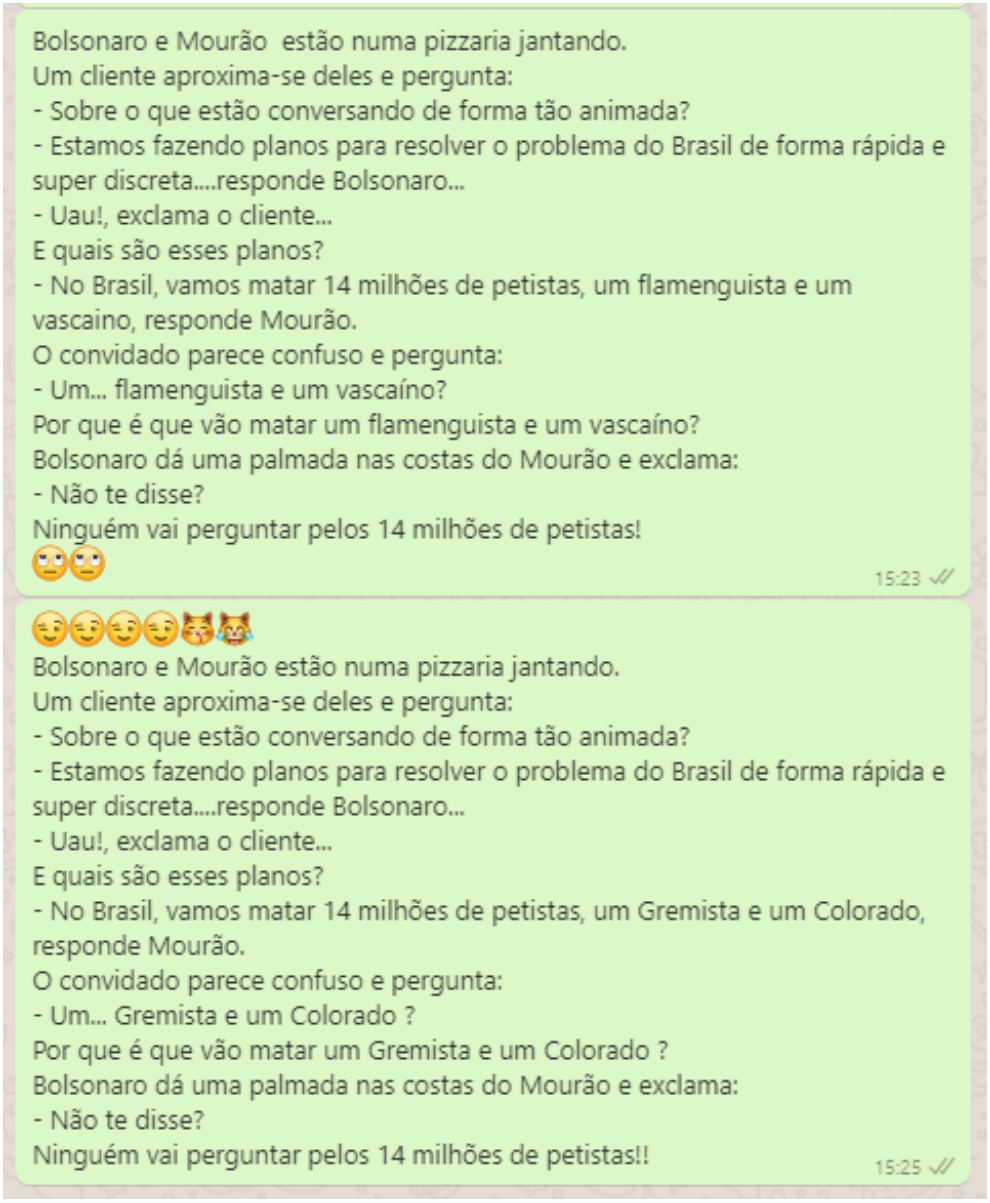
4.3.2 Atribuição automática de rótulos

Após a rotulação manual, os rótulos do subconjunto de mensagens únicas devem ser atribuídos ao conjunto completo de mensagens, enquanto que as mensagens não vistas no processo permanecem não-rotuladas, recebendo o valor -1. Contudo, observou-se de forma empírica que no subconjunto rotulado haviam mensagens muito similares, com a mesma semântica, mas com ligeiras alterações entre as palavras. Para exemplificar, a Figura 8 ilustra duas mensagens reais presentes no conjunto de dados onde essa variação ocorre em um mesmo conteúdo. Para melhor visualização, foi utilizada a renderização do WhatsApp, ilustrando como as mensagens são visualizadas em seu contexto original.

Observa-se que as mensagens são essencialmente a mesma, com a diferença de alguns *emojis* e um ponto de exclamação a mais. Por não serem idênticas, foram consideradas mensagens diferentes durante a rotulação. No caso da segunda mensagem deste exemplo, ela sequer é considerada uma mensagem viral, pois aparece com esses caracteres somente uma vez no conjunto de dados.

Para atribuir os rótulos criados ao conjunto de dados completo, considerando também as mensagens não idênticas com pequenas variações, utilizamos o Algoritmo 1. A lógica deste algoritmo é calcular a matriz *TF-IDF* para conjunto de dados completo D e para o subconjunto rotulado L. Para cada mensagem d_i do conjunto não-rotulado, é calculado a similaridade cosseno entre o seu vetor e os vetores l_i do subconjunto rotulado. O vetor de maior similaridade $l_{i^\star}$ é

Figura 8 – Exemplo de mensagens com a mesma informação, mas escritas de forma ligeiramente diferente.



Fonte: o autor.

recuperado.

Se a maior similaridade for maior que o limiar $t = 0,9$, atribui-se o rótulo de $l_{i^\star}$ para d_i. Caso contrário, d_i recebe o rótulo -1, que representa uma mensagem não-rotulada. Dessa forma, as mensagens que não entraram no processo de rotulação manual por serem ligeiramente diferentes das mensagens virais também recebem rótulos, adicionando assim uma variedade maior ao conjunto de dados rotulados.

Algoritmo 1: Atribuição de rótulos

Conjunto de mensagens não-rotuladas D, subconjunto de mensagens únicas rotuladas L

Conjunto D com mensagens rotuladas $t = 0,9$;

D_{tfidf} = matriz *TF-IDF* de D;

L_{tfidf} = matriz *TF-IDF* de L, considerando todos os documentos em D;

for d_i *em* D_{tfidf} **do**

  for l_j *em* L_{tfidf} **do**

    $\text{sim}_{i,j} = \cos(\theta_{d_i,l_j}) = \dfrac{\langle d_i, l_j \rangle}{||d_i|| \cdot ||l_j||}$.

  end

  $i^\star = \text{argmax}(\text{sim}_{i,j})$;

  $sim_{i^\star} = \max(\text{sim}_{i,j})$;

  if $sim_{i^\star} < t$ **then**

  end

  Atribui o rótulo -1 (não-rotulado) para d_i **else**

    Atribui o rótulo de $l_{i^\star}$ para d_i;

  end

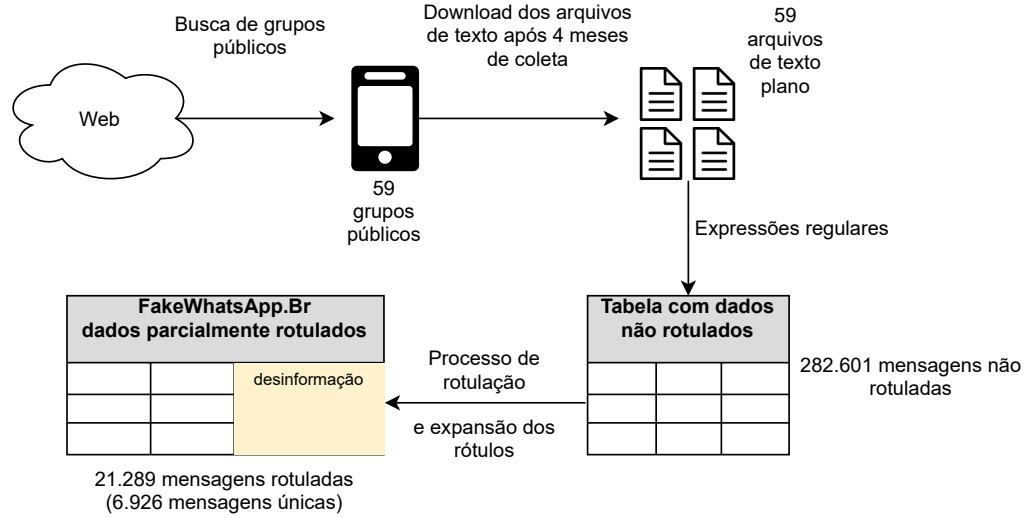
end

Após a rotulação automática, foi realizada uma inspeção manual da qualidade das mensagens rotuladas automaticamente e percebeu-se a existência de incoerências devido a mensagens compostas por repetição de palavras, como por exemplo "*#B17*" ou *emojis* repetidos. Assim, para remover esses ruídos, que não contém informação textual relevante, filtramos as mensagens com menos de 10 palavras únicas. Feito esse procedimento, observamos que todas as mensagens rotuladas automaticamente estavam coerentes.

Após esse processo, um total de 21.289 mensagens, incluindo repetições, foram rotuladas, representando 7,5% do total de mensagens. Destas, 11.412 (54%) mensagens foram rotuladas como desinformação, enquanto as 9.877 (46%) restantes foram rotuladas como não-desinformação. É importante ressaltar que, das mensagens rotuladas, muitas eram replicadas, sendo apenas 6.926 mensagens únicas. Todavia, após o processo de atribuição de rótulos por similaridade, houve um aumento de mensagens rotuladas de 1.642 acima das mensagens inicialmente rotuladas manualmente, representando um aumento de 31%. Isso indica que as mensagens com pequenas variações constituem uma parcela significativa das mensagens compartilhadas.

Após a atribuição dos rótulos, o processo de criação do FakeWhatsApp.Br está completo. A Figura 9 ilustra todas as etapas discutidas até o momento de forma resumida.

Figura 9 – Processo de criação do FakeWhatsApp.Br.



Fonte: o autor.

4.4 Análise exploratória

Nesta seção realizaremos uma análise das principais características do FakeWhatsApp.Br. A Tabela 4 descreve as características gerais dos dados, com quantidades de grupos, usuários, mensagens e período de coleta, além da quantidade de mensagens virais, mensagens de mídia e mensagens efetivamente rotuladas.

Observa-se nessa tabela uma proporção relativamente alta de mensagens contendo mídia, com cerca de 44,7%. Em outros trabalhos que analisavam conjuntos de dados de grupos públicos de WhatsApp, como de Garimella e Tyson (2018), observou-se uma proporção bem menor de mensagens de mídia. Entendemos que essa alta taxa deve-se aos grupos coletados serem grupos de campanha política, onde o propósito é principalmente fazer divulgação e propaganda ao invés de conversação. Essa alta taxa mostra a importância da análise dos arquivos de mídia no contexto da desinformação do WhatsApp. Embora esse trabalho se limite a analisar a desinformação textual, chamamos a atenção para o desafio de analisar a desinformação em outras mídias. Nota-se também uma alta taxa de mensagens virais, correspondendo a mais de 13% do total das mensagens de texto, mostrando que uma fração considerável das mensagens são replicadas.

Tabela 4 – Descrição geral do conjunto de dados.

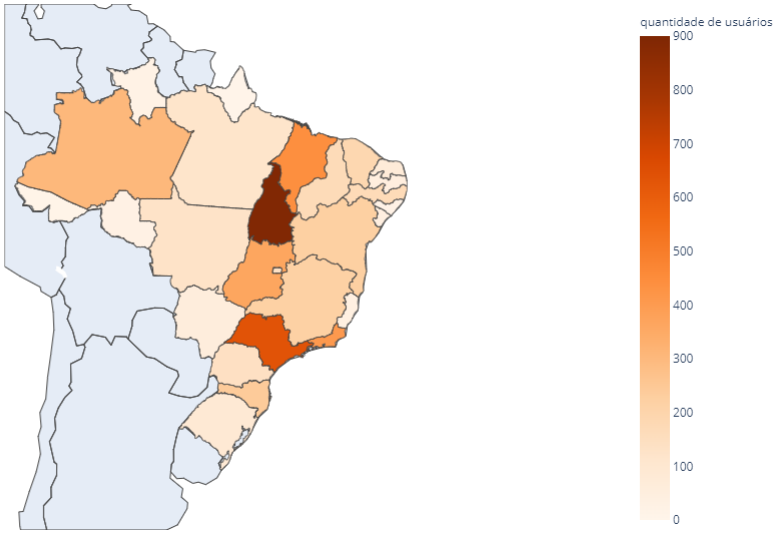
Variável	Valor
Período de coleta	02 de Julho até 29 de Outubro de 2018 (≈ 4 meses)
Grupos	59
Usuários	5.364
Mensagens	282.601
Mensagens com mídia	126.349
Mensagens virais	20.872
Mensagens únicas rotuladas	6.926

Fonte: o autor.

4.4.1 Análise geoespacial e temporal

Podemos analisar a distribuição geográfica dos usuários do conjunto de dados, uma vez que a partir do DDI e DDD pode-se descobrir o país e o estado (quando o país for o Brasil) do usuário. Dessa forma, observou-se que 128 usuários possuem um DDI estrangeiro. Embora correspondam a 2,4% do total de usuários, esse grupo foi responsável por 4,5% das mensagens, mostrando que, em média, foram mais ativos que usuários com DDI do Brasil, o que desperta atenção. No caso dos usuários com DDI brasileiro, podemos descobrir a sua unidade federativa através do DDD e quantificar a quantidade de usuários por estado. A Figura 10 ilustra a densidade de usuários no conjunto de dados por cada estado em um mapa temático. Pode-se observar que, embora hajam estados com maior representatividade, em destaque Tocantins e São Paulo, existem usuários de todos os estados, provendo uma representação espacial razoável.

Figura 10 – Quantidade de usuários por estado.

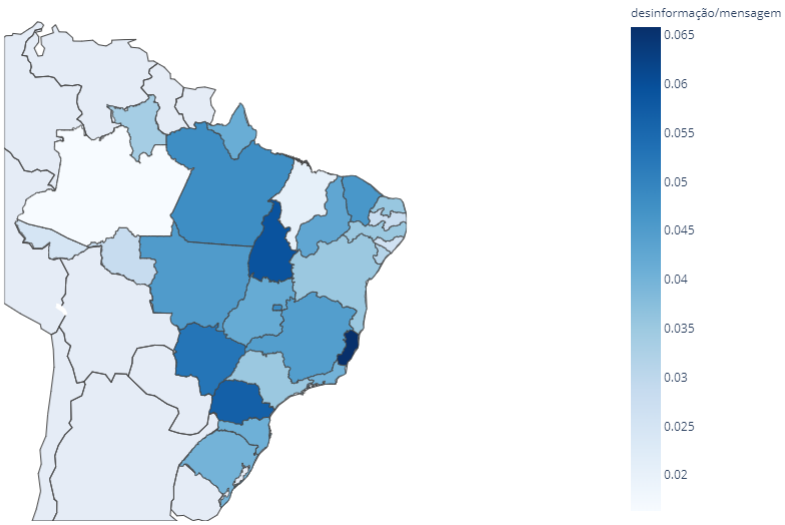


Fonte: o autor.

De forma complementar, podemos utilizar os rótulos atribuídos para analisar a quantidade relativa de desinformação em cada estado, normalizada pela quantidade total de

mensagens enviadas por cada estado. Essa análise é ilustrada na Figura 11. Destaca-se os estados do Espírito Santo, Paraná e Mato Grosso do Sul, que, mesmo possuindo uma quantidade baixa de usuários no conjunto de dados, possuem um valor alto de desinformação relativa. Uma possível causa seria a presença de desinformadores nesses estados, que compartilham uma quantidade de desinformação acima da média.

Figura 11 – Quantidade de desinformação por quantidade de mensagens em cada estado.



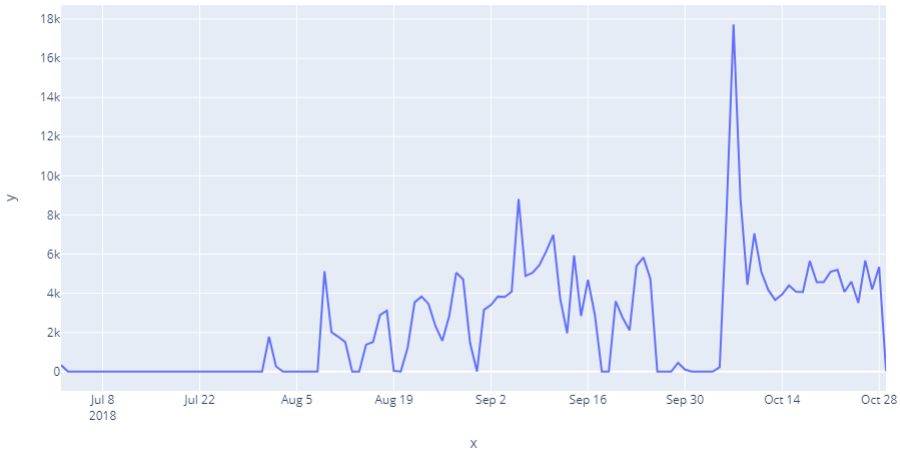
Fonte: o autor.

Com a variável de data/hora, podemos observar a variação da quantidade de mensagens enviadas ao longo do tempo. A Figura 12 ilustra a série temporal de mensagens por dia. Existem picos significativos de atividade em períodos que correspondem aos dias de votação no primeiro e segundo turno. Observa-se que em alguns dias não houve registro de mensagens. Acreditamos que isso ocorreu devido a uma falha no procedimento de coleta dos dados brutos, realizado de forma manual. Percebe-se que há lacunas nos dados, especialmente em alguns dias entre setembro e outubro, comprometendo parcialmente a representatividade da dimensão temporal. Esse problema de coleta passa a ser mitigado em trabalhos futuros devido à plataforma Farol Digital, desenvolvida no trabalho de Sá *et al.* (2021).

4.4.2 Análise das mensagens rotuladas

Nesta subseção são analisados os dados rotulados como contendo desinformação ou não contendo desinformação, buscando identificar padrões que diferenciem as duas classes. Para exemplificar o tipo de mensagem encontrada em cada classe, as Figuras 13 e 14 apresentam

Figura 12 – Quantidade de mensagens por dia ao longo da coleta. Observa-se lacunas na coleta. Destacam-se um pico de atividade durante o primeiro turno (7 de outubro).



Fonte: o autor.

amostras selecionadas aleatoriamente de cada classe, utilizando a renderização do WhatsApp. Percebe-se a grande variedade de mensagens em ambos os grupos. No grupo de desinformação, as mensagens apresentadas não caracterizam-se como *Fake News* por não utilizarem o estilo de escrita jornalístico, sendo mais próximas de rumores. Observa-se ainda mensagens curtas que fazem referência a arquivos de mídia. No caso do grupo de mensagens rotuladas como não-desinformação, as mensagens incluem humor, propaganda política e mesmo oração. Para fins de comparação, a Figura 15 ilustra mensagens não rotuladas (não-virais), que provém de diálogos entre os usuários dos grupos, ao invés de encaminhamento de mensagens.

4.4.2.1 Distribuições de variáveis linguísticas e de propagação

Sobre o conteúdo das mensagens, podemos analisar a quantidade de caracteres e a quantidade de palavras únicas no texto. Já em termos de propagação, podemos observar a quantidade de vezes que as mensagens foram compartilhadas. A distribuição dessas variáveis para cada grupo fornece informações relevantes sobre os padrões neles presentes. A Tabela 5 mostra a média, desvio padrão, mínimo, primeiro quartil, mediana, terceiro quartil e valor máximo para essas variáveis mencionadas.

Percebe-se que as distribuições em todas as variáveis diferem ligeiramente, sendo as mensagens rotuladas como desinformação formadas, em geral, por textos mais longos e mais compartilhados. A diferença entre os tamanhos de textos pode ser um problema para

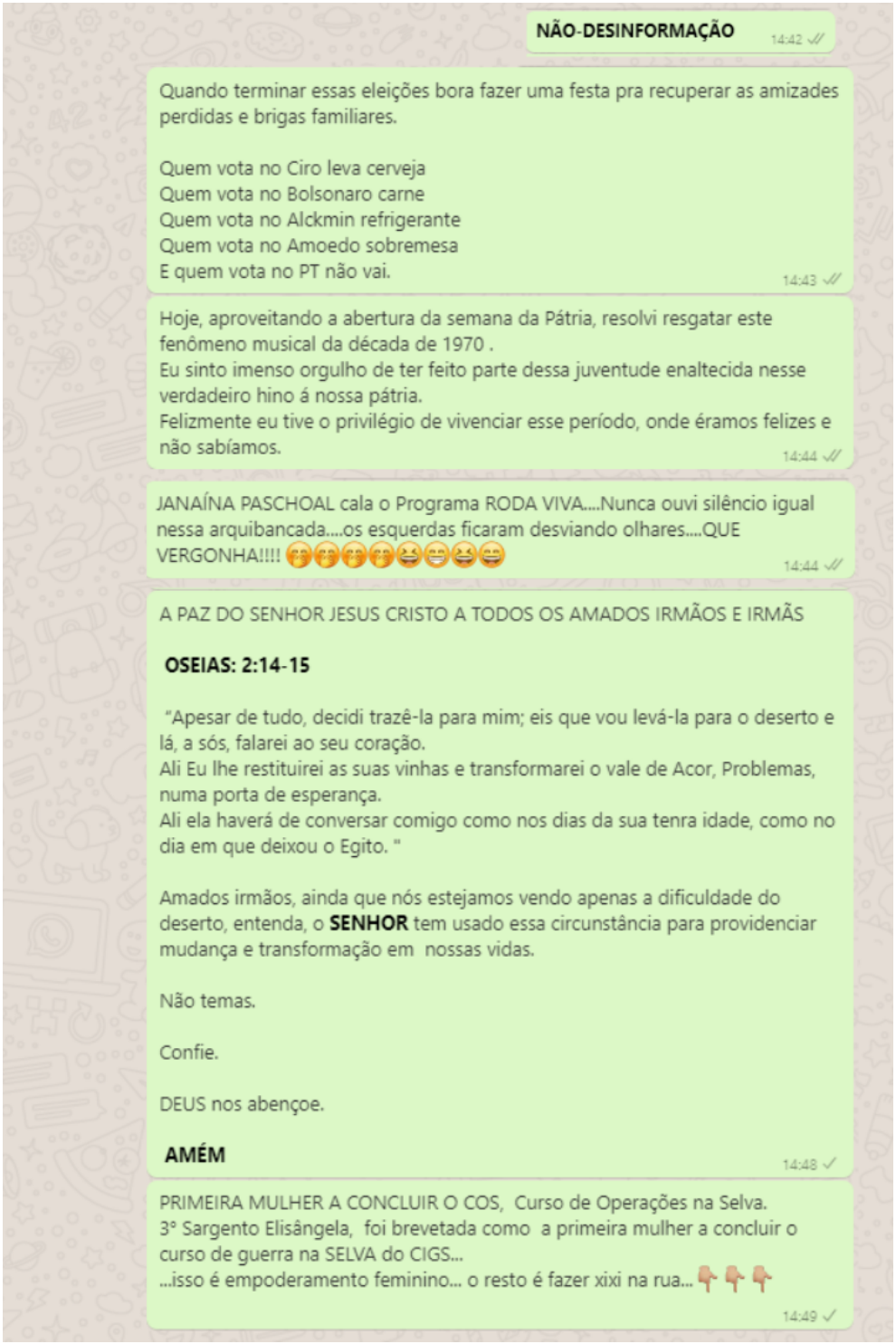
Figura 13 – Amostras das mensagens rotuladas como desinformação

Fonte: o autor.

um classificador baseado em conteúdo, que pode tornar-se enviesado pelo tamanho dos textos, classificando erroneamente textos curtos como não-desinformação. Além disso, conforme esperado, nota-se que a desinformação foi mais compartilhada, o que está de acordo com o afirmado por Vosoughi *et al.* (2018), e reforça a decisão desse trabalho de rotular as mensagens virais. Para melhor compreensão dos dados, a Figura 16 ilustra as distribuições, estimadas através do método *Kernel Density Estimation* / estimativa de densidade kernel (KDE) (DAVIS *et al.*, 2011). Observa-se que ambas classes possuem curvas semelhantes, com alta densidade de probabilidade em valores mais baixos e uma longa cauda. Contudo, as distribuições para não-desinformação possuem um pico mais alto em valores menores, indicando que a classe de desinformação possui maior probabilidade na extremidade da cauda.

Caracteres e palavras únicas são variáveis que descrevem a dimensão do conteúdo

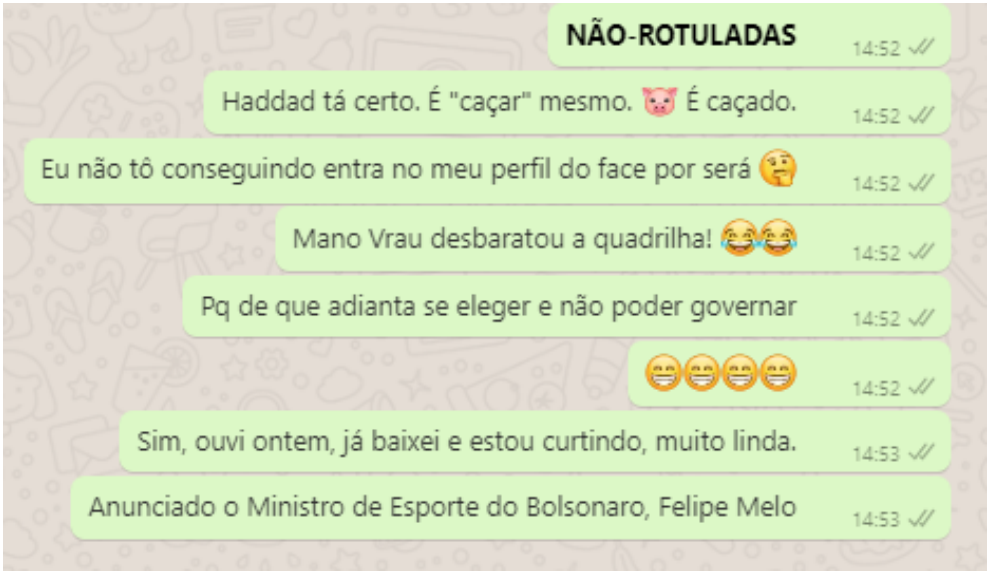
Figura 14 – Amostras das mensagens rotuladas como não-desinformação.



Fonte: o autor.

textual das mensagens rotuladas, enquanto a quantidade de compartilhamentos descrevem dados de propagação.

Figura 15 – Amostras das mensagens não-rotuladas.



Fonte: o autor.

Tabela 5 – Distribuições das quantidades de caracteres, palavras únicas e número de compartilhamentos para os dois grupos de dados rotulados. Em média, as mensagens rotuladas como desinformação são textos mais longos, com mais palavras únicas

	caracteres		**palavras únicas**		**compartilhamentos**	
	não-desinformação	**desinformação**	**não-desinformação**	**desinformação**	**não-desinformação**	**desinformação**
média	468.7	772.2	50.4	83.6	7.6	13.4
desvio-padrão	903.5	1129.2	71.7	102.3	12.1	18.3
mínimo	38.0	48.0	10.0	10.0	1.0	1.0
Q1	116.0	155.0	16.0	21.0	2.0	2.0
mediana	210.0	287.0	28.0	39.0	3.0	6.0
Q3	440.0	963.0	58.0	112.0	7.0	16.0
máximo	16257.0	13484.0	1194.0	1047.0	86.0	91.0

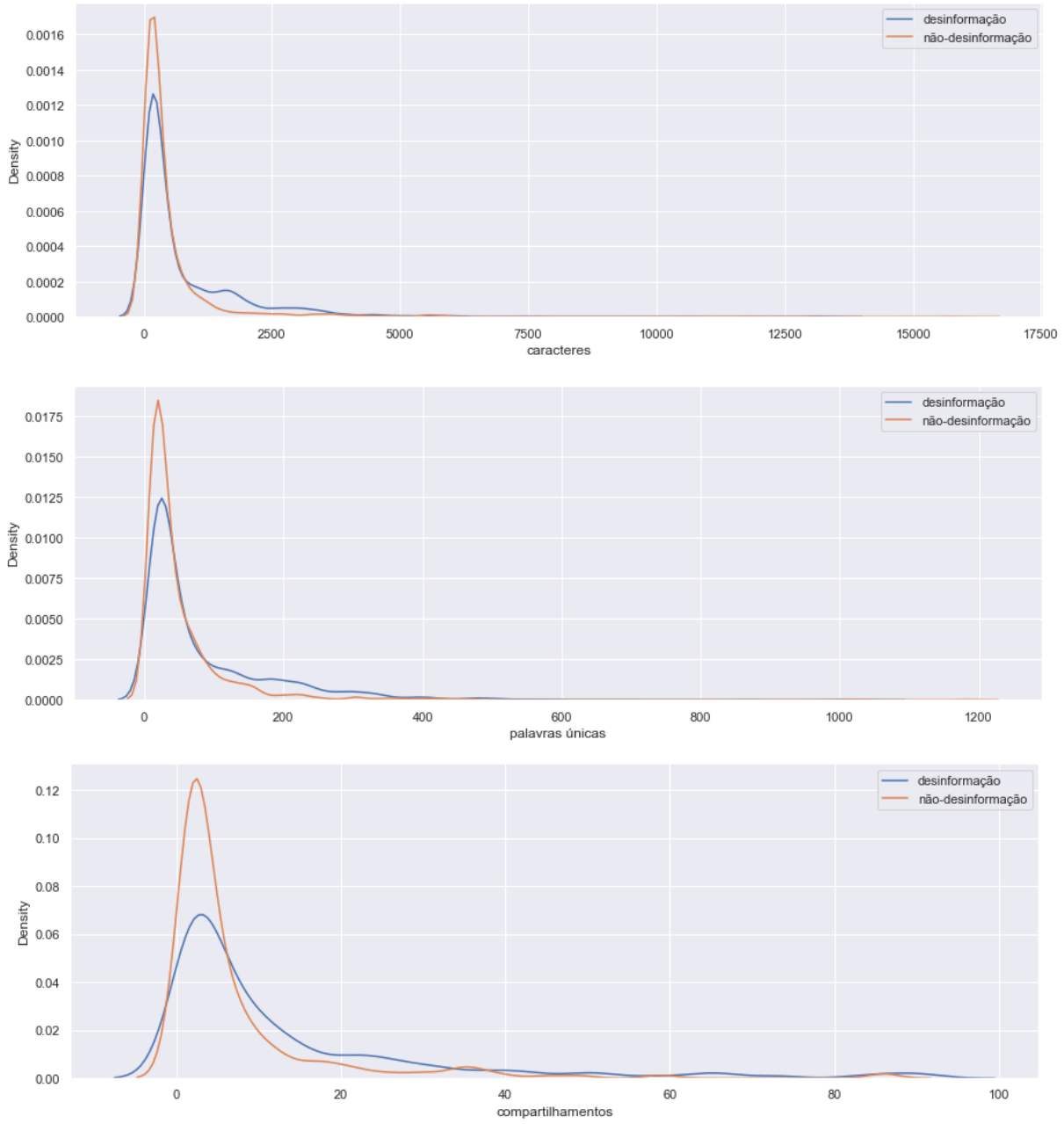
Fonte: o autor.

4.4.2.2 Termos mais frequentes e termos mais representativos

Pode-se aprofundar a análise do conteúdo contabilizando os termos mais frequentes em cada uma das classes, removendo termos irrelevantes como artigos e pronomes, as chamadas palavras de parada (*stop words*), e realizando a lematização do texto. As Figuras 17 e 18 ilustram os 15 bigramas mais frequentes em cada classe, incluindo *emojis*, representados pelo código parcial em português com o caractere sublinhado. Observa-se que os *emojis* são abundantes em ambas as classes, ocupando as primeiras colocações.

Percebe-se que existem bigramas que são frequentes nas duas classes, como "*dorso_da dorso_da*" (dorso da mão com dedo indicador apontando para baixo, utilizado para indicar que uma mídia ou link acompanha uma mensagem), "*jair bolsonaro*" e "*rosto_chorando rosto_chorando*". Dado o contexto do conjunto de dados, essa interseção é esperada. Esses termos podem não ser úteis para discriminar as classes. De forma complementar a análise de

Figura 16 – Distribuições das quantidades de caracteres, palavras únicas e compartilhamentos para as classes de desinformação e não-desinformação.



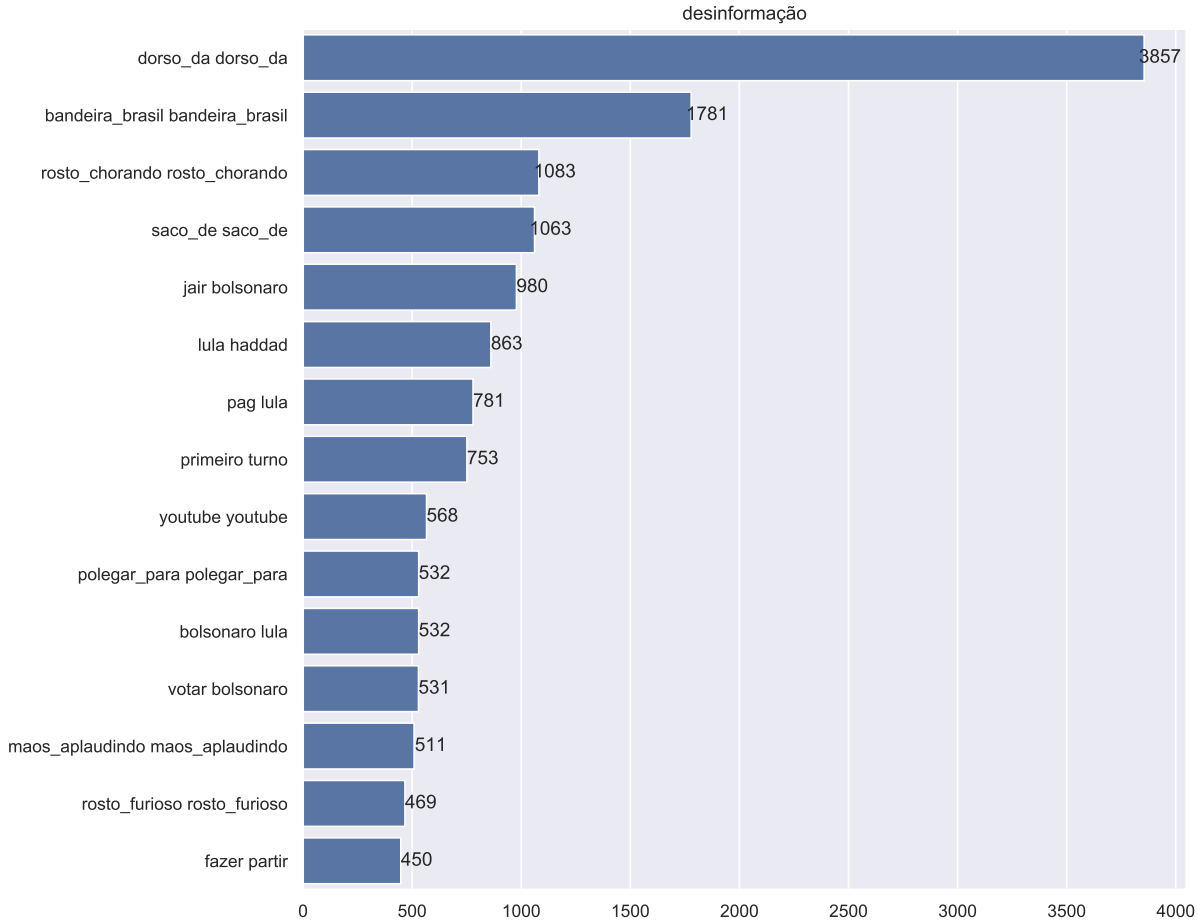
Fonte: o autor.

frequência dos termos dada a classe, podemos também utilizar a regra de Bayes para estimar a probabilidade de uma classe dado um termo. Assim, podemos encontrar os termos mais representativos de cada classe. Para isso, realizamos a modelagem formalizada na Equação 4.1:

$$P(C|t) = \frac{P(t|C)P(C)}{P(t)} \tag{4.1}$$

onde C é a classe (desinformação ou não-desinformação) e t é um termo (ou n-grama). $P(C)$ é a

Figura 17 – 15 bigramas mais frequentes da classe de desinformação.

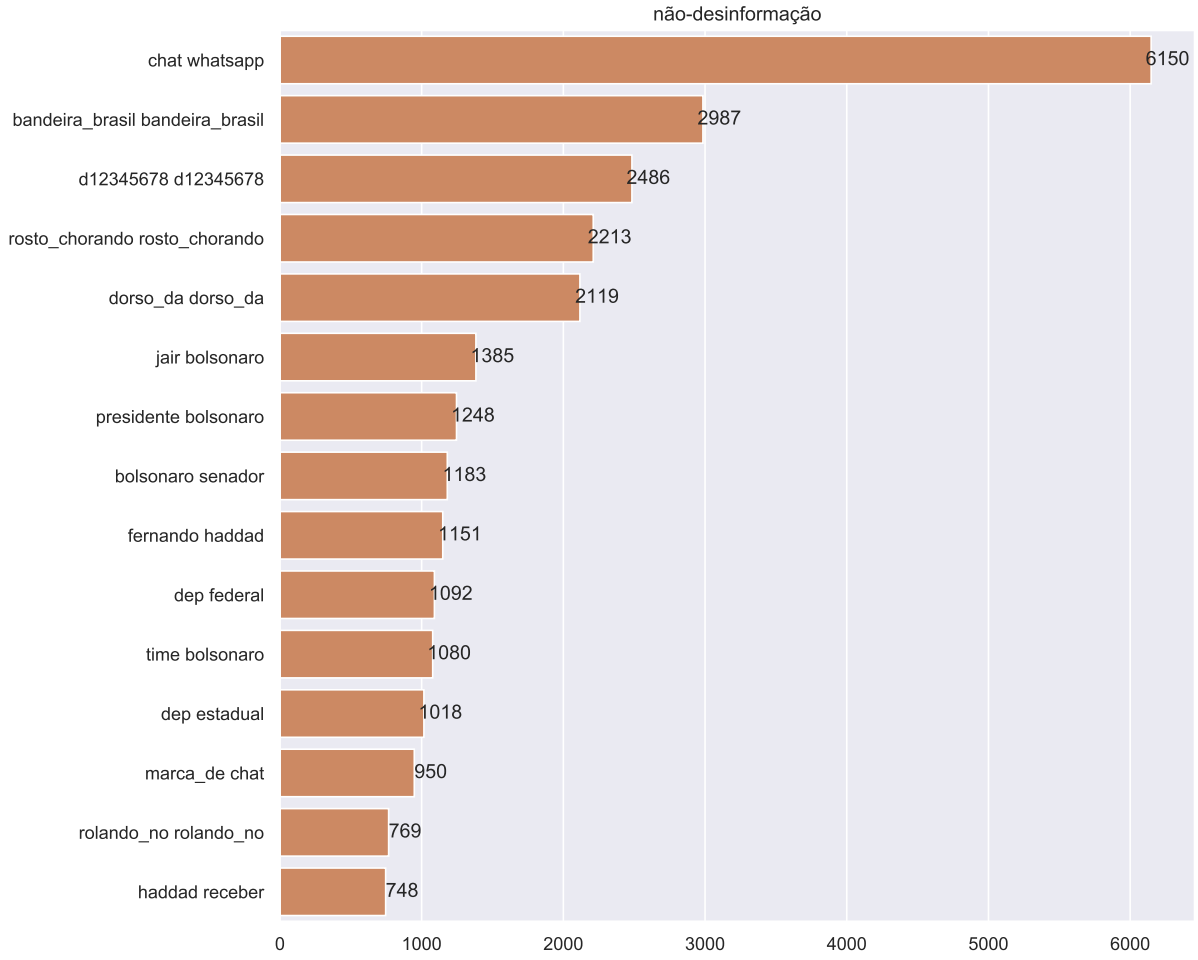


Fonte: o autor.

probabilidade a priori de um termo sorteado ao acaso no conjunto de dados pertencer a classe C, calculada como a quantidade total de termos da classe C dividido pela quantidade de termos totais. $P(t|C)$ é a verossimilhança, calculada como a frequência de t na classe C dividida pelo total de termos na classe C. $P(t)$, a verossimilhança marginal, é a probabilidade do termo t ocorrer, sorteando um termo aleatoriamente no conjunto de dados, calculada como a frequência de t dividida pelo total de termos. A partir disso, calculamos a posteriori $P(C|t)$. Observa-se que essa modelagem não leva em consideração as distribuições conjuntas, ou seja, a relação de t com outros termos presentes no texto. Ainda assim, os termos com valores mais altos da posteriori são atributos importantes para a discriminar as classes.

Calculando esse valor para os bigramas no conjunto de dados e ordenando do maior para o menor, obtemos os bigramas mais relevantes de cada classe, apresentados na Tabela 6. Todas as probabilidades calculadas estavam próximas de 99%, portanto apresentou-se somente as suas colocações. Observa-se que nessa análise os termos mais relevantes variam consideravelmente em cada classe, sem nenhuma interseção. Além disso, a probabilidade $p(t)$

Figura 18 – 15 bigramas mais frequentes da classe de não-desinformação.



Fonte: o autor.

de todos os termos apresentados é inferior a 0,0009, indicando que são termos pouco frequentes mas que discriminam bem as classes. Isso fortalece a intuição de que uma abordagem baseada em conteúdo pode separar satisfatoriamente as classes, dado às diferenças linguísticas entre elas e a presença de termos-chave para fazer essa separação em certos casos. Observa-se também que *emojis* estão inclusos nesses termos-chave, de modo que não devem ser ignorados.

Percebe-se que os primeiros colocados da classe de não-desinformação são referentes a propagandas políticas ("*dep federal*", "*atenção eleitor*", "*bolsonaro senador*") ou a propagandas de grupos de WhatsApp ("*acesse link*"). Os quatro últimos colocados da classe de não-desinformação são todos relacionados a uma mesma mensagem que circulou após o primeiro turno das eleições, cujo conteúdo era um pedido para eleitores não dirigissem ataques ao povo nordestino. Já na classe de desinformação, os termos remetem a ataques contra adversários políticos ("*bolsonaro lula*", "*ciro sobrar*") e acusações de corrupção ("*saco_de_dinheiro saco_de_dinheiro*", "*crime politico*", "*país onde*").

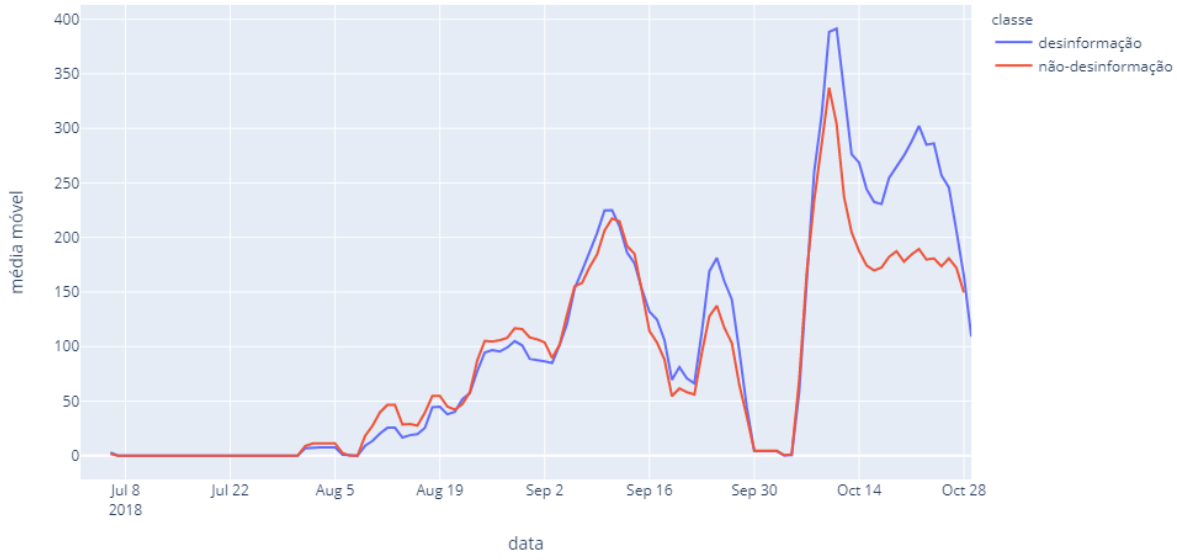
Tabela 6 – Termos com maior probabilidade à posteriori de cada classe, calculados com a regra de Bayes. As probabilidades calculadas estão todas próximas de 99%.

Colocação	Desinformação	Não-desinformação
1	saco_de_dinheiro saco_de_dinheiro	bolsonaro senador
2	ciro sobrar	dep federal
3	crime politico	haddad votos
4	politico organizado	marca_de_seleção marca_de_seleção
5	país onde	atenção eleitor
6	poder acontecer	bolsonaro voto
7	bolsonaro lula	compartilhar nada
8	dizer próprio	mensagens positivas
9	do mostrar	ofensivo nordeste
10	camisa candidato	nada ofensivo

4.4.2.3 Análise temporal

Aprofundando a análise de propagação, pode-se visualizar a quantidade de mensagens por dia de cada classe na forma de uma série temporal. A Figura 19 ilustra as curvas de média móvel de mensagens por dia, com uma janela de 5 dias. Observa-se que as curvas são fortemente correlacionadas, mas os picos são mais acentuados em quantidade bruta para a curva de desinformação.

Figura 19 – Média móvel de mensagens por dia com janela de 5 dias para cada classe.



Fonte: o autor.

4.5 Análise dos usuários

Cada linha do conjunto de dados representa uma mensagem enviada por um dado usuário em um grupo. Porém, a partir dessa modelagem podemos criar um conjunto de dados auxiliar de 5364 linhas, onde cada linha representa um usuário, com atributos calculados a partir do conjunto original e que descrevem seu comportamento global nos grupos e no período coletado. Como discutido anteriormente, obter informações sobre os usuários pode ser uma abordagem poderosa para detecção de desinformação, além de ser necessária para a detecção de desinformadores.

O desafio no qual esbarramos é que não foram encontrados na literatura propostas de atributos para descrever usuários no ambiente do WhatsApp. Dessa forma, uma das contribuições desse trabalho é a proposta de um conjunto de atributos e a sua exploração nos problemas de detecção de desinformação e desinformadores. Alguns desses atributos são inspirados em atributos de usuário extraídos do Twitter, enquanto outros são novos e específicos para o ambiente do WhatsApp. Além do identificador de cada usuário, categorizamos os seus atributos em dois grupos: atributos de atividade e atributos de rede.

É importante salientar que as informações contidas nos atributos de usuário são limitadas pelo recorte de grupos que podemos observar com a nossa coleta. Por exemplo, um usuário que em nossos dados possui pouca atividade, pode ter sido muito ativo em outros grupos que não tivemos acesso.

4.5.1 Atributos de atividade

Como o nome indica, atributos de atividade quantificam as ações tomadas pelos usuários nos grupos observados. Dividimos ainda em 3 subgrupos chamados de contagem, proporções e de atividade temporal. Os atributos de contagem são a contabilização bruta da atividade dos usuários. Os atributos desse grupo são descritos na lista a seguir:

- **Grupos**: quantidade de grupos que o usuário participa dentre os grupos coletados;
- **Número de mensagens**: quantidade total de mensagens enviadas pelo usuário;
- **Textos**: quantidade de mensagens de texto enviadas pelo usuário;
- **Mídia**: quantidade de mensagens de mídia enviadas pelo usuário;
- **Virais**: quantidade de mensagens virais enviadas pelo usuário;
- **Mensagens repetidas**: quantidade de mensagens duplicadas enviadas pelo usuário;

Tabela 7 – Medidas estatísticas das distribuições dos atributos de atividade do tipo contagem

	Grupos	Total de mensagens	Textos	Mídia	Virais	Mensagens repetidas	Desinformação
média	1.16	52.68	29.13	23.55	3.89	2.57	2.13
desvio-padrão	0.65	138.06	89.74	63.19	15.01	16.26	7.33
mínimo	1.00	1.00	0.00	0.00	0.00	0.00	0.00
Q1	1.00	3.00	2.00	1.00	0.00	0.00	0.00
mediana	1.00	13.00	6.00	4.00	0.00	0.00	0.00
Q3	1.00	45.00	23.00	19.00	2.00	1.00	1.00
máximo	11.00	4396.00	3742.00	1360.00	564.00	609.00	147.00

Fonte: o autor.

– **Desinformação**: quantidade de mensagens rotuladas como desinformação enviadas pelo usuário.

Valores altos nesses atributos, que estão próximos do valor máximo, indicam que o usuário foi muito ativo no escopo observado. Além disso, valores altos da quantidade de mídia, virais e mensagens repetidas indicam o comportamento propagandista. Valores altos de compartilhamento de desinformação indicam o comportamento de desinformador. Importante ressaltar que o atributo da quantidade desinformação só pôde ser obtido porque os dados passaram por um processo prévio de rotulação manual. Porém, os outros atributos poderiam ser obtidos de um conjunto de dados não rotulado.

A Tabela 7 mostra medias que descrevem as distribuições desses dados. Percebe-se que a maioria dos usuários não é altamente ativo. Todas as distribuições tem alta concentração em valores mais baixos, porém com um alto desvio padrão e um alto valor máximo, sendo distribuições de cauda longa. Especialmente no caso do atributo de desinformação, percebe-se que espalhar desinformação não é um comportamento normal entre os usuários, onde 75% dos usuários enviaram menos de uma mensagem contendo desinformação. No Capítulo 6 a análise das distribuições desses atributos e outros atributos de usuários será retomada com maior profundidade para caracterizar alguns tipos específicos de usuários.

Além da quantidade total, pode ser interessante observar as proporções, uma vez que alguns usuários podem não ter enviado uma grande quantidade de mensagens, mas possuem padrões identificáveis nas relações entre parte e todo de cada tipo de mensagem. Os atributos desse grupo são descritos na lista a seguir e suas distribuições são indicadas na Tabela 8:

– **Proporção de textos**: razão da quantidade de mensagens de texto pelo total de mensagens enviadas pelo usuário;
– **Proporção de mídia**: razão da quantidade de mensagens de mídia pelo total de mensagens enviadas pelo usuário. Por definição, é o complemento da proporção de textos;

Tabela 8 – Medidas estatísticas das distribuições dos atributos de atividade do tipo proporção

	Proporção				
	Textos	**Mídia**	**Virais**	**Mensagens repetidas**	**Desinformação**
média	0.567	0.433	0.069	0.039	0.041
desvio-padrão	0.317	0.317	0.140	0.107	0.107
mínimo	0.000	0.000	0.000	0.000	0.000
Q1	0.333	0.158	0.000	0.000	0.000
mediana	0.571	0.429	0.000	0.000	0.000
Q3	0.842	0.667	0.085	0.014	0.042
máximo	1.000	1.000	1.000	0.941	1.000

Fonte: o autor.

- **Proporção de virais**: razão da quantidade de mensagens virais pelo total de mensagens enviadas pelo usuário;
- **Proporção de mensagens repetidas**: razão da quantidade de mensagens virais pelo total de mensagens enviadas pelo usuário;
- **Proporção de desinformação**: razão da quantidade de mensagens rotuladas como desinformação pelo total de mensagens enviadas pelo usuário.

Observamos empiricamente que alguns desses atributos são especialmente indicativos de atividade incomum, como uma proporção alta de mensagens de mídia ou virais. Esses atributos destacam-se quando usuários não foram muito ativos. Para ilustrar esse ponto com um exemplo real, um usuário que compartilhou 17 mensagens virais pode não ser relevante quando analisado somente a quantidade bruta. Porém, foi observado que 100% das mensagens deste usuário correspondem a mensagens virais, o que é um comportamento que não corresponde ao de um usuário regular, utilizando o aplicativo para conversas, e sim de um propagandista, apenas repassando conteúdo. De forma análoga ao que ocorre nos atributos de contagem, os atributos de proporção de desinformação só são possíveis de se observar devido ao processo de rotulação.

Por último, analisamos também o comportamento temporal dos usuários, através da variável de mensagens por dia. A atividade dos usuários varia ao longo do tempo, portanto tomamos como atributos de atividade temporal as estatísticas descritivas da quantidade de mensagens por dia:

- **Dias ativos**: quantidade de dias em que o usuário enviou mensagens;
- **Média de mensagens por dia**;
- **Desvio padrão de mensagens por dia**;
- **Mediana de mensagens por dia**;
- **Máximo mensagens por dia**;

A Tabela 9 descreve as distribuições desses atributos. Atributos temporais podem ser

Tabela 9 – Medidas estatísticas das distribuições dos atributos de atividade do tipo temporal

	Dias ativo	Média diária de mensagens	Desvio padrão de mensagens diárias	Mediana das mensagens diárias	Máximo de mensagens diárias
média	33.3	2.1	2.8	1.4	10.3
desvio-padrão	30.0	5.0	4.7	4.8	16.9
mínimo	1.0	0.0	0.0	0.0	1.0
Q1	3.0	0.4	0.7	0.0	2.0
mediana	28.0	1.0	1.4	0.0	4.0
Q3	59.0	2.0	3.2	1.0	12.0
máximo	120.0	149.0	148.5	149.0	294.0

Fonte: o autor.

extraídos sem a necessidade de rótulos atribuídos manualmente e descrevem o comportamento do usuário ao longo do tempo em níveis de atividade. Atividades suspeitas incluem o usuário que tem picos de atividade muito acentuados alternados com dias sem nenhuma atividade. Uma baixa média, com alto desvio padrão e alto máximo pode ser um forte indicador.

4.5.2 Atributos de rede

Como visto no Capítulo 2, modelar as relações entre os usuários na forma de uma rede, ou grafo, pode fornecer informações relevantes sobre padrões de desinformação. Porém, diferente de redes sociais como Twitter ou Facebook, onde existem conexões bem definidas entre os usuários pela relação de seguir (Twitter) ou de amizade (Facebook), no WhatsApp essas conexões não são explícitas.

Assim, propomos uma modelagem das relações entre usuários do WhatsApp na forma de grafos direcionados e valorados, considerando o envio de mensagens em grupos. Nessa modelagem, cada nó representa um usuário e podemos considerar um grafo para cada tipo de mensagem: mensagem em geral, mensagem viral e mensagem com desinformação.

Considerando o grafo de mensagens gerais, onde cada nó representa um usuário, existe uma aresta direcionada entre o usuário i e o usuário j se o usuário i enviou uma mensagem para um grupo do qual o usuário j faz parte. O peso dessa aresta é a quantidade de mensagens enviadas pelo usuário i para aquele grupo. Um raciocínio análogo pode ser aplicado para criar um grafo apenas de mensagens virais: existe uma aresta direcionada entre o usuário i e o usuário j se o usuário i enviou uma mensagem viral para um grupo do qual o usuário j faz parte e o peso dessa aresta é quantidade de mensagens virais enviadas pelo usuário i para aquele grupo. O mesmo vale para o grafo de desinformação.

Percebe-se que nos três grafos, a quantidade de nós é a mesma, variando a quantidade de arestas. A contabilização da quantidade de arestas de cada tipo é apresentada na Tabela 10. Percebe-se que são grafos grandes, com muitas conexões. A Figura 20 exemplifica o formato

Tabela 10 – Quantidades de nós e arestas nos grafos gerados de relações entre os usuários

Grafos de usuários	
Número de nós	5.364
Arestas de mensagens gerais	1.125.326
Arestas de mensagens virais	551.069
Arestas de mensagens com desinformação	433.204

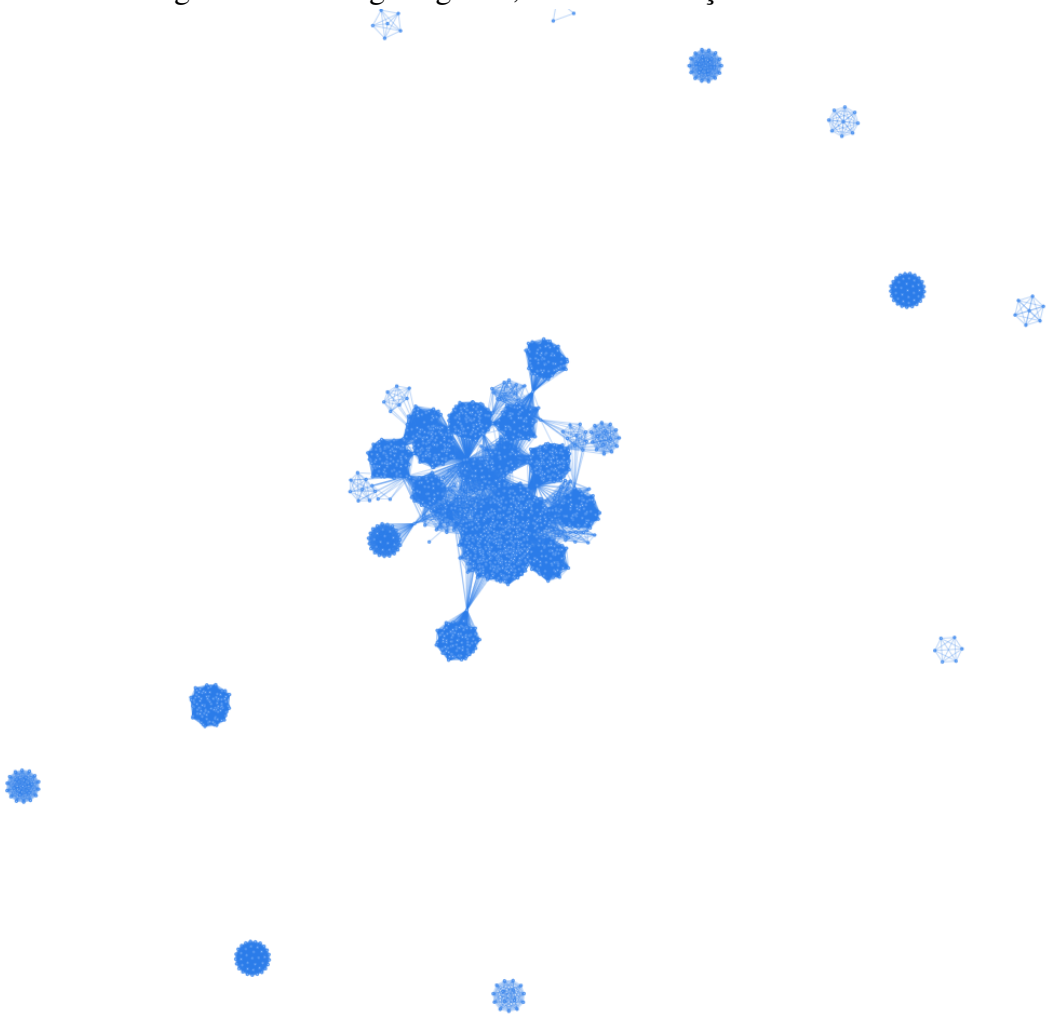
Fonte: o autor.

do grafo de mensagens gerais, através de uma amostra de 2000 usuários, uma vez que exibir o grafo completo não foi possível devido a limitação de recursos computacionais. Para fins de simplificação, nessa representação são ignorados os pesos e direcionamento das arestas. Pela figura, percebe-se que existem grupos isolados e um *cluster* de usuários fortemente conectados no centro. Isso ocorre pois existem usuários engajados que tem participação ativa em diversos grupos. Figura 21 ilustra esse *cluster* com mais detalhes. Percebe-se a existência de usuários com alta centralidade, ou seja, que interagiram com diversos outros usuários, e usuários que interagem apenas com seu grupo local.

Através dessa modelagem dos usuários, podemos obter algumas métricas básicas de nós em redes complexas (WEI *et al.*, 2013), que chamamos aqui de atributos de rede dos usuários, similar ao que foi feito por Benevenuto *et al.* (2008) no contexto do YouTube. São eles:

- **Grau de centralidade geral**: quantidade de arestas de mensagens gerais saindo do nó. Ou seja, para um dado usuário, mensura o número de usuários que receberam ao menos uma mensagem dele. Quanto maior, com mais usuários esse usuário teve contato;
- **Força geral**: somatório dos pesos de todas as arestas de mensagens gerais saindo do nó. É correlacionado com o grau de centralidade geral, porém a quantidade de mensagens enviadas também é levada em conta para calcular essa métrica. Um valor alto indica que o usuário enviou muitas mensagens que atingiram muitos usuários;
- **Grau de centralidade viral**: análogo ao grau de centralidade geral, mas relativo somente às mensagens virais, sendo a quantidade de arestas de mensagens virais saindo do nó. Quanto maior, mais usuários receberam mensagens virais desse usuário;
- **Força viral**: análogo à força geral, é somatório dos pesos de todas as arestas de mensagens virais saindo do nó. Um valor alto indica que o usuário enviou muitas mensagens virais que atingiram muitos usuários;
- **Grau de centralidade de desinformação**: análogo ao grau de centralidade geral, mas relativo somente às mensagens de desinformação, sendo a quantidade de arestas de de-

Figura 20 – Amostra do grafo de mensagens gerais, com uma seleção de 2000 usuários.



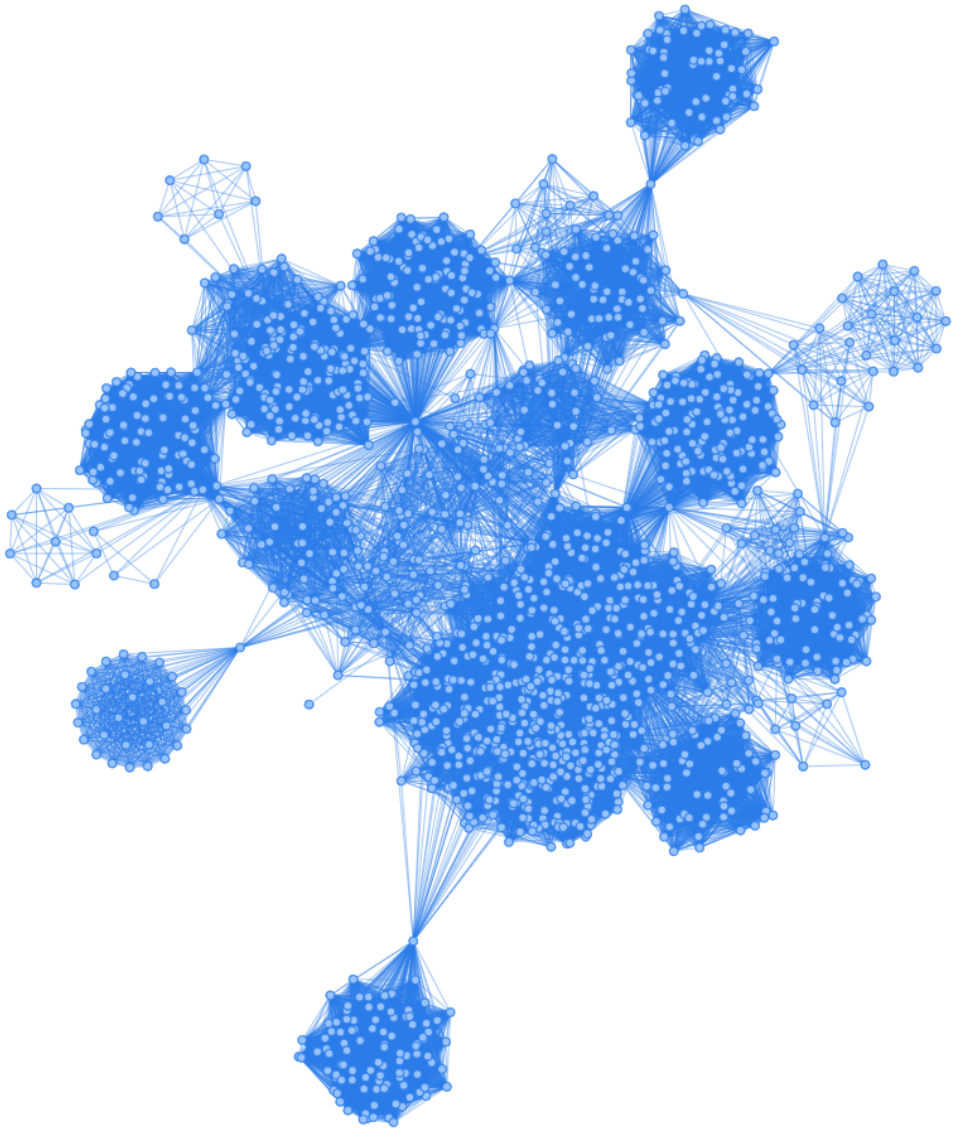
Fonte: o autor.

sinformação saindo do nó. Ou seja, a quantidade de usuários com quem esse usuário compartilhou desinformação. Quanto maior, mais usuários receberam desinformação desse usuário;

– **Força de desinformação**: análogo à força geral, é o somatório dos pesos de todas as arestas de desinformação saindo do nó. Um valor alto indica que o usuário enviou muita desinformação e que atingiu muitos usuários;

Percebe-se que as duas últimas podem ser calculadas devido aos rótulos manuais, enquanto que as outras podem ser calculadas a partir de dados não rotulados. Assim como fizemos para os outros atributos, apresentamos medidas descritivas das distribuições dos atributos de rede na Tabela 11. Percebe-se que, assim como a maioria dos outros atributos, os atributos de rede seguem uma distribuição de cauda longa, com a maioria dos usuários possuindo valores baixos nesses atributos e alguns raros usuários com valores muito altos, indicando um comportamento anormal desses usuários que merece investigação. Essa discussão será retomada no Capítulo 6.

Figura 21 – Detalhe do grafo mostrado na Figura 20, focando em grupos fortemente conectados.



Fonte: o autor.

Tabela 11 – Medidas estatísticas dos atributos de rede

	Centralidade geral	Força geral	Centralidade viral	Força viral	Centralidade de desinformação	Força de desinformação
média	215	10598	105	713	83	386
desvio-padrão	142	29226	151	2859	136	1392
mínimo	3	7	0	0	0	0
Q1	105	494	0	0	0	0
mediana	200	2114	0	0	0	0
Q3	278	8322	200	404	153	273
máximo	1710	672588	1681	96342	1506	28601

Fonte: o autor.

4.6 Limitações e vieses nos dados

Como observado na análise exploratória, o conjunto de dados proposto é bastante rico em informação, porém possui limitações que devem ser discutidas e consideradas no desenvolvimento de pesquisas baseadas nesses dados.

A princípio, podemos mencionar a forma como os dados brutos foram coletados. Por ser o objeto de estudo da pesquisa que originalmente realizou a coleta, os dados representam um período demarcado, com uma temática clara. Ou seja, os tópicos, posicionamentos, sentimentos e outros padrões linguísticos estão limitados a esse domínio específico, reduzindo o escopo da dimensão do conteúdo. Em outras palavras, um modelo de detecção de desinformação treinado com esses dados pode ter dificuldade de generalizar para textos de outros contextos, como saúde, ou mesmo de política mas em outro período temporal, uma vez que a linguagem é viva e padrões linguísticos são fortemente dependentes do domínio.

Além do escopo delimitado, o método de coleta em si possui falhas. Primeiramente, os grupos foram selecionados por uma amostra de conveniência. Ou seja, foram selecionados os que puderam ser encontrados através de buscas, e os links para alguns grupos foram encontrados em mensagens de outros grupos previamente adentrados. Logo, não constituem uma amostra aleatória, e de fato são grupos bastante correlacionados entre si. Segundo, o sistema utilizado para coletar as mensagens apresentou falhas, não registrando mensagens durante alguns dias. Essa lacuna prejudica análises de propagação, pois os padrões destas são incompletos. Não somente por esta falha, pois mesmo se as mensagens fossem coletadas em sua completude, ainda assim, seriam um recorte de uma quantidade relativamente pequena de grupos correlacionados. Diferente de redes sociais como o Twitter, onde podemos rastrear de forma quase completa a origem e propagação de determinada mensagem, no WhatsApp temos visibilidade da propagação da mensagem apenas no grupos monitorados, o que é uma restrição que pode inviabilizar a detecção de um padrão de propagação.

Por fim, por uma questão de falta de recursos humanos, o processo de rotulação foi feito com análise manual de mensagens por um único pesquisador, o que torna possível a presença de vieses nos rótulos. Embora muitas mensagens possam ser claramente definidas como desinformação, com alegações comprovadamente falsas, muitas outras entram em uma área cinza, com alegações que não podem ser comprovadas ou são uma emissão de opinião do autor do texto. Logo, apesar de um procedimento ter sido adotado para minimizar a presença de viés, a rotulação do que é ou não desinformação pode ser subjetivo e passível de debate.

Consideradas essas limitações, o conjunto de dados FakeWhatsApp.Br ainda pode ser bastante útil no estudo da desinformação no contexto do WhatsApp. Além disso, demarcadas essas falhas, pode-se buscar solucioná-las em trabalhos futuros, realizando a evolução contínua do conjunto de dados.

4.7 Conclusão

Nesse capítulo foi apresentado o conjunto de dados proposto nessa pesquisa, chamado FakeWhatsApp.Br. Foram apresentadas todas as etapas de construção, rotulação e realizada uma análise descritiva e das limitações desses dados. No próximo capítulo serão apresentados os experimentos para detecção de desinformação no conjunto de dados proposto.

5 DETECÇÃO DE DESINFORMAÇÃO NO WHATSAPP

Neste capítulo serão apresentados os resultados de uma série experimentos de detecção de desinformação utilizando aprendizado supervisionado no conjunto de dados rotulados descritos no Capítulo 4. Para essa tarefa são consideradas e comparadas abordagens baseadas em conteúdo, abordagens baseadas em contexto social e abordagens híbridas, buscando obter evidências para ajudar a responder as questões de pesquisa Q1, Q2 e Q3, discutidas no Capitulo 1. Também serão analisadas as limitações inerentes aos métodos utilizados, buscando responder a questão de pesquisa Q4. É importante ressaltar que os códigos utilizados para a implementação desses experimentos estão disponibilizados de forma pública no mesmo repositório onde encontram-se os dados, dando assim transparência aos resultados e permitindo a sua reprodutibilidade.

5.1 Separação dos conjuntos de treino e teste

Para obter resultados que representem de forma justa a capacidade de generalização dos métodos para dados não vistos, realizamos a separação dos dados rotulados em um conjunto de treino, usado para otimizar os parâmetros dos modelos, e um conjunto de teste, utilizado para efetivamente avaliar o desempenho dos modelos.

Comumente, a separação dos conjuntos de treino e teste é feita de forma randomizada. Contudo, o conjunto de dados criado possui uma particularidade que deve ser considerada ao realizar esta separação. Conforme detalhado no Capítulo 4, as mensagens rotuladas possuem variações quase idênticas, com pequenas alterações. Uma separação aleatória dessas mensagens em treino e teste possui uma chance considerável de que mensagens altamente similares estejam presentes nos dois grupos. Isso constitui um tipo de **vazamento de informação** (KAUFMAN *et al.*, 2012), pois uma mensagem muito similar do conjunto de teste já foi vista no conjunto de treino, levando a um melhor desempenho que não é necessariamente representativo de um cenário real. Ou seja, o desempenho do modelo no conjunto de teste não necessariamente representa a sua capacidade de generalizar para dados não vistos.

Buscando mitigar esse problema e para avaliar os métodos com maior justiça, realizamos uma separação de treino e teste em duas etapas. Primeiro, agrupamos as mensagens similares, definidas com nosso método de rotulação automático, em subconjuntos disjuntos. Segundo, realizamos uma separação aleatória e estratificada desses subconjuntos, mantendo uma

Tabela 12 – Quantidade de dados nos conjuntos de treino e teste. As variações referem-se a mensagens com alta similaridade a outras, com pequenas modificações.

	Conjunto de Treino	Conjunto de Teste
Total de dados	7807	1574
Total de variações	4503	747
Total positivos	3718	740
Total negativos	4089	834
Variações positivas	2319	399
Variações negativas	2184	348

Fonte: o autor.

proporção de 80% no conjunto de treino e 20% no conjunto de teste. Dessa forma, garantimos que mensagens muito similares às do conjunto de teste não sejam previamente vistas no conjunto de treino.

A Tabela 12 apresenta a quantidade de dados de cada classe em cada um dos conjuntos, bem como a quantidade de mensagens originais. Ou seja, mensagens que não são uma das variações quase idênticas. Conforme já discutido, a tarefa de detecção de desinformação é modelada nesse trabalho como um problema de classificação binária, onde o a classe positiva é formada pelas mensagens contendo desinformação e a classe negativa pelas mensagens sem desinformação. Em ambos os conjuntos, a proporção de dados de cada classe é balanceada, com um quantidade ligeiramente maior de dados da classe negativa em cada conjunto, contabilizando cerca de 52% dos dados tanto no conjunto de treino quanto no de teste. Observa-se também que existe uma grande quantidade de variações em ambos os conjuntos, contabilizando cerca de metade dos dados em ambos. A quantidade de variações são maiores na classe positiva, indicando que a desinformação sofreu mais compartilhamentos com modificações. A proporção de dados em cada conjunto é de 83% para o conjunto de treino e 17% para o conjunto de teste, mantendo-se próxima do pretendido de 80%-20%.

5.2 Métricas de desempenho

Para avaliar a performance dos métodos experimentados, escolhemos métricas de classificação binária bastante utilizadas na literatura. Conforme já discutido, em nossa modelagem as mensagens rotuladas como desinformação são positivas (P), enquanto as rotuladas como não-desinformação são negativas (N). Assim, o resultado de uma predição pode ser:

– Verdadeiro positivo (VP): desinformação que é corretamente classificada como desinformação.

- Verdadeiro negativo (VN): não-desinformação que é corretamente classificada como não-desinformação.
- Falso positivo (FP): não-desinformação que é incorretamente classificada como desinformação.
- Falso negativo (FN): desinformação que é incorretamente classificada como não-desinformação.

Assim, nossas métricas de desempenho são definidas de acordo com a lista abaixo e são calculadas a partir das predições pelos modelos no conjunto de teste:

- Acurácia (ACU): proporção de mensagens corretamente classificadas dentre todas as predições. Ou seja: $\frac{VP+VN}{P+N}$. É uma métrica geral, que não reflete os tipos de erros que ocorreram, mas uma vez que as classes são balanceadas fornece uma boa noção do desempenho.
- Precisão (PRE): proporção de mensagens preditas como desinformação e que realmente são desinformação. Ou seja: $\frac{VP}{VP+FP}$. Quanto maior a precisão, maior a confiança quando o modelo prediz que uma mensagem é desinformação. Uma baixa precisão indica que o modelo resulta em muitos falsos positivos.
- *Recall* (REC): ou taxa de verdadeiro positivos, é a proporção de desinformação corretamente classificada em relação ao total de desinformação existente. Ou seja: $\frac{VP}{VP+FN}$. Quanto maior o *recall*, maior a confiança que o modelo não ignora nenhuma desinformação. Um baixo *recall* indica que o modelo resulta em muitos falsos negativos.
- F1-*score* (F1): média harmônica entre precisão e o *recall*. Calculada como: $\frac{2}{PRE^{-1}+REC^{-1}}$. Assim como a acurácia, também é uma média geral, mas que leva em conta os falsos positivos e falsos negativos.
- AUC: área sobre a curva *Receiver Operator Characteristic Curve* / Curva Característica de Operação do Receptor (ROC). A curva ROC mostra a relação entre o *recall* e a taxa de falsos positivos ($\frac{VP}{VP+FP}$) para os possíveis limiares de decisão. Portanto, essa métrica leva em conta a probabilidade atribuída a cada predição, independente do limiar de decisão escolhido. O valor de 0,5 indica uma classificação aleatória (a curva é igual à diagonal dos eixos de *recall* e taxa de falsos negativos), enquanto o valor 1 indica uma classificação perfeita.

Embora todas as métricas sejam relevantes e deseje-se obter o maior valor possível em todas, em nossa análise consideramos que um falso negativo é um erro mais crítico do que um falso positivo. Isso se deve ao fato de que uma possível aplicação de um modelo de detecção

automática de desinformação seria alertar usuários humanos sobre o risco de uma mensagem conter desinformação, de modo que este usuário possa eventualmente realizar uma checagem de fatos e confirmar ou não a predição do modelo. No entanto, um falso negativo seria entendido como uma mensagem segura e ser propagada. Assim, o dano de bloquear uma mensagem incorretamente classificada como desinformação (que pode ser revertido quando contestado) é menor que permitir a propagação de desinformação. Dessa forma, damos atenção especial às métricas de *recall* e ao F1-*score*.

5.3 Algoritmos de classificação

Nestes experimentos, trabalharemos com dois métodos com grande utilização na literatura: **Regressão Logística** e **Rede Neural *MLP***.

Segundo Bishop (2006) a regressão logística (também conhecida por modelo logit e classificador de máxima entropia) é um método estatístico de classificação cuja superfície de decisão é linear. Ou seja, existe uma relação linear entre os atributos de entrada e a saída predita. Problemas cujas classes podem ser exatamente separadas dessa forma são ditos linearmente separáveis. Embora esse não seja o caso da grande parte dos problemas de classificação do mundo real, em especial problemas de detecção de desinformação, a regressão logística ainda é um modelo com forte adoção na literatura, devido a sua simplicidade e facilidade de uso, velocidade de treinamento e de predição, boa capacidade de generalização em muitos problemas e boa interpretabilidade quando comparado com outros métodos. A regressão logística utiliza um parâmetro para cada atributo, podendo, a partir do valor dos parâmetros, entender a importância que o modelo dá para cada atributo ao fazer sua predição. Portanto, é um modelo adequado para estabelecer um *baseline* de desempenho em um novo problema.

De acordo com Haykin (2010), redes neurais artificiais são uma grande família de modelos matemáticos vagamente inspirados no funcionamento do sistema nervoso animal. Esses modelos são capazes de aprender representações dos atributos que geram superfícies de decisão não-lineares, sendo capazes de solucionar problemas não linearmente separáveis com bom desempenho. Redes neurais *Multilayer Perceptron* são uma das arquiteturas de redes neurais artificiais *feedfoward* mais utilizadas devido a sua facilidade de implementação, grande flexibilidade e bom poder de generalização em muitos problemas. Porém, por serem modelos mais complexos, podendo conter uma grande quantidade de parâmetros em relação aos atributos de entrada, em geral possuem menor interpretabilidade. Além disso, quando o treinamento não é

regularizado, essas redes podem levar a um sobreajuste (*overfitting*) sobre os dados de treino, prejudicando a sua generalização.

Uma característica em comum entre regressão logística e *MLP* é que ambos são modelos probabilísticos. Ou seja, a predição desses modelos é uma probabilidade, associada ao pertencimento de uma classe. Se a probabilidade está acima de um dado limiar, o objeto é classificado como pertencente à classe. Em muitas aplicações isso é desejado, pois podemos estar interessados não somente em uma previsão de classe, mas em uma valor que diga o nível de confiança que o modelo tem naquela previsão.

Isso é especialmente interessante quando um usuário humano utiliza o resultado das predições em um processo de tomada de decisão, como pode ser o caso de um sistema de detecção de desinformação, onde o usuário final irá decidir se uma dada informação é ou não falsa. Além disso, o limiar de decisão pode ser ajustado de acordo com a sensibilidade desejada para o método. Por exemplo, se for mais importante que o método consiga detectar todas as desinformações, mesmo que prediga incorretamente que muitas mensagens sejam desinformação quando não são, basta diminuir o limiar de decisão. Neste trabalho, obtemos um limiar de decisão ótimo, em termos de acurácia, separando um subconjunto de validação do conjunto de treino e buscando o valor de limiar que forneça o melhor resultado nesse subconjunto.

Utilizar esses dois modelos permite uma avaliação da dificuldade e o estabelecimento de uma *baseline* de desempenho, pois utilizamos um modelo linear, mais simples, e um não-linear, com maior flexibilidade, ajudando a responder a nossa questão de pesquisa Q1. Durante a pesquisa dessa dissertação foram também realizados experimentos com outros classificadores de diferentes famílias de aprendizado de máquina, avaliando também abordagens de *boosting*, *bagging*, Naive-Bayes, vizinhos mais próximos e *SVM*, cujos resultados foram reportados em Cabral *et al.* (2021). Porém, não houve diferenças significativas de performance, exceto para os classificadores Naive-Bayes e *K-Nearest Neighbors* / K-vizinhos mais próximos (KNN), que ficaram abaixo dos outros em quase todos os cenários avaliados. Por questão de simplificação e por descrever bem o problema, além dos argumentos já discutidos, apresentamos somente resultados com os classificadores Regressão Logística (*Logit*) e *MLP*.

Tabela 13 – Espaços de busca de hiperparâmetros candidatos para busca aleatória. U representa a distribuição uniforme contínua enquanto que U_d representa distribuição uniforme discreta.

Hiperparâmetro	Espaço de busca
Camadas ocultas	$U_d(1,4)$
Neurônios em cada camada oculta	$25 \cdot U_d(1,15)$
Taxa de aprendizagem inicial	$10^{U(-4,-1)}$
Coeficiente de regularização L2	$10^{U(-6,-2)}$
Tamanho de *minibatch*	$50 \cdot U_d(1,7)$

Fonte: o autor.

5.3.1 Otimização de hiperparâmetros

Ambos os modelos são implementados utilizando a biblioteca Python scikit-learn[1] (PEDREGOSA *et al.*, 2011). A regressão logística é um modelo com uma quantidade relativamente pequena de hiperparâmetros. Como esse modelo foi utilizado para se obter um *baseline*, foram utilizados o conjunto padrão de hiperparâmetros utilizados na biblioteca.

Já a rede *MLP* possui uma grande quantidade de hiperparâmetros, cuja correta seleção em cada problema afeta diretamente o desempenho do modelo. Dentre alguns dos hiperparâmetros importantes podemos citar a própria arquitetura da rede (quantidade de camadas ocultas e quantidade de neurônios em cada camada), o passo de aprendizagem, o coeficiente de regularização L2, a função de ativação das camadas ocultas e o tamanho do *minibatch* de treinamento.

Uma abordagem efetiva para encontrar bons hiperparâmetros dentre uma vasta gama de possibilidades é a busca aleatória (*random search*) (BERGSTRA; BENGIO, 2012). Realizamos a busca aleatória para selecionar os hiperparâmetros da *MLP*, subdividindo o conjunto de treino em um subconjunto de treino e de validação na proporção 90%-10% e realizando 10 experimentos de avaliação. Os hiperparâmetros candidatos são amostrados aleatoriamente de acordo com espaços de busca apresentados na Tabela 13. Em cada experimento uma rede é treinada com os hiperparâmetros amostrados nos dados de treino e avaliada sobre os dados de validação, tendo como critério de parada que a acurácia sobre o conjunto de validação não melhora mais do que 10^{-3} após três épocas de treinamento consecutivas. Após a busca aleatória, o melhor conjunto de hiperparâmetros, em termos de acurácia, é utilizado para treinar um modelo com os dados de validação e treino. Este modelo é então utilizado para fazer predições sobre os dados de teste.

[1] https://scikit-learn.org/

5.4 Extração de atributos

A engenharia de atributos para detecção de desinformação no WhatsApp é uma importante contribuição deste trabalho, visto que é um problema pouco explorado e as informações obtidas com esses experimentos ajudam a responder as questões de pesquisa Q2 e Q3, e a traçar possíveis pesquisas futuras. Nessa dissertação, exploramos atributos de conteúdo, atributos sociais e uma abordagem híbrida, sendo as duas últimas atributos propostos nessa pesquisa.

Atributos temporais também podem ser extraídos do nosso conjunto de dados, mas optamos por não explorá-los por alguns motivos. Primeiro, o foco dessa dissertação é na detecção precoce da desinformação, e atributos temporais exigem que uma mensagem tenha um certo nível de propagação para se detectar um padrão discernível. Segundo, devido a falhas de coleta mencionadas no Capítulo 4, os atributos temporais possuem lacunas de alguns dias, além da própria lacuna existente devido ao recorte observável de uma amostra de grupo de WhatsApp. Em nossa análise exploratória, o padrão de propagação em termos de mensagens por dia, de mensagens com desinformação e mensagens sem desinformação se mostrou similar. Apesar disso, atributos temporais podem ser bastante informativos quando utilizados corretamente e são uma abordagem interessante de ser explorada no futuro, uma vez que hajam dados mais completos.

Por fim, destaca-se que foram realizados experimentos utilizando como atributos a quantificação de pontuação, *emojis* e caracteres de formatação do WhatsApp, mas devido ao baixo desempenho obtido, não foram apresentados nesse texto. A seguir são descritos cada um dos conjuntos de atributos utilizados nos experimentos.

5.4.1 Atributos de conteúdo

Conforme discutido anteriormente nos Capítulos 2 e 3, os atributos de conteúdo são os mais utilizados e mais diretos na detecção de desinformação, alcançando bons resultados na literatura. Em nossos experimentos serão utilizados atributos do tipo lexicais, baseados nas palavras presentes em cada mensagens e nas relações entre essas palavras em todo o *corpus*.

5.4.1.1 Pré-processamento textual

O processo de extração de atributos lexicais para classificação de texto comumente envolve etapas de pré-processamento. Uma dessas etapas é a tokenização, também conhecida

como segmentação de palavras. Essa etapa realiza a quebra das sequências de caracteres em um texto localizando o limite de cada palavra, ou seja, os pontos onde uma palavra termina e outra começa (BARBOSA *et al.*, 2017). As palavras assim identificadas são chamadas de *tokens*. Outras etapas de processamento podem ser específicas de acordo com o domínio e devem ser realizadas para que os *tokens* extraídos revelem padrões distinguíveis para a classificação. No caso de mensagens de WhatsApp, observou-se que os textos são bastante ruidosos, contendo uma profusão de erros de escrita, abreviações, gírias, *emojis*, caracteres de formatação e URLs. Considerando esses fatores, nosso pré-processamento engloba os seguintes passos:

1. Converter todas as letras para minúsculas;
2. Substituir todas as URLs por apenas a raiz do domínio. Por exemplo, a URL https://m.facebook.com/story.php?story_fbid=525897467858283&id=243021376285804 torna-se apenas "facebook";
3. Substituir sequências de *emojis* pelos mesmos *emojis* separados por espaço. Por exemplo, uma sequência de três *emojis* da bandeira do Brasil, representado por ":brazil::brazil::brazil:", torna-se ":brazil: :brazil: :brazil:". Isso permite que cada *emoji* seja tratado como um *token*, evitando assim que sequências contíguas de *emojis* aumentem o vocabulário desnecessariamente;
4. Substituir sequências de mais de três caracteres repetidos por somente três desses caracteres. Por exemplo, "kkkkkkkkkkkk" torna-se "kkk". Isso é feito devido ao mesmo princípio do passo anterior;
5. Remover caracteres que não sejam letras, *emojis* ou os caracteres de formatação "*" (marca texto como negrito) e "_" (marca texto como itálico);
6. Remover palavras de parada, ou seja, palavras pouco informativas, como artigos e pronomes;
7. Lematizar as palavras, ou seja, substituir as palavras pelo seu lema, ignorando flexões e tempo verbal. Por exemplo os termos "tiver", "tenho" e "tinha" tornam-se "ter";
8. Remover acentuação.

O processo de pré-processamento é ilustrado na forma de um fluxograma na Figura 22. Percebe-se que nosso pré-processamento considera a presença de *emojis* e dos caracteres de marcação como *tokens* ao invés de removê-los. Isso se deve ao nosso conhecimento do domínio, obtido durante a rotulação dos dados, onde foi observado que a presença desses *tokens* pode ser um indicativo relevante de desinformação. Além disso, segundo Novak *et al.* (2015), *emojis* são

fortes indicadores de sentimento, e, como já discutido anteriormente, os sentimentos podem ser relevantes na detecção de desinformação.

Após o pré-processamento, é necessário utilizar uma representação vetorial do texto, que serão os atributos de entrada de fato utilizados pelos modelos. Utilizamos para isso o *Bag of Words* binário (*BoW*), o *TF-IDF* e o modelo de *word embedding* Word2Vec.

5.4.1.2 Bag of Words e TF-IDF

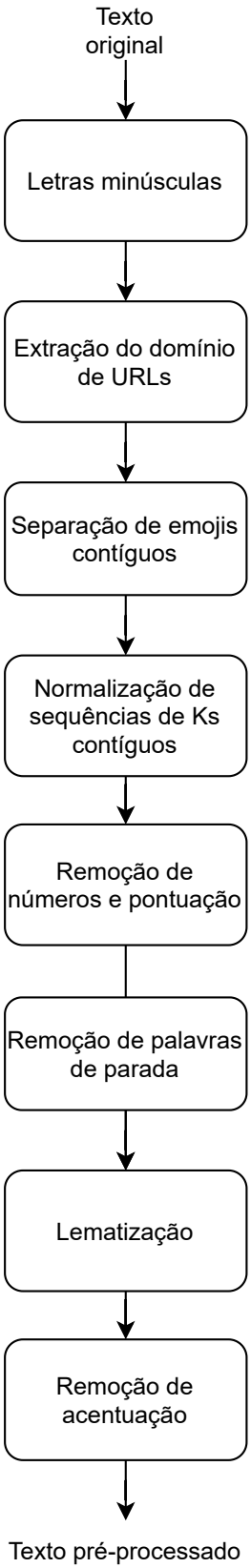
Tanto para o *BoW* quanto para o *TF-IDF*, utilizamos unigramas e bigramas como *tokens*. Essa decisão, assim como as etapas de pré-processamento, são sustentadas pelos resultados apresentados em Cabral *et al.* (2021), onde foram realizados testes com diferentes combinações de pré-processamento e de parâmetros de atributos *BoW* e *TF-IDF*, observando-se que o uso de bigramas é relevante para a detecção de desinformação.

Devido à combinação de unigramas e bigramas, a dimensão dos vetores para esses atributos é consideravelmente elevada, superando 130 mil dimensões. Essa alta dimensionalidade pode ser problemática para modelos de aprendizado de máquina devido a chamada "maldição da dimensionalidade" (VERLEYSEN; FRANÇOIS, 2005). Além da alta dimensionalidade, os vetores são extremamente esparsos. Assim, foram experimentados métodos para reduzir a dimensão desses vetores.

Uma primeira abordagem foi utilizar somente os termos mais frequentes como atributos, ignorando os menos frequentes. Dessa forma, pode-se escolher uma quantidade de atributos significativamente menor, como mil, 5 mil, 10 mil, 25 mil, ou outro valor. Porém, após diversos testes variando essa quantidade de termos, até no máximo 50 mil, todos os resultados foram piores quando comparados com a utilização de todos os termos. Isso deve-se possivelmente a presença de termos-chave pouco frequentes mas que tem um impacto considerável para identificar uma classe, como foi observado na análise dos termos com maior probabilidade a posteriori em cada classe no Capítulo 4.

Outra estratégia foi realizar uma redução de dimensionalidade com *Principal Component Analysis* / Análise dos Componentes Principais (PCA)(JOLLIFFE, 2005). A redução da dimensão do espaço original para um espaço vetorial de dimensão 100 resultou em uma perda de informação que piorou a classificação. Embora seja relevante mencionar que foram feitos e tiveram pior desempenho, optamos por não apresentar os resultados dos experimentos de redução de dimensionalidade neste trabalho, focando no desempenho dos atributos na forma base.

Figura 22 – Etapas de pré-processamento de texto.



Fonte: o autor.

5.4.1.3 Word2Vec

O modelo Word2Vec foi treinado utilizando unigramas pré-processados das mensagens do conjunto de treino e as mensagens não-rotuladas, totalizando 114.444 mensagens. As mensagens do conjunto de teste não foram utilizadas para treinamento. Foram utilizadas 15 épocas de treinamento, uma janela de 5 palavras e o vetor resultante possui dimensão igual a 100. Após o treino do modelo, cada mensagem é representada obtendo o vetor de cada palavra presente na mensagem, ignorando palavras fora do vocabulário, e obtendo o vetor médio desses vetores. Ou seja, o centroide dos *embeddings*. Esse vetor resultante é então normalizado utilizando o *z-score*, subtraindo o valor inicial pela média do atributo e dividindo pela variância, considerando somente o conjunto de treino. Isso faz com que cada atributo do vetor tenha média zero e variância unitária.

5.4.2 Atributos sociais

Atributos sociais utilizam informações sobre os usuários que interagiram com as mensagens para detectar desinformação. Enquanto os atributos de conteúdo utilizados nesse trabalho tem ampla utilização na literatura, atributos sociais costumam ser dependentes do domínio. No caso do WhatsApp, não foram encontrados atributos na literatura para descrever o usuário, por isso utilizamos os atributos propostos no Capítulo 4.

Primeiramente, é necessário definir como os atributos de usuários serão atrelados às mensagens. Sabe-se que vários usuários diferentes podem compartilhar a mesma mensagem e também o mesmo usuário pode compartilhar a mesma mensagem diversas vezes. Uma vez que estamos interessados na detecção precoce da desinformação, representamos a mensagem na dimensão social pelos atributos do usuário que compartilhou a mensagem pela primeira vez no conjunto de dados. A hipótese subjacente a essa escolha é que esses usuários seriam a fonte da desinformação e, portanto, teriam características que os identificassem como suspeitos, assim como as mensagens que publicam.

Conforme apresentado no Capítulo 4, os atributos que foram extraídos dos usuários incluem: número de grupos, número de mensagens; textos; mídia; virais; mensagens repetidas; proporção de textos; proporção de mídia; proporção de virais; proporção de mensagens repetidas; dias ativos; média de mensagens por dia; desvio padrão de mensagens por dia; mediana de mensagens por dia; máximo de mensagens por dia; grau de centralidade geral; força geral; grau

de centralidade viral; e força viral. Portanto, ao todo, foram extraídos 19 atributos. É importante ressaltar que os atributos que descrevem o comportamento do usuário em relação a publicação de desinformação não são utilizados, pois só podem ser calculados devido aos rótulos, não existindo a *priori*.

Como não sabe-se de antemão quais atributos são mais relevantes para a detecção de desinformação, realizamos uma etapa de seleção de atributos, utilizando uma Árvore de Decisão (DENG; RUNGER, 2012). Para isso, treinamos o modelo com o conjunto de treino e obtemos a Importância Gini ou *Mean Decrease in Impurity* / decaimento médio de impureza (MDI), que conta as vezes que um atributo é usado para dividir um nó, ponderado pelo número de amostras que ele divide. Assim, selecionamos somente os 10 atributos com maior importância para a classificação.

Essa abordagem possui algumas limitações que valem a pena ressaltar. Primeiro, sabe-se que nossos dados representam um recorte de alguns grupos, portanto o primeiro usuário que publica uma mensagem não é necessariamente a fonte original desta mensagem, mas apenas o primeiro usuário que podemos observar que publicou essa mensagem. Segundo, um usuário propagandista que compartilha muita desinformação pode também publicar em igual quantidade mensagens que não são rotuladas como desinformação. Terceiro, os atributos que extraímos podem não ser completos (devido a só podermos observar as ações de um usuário nos grupos que coletamos) ou representativos o suficiente. Apesar dessas limitações, o experimento é válido para entender o peso que os atributos sociais propostos podem ter nesse contexto.

5.4.3 Atributos híbridos

Buscando oferecer mais informações que melhorem o poder de classificação dos modelos, os atributos híbridos combinam os atributos sociais e de conteúdo concatenando os vetores. A intuição por trás dessa abordagem é que ela é vagamente similar ao procedimento humano de detecção de desinformação, verificando se o conteúdo da mensagem é suspeito e se a fonte da mensagem também é suspeita. Utilizamos atributos de usuários concatenados com atributos Word2Vec e atributos de usuários concatenados com atributos *TF-IDF*.

A combinação de atributos de usuário com vetores Word2Vec é bastante direta, pois tratam-se de vetores densos. Já a combinação dos vetores de usuário com vetores *TF-IDF* pode ser problemática, pois os vetores *TF-IDF* são extremamente esparsos, dado o tamanho do vocabulário, principalmente quando utilizados bigramas. Assim, os atributos de usuário podem

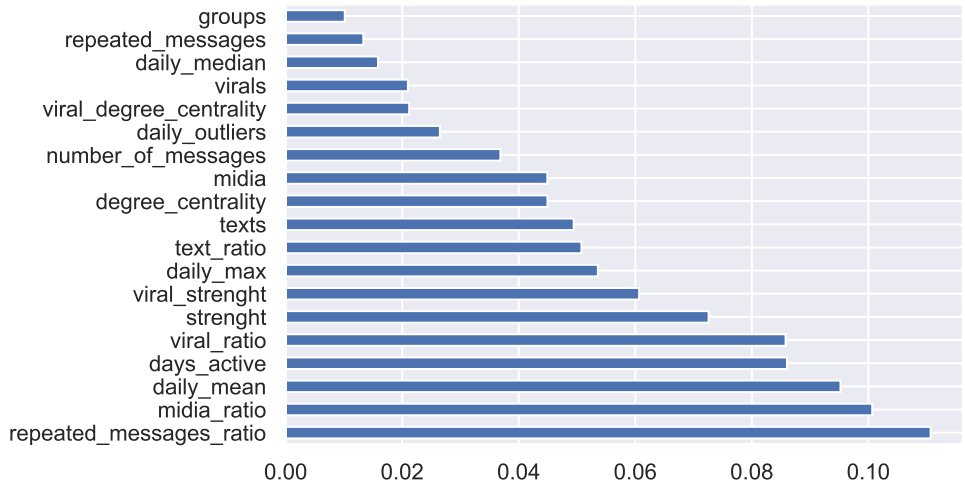
ganhar mais importância do que os atributos *TF-IDF*. Para mitigar esse problema, são utilizados somente os unigramas para representar o vetor *TF-IDF*. Apesar de não solucionar a esparsidade, isso reduz a dimensão do vetor consideravelmente.

5.5 Resultados e discussão

Nesta seção, são apresentados e discutidos os resultados dos experimentos previamente apresentados.

Os atributos sociais escolhidos através do método de seleção de atributos baseado em Árvore de Decisão foram, em ordem de importância, a proporção de mensagens repetidas, média de mensagens por dia, proporção de mídia, proporção de virais, número de dias ativo, força geral, força viral, proporção de textos, número de textos e máximo de mensagens por dia. A Figura 23 ilustra a importância atribuída a cada atributo.

Figura 23 – Importância Gini dos atributos sociais, calculados por uma Árvore de Decisão.



Fonte: o autor.

A Tabela 14 apresenta os resultados de classificação para todos os experimentos, bem como a quantidade de atributos. Observa-se que as abordagens baseadas exclusivamente em conteúdo obtiveram o melhor desempenho, com destaque para os atributos *TF-IDF*. O uso *word embeddings* ficou abaixo tanto dos atributos *BoW* como *TF-IDF*, possivelmente devido à alta quantidade de ruído nos textos e à perda de informação ao realizar a média dos vetores de cada palavra. Observa-se que no trabalho de Silva *et al.* (2020b), que realiza classificação de *fake news* no *corpus* Fake.Br, os métodos que utilizaram *word embeddings* também performaram pior que a representação *BoW*. Isso pode ocorrer devido a presença de termos-chave que discriminam

bem as classes e são bem representados por *BoW* e *TF-IDF*.

No caso dos atributos *BoW* e *TF-IDF*, a regressão logística performou melhor que a *MLP*, com o melhor resultado de F1-*score* de **0,821** para a regressão logística com *TF-IDF*. Esse resultado é interessante, pois um modelo linear, mais simples que a *MLP*, conseguiu generalizar melhor para dados não vistos nesse problema, indicando um possível caso de *overfitting* da *MLP*. Além disso, pelo princípio da Navalha de Occam, uma solução mais simples é preferível a uma mais complexa. Como já mencionado, a regressão logística possui a vantagem de ser mais interpretável do que a *MLP* e redes neurais de forma geral.

O melhor resultado de *recall* atingiu a marca de **0,87**, com uma precisão de **0,77** e F1 de **0,82**, indicando que houve uma proporção maior de falsos positivos do que de falsos negativos. Esse é um bom resultado de acordo com os critérios estabelecidos, mas que possui espaço para melhoria, principalmente quando comparado com outros resultados apresentados na literatura para problemas similares.

Ainda assim, uma comparação direta com esses resultados pode não ser justa, devido ao caráter ruidoso dos dados tratados no nosso problema. Por exemplo, no já mencionado trabalho de Silva *et al.* (2020b), os autores reportaram um F1-*score* máximo de **0,971** de classificação sobre o *corpus* Fake.Br, um resultado quase perfeito. Porém, deve-se considerar que esse *corpus* é formado por textos em formato de notícia jornalística, sendo consideravelmente mais padronizados, além de mais longos, uma vez que os textos coletados nesse corpus possuem no mínimo 100 palavras. Em oposição a isso, já demonstramos que no FakeWhatsApp.Br há uma grande quantidade de mensagens curtas e ruidosas.

Observa-se que as abordagens que utilizaram atributos sociais tiveram um pior desempenho, enquanto que as abordagens que utilizaram somente atributos de conteúdo performaram melhor. Em particular, as abordagens baseadas exclusivamente em atributos sociais resultaram em um desempenho comparável a uma classificação aleatória. Mais ainda, a combinação de atributos sociais com os de conteúdo, piorou o desempenho quando comparada ao uso exclusivo atributos de conteúdos, indicando que os atributos sociais acrescentaram ruído ao classificador ao invés de informação útil.

Algumas hipóteses podem ser levantadas sobre o desempenho dos atributos sociais. Além da já conhecida incompletude dos dados, é possível que esses atributos sozinhos simplesmente não sejam bons preditores se uma mensagem é ou não desinformação. Afinal, mesmo um usuário que possua um padrão de atividade que possa ser identificado como suspeito, isso

Tabela 14 – Resultados obtidos com os experimentos de classificação. Os melhores resultados em cada métrica estão destacados em negrito. Observa-se que as melhores abordagens foram as baseadas em conteúdo, onde o uso dos atributos sociais piorou o desempenho dos modelos na abordagem híbrida. A combinação do classificador logit com atributos TF-IDF obteve os melhores resultados.

Dimensão	Atributos	Dimensões	Modelo	ACU	PRE	REC	F1	AUC
Conteúdo	BoW	130588	Logit	0.812	**0.804**	0.794	0.799	**0.904**
			MLP	0.794	0.776	0.791	0.783	0.893
	TF-IDF		Logit	**0.822**	0.777	**0.871**	**0.821**	0.902
			MLP	0.805	0.756	0.864	0.807	0.902
	Word2Vec	100	Logit	0.736	0.696	0.779	0.735	0.804
			MLP	0.792	0.785	0.77	0.777	0.872
Social	Usuário	10	Logit	0.560	0.537	0.470	0.501	0.605
			MLP	0.587	0.555	0.618	0.584	0.622
Híbrido	Usuário + TF-IDF (unigramas)	17525	Logit	0.802	0.784	0.799	0.791	0.865
			MLP	0.776	0.783	0.726	0.753	0.859
	Usuário + Word2Vec	110	Logit	0.734	0.696	0.769	0.731	0.807
			MLP	0.764	0.786	0.685	0.732	0.830

Fonte: o autor.

não implica que toda mensagem que ele compartilhar será desinformação. Mais ainda, usuários crédulos, que não são desinformadores, mas ocasionalmente compartilham desinformação podem também não ter atributos que sejam identificáveis como suspeitos.

Além disso, foram utilizados os atributos de apenas um usuário para classificar a mensagem. Uma alternativa que pode prover mais informação útil é utilizar as informações de mais usuários que compartilharam a mensagem, levando em conta a média dos atributos ou tratando-os como uma sequência temporal. Mais ainda, poderia ser realizado um *embedding* de usuário a partir do grafo de interações entre usuários, onde este *embedding* conteria informação da relação de um usuário com outros. Contudo, apesar da falha para detectar desinformação, os atributos propostos foram úteis para a detecção de desinformadores, que será discutida no Capítulo 6.

5.6 Interpretabilidade da melhor abordagem

Embora o estudo da interpretação de como os modelos realizam predições esteja fora do escopo deste trabalho, podemos aproveitar a simplicidade da abordagem de melhor desempenho (regressão logística com atributos *TF-IDF*) para analisar o que foi aprendido pelo modelo sobre a tarefa. Para isso, pode-se observar os valores dos parâmetros do modelo e os termos aos quais eles se referem. Os pesos com os maiores valores positivos para os atributos

TF-IDF em ordem decrescente são referentes aos seguintes termos (unigramas e bigramas):

1. *video*
2. *audio*
3. *pt*
4. *compartilhar*
5. *repassar*
6. *tentar*
7. *governar*
8. *youtube*
9. *rosto_furioso*
10. *ver*

Ou seja, pode-se entender que o modelo dá alta importância a presença desses termos ao fazer uma predição. Um texto com um alto valor *TF-IDF* para estes termos (com uma frequência alta na mensagem) teria maior probabilidade de ser classificado como desinformação. É interessante observar a presença de termos como "*compartilhar*" e "*repassar*", comumente associados à desinformação. Além disso, os termos "*video*", "*audio*" e "*youtube*" são associados à arquivos de mídia ou vídeos externos. Também nota-se a presença do código "*rosto_furioso*" referente a um *emoji*. Já considerando as mensagens negativas com maior valor absoluto, temos:

1. *bolsonaro*
2. *chat*
3. *chat whatsapp*
4. *dia*
5. *rolando_no*
6. *rosto_sorridente*
7. *postar*
8. *feirar*
9. *whatsapp*
10. *bandeira_brasil*

Um valor alto destes atributos aumenta a chance de que a mensagem seja classificada como não-desinformação. Notam-se os termos "*chat*", "*chat whatsapp*" e "*whatsapp*", que são comuns em mensagens de divulgação de grupos públicos, rotuladas como não desinformação. O termo "*dia*" refere-se possivelmente a mensagens de bom dia, que são rotuladas como não

desinformação. Já os *emojis* desta lista são relacionados a textos de humor ("*rolando_no*" e "*rosto_sorridente*") ou propaganda ("*bandeira_brasil*").

Esse resultado é coerente com o observado empiricamente durante a rotulação das mensagens, indicando que o modelo identificou palavras-chave associadas as classes no contexto desses dados. Por outro lado, os termos também aparentam ser muito específicos de um domínio restrito: as eleições presidenciais no Brasil de 2018. Indicando que esse modelo teria um desempenho baixo em detectar desinformação em outros contextos, como por exemplo da pandemia de Covid-19, pois os termos utilizados podem seguir distribuições de probabilidade diferentes. Iremos validar essa ideia na Seção 5.9.

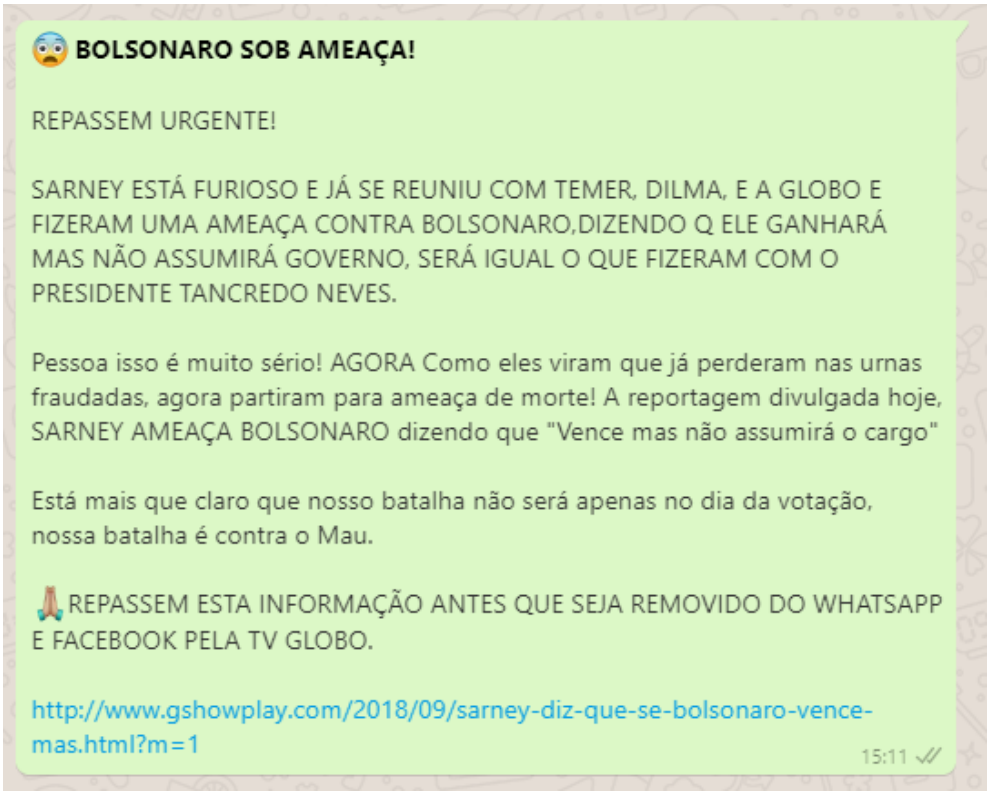
Pode-se também interpretar predições individuais observando a contribuição de cada termo na predição. Ou seja, o produto do peso do modelo pelo valor *TF-IDF* do termo. Por exemplo, considerando a mensagem rotulada como desinformação ilustrada na Figura 24. Nosso modelo classificou esta mensagem como desinformação, com probabilidade associada de 63%. Podemos observar os termos (pré-processados) que mais contribuíram positivamente para essa predição, e o valor dessa contribuição, ilustrados na Figura 25. De maneira similar, podemos observar os termos que contribuíram negativamente, reduzindo a probabilidade estimada de a mensagem ser desinformação, ilustrados na Figura 26

Observamos na Figura 25 que os termos "*repassar*" e "*urgente*" são fortes preditores, o que está de acordo com nosso conhecimento do domínio, onde sabe-se que a desinformação usualmente adota tom alarmista e apelo para compartilhamento. Já na Figura 26 percebe-se que os termos "*dia*" e "*whatsapp*" aparecem como preditores negativos, o que não condiz com o que um leitor humano se atentaria para identificar uma mensagem como não suspeita. O valor negativo desses termos é possivelmente devido a suas frequência em mensagens rotuladas como não-desinformação, como mensagens de bom dia ou mensagens de divulgação de grupos públicos. Apesar disso, o modelo conseguiu predizer corretamente, com uma margem razoável acima de 50%.

5.7 Análise de erros

Além das métricas quantitativas de desempenho, nos interessa também analisar qualitativamente os casos de predições incorretas e obter percepções sobre as limitações de nossos métodos e apontar possíveis soluções futuras. Buscamos assim, obter informações para responder a questão de pesquisa Q4.

Figura 24 – Exemplo de uma mensagem rotulada como desinformação em renderização do WhatsApp. O modelo de regressão logística com atributos *TF-IDF* atribuiu a essa mensagem uma probabilidade de 63% de ser desinformação.



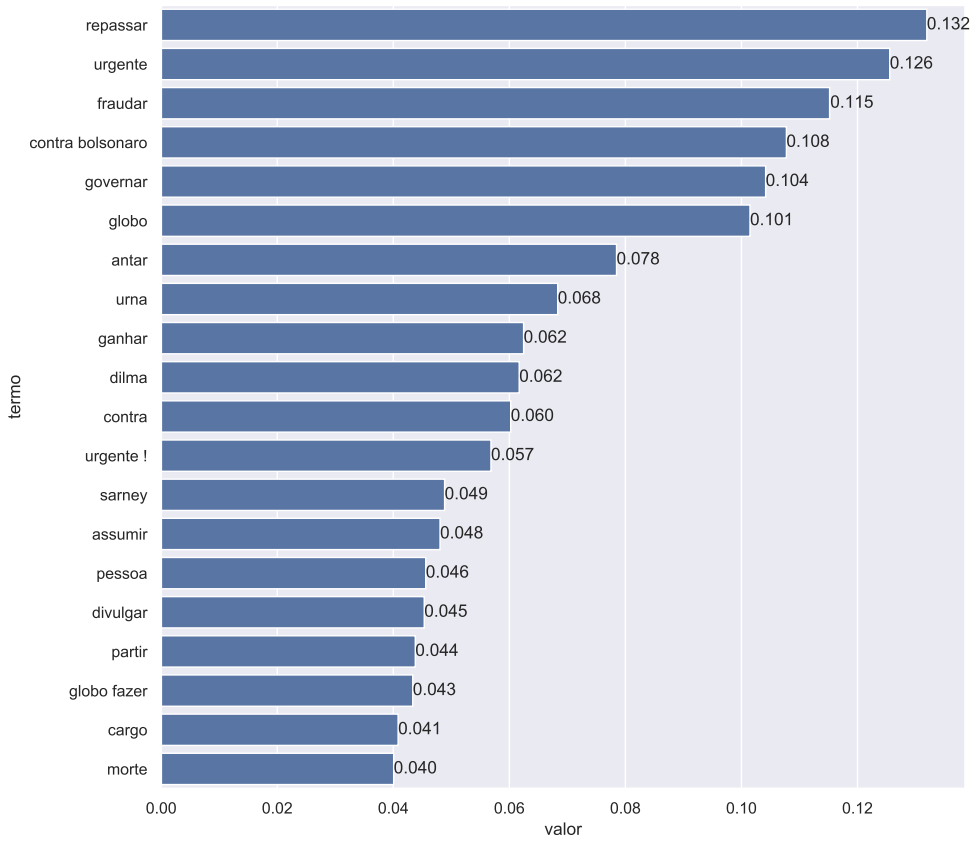
Fonte: o autor.

Para este fim, consideramos todas as predições incorretas feitas pela abordagem que apresentou o melhor desempenho, a regressão logística com atributos *TF-IDF*. A Figura 27 ilustra o resultado desta classificação através de uma matriz de confusão, contabilizando os erros e acertos em cada tipo. Para a análise de erro, foram analisados individualmente todos os falsos positivos e falsos negativos. A partir dos padrões observados classificamos os textos por dois critérios: tamanho e referência a informação externa.

Conforme discutido anteriormente, muitas das mensagens fazem referência a alguma informação externa, seja um arquivo de mídia ou a uma página web por meio de uma URL. Ou seja, toda a informação a que a mensagem faz referência não encontra-se exclusivamente no texto. Isso pode ser uma dificuldade para um modelo baseado em conteúdo, pois muitas vezes o que permite que um especialista humano identifique que uma mensagem desse tipo é uma desinformação é o conteúdo que está presente na informação externa. Conteúdo este que nossos modelos não tem acesso.

Em termos de tamanho, como comentado anteriormente, a maioria dos textos são curtos. Para obter uma diferenciação entre textos longos e curtos, estabelecemos um limiar

Figura 25 – Top 20 atributos que mais contribuíram positivamente para predição de desinformação e o valor da contribuição.



Fonte: o autor.

de 50 palavras, obtido através da distribuição de quantidade de palavras do conjunto de teste, correspondendo aproximadamente ao 55 percentil. Assim, textos com mais de 50 palavras são considerados longos, enquanto textos com 50 palavras ou menos são considerados curtos.

Combinando os dois critérios, obtivemos 4 categorias, que listamos abaixo com exemplos de falsos negativos:

– **Texto curto, com informação externa**. Exemplo:

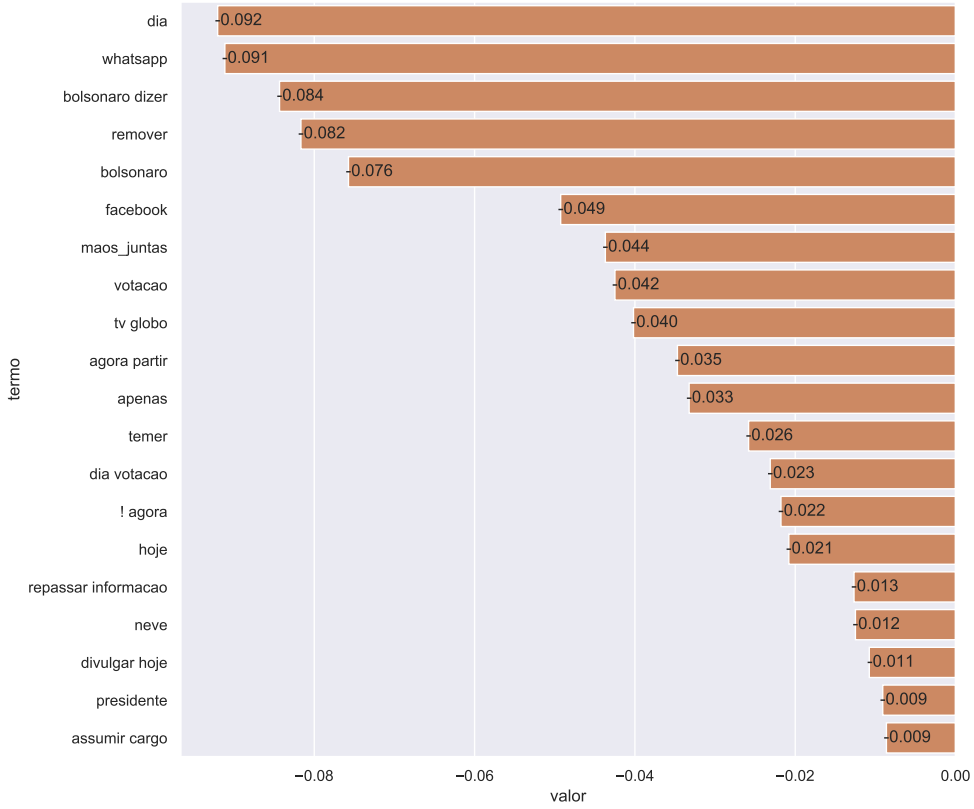
"*Observe na filmagem que ele dá um soco no ferimento de Bolsonaro que Bolsonaro chega a gritar.*";

– **Texto longo, com informação externa**. Exemplo:

"*PESSOAL ASSISTA O ÚLTIMO VÍDEO DE JAIR BOLSONARO QUE FOI FEITO ONTEM , AO VIVO . ELE MESMO ESTÁ ORIENTANDO. E PARA IR UM POUCO ANTES DAS 17HS NO LOCAL ONDE VOCE VOTOU ESPERA TERMINAR TUDO, ESPERA O FISCAL OU MESARIO COLOCAR NO MURAL O PAPEL DO RESULTADO TOTAL DA VOTAÇÃO. AÍ VOCE TIRA A FOTO E ENVIA PARA fiscaldojair.com.br .*";

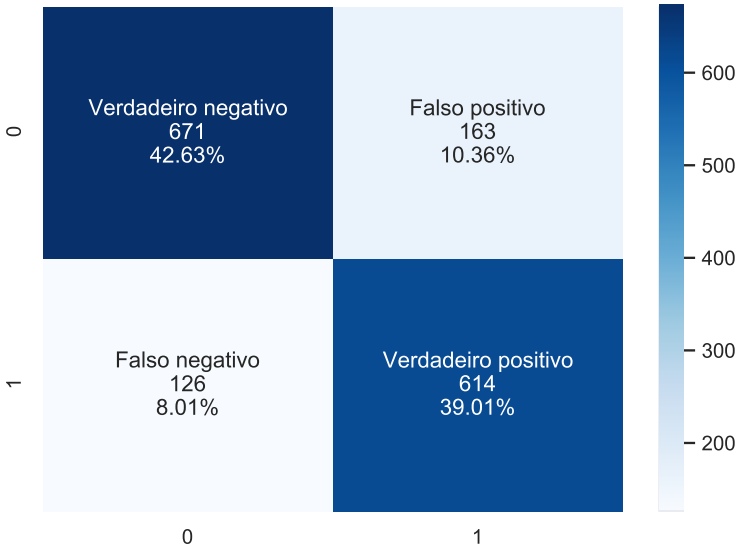
– **Texto curto, sem informação externa**. Exemplo:

Figura 26 – Top 20 atributos que mais contribuíram negativamente para predição de desinformação e o valor da contribuição.



Fonte: o autor.

Figura 27 – Matriz de confusão da classificação da regressão logística com *TF-IDF*.



Fonte: o autor.

"*Cuidado *Petistas* estão postando mensagens xingando os Nordestinos como se fossem eleitores do Bolsonaro de outros estados do Brasil para causar raiva nos eleitores bolsonarianos no Nordeste. Espalhem essa notícia*";

Tabela 15 – Quantidade de mensagens de cada categoria para cada tipo de erro e suas respectivas proporções aproximadas. Nota-se que a principal causa de erros tanto para falsos positivos quanto para falsos negativos são textos curtos, com informação externa.

	Falsos positivos	Falsos negativos
Texto curto, com informação externa	89 (47%)	75 (60%)
Texto longo, com informação externa	11 (6%)	14 (11%)
Texto curto, sem informação externa	12 (7%)	16 (13%)
Texto longo, sem informação externa	76 (40%)	20 (16%)
Total	188	125

Fonte: o autor.

– **Texto longo, sem informação externa**. Exemplo:

"**NÃO PODEMOS DESISTIR: AINDA HÁ TEMPO* *DE LUTAR POR ELEIÇÕES LIMPAS E ABERTAS* Vamos pressionar pela aprovação do Decreto Legislativo que OBRIGA a adoção de urnas de lona e voto em cédula de papel ainda nesta Eleições 2018! Faça sua parte, ajude a pressionar, veja a lista de contatos dos senadores no texto que acompanha este banner: #ForaUrnasEletrônicas #VotoNaCédulaJá #Eleições2018*";

Categorizamos todos os falsos positivos e falsos negativos e contamos a quantidade de cada categoria nos dois grupos. O resultado dessa contabilização é apresentado na Tabela 15. Conforme observado, os textos curtos com informação externa predominam como categoria onde mais ocorrem erros tanto para falsos positivos quanto para falsos negativos, com uma maior predominância nos falsos negativos. Ou seja, desinformação dessa categoria é mais difícil de detectar pelo nosso modelo. Isso é esperado, uma vez que essa abordagem é baseada nas correlações entre frequências de palavras, e textos curtos provém pouca informação nesse sentido, gerando vetores *TF-IDF* muito esparsos e com pouca informação. Essa é uma limitação relevante dessa abordagem, uma vez que grande parte das mensagens virais no contexto do WhatsApp são textos curtos e que fazem referência a informações externas, demandando o desenvolvimento de soluções que considerem essa particularidade dos dados. Analisaremos mais o impacto das mensagens curtas na classificação na Subseção 5.8.

Observa-se também que os falsos positivos tiveram uma proporção alta de textos longos sem informação externa, enquanto para os falsos negativos essa proporção é bem menor. Isso parece indicar que o modelo está enviesado a classificar mensagens longas como desinformação. De fato, no Capítulo 4, mostramos que o conjunto de mensagens rotuladas como desinformação continha uma quantidade maior de mensagens longas. Ou seja uma distribuição de quantidade de palavras de cauda mais longa, o que pode ter levado a este enviesamento.

Tabela 16 – Quantidade de dados nos conjuntos de treino e teste para mensagens com mais de 50 palavras.

	Treino	Teste
Total de dados	2682	713
Positivos	1569	413
Negativos	1113	300

Fonte: o autor.

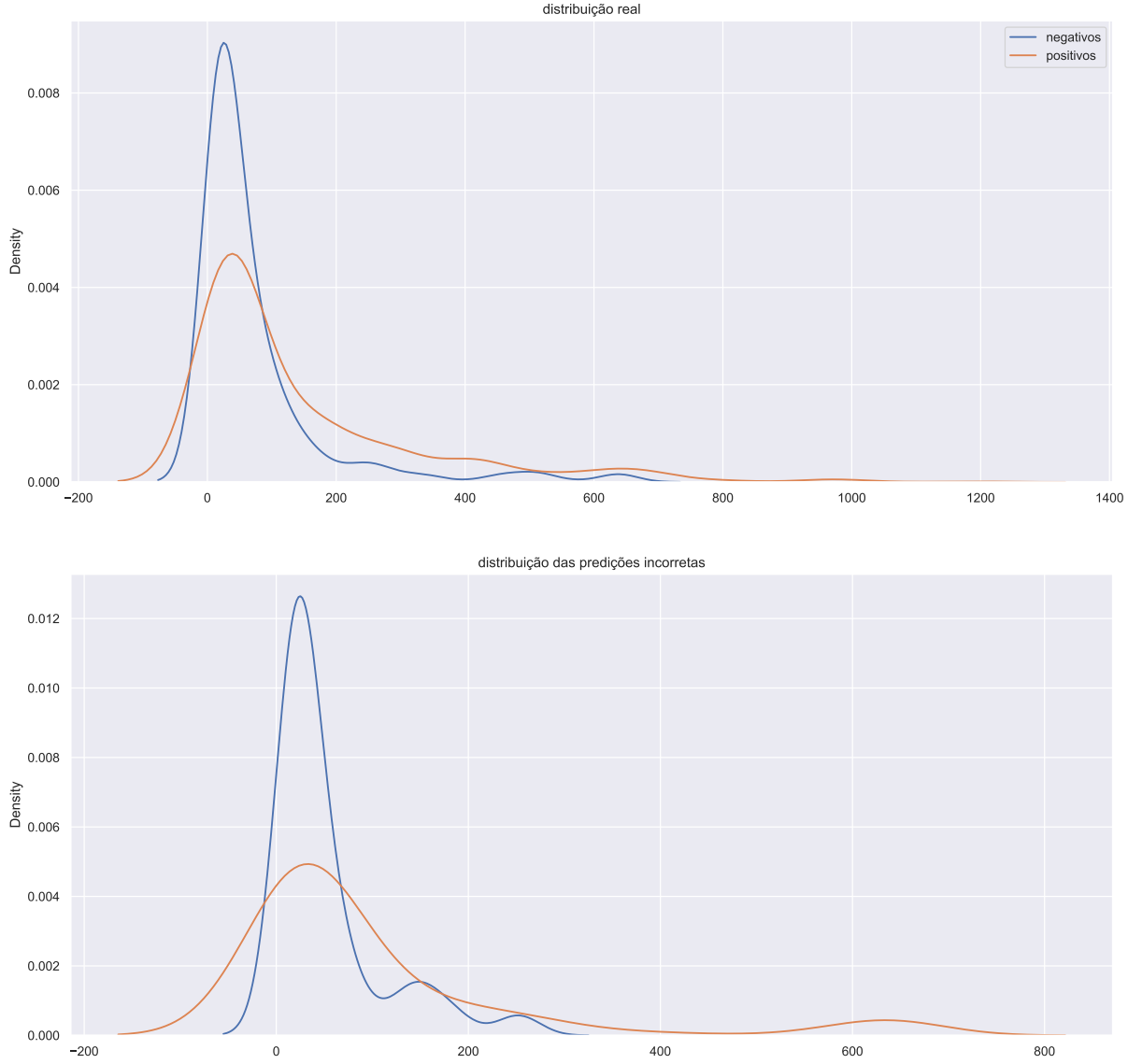
Reforçamos essa hipótese analisando as distribuições da quantidade de palavras nas mensagens do conjunto de teste. A Figura 28 ilustra a diferença entre as distribuições das mensagens positivas e negativas e dos falsos positivos e falsos negativos do conjunto de teste através do KDE. De fato, comparando as distribuições de positivos e negativos, percebe-se que a cauda da distribuição de positivos é mais longa, com mensagens com mais de 800 palavras, enquanto o texto mais longo das mensagens negativas possui 647 palavras. Observa-se que embora a maior densidade de ambos os tipos de erros se concentrem em mensagens curtas, existe um aumento de densidade de falsos positivos em mensagens em torno de 600 palavras. Observamos nos resultados que todos os textos com mais 550 palavras foram preditos como positivo. Observando as mensagens rotuladas, haviam 11 mensagens rotuladas como negativas e 42 como positivas com mais de 550 palavras. Essa tendência é problemática e deve ser considerada em trabalhos futuros.

5.8 Análise de classificação somente com mensagens longas

Buscando compreender melhor o impacto das mensagens curtas na detecção de desinformação, realizamos um novo experimento de classificação, filtrando todas as mensagens com mais de 50 palavras, reduzindo os dados a cerca de metade do tamanho original. Assim, realizamos uma nova rodada de treino e teste com esse subconjunto de dados, utilizando o nosso método de melhor desempenho. A quantidade de dados de cada classe nos conjuntos de treino e teste filtrados são apresentadas na Tabela 16. Observa-se que os dados ainda são aproximadamente balanceados em ambas as classes, mas com mais exemplos da classe positiva. Isso é esperado, uma vez que há mais desinformação com textos longos nos nossos dados.

Após realizar o mesmo procedimento de treino e teste com regressão logística e *TF-IDF*, os resultados obtidos nesse subconjunto de textos longos demonstram uma melhoria de performance, que são apresentados na Tabela 17. Percebe-se que quase todas as desinformações foram corretamente classificadas, com um *recall* superior a 0,90, além de uma alta precisão,

Figura 28 – Estimativa de densidade de probabilidade das quantidades de palavras em mensagens do conjunto de teste, considerando as mensagens rotuladas como positiva e negativas (acima) e as predições incorretas (abaixo). Observa-se que a distribuição de falsos positivos possui um aumento de densidade em textos muito longos, com cerca de 600 palavras.



Fonte: o autor.

quando comparado à performance com todos os dados. Em particular, o F1-score teve uma melhoria de aproximadamente 9%.

Esse resultado reforça a problemática da classificação de textos curtos com os métodos utilizados. Se considerarmos apenas os textos longos, temos um problema mais fácil de ser resolvido e isso pode ser útil para aplicações práticas. Por exemplo, um dos resultados do trabalho de Monteiro *et al.* (2018a) foi a criação de um serviço no WhatsApp[2] onde os usuários podem enviar mensagens para detecção de *Fake News* através de modelos desenvolvidos com

[2] https://api.whatsapp.com/send?phone=5516988212457&text=Nilc-FakeNews

Tabela 17 – Desempenho da regressão logística com representação TF-IDF quando treinado e testado somente com textos longos, com 50 ou mais palavras. Observa-se o salto de desempenho em relação a quando considerados os textos curtos.

Métrica	Resultado
Acurácia	0.872
Precisão	0.869
Recall	0.918
F1	0.893
AUC	0.918

Fonte: o autor.

o *corpus* Fake.Br. Esse serviço possui uma limitação de que os textos enviados devem ter no mínimo 100 palavras. Essa é uma limitação razoável quando consideramos textos de notícias, do qual o Fake.Br é composto, mas no contexto de mensagens de WhatsApp essa limitação reduz bastante o escopo do serviço. Nosso modelo, já consegue um bom desempenho com um mínimo de 50 palavras, permitindo uma maior cobertura em um serviço como este, além de ser especificamente treinado com mensagens que circulam em WhatsApp.

Porém, os textos curtos também fazem parte do cenário de desinformação no WhatsApp. De fato, textos curtos compõem a maioria das mensagens rotuladas. Portanto, é necessário o desenvolvimento de métodos que contemplem esse desafio. No caso de mensagens curtas que fazem referência a conteúdo externo, como imagens, áudios, vídeos ou páginas da web, pode-se buscar extrair informação desses conteúdos além do texto em si da imagem, abrindo margem para abordagens baseadas em conteúdo hibridas, que mesclam atributos textuais, de imagem, ou de vídeo para se chegar a um classificação. Além, é claro, de novos métodos baseados em propagação.

5.9 Avaliação de detecção de desinformação no contexto da pandemia do Covid-19

Ao criar um modelo de detecção de desinformação, é desejável que ele possa ser aplicado para diferentes tipos de desinformação. Por sua natureza, o conteúdo de desinformações muda com o tempo, uma vez que novos tópicos, notícias e contextos sociais surgem. Conforme discutido anteriormente, uma limitação de abordagens baseadas em conteúdo é a forte dependência do contexto dos dados, especialmente para uma abordagem baseada em frequência de termos como o *TF-IDF*. As distribuições de probabilidade dos termos podem variar de acordo com o tempo, o domínio do discurso, a plataforma de onde os dados foram coletados, dentre outros, de

Tabela 18 – Comparação da performance do nosso melhor modelo quando aplicado no contexto de mensagens de WhatsApp sobre a pandemia do Covid-19, coletadas em 2020. A comparação se restringiu a precisão, *recall* e *F1-score* pois estas foram as métricas apresentadas no trabalho original.

	PRE	**REC**	**F1**
(MARTINS *et al.*, 2021)	0.771	0.791	0.778
Nosso modelo	0.517	0.795	0.627

Fonte: o autor.

modo que um modelo treinado com dados de um contexto em particular pode ter problemas de performance quando utilizado em dados de outro contexto.

Para confirmar essa deficiência, utilizamos nosso melhor método para realizar predições sobre os dados publicados em Martins *et al.* (2021). Esses dados são mensagens de WhatsApp coletados pelo mesmo método descrito nessa dissertação, mas durante os meses de Abril e Junho de 2020 e são todas referentes a pandemia do Covid-19. Esse conjunto de dados possui 2898 mensagens, das quais 1985 são rotuladas como positivas (69%) e 913 são rotuladas como negativas (31%), possuindo portanto um desbalanceamento considerável entre as classes. Utilizamos nosso modelo treinado com dados de 2018 e realizamos predições sobre este conjunto de dados de 2020 e comparamos com o melhor resultado obtido na publicação original, utilizando um classificador *SVM* linear e representação *BoW*. Esses resultados estão apresentados na Tabela 18.

Observa-se que nosso modelo manteve um bom desempenho em termos de *recall*, comparável com o *recall* do artigo, porém teve um baixo desempenho em precisão e, consequentemente, em termos de *F1-score*, enquanto o método proposto por Martins *et al.* (2021) teve um melhor desempenho em precisão. Ou seja, nosso modelo possui uma tendência a predizer que as mensagens nesse contexto são desinformação, conseguindo classificar corretamente a maior parte da desinformação, mas com baixo desempenho para identificar mensagens que não são desinformação, errando em cerca de metade destas.

Isso possivelmente deve-se a uma mudança nas distribuições de probabilidade da frequência de palavras, o que é esperado, uma vez que surgem novos tópicos não vistos por nosso modelo, como isolamento social, uso de máscaras, vacinação, transmissão, mortalidade, capacidade dos hospitais, ação da China, dentre outros temas ligados a este domínio.

Assim, uma aplicação de detecção de desinformação em WhatsApp baseada em conteúdo exige um retreinamento contínuo de modo a evitar a queda de performance com o tempo, uma vez que é esperado que a distribuição de probabilidade dos dados varie. Esse

retreinamento exige dados rotulados, que, como já foi discutido, são custosos de se obter. Dessa forma, apesar do bom desempenho obtido nos experimentos nesse trabalho, modelos baseados em conteúdo possuem restrições práticas que levam a demanda por técnicas baseadas em propagação.

5.10 Conclusão

Neste capítulo, descrevemos uma série extensa de experimentos em torno do problema de detecção de desinformação no nosso conjunto de dados. Os resultados forneceram achados relevantes para responder parte de nossas questões de pesquisa, proporcionando informações sobre a dificuldade deste problema, bem como acerca dos melhores métodos, dentre os que foram avaliados, e de suas inerentes limitações. No próximo capitulo, discutiremos algumas estratégias para a detecção de desinformadores.

6 DETECÇÃO DE DESINFORMADORES NO WHATSAPP

Este capítulo é dedicado à análise do problema de detecção de desinformadores no WhatsApp. Ou seja, usuários que disseminam desinformação em grande escala. Conforme já discutido, a identificação desses usuários é um ponto-chave para mitigar a propagação da desinformação. Contudo, não existe na literatura uma definição específica do que caracteriza um desinformador no contexto do WhatsApp. Por isso, neste capítulo, realizamos uma análise do comportamento dos usuários com base nos atributos propostos no Capítulo 4. A partir dessa análise, propomos uma definição de desinformadores baseada nos dados rotulados observados e propomos dois métodos de detecção automática desses usuários. Esses experimentos nos fornecem informação para responder a questão de pesquisa Q5.

6.1 Definição de desinformadores

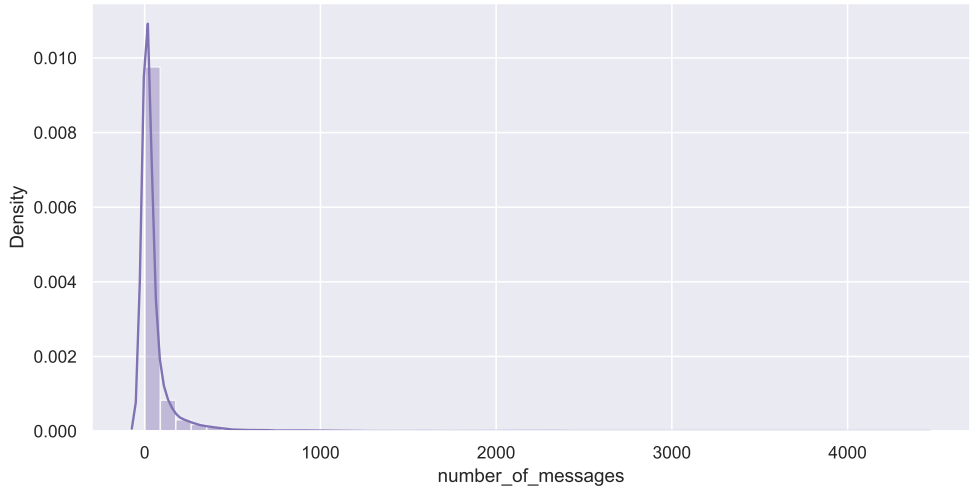
Inspirados pelo trabalho de Uddin *et al.* (2014), que propõe categorias de usuários no Twitter com base em suas atividade nessa plataforma, nós propomos uma definição de desinformadores baseada nos dados coletados. Essa definição leva em consideração o comportamento dos usuários, mensurado pelos atributos propostos anteriormente.

Diferentemente de Uddin *et al.* (2014), não estamos interessados em categorizar todos os tipos de usuários, nem estamos interessados no usuário regular, que utiliza o WhatsApp de forma casual. Estamos interessados em usuários com valores extremos em certos atributos, que indicam atividades suspeitas, e como esses conjuntos de usuários se relacionam com a desinformação propagadas nos grupos.

Para isso, primeiramente, analisamos a atividade geral dos 5364 usuários do conjunto de dados, através da distribuição da quantidade total de mensagens que cada usuário enviou nos grupos analisados. As medidas estatísticas desse atributo foram apresentadas na Tabela 7 do Capítulo 4. Desta tabela, notamos que a distribuição do total de mensagens, assim como de outros atributos de usuários, possui uma cauda longa, com a grande maioria dos usuários tendo baixa atividade, ilustrado na Figura 29. De fato, apenas 25% dos usuários enviaram mais de 45 mensagens.

Como estamos interessados em usuários que tiveram atividade relevante, para estabelecer a definição de desinformadores iremos analisar apenas a metade de usuários que mais enviou mensagens. Ou seja, usuários que mandaram uma quantidade de mensagens maior do

Figura 29 – Distribuição da quantidade de mensagens enviadas por usuários no conjunto de dados.



Fonte: o autor.

que a mediana, que corresponde a 13 mensagens. Assim, esse recorte foi de 2633 usuários, chamados de usuários ativos.

O próximo passo foi observar as distribuições dos atributos desse subconjunto de usuários. Como já mencionado, estamos interessados no comportamento anômalo para definir usuários-chave, por isso, cada grupo é definido através do conhecido método de detecção de *outliers* baseado na distância interquartil (WALFISH, 2006), onde um *outlier* é definido como o valor igual a $Q3 + 1,5 \cdot IQ$, onde $Q3$ é o terceiro quartil e IQ é a distância interquartil da distribuição, considerando somente o subconjunto de usuários ativos.

Propomos então a definição de desinformadores a partir do atributo força de desinformação. Nesse trabalho, os desinformadores são o grupo que, dentre os usuários ativos, possui um valor anômalo de força de desinformação. Um valor anômalo está acima do limiar de 28,601, correspondente ao valor de *outlier* pela distância interquartil. Ou seja, são usuários que disseminaram uma grande quantidade de desinformação e/ou alcançaram muitos usuários. Abaixo seguem alguns exemplos de mensagens do desinformador de maior atividade:

– "*https://youtu.be/iXi3X2XDg6A ..... *URGENTE* !! multipliquem este vídeo ao máximo!!*"
– "*https://youtu.be/WcXXsERafNA. *MAIS UMA FAKE NEWS do HADDAD DESMACARADA!!!* *COMPARTILHEM com todos os seus contatos!!!* vamos colocar este vídeo *EM ALTA* no YouTube!!!!*"
– "*Mais uma fake News da mídia.....o assassinato do capoeirista não teve nada a ver com política ou muito menos com apoiador de Bolsonaro..... *CANALHAS!! ..... Divulgue*

Tabela 19 – Descrição da categoria de desinformadores em termos de quantidade de usuários, porcentagem desses usuários em relação ao total, quantidade de desinformação enviada por usuários dessa categoria e porcentagem de desinformação em relação a desinformação total.

Categoria	Atributo	Limiar	# de usuários	% de usuários	# de desinformação	% de desinformação
Desinformadores	Força de desinformação	28601	132	2.5%	4533	39.7%

Fonte: o autor.

*este vídeo para todos os seus contatos e grupos do WhatsApp**"

– "**No Ceará, o Comando Vermelho(CV) PROIBIU propaganda de BOLSONARO nos territórios que* *"administra"* *Somente LULA E CIRO Podem. Por serem aliados do CRIME.* Alguém tem dúvida agora da quadrilha?*"

A categoria de desinformadores poderia também ter sido definida pelo atributo de contagem de desinformação, proporção de desinformação ou grau de centralidade de desinformação. Optamos por utilizar o atributo de rede pois nos interessa não somente a quantidade de desinformações compartilhadas, mas o alcance que elas tiveram, e esse atributo encapsula ambas informações. Ou seja, nossa definição de desinformadores engloba usuários que podem ser caracterizados como "espalhadores" devido ao alcance e frequência de suas ações, que causam mais danos que usuários crédulos de baixo alcance.

A Tabela 19 apresenta informações sobre os usuários categorizados como desinformadores. Pode-se observar o total de usuários pertencentes a categoria, a proporção de usuários dessa categoria comparando com o total de usuários do conjunto de dados, o total de desinformação compartilhada por estes usuários e a proporção dessa quantidade considerando toda desinformação no conjunto de dados.

Destaca-se que a categoria de desinformadores é formada por apenas 2,5% dos usuários, mas estes são responsáveis por um grande de volume do total de desinformação, alcançando quase 40%. Isso evidencia que a maior parte da desinformação é propagada por uma pequena quantidade de usuários desinformadores, agindo de forma maliciosa ou não, o que reforça a necessidade de identificar esses usuários como forma de mitigar a propagação da desinformação.

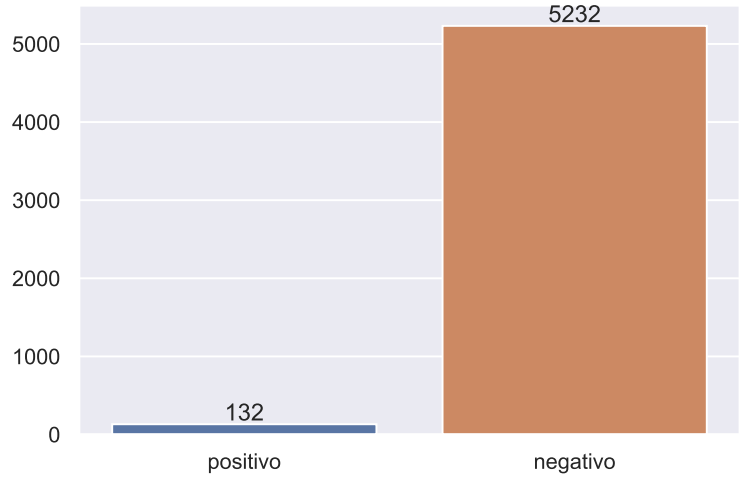
6.2 Experimentos de detecção de desinformadores

Conforme já discutido, a detecção de desinformadores é uma tarefa essencial para mitigar a propagação de desinformação. A partir da definição baseada em dados feita na seção anterior, foram realizados experimentos de classificação binária para identificar se um usuário é

um desinformador (positivo) ou não é desinformador (negativo). É importante ressaltar que o rótulo de desinformador só pode ser atribuído devido ao processo de rotulação das mensagens, que são utilizadas para calcular a métrica de força de desinformação que define a classe de um usuário. Porém, existem outros atributos de usuário que são conhecidos à priori, sem nenhum processo de rotulação manual, e estes podem ser utilizados para identificar os desinformadores.

Com esta definição, podemos observar o balanceamento de classes em nosso conjunto de dados de usuários, ilustrado na Figura 30. Observa-se que há um grande desbalanceamento entre as classes, onde a classe de desinformadores é minoritária, o que costuma aumentar a dificuldade de classificação, pois classificadores podem tender a reconhecer apenas os padrões da classe majoritária.

Figura 30 – Balanceamento entre as classes de usuário. Percebe-se que é um problema de classes extremamente desbalanceadas, onde a classe positiva, de desinformadores, é minoritária.



Fonte: o autor.

Para validar o desempenho das abordagens experimentadas, realizamos a separação aleatória de dados nos conjuntos de treino e de teste de forma estratificada. Ou seja, mantendo a proporção das classes em cada conjunto. Assim, o total de usuários é dividido, com 80% para o conjunto de treino e 20% para o conjunto de teste. A Tabela 20 apresenta a quantidade de dados de cada classe presentes em cada conjunto.

Nossa abordagem de classificação inicial parte da suposição de que existe uma forte correlação entre as variáveis de força de desinformação e força viral, uma vez que toda desinformação também é uma mensagem viral em nossa rotulação. De fato, ao analisar a correlação da força viral com outras variáveis que podem ser obtidas em dados não rotulados, a

Tabela 20 – Quantidade de dados negativos e positivos nas classes de treino e teste para detecção de desinformadores.

	Treino	Teste
Positivos (desinformadores)	106	26
Negativos (não-desinformadores)	4185	1047
Total	4291	1073

Fonte: o autor.

variável mais fortemente correlacionada é a força viral, com índice de correlação de 0,87. Assim, um usuário comprometido em divulgar e espalhar mensagens virais tem boas chances de espalhar mensagens rotuladas como desinformação. Utilizaremos essa informação como uma abordagem para detectar os desinformadores.

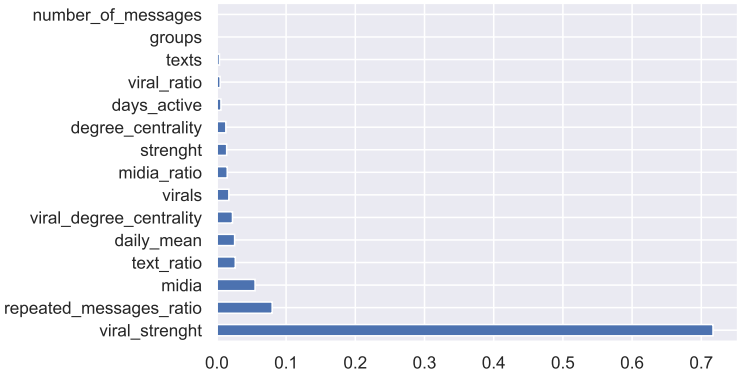
Nossa abordagem inicial é classificar todo usuário com valor igual ou acima do limiar de *outlier* no atributo de força viral como um desinformador, de forma análoga ao limiar calculado para a definição de desinformadores. Ou seja, todo usuário com valor anômalo de força viral, que dissemina mensagens virais em larga escala, é classificado como desinformador. O limiar observado nos dados é de 5.675.

A segunda abordagem utiliza o modelo de regressão logística. Os atributos de entrada são os atributos de usuário normalizados pelo método *z-score*, utilizando a média e variância do conjunto de treino. Um subconjunto dos atributos são selecionados pelo mesmo método de Árvore de Decisão, utilizado no Capítulo 5. O resultado da importância dos atributos é ilustrado na Figura 31. Observa-se que o atributo de maior importância é, de fato, a força viral. Mas outros atributos como a proporção de mensagens repetidas, a quantidade de mídia, a média de mensagens diárias e a quantidade de virais também adicionam informação para o classificador. Assim, os atributos escolhidos são estes 5 com maior importância.

Além da escolha dos atributos, outra etapa importante é a otimização do limiar de decisão. Devido ao desbalanceamento, o modelo pode tender a estimar probabilidades mais baixas para a classe positiva, portanto é necessário escolher um limiar de decisão adequado, que nesse caso pode ser inferior a 0,5. Para otimizar a escolha do limiar de decisão, utilizamos a mesma metodologia citada no Capítulo 5, obtendo o valor ótimo de acurácia em um subconjunto de validação, separado do conjunto de treino. Afim de evitar a inserção de ruído nos dados de treino, não foram utilizadas técnicas de sobreamostragem, de modo que o desbalanceamento é tratado pela escolha de um limiar de decisão adequado.

O resultado da classificação é apresentado na Tabela 21, utilizando as mesmas métri-

Figura 31 – Importância dos atributos na classificação de desinformadores.



Fonte: o autor.

Tabela 21 – Resultado da classificação de desinformadores a partir da limiarização da força viral e da regressão logística, com limiar de decisão de 0,24.

Método	**ACU**	**PRE**	**REC**	**F1**	**AUC**
Limiarização	0.994	0.916	0.846	0.879	-
Logit	**0.997**	**0.925**	**0.961**	**0.943**	0.998

Fonte: o autor

cas de desempenho já utilizadas para classificação de desinformação. Embora a acurácia não seja uma métrica adequada para este problema, uma vez que as classes são muito desbalanceadas, ela também foi apresentada. Observa-se que o método baseado em uma simples limiarização de um atributo, a partir da definição de *outliers* pela distância interquartil, já trouxe um resultado razoável em termos de precisão e *recall*, com um F1-score aproximado de 0,88. Aproximadamente 85% de todos os desinformadores do conjunto de teste foram identificados por esse método.

Os resultados obtidos pela regressão logística foram alcançados com um limiar de decisão de 0,24. Ou seja, são classificados como positivos os usuários que o modelo estima que possuam probabilidade maior do que 24% de serem positivos. O desempenho obtido pela regressão logística foi superior em todas as métricas, em particular no *recall*, onde aproximadamente 96% dos desinformadores foram identificados, e com uma alta precisão, o que significa uma baixa taxa de falso positivos.

Consideramos esse resultado satisfatório e ressaltamos que foi alcançado utilizando um simples classificador linear, indicando que o problema não é especialmente complexo para esses dados. Assim como no problema de detecção de desinformação, a regressão logística permite que técnicas de interpretabilidade sejam aplicadas tanto para compreensão do que o modelo aprendeu quanto das contribuições de cada atributo para predições individuais.

6.3 Limitações

Apesar do bom resultado obtido com a regressão logística na detecção dos desinformadores, algumas considerações devem ser feitas sobre as restrições desses métodos.

Nossa definição de desinformadores leva em consideração a distribuição de um atributo de usuário observado em nosso conjunto de dados. Essa distribuição, por sua vez, depende do tempo de coleta e da quantidade de grupos que foram observados. Por exemplo, se mais grupos fossem observados e durante um tempo mais prolongado, é possível que a distribuição da quantidade de desinformação publicadas pelos usuários fosse diferente, pois haveriam mais desinformadores e seria contabilizada mais desinformação, em mais grupos, durante um período maior, fazendo com que o limiar de *outlier* aumentasse. O mesmo vale para as distribuições dos atributos utilizados para classificação, o que dificultaria a generalização para conjuntos de dados maiores, para modelos supervisionados, como a regressão logística.

Um modelo treinado com um recorte menor de dados, quando aplicado em um conjunto de dados maior poderia tender classificar mais usuários como desinformadores, gerando uma maior quantidade de falso positivos e diminuindo sua precisão. Uma alternativa para mitigar esse possível problema seria utilizar somente atributos de proporção como entrada para o modelo. Outra possibilidade seria utilizar o conteúdo textual das mensagens publicadas pelos usuários como atributos de entrada, conforme feito por Braz e Goldschmidt (2018) para classificação de *bots*.

O método baseado em limiarização da força viral sofreria menos com esse problema, pois este é não-supervisionado, não necessitando de dados rotulados para treino. O limiar de decisão desse método é simplesmente ajustado de acordo com a distribuição dos dados observados. Embora tenha apresentado um desempenho menor, esse método é um bom ponto de partida para investigar um conjunto de dados de WhatsApp não rotulados.

Outro problema que surge seria a questão ética em uma possível aplicação. Uma vez que estamos realizando a classificação de usuários, presumivelmente seres humanos, é necessário cuidado quanto às ações que seriam tomadas quando o modelo identificasse uma conta como desinformadora. O bloqueio ou banimento de contas pode ser encarado como uma restrição à liberdade de expressão de usuários. Nesse contexto, um falso positivo possui um impacto maior quando comparado ao problema de detecção de desinformação, uma vez que o objeto da classificação são pessoas e não mensagens. É necessário que modelos que façam essa predição sejam interpretáveis e auditáveis, dando transparência ao processo. Além disso, é necessário

tratar possíveis vieses negativos sobre atributos sensíveis dos usuários.

6.4 Conclusão

Neste capítulo, apresentamos dois métodos, sendo um supervisionado e um não-supervisionado, que podem ser utilizadas para a detecção automática de desinformadores no WhatsApp. Os resultados experimentais indicam a viabilidade de ambos. As limitações dos métodos propostos e de suas possíveis aplicações foram discutidas e possíveis melhorias foram apontadas. Adicionalmente, propomos uma definição para o conceito de "desinformadores". Vale destacar ainda que a detecção automática de desinformadores é tão importante quanto a detecção da desinformação em si. Diferentes soluções para esses dois problemas podem ser utilizadas, separadamente ou em conjunto, com a finalidade de mitigar o espalhamento de desinformação nas redes sociais. O próximo capítulo conclui essa dissertação, revisitando os desafios apontados, as contribuições alcançadas e os trabalhos futuros.

7 CONCLUSÕES E TRABALHOS FUTUROS

Esta dissertação abordou os problemas de detecção de desinformação e de desinformadores no ambiente do WhatsApp. Com esta premissa, foi realizando um amplo estudo que envolveu a criação e disponibilização de um conjunto de dados anonimizados, coletados de grupos públicos de WhatsApp durante a campanha das eleições presidenciais brasileiras de 2018, o FakeWhatsApp.Br. Foi realizada a análise e rotulação desses dados, bem como a condução de uma série de experimentos de classificação. Buscamos assim traçar diretrizes acerca desses problemas, onde nossos experimentos e análises trouxeram novos conhecimentos e forneceram insumos para traçar hipóteses sobre as questões de pesquisa propostas:

Q1. *O quão desafiadora é a detecção de desinformação textual no WhatsApp utilizando técnicas de NLP e Aprendizado de Máquina Supervisionado?*

Foram realizados experimentos de classificação combinando atributos textuais, sociais e híbridos, com um classificador linear, a regressão logística, e um não linear, uma rede neural *MLP*, com o último realizando otimização de hiperparâmetros por busca aleatória. O melhor resultado obtido alcançou um *F1-score* de 0,82. Esse valor pode servir de *baseline* de desempenho para trabalhos futuros. O resultado obtido mostra que a detecção de desinformação robusta no WhatsApp ainda é um problema em aberto;

Q2. *Quais atributos podem ser extraídos para descrever o comportamento dos usuários no contexto do WhatsApp e como podem ser explorados para auxiliar a detecção de desinformação?*

Foram propostos e avaliados 22 atributos de usuários, incluindo os atributos que descrevem o comportamento do usuário acerca da quantidade e do tipo de mensagens enviadas, a proporção dos tipos de mensagens, o comportamento temporal e atributos de grafo que descrevem a centralidade e nível de atividade do usuário para cada tipo de mensagem em uma análise de rede. Porém, esses atributos não foram informativos para a detecção de desinformação, nem mesmo quando combinados com atributos textuais, piorando a performance da classificação;

Q3. *Quais combinação de métodos de pré-processamento, extração de atributos e algoritmos de classificação podem ser adequadamente explorados para a tarefa de detecção de desinformação no WhatsApp?*

Nossos experimentos mostraram que os atributos de conteúdo performaram consideravelmente melhor que os sociais e os híbridos. Em particular, os atributos *BoW* e *TF-IDF*,

utilizando bigramas e unigramas, foram melhores que os *word embeddings*. Em termos de classificadores, a regressão logística, um classificador linear, performou melhor que a *MLP*, um classificador não-linear, mesmo com a otimização de hiperparâmetros, o que indica um possível caso de *overfitting*. O melhor resultados foi alcançado com a regressão logística e atributos *TF-IDF*;

Q4. *Quais as limitações das melhores abordagens de detecção de desinformação avaliadas para responder a Q3?* Através de uma análise dos erros do melhor experimento de classificação, observou-se que a maioria das predições incorretas ocorriam em textos curtos, que frequentemente fazem referência a informações externas ao texto, como arquivos de mídia ou páginas Web. Logo, há uma lacuna de informação para o modelo, pois informações que possivelmente caracterizam a desinformação são ignoradas por este, que utiliza apenas o texto contido na mensagem. Esta é uma limitação relevante de abordagens baseadas em *NLP* para detecção de desinformação no WhatsApp, uma vez que uma parcela considerável de mensagens são desse tipo. Além disso, devido a diferença de distribuição de quantidade de palavras em cada classe, observou-se que o modelo teve uma tendência a classificar mensagens muito longas como desinformação, sendo esta uma fonte de falsos positivos. Porém, quando realizamos um experimento considerando somente mensagens longas, o F1 aumenta para aproximadamente 0,9. Finalmente, quando testamos utilizar nosso modelo para classificar mensagens acerca de outro contexto em outro período (mensagens sobre a pandemia do Covid-19), observamos uma queda considerável de desempenho, com F1 de 0,63, tornando evidente a necessidade de treinamento contínuo para métodos baseados em conteúdo;

Q5. *Que atributos e métodos podem ser adequadamente explorados para detecção de usuários desinformadores no contexto do WhatsApp?* Embora não tenham se mostrado úteis para a detecção de desinformação, os atributos sociais propostos nesse trabalho foram fundamentais para a definição e detecção de desinformadores no contexto do WhatsApp. Observou-se que o atributo de grafo denominado força viral, que indica o volume e alcance de desinformação propagada por um usuário, tem uma alta correlação com o atributo de força de desinformação, sendo portanto um importante preditor de desinformadores. Foram avaliados um método não-supervisionado, baseado na detecção de *outliers* em força viral e um método supervisionado, treinando um modelo de regressão logística, onde também foram utilizados atributos de proporção de mensagens repetidas, quantidade de

mídia, média de mensagens diárias e a quantidade de mensagens virais. O último método obteve um bom desempenho, com F1 de 0,95.

Além das questões de pesquisa, as seguintes inovações deste trabalho podem ser destacadas:

- A criação e validação do FakeWhatsApp.Br. Até onde sabemos, é o primeiro conjunto de dados coletado do WhatsApp em português brasileiro que é público, rotulado e com dados de propagação;
- A proposta de um conjunto de atributos sociais para descrever o comportamento dos usuários no contexto do WhatsApp;
- A proposta de uma definição de desinformadores no contexto da WhatsApp baseada em dados e a detecção desses usuários nesta definição.

7.1 Trabalhos futuros

Os achados desta dissertação desenham múltiplas possibilidades de pesquisas futuras, com estes ou novos dados que sejam coletados do WhatsApp. Em particular, destaca-se que os dados coletados pela plataforma Farol Digital (SÁ *et al.*, 2021), que teve contribuição direta desta pesquisa, solucionam os problemas de falhas de coletas encontrados no FakeWhatsApp.Br. Esta plataforma permite obter bases de dados maiores e mais completas, permitindo também a coleta dos arquivos de mídia, que não estavam presentes neste trabalho. Dentre os trabalhos futuros, elencamos:

- Explorar técnicas de análise de grafos, como detecção de comunidades e análise de fluxo, e novas métricas como PageRank, coeficiente de clusterização, *betweenness*, reciprocidade e assortatividade, para auxiliar a descrever e detectar desinformação e desinformadores. Explorar diferentes modelagens de grafos, como a utilizada por Nobre *et al.* (2022), na qual existe uma aresta entre usuários caso eles tenham compartilhado o mesmo conteúdo;
- Explorar a criação de *embeddings* de grafos, como DeepWalk (PEROZZI *et al.*, 2014), para representar usuários. Essa representação poderia ser utilizada tanto para auxiliar na detecção de desinformação como na detecção de desinformadores. Outra alternativa seria o uso de *Graph Neural Networks* / Redes Neurais para Grafos (GNN);
- Análises linguísticas mais aprofundadas das particularidades e modos de expressão observados no WhatsApp, em particular comparando quantitativamente com textos de outras plataformas como Twitter, Youtube e Facebook. Realizar análises linguísticas comparando

as classes de desinformação e não-desinformação;

- Detecção de desinformação utilizando os arquivos de mídia, através de classificação de imagens, vídeo, áudio ou de classificação multi-tarefa, extraindo e combinando atributos de diferentes tipo de mídia e de texto;
- Explorar técnicas de análise de sentimentos para detecção de desinformação;
- Enriquecimento semântico dos textos através da criação de uma base de conhecimento que permita associar mais informação à mensagens curtas. Essa base poderia ser construída através de uma coleta de fontes como a Wikipedia ou de artigos de agências de checagem de fatos;
- Explorar técnicas mais aprofundadas de interpretabilidade como LIME (RIBEIRO *et al.*, 2016) e SHAPE (LUNDBERG; LEE, 2017);
- Explorar métodos que utilizem a informação temporal das mensagens, como modelos epidemiológicos, ou a representação da mensagem como uma série temporal dos atributos de usuários que interagiram como a mesma;
- Explorar o conteúdo textual publicado por usuários para detecção de desinformadores;
- Estudo de métodos semi-automáticos para reduzir o custo de rotular dados para atualização contínua de modelos. Explorar o uso de *active learning* (SETTLES, 2009). Explorar métodos baseados em *crowdsourcing*. Considerar em uma nova rotulação o uso de multiclasses que expressem diferentes graus ou tipos de desinformação;
- Por fim, pesquisar e desenvolver técnicas de mitigação da propagação de desinformação, como o uso de *bots* educativos.

REFERÊNCIAS

ABDIN, L. Bots and fake news: the role of whatsapp in the 2018 brazilian presidential election. **Casey Robertson**, v. 41, n. 1, 2019.

AGÊNCIA BRASIL. **WhatsApp é principal fonte de informação do brasileiro, diz pesquisa**. 2019. Disponível em: https://agenciabrasil.ebc.com.br/geral/noticia/2019-12/whatsapp-e-principal-fonte-de-informacao-do-brasileiro-diz-pesquisa. Acesso em: 17 mai. 2021.

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. **Journal of economic perspectives**, v. 31, n. 2, p. 211–36, 2017.

AN, G. Literature review for deception detection. **The City University of New York, Report**, 2015.

BANAJI, S.; BHAT, R.; AGARWAL, A.; PASSANHA, N.; PRAVIN, M. S. Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india. Department of Media and Communications, London School of Economics and . . . , 2019.

BARBERÁ, P.; JOST, J. T.; NAGLER, J.; TUCKER, J. A.; BONNEAU, R. Tweeting from left to right: Is online political communication more than an echo chamber? **Psychological science**, Sage Publications Sage CA: Los Angeles, CA, v. 26, n. 10, p. 1531–1542, 2015.

BARBOSA, J.; VIEIRA, J. P. A.; SANTOS, R.; JUNIOR, G. V. M.; MUNIZ, M. d. S.; MOURA, R. S. Introdução ao processamento de linguagem natural usando python. **III Escola Regional de Informatica do Piauí**, v. 1, p. 336–360, 2017.

BENEVENUTO, F.; RODRIGUES, T.; ALMEIDA, V.; ALMEIDA, J.; GONÇALVES, M. Detectando usuários maliciosos em interações via vídeos no youtube. In: **Proceedings of the 14th Brazilian Symposium on Multimedia and the Web**. [*S. l.*: *s. n.*], 2008. p. 138–145.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **The Journal of Machine Learning Research**, JMLR. org, v. 13, n. 1, p. 281–305, 2012.

BISHOP, C. M. Pattern recognition. **Machine learning**, v. 128, n. 9, 2006.

BRATU, S. *et al.* The fake news sociology of covid-19 pandemic fear: Dangerously inaccurate beliefs, emotional contagion, and conspiracy ideation. **Linguistic and Philosophical Investigations**, Addleton Academic Publishers, n. 19, p. 128–134, 2020.

BRAZ, P. A.; GOLDSCHMIDT, R. R. Redes neurais convolucionais na detecção de bots sociais: Um método baseado na clusterização de mensagens textuais. In: SBC. **Anais do XVIII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais**. [*S. l.*], 2018. p. 323–336.

BUCKELS, E. E.; TRAPNELL, P. D.; PAULHUS, D. L. Trolls just want to have fun. **Personality and individual Differences**, Elsevier, v. 67, p. 97–102, 2014.

CABRAL, L.; MONTEIRO, J. M.; SILVA, J. W. F. da; MATTOS, C. L.; MOURAO, P. J. C. Fakewhastapp. br: Nlp and machine learning techniques for misinformation detection in brazilian portuguese whatsapp messages. 2021.

CHARLES, A. C.; SAMPAIO, J. de O. Checking fake news on web browsers: an approach using collaborative datasets. In: **Workshop on Big Social Data and Urban Computing**. [*S. l.*: *s. n.*], 2018.

CHENG, J.; BERNSTEIN, M.; DANESCU-NICULESCU-MIZIL, C.; LESKOVEC, J. Anyone can become a troll: Causes of trolling behavior in online discussions. In: **Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing**. [*S. l.*: *s. n.*], 2017. p. 1217–1230.

CHU, Z.; GIANVECCHIO, S.; WANG, H.; JAJODIA, S. Who is tweeting on twitter: human, bot, or cyborg? In: **Proceedings of the 26th annual computer security applications conference**. [*S. l.*: *s. n.*], 2010. p. 21–30.

CONROY, N. J.; RUBIN, V. L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. **Proceedings of the Association for Information Science and Technology**, v. 52, n. 1, p. 1–4, 2015. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010082.

CORDEIRO, P. R.; PINHEIRO, V. Um corpus de notícias falsas do twitter e verificação automática de rumores em lingua portuguesa. In: **STIL-Brazilian Symposium in Information and Human Language Technology. IEEE, Salvaldor, BA, Brazil**. [*S. l.*: *s. n.*], 2019. p. 220–228.

CRESCI, S.; PIETRO, R. D.; PETROCCHI, M.; SPOGNARDI, A.; TESCONI, M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In: **Proceedings of the 26th international conference on world wide web companion**. [*S. l.*: *s. n.*], 2017. p. 963–972.

DAVIS, R. A.; LII, K.-S.; POLITIS, D. N. Remarks on some nonparametric estimates of a density function. In: **Selected Works of Murray Rosenblatt**. [*S. l.*]: Springer, 2011. p. 95–100.

DENG, H.; RUNGER, G. Feature selection via regularized trees. In: IEEE. **The 2012 International Joint Conference on Neural Networks (IJCNN)**. [*S. l.*], 2012. p. 1–8.

ÉPOCA NEGÓCIOS. **WhatsApp diz como tenta combater fake news no Brasil**. 2018. Disponível em: https://epocanegocios.globo.com/Tecnologia/noticia/2018/10/whatsapp-diz-como-tenta-combater-fake-news-no-brasil.html. Acesso em: 31 jul. 2021.

EXAME. **Em guerra contra fake news, WhatsApp diz que reenvios em massa caíram 70%**. 2020. Disponível em: https://exame.com/tecnologia/whatsapp-anuncia-reducao-de-70-em-encaminhamento-de-mensagens/. Acesso em: 31 jul. 2021.

FALLIS, D. A functional analysis of disinformation. **iConference 2014 Proceedings**, iSchools, 2014.

Faustini, P.; Covões, T. Fake news detection using one-class classification. In: **2019 8th Brazilian Conference on Intelligent Systems (BRACIS)**. [*S. l.*: *s. n.*], 2019. p. 592–597. ISSN 2643-6256.

FREIRE, P.; GOLDSCHMIDT, R. Combatendo fake news nas redes sociais via crowd signals implícitos. In: **Anais do XVI Encontro Nacional de Inteligência Artificial e**

Computacional. Porto Alegre, RS, Brasil: SBC, 2019. p. 424–435. ISSN 0000-0000. Disponível em: https://sol.sbc.org.br/index.php/eniac/article/view/9303.

FREIRE, P. M.; GOLDSCHMIDT, R. Uma introdução ao combate automático às fake news em redes sociais virtuais. **Sociedade Brasileira de Computação**, 2019.

GAGLANI, J.; GANDHI, Y.; GOGATE, S.; HALBE, A. Unsupervised whatsapp fake news detection using semantic search. In: IEEE. **2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)**. [*S. l.*], 2020. p. 285–289.

GALHARDI, C. P.; FREIRE, N. P.; MINAYO, M. C. d. S.; FAGUNDES, M. C. M. Fato ou fake? uma análise da desinformação frente à pandemia da covid-19 no brasil. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 25, p. 4201–4210, 2020.

GALLOTTI, R.; VALLE, F.; CASTALDO, N.; SACCO, P.; DOMENICO, M. D. Assessing the risks of 'infodemics' in response to covid-19 epidemics. **Nature Human Behaviour**, Nature Publishing Group, v. 4, n. 12, p. 1285–1293, 2020.

GARIMELLA, K.; TYSON, G. Whatsapp, doc? a first look at whatsapp public group data. **arXiv preprint arXiv:1804.01473**, 2018.

GRANIK, M.; MESYURA, V. Fake news detection using naive bayes classifier. In: IEEE. **2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)**. [*S. l.*], 2017. p. 900–903.

GUACHO, G. B.; ABDALI, S.; SHAH, N.; PAPALEXAKIS, E. E. Semi-supervised content-based detection of misinformation via tensor embeddings. In: IEEE. **2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)**. [*S. l.*], 2018. p. 322–325.

GUO, B.; DING, Y.; YAO, L.; LIANG, Y.; YU, Z. **The Future of Misinformation Detection: New Perspectives and Trends**. 2019.

GUO, C.; CAO, J.; ZHANG, X.; SHU, K.; YU, M. Exploiting emotions for fake news detection on social media. **arXiv preprint arXiv:1903.01728**, 2019.

HAMDI, T.; SLIMI, H.; BOUNHAS, I.; SLIMANI, Y. A hybrid approach for fake news detection in twitter based on user features and graph embedding. In: SPRINGER. **International conference on distributed computing and internet technology**. [*S. l.*], 2020. p. 266–280.

HAYKIN, S. **Neural Networks and Learning Machines, 3/E**. [*S. l.*]: Pearson Education India, 2010.

INDUMATHI, J.; GITANJALI, J. The avant-garde ways to prevent the whatsapp fake news. In: **Emerging Research in Data Engineering Systems and Computer Communications**. [*S. l.*]: Springer, 2020. p. 487–498.

INSIGHT), N. N. de Marketing e C. Whatsapp, instagram e youtube são os apps mais usados na pandemia. **Exame**, 2020. Disponível em: https://exame.com/tecnologia/whatsapp-instagram-e-youtube-sao-os-apps-mais-usados-na-pandemia/.

INTERVOZES. **Desinformação: ameaça ao direito à comunicação muito além das fake news. Disponível**. 2019. Disponível em: http://intervozes.org.br/publicacoes/desinformacao-ameaca-ao-direito-a-comunicacao-muito-alem-das-fake-news/. Acesso em: 29 jun. 2021.

JIN, F.; DOUGHERTY, E.; SARAF, P.; CAO, Y.; RAMAKRISHNAN, N. Epidemiological modeling of news and rumors on twitter. In: **Proceedings of the 7th workshop on social network mining and analysis**. [*S. l.*: *s. n.*], 2013. p. 1–9.

JINDAL, N.; LIU, B. Opinion spam and analysis. In: **Proceedings of the 2008 international conference on web search and data mining**. [*S. l.*: *s. n.*], 2008. p. 219–230.

JOLLIFFE, I. Principal component analysis. **Encyclopedia of statistics in behavioral science**, Wiley Online Library, 2005.

KAUFMAN, S.; ROSSET, S.; PERLICH, C.; STITELMAN, O. Leakage in data mining: Formulation, detection, and avoidance. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, ACM New York, NY, USA, v. 6, n. 4, p. 1–21, 2012.

KSHETRI, N.; VOAS, J. The economics of "fake news". **IT Professional**, IEEE, v. 19, n. 6, p. 8–12, 2017.

LAZER, D. M. J.; BAUM, M. A.; BENKLER, Y.; BERINSKY, A. J.; GREENHILL, K. M.; MENCZER, F.; METZGER, M. J.; NYHAN, B.; PENNYCOOK, G.; ROTHSCHILD, D.; SCHUDSON, M.; SLOMAN, S. A.; SUNSTEIN, C. R.; THORSON, E. A.; WATTS, D. J.; ZITTRAIN, J. L. The science of fake news. **Science**, American Association for the Advancement of Science, v. 359, n. 6380, p. 1094–1096, 2018. ISSN 0036-8075. Disponível em: https://science.sciencemag.org/content/359/6380/1094.

LEE, K.; EOFF, B. D.; CAVERLEE, J. Seven months with the devils: A long-term study of content polluters on twitter. In: **Fifth international AAAI conference on weblogs and social media**. [*S. l.*: *s. n.*], 2011.

LEITE, M. A. G. L.; GUELPELI, M. V. C.; SANTOS, C. Q. Um modelo baseado em regras para a detecção de bots no twitter. In: SBC. **Anais do IX Brazilian Workshop on Social Network Analysis and Mining**. [*S. l.*], 2020. p. 37–48.

LÊU, M. de O.; MORAIS, D. M. G. de; XAVIER, F.; DIGIAMPIETRI, L. A. Detecção automática de bots em redes sociais: um estudo de caso no segundo turno das eleições presidenciais brasileiras de 2018. **Revista de Sistemas de Informação da FSMA**, v. 24, p. 31–39, 2019.

LI, Q.; ZHANG, Q.; SI, L. Rumor detection by exploiting user credibility information, attention and multi-task learning. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [*S. l.*: *s. n.*], 2019. p. 1173–1179.

LIN, P.; SONG, Q.; SHEN, J.; WU, Y. Discovering graph patterns for fact checking in knowledge graphs. In: SPRINGER. **International Conference on Database Systems for Advanced Applications**. [*S. l.*], 2018. p. 783–801.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: **Proceedings of the 31st international conference on neural information processing systems**. [*S. l.*: *s. n.*], 2017. p. 4768–4777.

MA, J.; GAO, W.; WEI, Z.; LU, Y.; WONG, K.-F. Detect rumors using time series of social context information on microblogging websites. In: **Proceedings of the 24th ACM international on conference on information and knowledge management**. [*S. l.*: *s. n.*], 2015. p. 1751–1754.

MAALEJ, Z. **Discourse Studies**, Sage Publications, Ltd., v. 3, n. 3, p. 376–378, 2001. ISSN 14614456, 14617080. Disponível em: http://www.jstor.org/stable/24047513.

MACHADO, C.; KIRA, B.; HIRSCH, G.; MARCHAL, N.; KOLLANYI, B.; HOWARD, P. N.; LEDERER, T.; BARASH, V. News and political information consumption in brazil: Mapping the first round of the 2018 brazilian presidential election on twitter. **The computational propaganda project. Algorithms, automation and digital politics. https://comprop. oii. ox. ac. uk/research/brazil2018**, 2018.

MACHADO, C.; KIRA, B.; NARAYANAN, V.; KOLLANYI, B.; HOWARD, P. A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. In: . New York, NY, USA: Association for Computing Machinery, 2019. (WWW '19), p. 1013–1019. ISBN 9781450366755. Disponível em: https://doi.org/10.1145/3308560.3316738.

MARTINS, A. D. F.; CABRAL, L.; MOURÃO, P. J. C.; MONTEIRO, J. M.; MACHADO, J. Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages. In: SPRINGER. **International Conference on Applications of Natural Language to Information Systems**. [*S. l.*], 2021. p. 199–206.

MIHAILIDIS, P.; VIOTTY, S. Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. **American behavioral scientist**, SAGE Publications Sage CA: Los Angeles, CA, v. 61, n. 4, p. 441–454, 2017.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MITRA, T.; GILBERT, E. Credbank: A large-scale social media corpus with associated credibility annotations. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [*S. l.*: *s. n.*], 2015. v. 9, n. 1.

MOHSENI, S.; RAGAN, E.; HU, X. Open issues in combating fake news: Interpretability as an opportunity. **arXiv preprint arXiv:1904.03016**, 2019.

MONTEIRO, R.; SANTOS, R.; PARDO, T.; ALMEIDA, T.; RUIZ, E.; VALE, O. Contributions to the study of fake news in portuguese: New corpus and automatic detection results: 13th international conference, propor 2018, canela, brazil, september 24–26, 2018, proceedings. In: ____. [*S. l.*: *s. n.*], 2018. p. 324–334. ISBN 978-3-319-99721-6.

MONTEIRO, R. A.; SANTOS, R. L.; PARDO, T. A.; ALMEIDA, T. A. D.; RUIZ, E. E.; VALE, O. A. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [*S. l.*], 2018. p. 324–334.

MORAES, M. P.; SAMPAIO, J. de O.; CHARLES, A. C. Data mining applied in fake news classification through textual patterns. In: **Proceedings of the 25th Brazillian Symposium on Multimedia and the Web**. [*S. l.*: *s. n.*], 2019. p. 321–324.

MORENO, J.; BRESSAN, G. Factck.br: A new dataset to study fake news. In: **Anais do XXV Simpósio Brasileiro de Multimídia e Web**. Porto Alegre, RS, Brasil: SBC, 2019. p. 525–527. Disponível em: https://sol.sbc.org.br/index.php/webmedia/article/view/8073.

MOURÃO, P. J. C. **A República do Ódio ou do Método Etnodata**. 2020. Disponível em: https://alicenews.ces.uc.pt/index.php?lang=1&id=30865. Acesso em: 29 jul. 2021.

NAKASHOLE, N.; MITCHELL, T. M. Language-aware truth assessment of fact candidates. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 1009–1019. Disponível em: https://aclanthology.org/P14-1095.

NEWMAN, N.; FLETCHER, R.; SCHULZ, A.; ANDI, S.; NIELSEN, R.-K. Reuters institute digital news report 2020. **Report of the Reuters Institute for the Study of Journalism**, 2020.

NICKERSON, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. **Review of general psychology**, SAGE Publications Sage CA: Los Angeles, CA, v. 2, n. 2, p. 175–220, 1998.

NOBRE, G. P.; FERREIRA, C. H.; ALMEIDA, J. M. A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp. **Information Processing & Management**, Elsevier, v. 59, n. 1, p. 102757, 2022.

NOVAK, P. K.; SMAILOVIĆ, J.; SLUBAN, B.; MOZETIČ, I. Sentiment of emojis. **PloS one**, Public Library of Science San Francisco, CA USA, v. 10, n. 12, p. e0144296, 2015.

ORGANIZATION, W. H. *et al.* Who public health research agenda for managing infodemics. World Health Organization, 2021.

ORLOV, M.; LITVAK, M. Using behavior and text analysis to detect propagandists and misinformers on twitter. In: SPRINGER. **Annual International Symposium on Information Management and Big Data**. [*S. l.*], 2018. p. 67–74.

PAN, J. Z.; PAVLOVA, S.; LI, C.; LI, N.; LI, Y.; LIU, J. Content based fake news detection using knowledge graphs. In: VRANDEČIĆ, D.; BONTCHEVA, K.; SUÁREZ-FIGUEROA, M. C.; PRESUTTI, V.; CELINO, I.; SABOU, M.; KAFFEE, L.-A.; SIMPERL, E. (Ed.). **The Semantic Web – ISWC 2018**. Cham: Springer International Publishing, 2018. p. 669–683. ISBN 978-3-030-00671-6.

PARISER, E. **The filter bubble: How the new personalized web is changing what we read and how we think**. [*S. l.*]: Penguin, 2011.

PAUL, C.; MATTHEWS, M. The russian "firehose of falsehood" propaganda model. **Rand Corporation**, JSTOR, v. 2, n. 7, p. 1–10, 2016.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEROZZI, B.; AL-RFOU, R.; SKIENA, S. Deepwalk: Online learning of social representations. In: **Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining**. [*S. l.*: *s. n.*], 2014. p. 701–710.

POSETTI, J.; MATTHEWS, A. A short guide to the history of'fake news' and disinformation. **International Center for Journalists**, v. 7, p. 1–19, 2018.

POTTHAST, M.; KIESEL, J.; REINARTZ, K.; BEVENDORFF, J.; STEIN, B. A stylometric inquiry into hyperpartisan and fake news. **arXiv preprint arXiv:1702.05638**, 2017.

QAZVINIAN, V.; ROSENGREN, E.; RADEV, D.; MEI, Q. Rumor has it: Identifying misinformation in microblogs. In: **Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing**. [*S. l.*: *s. n.*], 2011. p. 1589–1599.

QIAN, F.; GONG, C.; SHARMA, K.; LIU, Y. Neural user response generator: Fake news detection with collective user intelligence. In: **IJCAI**. [*S. l.*: *s. n.*], 2018. v. 18, p. 3834–3840.

QIU, X.; OLIVEIRA, D. F.; SHIRAZI, A. S.; FLAMMINI, A.; MENCZER, F. Limited individual attention and online virality of low-quality information. **Nature Human Behaviour**, Nature Publishing Group, v. 1, n. 7, p. 0132, 2017.

QUANDT, T.; FRISCHLICH, L.; BOBERG, S.; SCHATTO-ECKRODT, T. Fake news. **The international encyclopedia of journalism studies**, John Wiley & Sons, Inc. Hoboken, NJ, USA, p. 1–6, 2019.

RESENDE, G.; MELO, P.; SOUSA, H.; MESSIAS, J.; VASCONCELOS, M.; ALMEIDA, J.; BENEVENUTO, F. (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In: . [*S. l.*: *s. n.*], 2019.

RESENDE, G.; MESSIAS, J.; SILVA, M.; ALMEIDA, J.; VASCONCELOS, M.; BENEVENUTO, F. A system for monitoring public political groups in whatsapp. In: **Proceedings of the 24th Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: Association for Computing Machinery, 2018. (WebMedia '18), p. 387–390. ISBN 9781450358675. Disponível em: https://doi.org/10.1145/3243082.3264662.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Model-agnostic interpretability of machine learning. **arXiv preprint arXiv:1606.05386**, 2016.

ROSENFELD, A.; SINA, S.; SARNE, D.; AVIDOV, O.; KRAUS, S. A study of whatsapp usage patterns and prediction models without message content. **arXiv preprint arXiv:1802.03393**, 2018.

ROSS, L.; WARD, A. *et al.* Naive realism in everyday life: Implications for social conflict and misunderstanding. **Values and knowledge**, v. 103, p. 135, 1996.

ROTTER, J. B. Interpersonal trust, trustworthiness, and gullibility. **American psychologist**, American Psychological Association, v. 35, n. 1, p. 1, 1980.

RUBIN, V. L.; CHEN, Y.; CONROY, N. K. Deception detection for news: three types of fakes. **Proceedings of the Association for Information Science and Technology**, Wiley Online Library, v. 52, n. 1, p. 1–4, 2015.

RUBIN, V. L.; LUKOIANOVA, T. Truth and deception at the rhetorical structure level. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 66, n. 5, p. 905–917, 2015.

RUBIN, V. L.; VASHCHILKO, T. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. In: **Proceedings of the Workshop on Computational Approaches to Deception Detection**. [*S. l.*: *s. n.*], 2012. p. 97–106.

RUCHANSKY, N.; SEO, S.; LIU, Y. Csi: A hybrid deep model for fake news detection. In: **Proceedings of the 2017 ACM on Conference on Information and Knowledge Management**. [*S. l.*: *s. n.*], 2017. p. 797–806.

RUIZ, E. E. S. Fake news detection on fake. br using hierarchical attention networks. In: SPRINGER NATURE. **Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings**. [*S. l.*], 2020. v. 12037, p. 143.

SÁ, I. C. de; MONTEIRO, J. M.; SILVA, J. W. F. da; MEDEIROS, L. M.; MOURAO, P. J. C.; CUNHA, L. C. C. da. Digital lighthouse: A platform for monitoring public groups in whatsapp. 2021.

SANTIA, G.; WILLIAMS, J. Buzzface: A news veracity dataset with facebook user commentary and egos. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [*S. l.*: *s. n.*], 2018. v. 12, n. 1.

SANTOS, B. L.; FERREIRA, G. E.; BRAZ, R. R.; DIGIAMPIETRI, L. A. *et al.* Comparação de algoritmos para detecção de bots sociais nas eleições presidenciais no brasil em 2018 utilizando características do usuário. **Revista Brasileira de Computação Aplicada**, v. 13, n. 1, p. 53–64, 2021.

SCHUSTER, T.; SCHUSTER, R.; SHAH, D. J.; BARZILAY, R. The limitations of stylometry for detecting machine-generated fake news. **Computational Linguistics**, MIT Press, n. Just Accepted, p. 1–18, 2020.

SETTLES, B. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences, 2009.

SHI, B.; WENINGER, T. Fact checking in large knowledge graphs: A discriminative predict path mining approach. **arXiv preprint arXiv:1510.05911**, 2015.

SHU, K.; BERNARD, H. R.; LIU, H. Studying fake news via network analysis: detection and mitigation. In: **Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining**. [*S. l.*]: Springer, 2019. p. 43–65.

SHU, K.; SLIVA, A.; WANG, S.; TANG, J.; LIU, H. **Fake News Detection on Social Media: A Data Mining Perspective**. 2017.

SHU, K.; WANG, S.; LIU, H. Understanding user profiles on social media for fake news detection. In: IEEE. **2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)**. [*S. l.*], 2018. p. 430–435.

SHU, K.; WANG, S.; LIU, H. Beyond news contents: The role of social context for fake news detection. In: **Proceedings of the twelfth ACM international conference on web search and data mining**. [*S. l.*: *s. n.*], 2019. p. 312–320.

SILVA, F. R. M. da; FREIRE, P. M. S.; SOUZA, M. P. de; PLENAMENTE, G. de A. B.; GOLDSCHMIDT, R. R. Fakenewssetgen: A process to build datasets that support comparison among fake news detection methods. In: **Proceedings of the Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: Association for Computing Machinery, 2020. (WebMedia '20), p. 241–248. ISBN 9781450381963. Disponível em: https://doi.org/10.1145/3428658.3430965.

SILVA, R. M.; SANTOS, R. L.; ALMEIDA, T. A.; PARDO, T. A. Towards automatically filtering fake news in portuguese. **Expert Systems with Applications**, Elsevier, v. 146, p. 113199, 2020.

SU, Q.; WAN, M.; LIU, X.; HUANG, C.-R. Motivations, methods and metrics of misinformation detection: An nlp perspective. **Natural Language Processing Research**, v. 1, p. 1–13, 2020. ISSN 2666-0512. Disponível em: https://doi.org/10.2991/nlpr.d.200522.001.

SUNDAR, S. S. There'sa psychological reason for the appeal of fake news. **New Republic**, 2016.

TCHECHMEDJIEV, A.; FAFALIOS, P.; BOLAND, K.; GASQUET, M.; ZLOCH, M.; ZAPILKO, B.; DIETZE, S.; TODOROV, K. Claimskg: A knowledge graph of fact-checked claims. In: SPRINGER. **International Semantic Web Conference**. [*S. l.*], 2019. p. 309–324.

UDDIN, M. M.; IMRAN, M.; SAJJAD, H. Understanding types of users on twitter. **arXiv preprint arXiv:1406.1335**, 2014.

VEDOVA, M. L. D.; TACCHINI, E.; MORET, S.; BALLARIN, G.; DIPIERRO, M.; ALFARO, L. de. Automatic online fake news detection combining content and social signals. In: IEEE. **2018 22nd Conference of Open Innovations Association (FRUCT)**. [*S. l.*], 2018. p. 272–279.

VERLEYSEN, M.; FRANÇOIS, D. The curse of dimensionality in data mining and time series prediction. In: SPRINGER. **International work-conference on artificial neural networks**. [*S. l.*], 2005. p. 758–770.

VICARIO, M. D.; VIVALDO, G.; BESSI, A.; ZOLLO, F.; SCALA, A.; CALDARELLI, G.; QUATTROCIOCCHI, W. Echo chambers: Emotional contagion and group polarization on facebook. **Scientific reports**, Nature Publishing Group, v. 6, n. 1, p. 1–12, 2016.

VOSOUGHI, S.; MOHSENVAND, M. N.; ROY, D. Rumor gauge: Predicting the veracity of rumors on twitter. **ACM transactions on knowledge discovery from data (TKDD)**, ACM New York, NY, USA, v. 11, n. 4, p. 1–36, 2017.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. **Science**, v. 359, p. 1146–1151, 03 2018.

WALFISH, S. A review of statistical outlier methods. **Pharmaceutical technology**, ASTER PUBLISHING CORPORATION, v. 30, n. 11, p. 82, 2006.

WANG, P.; ANGARITA, R.; RENNA, I. Is this the era of misinformation yet: Combining social bots and fake news to deceive the masses. In: **Companion Proceedings of the The Web Conference 2018**. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018. (WWW '18), p. 1557–1561. ISBN 9781450356404. Disponível em: https://doi.org/10.1145/3184558.3191610.

WANG, W. Y. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 422–426. Disponível em: https://aclanthology.org/P17-2067.

WATERLOO, S. F.; BAUMGARTNER, S. E.; PETER, J.; VALKENBURG, P. M. Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp. **new media & society**, SAGE Publications Sage UK: London, England, v. 20, n. 5, p. 1813–1831, 2018.

WEI, D.; DENG, X.; ZHANG, X.; DENG, Y.; MAHADEVAN, S. Identifying influential nodes in weighted networks based on evidence theory. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 392, n. 10, p. 2564–2575, 2013.

WU, L.; LIU, H. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In: **Proceedings of the eleventh ACM international conference on Web Search and Data Mining**. [*S. l.*: *s. n.*], 2018. p. 637–645.

YANG, S.; SHU, K.; WANG, S.; GU, R.; WU, F.; LIU, H. Unsupervised fake news detection on social media: A generative approach. In: **Proceedings of the AAAI conference on artificial intelligence**. [*S. l.*: *s. n.*], 2019. v. 33, n. 01, p. 5644–5651.

ZERVOPOULOS, A.; ALVANOU, A. G.; BEZAS, K.; PAPAMICHAIL, A.; MARAGOUDAKIS, M.; KERMANIDIS, K. Hong kong protests: Using natural language processing for fake news detection on twitter. In: SPRINGER. **IFIP International Conference on Artificial Intelligence Applications and Innovations**. [*S. l.*], 2020. p. 408–419.

ZHANG, Y.; HARA, T. A probabilistic model for malicious user and rumor detection on social media. In: **HICSS**. [*S. l.*: *s. n.*], 2020. p. 1–10.

ZHOU, X.; ZAFARANI, R. Fake news: A survey of research, detection methods, and opportunities. **arXiv preprint arXiv:1812.00315**, v. 2, 2018.

ZHOU, Z.; GUAN, H.; BHAT, M. M.; HSU, J. Fake news detection via nlp is vulnerable to adversarial attacks. **ArXiv**, abs/1901.09657, 2019.

ZUBIAGA, A.; LIAKATA, M.; PROCTER, R.; HOI, G. W. S.; TOLMIE, P. Analysing how people orient to and spread rumours in social media by looking at conversational threads. **PloS one**, Public Library of Science San Francisco, CA USA, v. 11, n. 3, p. e0150989, 2016.

ZUCKERMAN, M.; DEPAULO, B. M.; ROSENTHAL, R. Verbal and nonverbal communication of deception. In: **Advances in experimental social psychology**. [*S. l.*]: Elsevier, 1981. v. 14, p. 1–59.