Universidade Federal do ABC
Centro de Matemática, Computação & Cognição
Bacharelado em Ciência da Computação

Revisão de abordagens baseadas em sumarização extrativa, segmentação textual e classificação factual para apoio à detecção de desinformação em postagens de mídias sociais

Carlos Henrique Alencar Lima

Santo André - SP, novembro de 2025

Carlos Henrique Alencar Lima

Revisão de abordagens baseadas em sumarização extrativa, segmentação textual e classificação factual para apoio à detecção de desinformação em postagens de mídias sociais

Projeto de Graduação em Computação apresentado como parte dos requisitos necessários para a obtenção do Título de Bacharel em Ciência da Computação.

Universidade Federal do ABC – UFABC

Centro de Matemática, Computação & Cognição

Bacharelado em Ciência da Computação

Orientadora: Profª Drª Denise Hideko Goya

Coorientadora: Drª Patrícia Dias dos Santos

Santo André - SP

novembro de 2025

Resumo

A crescente disseminação de desinformação em ambientes digitais, aliada à velocidade com que conteúdos circulam em plataformas e redes sociais, transformou-se em um desafio central para a sociedade contemporânea. Neste contexto, o trabalho analisa empiricamente um conjunto de processos de pré-tratamento e filtragem textual, em particular, sumarização automática, segmentação guiada por conjunções e detecção de afirmações factuais, aplicados à análise de desinformação. No primeiro estágio, três modelos de sumarização baseada em *transformers* (BERT-base, DistilBERT e SBERT-MiniLM) foram comparados quanto à capacidade de reduzir ruído preservando trechos factuais. Os resultados mostram que todos produziram resumos em torno de 25%–27% do texto original, mas o SBERT-MiniLM apresentou melhor equilíbrio entre retenção de tópicos centrais, similaridade semântica e custo computacional, enquanto o DistilBERT se mostrou mais conservador na redução de conteúdo factual. Em seguida, avaliou-se uma estratégia de segmentação baseada na conjunção "que", aplicada a notícias de checagem e a *clusters* de *tweets*. Essa abordagem recuperou conjuntos maiores e mais pertinentes de trechos do que termos-chave definidos manualmente e superou métodos baseados apenas em aspas, que retornaram poucos segmentos relevantes. Por fim, investigou-se o uso de modelos para identificação automática de afirmações factuais, com o XLM-R-Large-*ClaimDetection* e um classificador SVM atuando como filtros sobre grandes volumes de texto. O XLM-R-Large obteve acurácia de 0,88 em português e elevado *recall* para sentenças factuais importantes, enquanto o SVM apresentou desempenho moderado, contribuindo para reduzir o volume de conteúdo a ser inspecionado manualmente. Ademais, esses processos permitiram comparar diferentes estratégias de rotulagem de desinformação em *tweets* e retuítes, evidenciando o *trade-off* entre métodos mais conservadores, que se aproximam do padrão manual, e abordagens mais assertivas, que ampliam a cobertura de conteúdo potencialmente desinformativo ao custo de maior risco de super-rotulagem. Os resultados obtidos oferecem subsídios para o aperfeiçoamento de ferramentas computacionais voltadas ao enfrentamento de usos maliciosos de dados textuais.

Palavras-chaves: Desinformação, sumarização automática, extração de termos-chave, segmentação textual, aprendizado de máquina.

Abstract

The growing spread of disinformation in digital environments, together with the speed at which content circulates on platforms and social networks, has become a central challenge for contemporary society. In this context, this work empirically analyzes a set of text preprocessing and filtering procedures, in particular, automatic summarization, conjunction-based segmentation, and factual claim detection, applied to disinformation analysis. In the first stage, three transformer-based summarization models (BERT-base, DistilBERT, and SBERT-MiniLM) were compared in terms of their ability to reduce noise while preserving factual segments. The results show that all models produced summaries of about 25%–27% of the original text, but SBERT-MiniLM achieved the best balance between topic retention, semantic similarity, and computational cost, whereas DistilBERT behaved more conservatively in the reduction of factual content. Next, a segmentation strategy based on the conjunction "that" was evaluated on fact-checking news articles and on *tweet* clusters. This approach recovered larger and more pertinent sets of segments than manually defined keywords and outperformed quote-based methods, which returned few relevant fragments. Finally, we investigated the use of models for automatic detection of factual claims, with XLM-R-Large-*ClaimDetection* and an SVM classifier acting as filters over large text collections. XLM-R-Large achieved an accuracy of 0.88 in Portuguese with high recall for important factual sentences, while the SVM model obtained moderate performance, helping to reduce the amount of content that requires manual inspection. Taken together, these processes enabled the comparison of different labeling strategies for disinformation in tweets and retweets, highlighting the trade-off between more conservative methods, closer to manual annotation, and more assertive approaches that expand coverage at the cost of a higher risk of over-labeling. The results provide input for improving computational tools aimed at countering malicious uses of textual data.

Keywords: Disinformation, automatic summarization, key term extraction, text segmentation, machine learning.

Sumário

1 Introdução

A criação e disseminação da desinformação não são fenômenos recentes. Desde os tempos em que a comunicação se dava por jornais impressos, rádios, televisões ou até mesmo pela transmissão oral, já existiam tentativas de manipular ou distorcer fatos com diferentes intenções, podendo ser políticas, econômicas ou sociais. No entanto, o avanço tecnológico transformou profundamente esse cenário. A internet e, especialmente, as redes sociais, ampliaram exponencialmente o alcance e a velocidade com que conteúdos são compartilhados.

Com a popularização e o barateamento do acesso à internet, a sociedade passou a viver em um ambiente globalizado e hiperconectado, no qual o fluxo de informações é constante e instantâneo. Plataformas como *WhatsApp*, *Messenger*, *blogs* e portais de notícias tornaram-se canais centrais para o consumo e a troca de conteúdo. Essa democratização da informação, embora traga benefícios, também intensificou o desafio de distinguir o que é verdadeiro do que é manipulador ou enganoso. O excesso de informações, algumas provenientes de fontes jornalísticas verificadas e outras de origem duvidosa, cria um ambiente em que o discernimento crítico se torna fundamental para evitar a propagação da desinformação e para possibilitar a identificação de termos, expressões e padrões linguísticos associados a esse fenômeno.

A situação torna-se ainda mais preocupante à medida que empresas, governos e até indivíduos começam a gerar e disseminar informações por meio de redes sociais para benefícios próprios. Desse modo, analisar e filtrar todo dado recebido, destacando termos e construções mais recorrentes em conteúdos desinformativos, pode-se tornar demorado e inviável, já que, aproximadamente, 6 bilhões de pessoas, o que representa cerca de 74% da população mundial, possuem acesso à internet de acordo com os dados levantados pela União Internacional de Telecomunicações e representado na Figura 1, ou seja, são possíveis disseminadores de desinformação, mesmo que não intencionalmente. Para lidar com tal situação, além da implementação de ferramentas capazes de discernir quanto às mensagens transmitidas, é necessário que haja um auxílio governamental de modo a limitar o compartilhamento de informações não verificadas, bem como incentivar o desenvolvimento de métodos de extração de termos e indícios linguísticos que ajudem a mapear e compreender a difusão da desinformação.

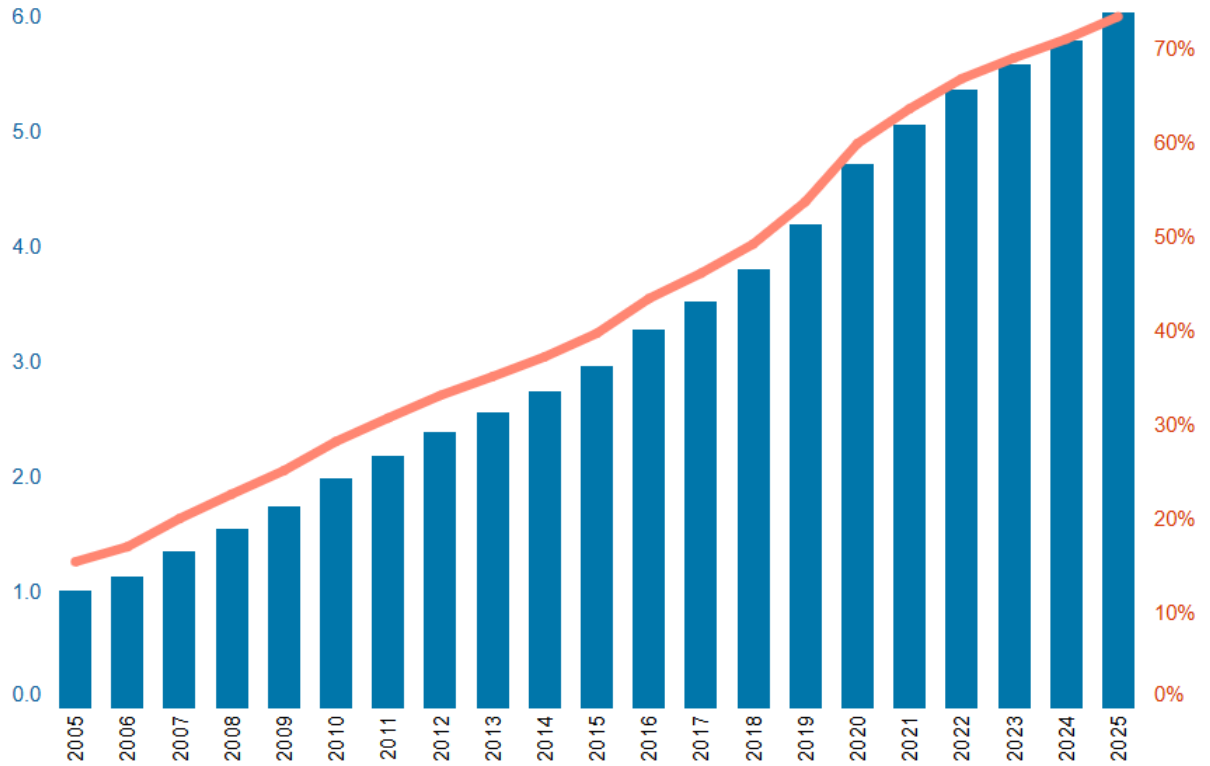

Figura 1 – Indivíduos usando a internet - Número de pessoas em bilhões por ano.

Fonte: (ITU, 2025)

Além do enfoque computacional, a disseminação de desinformação *online* apresenta um desafio para diversas outras áreas científicas, sendo uma delas a área científica que envolve questões psicológicas. Muitas vezes, a dificuldade para diferenciar a verdade da ficção está relacionada à falta de precisão com que os indivíduos refletem sobre determinado assunto, podendo utilizar menos raciocínio caso a notícia confirme suas próprias ideias (PENNYCOOK; RAND, 2021). Dessa forma, é notório que ocorre uma manipulação emocional para incentivar certos tipos de comportamentos que beneficiam o emissor de tais desinformações, tendo como exemplo manipulações com finalidades políticas capazes de proliferarem comportamentos agressivos e preconceituosos (SOUZA, 2019). Portanto, é possível notar que a vulnerabilidade do pensamento não racional humano pode ser explorada para influenciar comportamentos. Ao mesmo tempo, essa vulnerabilidade pode servir de indício para identificar padrões linguísticos e extrair termos característicos de mensagens manipulativas, já que conteúdos manipulativos tendem a conter informações que exploram o lado sentimental dos leitores.

A desordem informacional tornou-se um elemento estruturante da comunicação digital contemporânea, marcada pela crescimento acelerada de conteúdos verdadeiros, falsos e distorcidos em plataformas como o *Twitter* (SILVA, 2020). Durante a pandemia de COVID-19, essa dinâmica se intensificou e configurou a chamada infodemia, entendida como uma superabundância de informações que dificulta o reconhecimento de orientações

confiáveis (Organização Pan-Americana da Saúde (OPAS/OMS), 2020).

A classificação proposta por Wardle e Derakhshan (WARDLE, 2017), posteriormente ampliada pela UNESCO (IRETON; POSETTI, 2018), estabelece três formas principais de distorção, informação incorreta, desinformação e má-informação, que ajudam a compreender a complexidade do fenômeno. No Brasil, o Tribunal Superior Eleitoral passou a empregar o termo "desinformação" de modo abrangente para abarcar diversas estratégias de manipulação de conteúdo, não restritas ao formato jornalístico característico das chamadas "*fake news*" (Tribunal Regional Eleitoral de São Paulo, 2023).

Pesquisas empíricas reforçam essa preocupação ao demonstrar que conteúdos enganosos tendem a se espalhar com maior velocidade e alcance do que informações verificadas (VOSOUGHI; ROY; ARAL, 2018; CINELLI et al., 2020), evidenciando que a lógica de funcionamento das redes sociais favorece sua difusão. Diante desse cenário, este trabalho adota "desinformação" como eixo central de análise, alinhando-se à literatura e às instituições que tratam o tema como parte de um desafio amplo para a comunicação pública e a governança democrática (GARCÍA-SAISÓ et al., 2021).

Com base na análise dos métodos de avaliação de veracidade atual, duas abordagens principais se destacam: a Abordagem Linguística e a Abordagem de Redes. A Abordagem Linguística foca na análise do conteúdo de mensagens enganosas, identificando padrões de linguagem que indicam engano, como o uso de pronomes, conjunções e palavras de emoção negativa. Por outro lado, a Abordagem de Redes utiliza informações de rede, como metadados de mensagens ou consultas estruturadas em redes de conhecimento, para calcular medidas agregadas de engano. (CONROY; RUBIN; CHEN, 2015) Ambas as abordagens geralmente incorporam técnicas de aprendizado de máquina para treinar classificadores adaptados à análise. No entanto, dado o foco deste trabalho, a Abordagem de Redes não será explorada, tendo ênfase na exploração das técnicas de Processamento de Linguagem Natural (PLN) para a detecção de desinformação intencionalmente falsa.

Também é essencial a elaboração de uma definição para fato. Para este estudo, será considerado uma afirmação ou proposição objetiva que pode ser verificada como verdadeira ou falsa com base em evidências disponíveis, sem depender de julgamentos subjetivos como relevância, importância editorial ou interpretações pessoais. Para ser classificada como fato, a afirmação deve ser suficientemente clara, específica e independente de contextos variáveis que possam influenciar sua interpretação. A verificabilidade de um fato depende da existência de métodos ou dados concretos que permitam sua avaliação por diferentes indivíduos ou organizações de maneira consistente e replicável (KONSTANTINOVSKIY et al., 2021).

1.1 Objetivos

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é extrair, de forma automatizada, termos factuais presentes em notícias relevantes com conteúdos verificados por "agências" como Aos Fatos, Fatos ou fake, Uol confere e Estadão Verifica com conteúdos envolvendo urnas eletrônicas no Brasil, de modo a otimizar o processo de classificação desses conteúdos. Para isso, serão utilizadas técnicas de PLN, como modelos de *transformers*, BERT, SVM e *regex*. A partir dessa análise, serão apontados fatores que podem contribuir para o avanço da área de PLN e para a melhoria das ferramentas de detecção de desinformação.

1.1.2 Objetivos Específicos

Com o objetivo de delinear e alcançar o objetivo geral proposto, os seguintes objetivos específicos foram estabelecidos:

- Analisar modelos relevantes por meio de revisão dos modelos de PLN usados em detecção de fatos relevantes.
- Aplicar modelos em dados verificados e ajustá-los para analisar textos em Português e adaptar para o contexto de desinformação.
- Definir e utilizar métricas apropriadas, como precisão, *recall* e *F1-score*, para avaliar o desempenho dos modelos aplicados.
- Coletar, analisar e avaliar os resultados obtidos, utilizando métricas de desempenho e parametrizando diferentes configurações.
- Comparar os resultados com pesquisas existentes, identificar diferenças.
- Sugerir possíveis melhorias e avanços com base nas comparações realizadas.

1.2 Justificativa

De acordo com uma pesquisa realizada pela *PR Newswire*, aproximadamente 65% das gerações *Millennials* e Z preferem se comunicar com mais frequência por meio de ambientes digitais do que pessoalmente (FRANZESE, 2017). Este dado revela uma transformação significativa nos padrões de interação social, impulsionada pela popularização dos *smartphones*, redes sociais e aplicativos de mensagens. Nesse contexto, essas gerações estão mais expostas a conteúdos que podem ser potencialmente falsos, especialmente na forma de textos. A ampla exposição a informações digitais e a velocidade com que elas se disseminam tornam a identificação de desinformação um desafio ainda maior.

Outro fator agravante é o fato de que, apesar de possuírem a obrigação de se atentarem à veracidade das notícias, nem mesmo especialistas estão isentos de acreditarem em informações falsas. Em 2013, em um *tweet* da agência de notícias Associated Press, foi dito que houve uma suposta explosão que teria ferido Barack Obama, o que levou a uma perda de 130 bilhões de dólares em ações, tal valor sendo recuperado rapidamente após ser anunciado que a notícia era falsa (RAPOZA, 2017). Entretanto, mesmo com a recuperação econômica rápida, pode-se notar que a disseminação de desinformação tem impactos em toda economia de um país. Desse modo, é inegável que a propagação de informação na internet deve receber atenção de autoridades públicas, empresariais e da comunidade científica.

Diante dessa realidade, é de extrema importância o desenvolvimento de ferramentas que possam tornar esses ambientes digitais mais seguros e menos suscetíveis à desinformação. A revisão e aprimoramento de técnicas para mitigar a disseminação de desinformação são essenciais para superar os desafios atuais. Em particular, o avanço das ferramentas de PLN desempenha um papel fundamental nesse contexto, já que permitem facilitar e otimizar processos que, sem elas, seriam mais custosos e menos precisos.

O processo de checagem de fatos em uma organização pode ser dividido em quatro etapas principais: (1) monitorar a mídia, capturando conteúdos como artigos, vídeos e imagens; (2) detectar afirmações verificáveis; (3) verificar as afirmações por meio de pesquisa detalhada; e (4) publicar os resultados das verificações. No entanto, a maior parte da pesquisa científica tem-se concentrado na etapa de verificação da veracidade, também conhecido como *checking claims*, com estudos rotulados como detecção de *fake news* ou desinformação, colocando em segundo plano etapas iniciais igualmente essenciais, como a detecção de afirmações (KONSTANTINOVSKIY et al., 2021).

Observa-se que grande parte das pesquisas em detecção automática de desinformação permanece centrada na formulação do problema como uma tarefa de classificação supervisionada, em geral binária ou multiclasse, aplicada diretamente à veracidade das notícias (OSHIKAWA; QIAN; WANG, 2020). Em contrapartida, etapas intermediárias do *pipeline* de checagem, como a identificação sistemática de afirmações verificáveis, a seleção de trechos relevantes e a otimização dos modelos de linguagem utilizados nessa filtragem inicial, ainda recebem atenção mais limitada na literatura, surgindo apenas mais recentemente como objetos de estudo específicos em tarefas de *claim detection* e normalização de afirmações (KONSTANTINOVSKIY et al., 2021). Essa lacuna torna-se ainda mais evidente no contexto de línguas de poucos recursos, como o português, em que há escassez de recursos anotados e de estudos sistemáticos voltados à detecção de desinformação e às etapas prévias de seleção de conteúdo. Portanto, é notória a necessidade de investigar precisamente esse elo intermediário do processo, propondo e avaliando estratégias de PLN para detecção automática de afirmações em português, com o objetivo de tornar o fluxo de

fact-checking mais eficiente, escalável e alinhado às necessidades práticas de organizações de checagem de fatos.

Automatizar essas etapas iniciais, especialmente a detecção de afirmações, é fundamental para alimentar o processo de verificação com entradas relevantes e gerenciar o volume de informações. Sem uma lista bem apurada de afirmações verificáveis, a etapa de determinação da veracidade não pode funcionar de forma eficaz. (KONSTANTINOVSKIY et al., 2021) Portanto, integrar ferramentas automatizadas pode tornar o processo mais eficiente, reduzir a carga de trabalho dos verificadores e aumentar a qualidade e a agilidade na disseminação de informações confiáveis.

2 Fundamentação Teórica

O avanço das tecnologias de Processamento de Linguagem Natural (PLN) e de Aprendizado de Máquina (*Machine Learning-ML*) tem transformado profundamente a forma como se compreende, organiza e interpreta grandes volumes de informação textual. Essas áreas, situadas na intersecção entre a Linguística Computacional (LC) e a Inteligência Artificial (IA), fornecem as bases para o desenvolvimento de sistemas capazes de compreender o conteúdo semântico de textos, extrair padrões linguísticos e realizar tarefas de classificação, sumarização e detecção de desinformação. Neste capítulo, são abordados os principais conceitos e fundamentos teóricos que sustentam as técnicas utilizadas nesta pesquisa, com ênfase nas metodologias de extração de termos factuais e nos modelos de classificação textual aplicados ao combate à disseminação de desinformação nos meios digitais.

2.1 Processamento de Linguagem Natural

O PLN é uma área da Inteligência Artificial dedicada ao desenvolvimento de métodos que possibilitam que computadores compreendam e manipulem a linguagem humana de forma automatizada (JURAFSKY; MARTIN, 2025). O PLN combina aspectos da linguística, estatística e ciência da computação, permitindo o tratamento de dados textuais em larga escala e a extração de informações de maneira sistemática. Entre suas principais tarefas estão a tokenização, remoção de *stopwords*, lematização, *stemming*, análise sintática e semântica, reconhecimento de entidades nomeadas (*NER*) e análise de sentimentos.

No contexto desta pesquisa, o PLN é utilizado para o pré-processamento de textos, etapa que inclui a normalização textual, a segmentação de sentenças e a identificação de termos relevantes. Essas técnicas garantem que o texto seja representado de maneira estruturada, possibilitando a aplicação de algoritmos de *ML* e aprendizado profundo. A capacidade do PLN de identificar padrões linguísticos e semânticos possui grande relevância para a detecção de desinformação, uma vez que esse fenômeno se manifesta por meio de estruturas discursivas específicas, escolhas lexicais tendenciosas e manipulação intencional de contextos.

2.2 Representação de Texto

A representação de texto constitui um dos pilares do PLN, pois define a forma como as palavras e sentenças são convertidas em formatos compreensíveis para os algoritmos

computacionais. Entre os métodos mais tradicionais, destacam-se o modelo de *Bag of Words* (*BoW*) e o *Term Frequency - Inverse Document Frequency* (*TF-IDF*) (SALTON; WONG; YANG, 1975).

O modelo *BoW* representa cada documento como um vetor em $\mathbb{R}^V$, em que V é o tamanho do vocabulário, e cada componente corresponde tipicamente à frequência (ou presença/ausência) de um termo no documento. Nessa representação, assume-se independência entre termos e ignora-se completamente a ordem das palavras, de modo que dois textos com o mesmo multiconjunto de termos terão a mesma representação vetorial (HACOHEN-KERNER; MILLER; YIGAL, 2020).

Já o *TF-IDF* aprimora o modelo *BoW* ao ponderar a importância de cada termo com base em sua frequência no documento e na raridade no corpus (MANNING; SCHÜTZE, 1999; SALTON; WONG; YANG, 1975).

Formalmente, seja f_{ik} a frequência do termo k no documento i (com $i = 1, \ldots, n$ documentos e vocabulário indexado por k). Seguindo o modelo vetorial clássico de Salton, Wong e Yang (SALTON; WONG; YANG, 1975), a componente de frequência de termo (*term frequency*, TF) é dada por

$$\mathrm{TF}_{ik} = f_{ik}. \tag{2.1}$$

Seja ainda d_k o número de documentos em que o termo k aparece (*document frequency*) e n o número total de documentos da coleção. A componente de frequência inversa de documento (*inverse document frequency*, IDF) pode ser escrita como

$$\mathrm{IDF}_k = \lceil \log_2(n) \rceil - \lceil \log_2(d_k) \rceil + 1, \tag{2.2}$$

de modo que o peso TF-IDF atribuído ao termo k no documento i é

$$w_{ik} = \text{TF-IDF}_{ik} = f_{ik} \cdot \mathrm{IDF}_k. \tag{2.3}$$

Com o avanço da área, surgiram métodos mais sofisticados de representação vetorial, conhecidos como *word embeddings*, que mapeiam palavras em vetores contínuos em um espaço de alta dimensionalidade. Técnicas como *Word2Vec* (MIKOLOV et al., 2013), *GloVe* (PENNINGTON; SOCHER; MANNING, 2014) e *FastText* (BOJANOWSKI et al., 2017) capturam relações semânticas e contextuais entre palavras com base em sua coocorrência. Os *embeddings* contextuais, derivados de modelos mais complexos como *BERT* e *SBERT*, passaram a representar o significado das palavras considerando o contexto em que aparecem, oferecendo resultados superiores em tarefas semânticas complexas.

2.3 Sumarização Extrativa

A sumarização extrativa de texto é uma técnica de PLN que visa gerar versões reduzidas de documentos, preservando as informações mais relevantes e essenciais para o

entendimento do conteúdo original. Dentro dessa abordagem, as sentenças mais significativas são selecionadas diretamente do texto-fonte, sem a necessidade de reformulação ou geração de novas frases, ou seja, apenas reutiliza trechos existentes no texto original.

A tarefa pode ser vista como um problema de seleção de unidades textuais (tipicamente sentenças) a partir de um documento ou de um conjunto de documentos, de modo a maximizar a cobertura do conteúdo informativo e minimizar a redundância, produzindo um sumário curto, coerente e informativo. Segundo Nenkova e McKeown, a sumarização automática consiste justamente em selecionar e integrar o conteúdo-chave de um ou mais textos em uma saída condensada, sendo as abordagens extrativas a forma mais tradicional e amplamente estudada (NENKOVA; MCKEOWN, 2011).

Do ponto de vista metodológico, pesquisas como a de Gupta e Lehal (2010) mostram que as técnicas extrativas podem ser agrupadas em abordagens baseadas em características superficiais do texto (frequência de termos, posição da sentença), em modelos de aprendizado supervisionado que classificam sentenças como relevantes ou não e em métodos baseados em grafos, que avaliam a importância de uma sentença pela sua centralidade em uma rede de similaridades (GUPTA; LEHAL, 2010).

Já para o contexto da língua portuguesa, Costa e Martins (2015) comparam sistematicamente diferentes estratégias de sumarização automática extrativa, incluindo abordagens de frequência, centróides e métodos grafo-baseados, e mostram que, mesmo sem geração de novas frases, essas técnicas são capazes de produzir sumários competitivos, desde que combinadas com boas estratégias de seleção e avaliadas por métricas apropriadas (COSTA; MARTINS, 2015).

Portanto, a sumarização extrativa representa uma ferramenta valiosa para a redução de textos longos, facilitando a compreensão rápida de conteúdos postados em redes sociais ou de notícias, sem perder a essência informativa. No presente estudo, essa técnica foi empregada para identificar e extrair sentenças factuais de notícias, contribuindo para a análise e classificação de informações relevantes no combate à desinformação.

2.4 Aprendizado de Máquina na Classificação Textual

O Aprendizado de Máquina (*ML*) é uma subárea da IA voltada para o desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e realizar previsões ou classificações sem intervenção humana direta (MITCHELL, 1997). No contexto do PLN, o *ML* é amplamente utilizado para classificar textos, identificar tópicos, analisar sentimentos e detectar desinformação.

Os modelos tradicionais de *ML* aplicados a textos incluem algoritmos supervisionados como *Support Vector Machines* (*SVM*), *Naive Bayes* e *Random Forests*. Esses

métodos dependem de representações vetoriais do texto, como *BoW* e *TF-IDF*, para extrair características numéricas. O *SVM*, em particular, tem se mostrado eficiente na separação de classes lineares e foi utilizado neste estudo para a classificação de sentenças factuais e não factuais, alcançando resultados satisfatórios em acurácia, embora com limitações quanto à relevância prática das classificações.

Para a avaliação dos modelos de *ML*, métricas como precisão, *recall*, *F1-score* e acurácia são amplamente utilizadas. A precisão mede a proporção de classificações corretas entre as predições positivas. O *recall* avalia a capacidade do modelo em identificar todas as instâncias relevantes. O *F1-score* combina ambas as métricas, oferecendo uma medida harmônica de desempenho. Tais métricas são fundamentais para compreender a eficiência dos algoritmos de classificação textual aplicados à detecção de desinformação.

2.5 Aprendizado Profundo (*Deep Learning*)

O Aprendizado Profundo (*Deep Learning - DL*) representa uma evolução do *ML*, caracterizando-se pela utilização de redes neurais artificiais com múltiplas camadas de processamento. Essas redes são capazes de modelar relações complexas e não lineares nos dados, tornando-se particularmente eficazes em tarefas envolvendo linguagem natural.

As Redes Neurais Recorrentes (*RNNs*), especialmente suas variantes *LSTM* (*Long Short-Term Memory*) e *GRU* (*Gated Recurrent Unit*), foram amplamente empregadas para capturar dependências temporais em sequências textuais. No entanto, apesar de seu sucesso, essas arquiteturas apresentam limitações quanto à paralelização e ao tratamento de contextos longos.

Essas limitações foram superadas com o surgimento dos *Transformers* (VASWANI et al., 2017), uma arquitetura baseada em mecanismos de atenção que permite o aprendizado bidirecional do contexto textual, conforme ilustrado na Figura 2. Modelos derivados, como *BERT* (*Bidirectional Encoder Representations from Transformers*), *DistilBERT*, *SBERT* (*Sentence-BERT*) e *XLM-RoBERTa*, aprimoraram o PLN ao oferecer representações contextuais mais detalhadas e precisas. No presente estudo, esses modelos foram empregados para sumarização e extração de termos factuais, com destaque para o *SBERT*, que apresentou melhor equilíbrio entre desempenho e tempo de processamento.

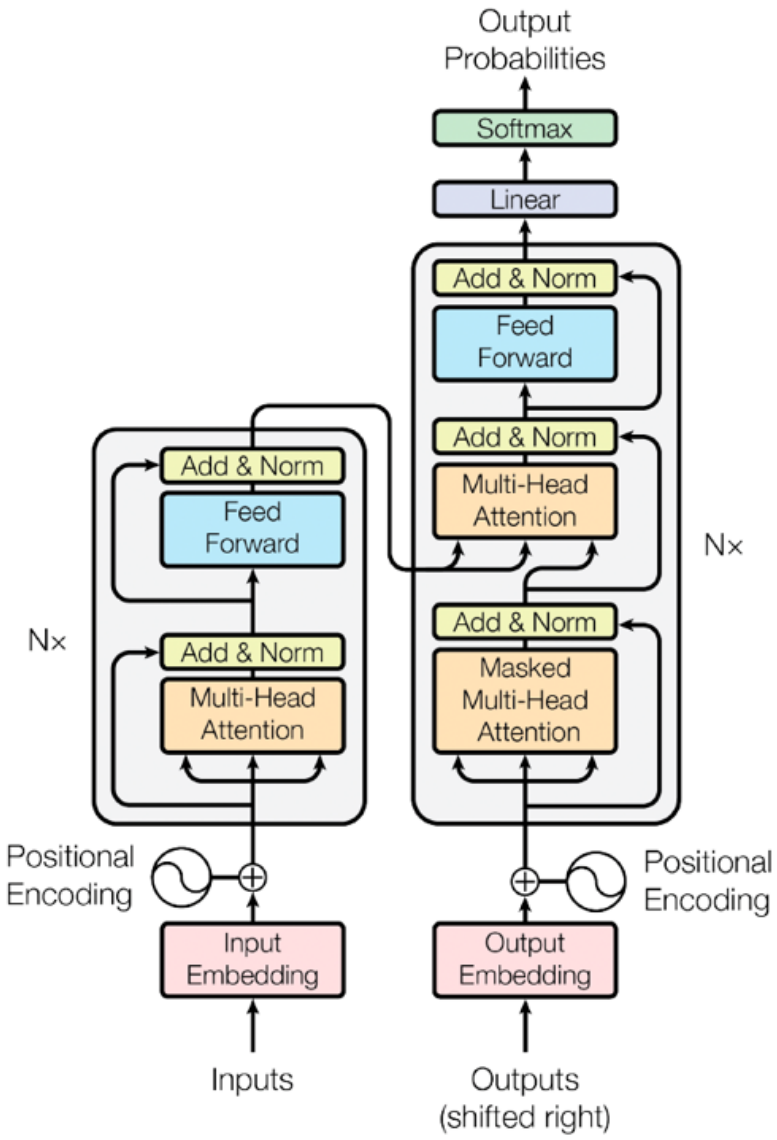

Figura 2 – *Transformer - Modelo de Arquitetura*

Fonte: (VASWANI et al., 2017)

A capacidade dos modelos baseados em *Transformers* de compreender o contexto global das sentenças e aprender relações semânticas complexas os torna ferramentas fundamentais para o combate à desinformação, especialmente em ambientes multilíngues, como o português.

Além disso, a formulação original dos *Transformers* parte da ideia de que uma sequência pode ser decomposta em um conjunto de *tokens* sobre os quais se aplica, de forma iterativa, um bloco composto por autoatenção *multi-head* (*MHSA*) e uma rede neural *perceptron* multicamada (*MLP*) posicionada por *token*. Em cada bloco, o mesmo vetor de entrada x_i é projetado em três espaços lineares distintos, gerando $q_i = x_i W_Q$, $k_i = x_i W_K$ e $v_i = x_i W_V$. O parâmetro q_i passa a representar a demanda informacional do *token*, o parâmetro k_i os atributos sob os quais esse *token* pode ser recuperado pelos

demais, e o parâmetro v_i o conteúdo semântico efetivamente agregável.

A *MLP* posicionada por *token*, definida por

$$\mathrm{MLP}(x) = \phi(xW_1 + b_1)W_2 + b_2,$$

em que $\phi(\cdot)$ denota uma função de ativação não linear (tipicamente *ReLU* ou *GELU*). Explicitando essa operação por *token* i, temos $h_i = \phi(\tilde{x}_i W_1 + b_1)$ e $z_i = h_i W_2 + b_2$, com posterior aplicação da conexão residual e da normalização em camadas (*Layer Normalization*), de modo que $x'_i = \mathrm{LayerNorm}(\tilde{x}_i + z_i)$ (VASWANI et al., 2017).

A autoatenção calcula, para cada posição, coeficientes de ponderação assimétricos a partir dos produtos escalares direcionados $q_i k_j^\top$, o que permite selecionar, de todo o enunciado, justamente os *tokens* mais relevantes e, assim, modelar dependências de longo alcance sem recorrer à recorrência. Apenas os valores são combinados $z_i = \sum_j \alpha_{ij} v_j$ e o resultado não substitui a representação original: ele é reinjetado por meio de uma conexão residual e estabilizado por *Layer Normalization*, garantindo que cada camada aprenda apenas um refinamento suave da representação anterior e que o empilhamento profundo não degrade o contexto (TURNER, 2024).

Essa combinação de atenção global, paralelizável e ligada por conexões residuais explica por que *Transformers* superam *RNNs* e Redes Neurais Convolucionais *CNNs* em tarefas de PLN que exigem contexto amplo e comparações cruzadas entre *tokens*.

2.6 Extração de Termos-Chave

A extração de termos-chave é um processo com potencial significativo de otimizar o pré-processamento textual, cujo objetivo é identificar automaticamente as palavras ou expressões mais representativas de um texto. Essa atividade permite resumir o conteúdo, facilitar a indexação e apoiar processos de classificação e análise semântica.

As abordagens tradicionais baseiam-se em métodos estatísticos, como *TF-IDF*, *RAKE* (*Rapid Automatic Keyword Extraction*), *TextRank* (MIHALCEA; TARAU, 2004) e *YAKE* (*Yet Another Keyword Extractor*). Tais métodos consideram a frequência e coocorrência de termos, oferecendo soluções rápidas e interpretáveis, embora limitadas no entendimento de contexto.

Em contrapartida, abordagens modernas utilizam embeddings semânticos e modelos supervisionados para capturar significados contextuais. Técnicas baseadas em *sentence embeddings*, como as fornecidas pelo SBERT (REIMERS; GUREVYCH, 2019), avaliam a similaridade de cosseno entre frases e palavras, permitindo identificar termos de maior relevância sem depender exclusivamente de frequência estatística. Além disso, modelos supervisionados e híbridos, que combinam embeddings com classificadores como SVMs

ou redes neurais, apresentam resultados expressivos em tarefas de extração semântica de palavras-chave (UMAIR; SULTANA; LEE, 2024).

2.7 Desinformação

A desinformação é um fenômeno multifacetado que se refere à disseminação intencional de informações falsas ou enganosas com o objetivo de manipular percepções e comportamentos (ZHOU; ZAFARANI, 2020). No ambiente digital, sua propagação é potencializada pela velocidade das redes sociais e pela personalização algorítmica, que cria bolhas informacionais e amplia o impacto de narrativas falsas.

É importante ter em mente o tipo específico de informação que será tratado neste trabalho, uma vez que o fenômeno da desordem informacional representa uma das expressões mais complexas da crise contemporânea da comunicação. Conforme analisa Silva (SILVA, 2020), a desordem informacional manifesta-se no ambiente digital como um caos comunicacional, em que conteúdos verdadeiros, falsos e distorcidos se misturam em alta velocidade, especialmente nas redes sociais como o *Twitter*. Essa dinâmica, acelerada pela pandemia de COVID-19, caracteriza a chamada infodemia, marcada pelo excesso de informações, algumas corretas, outras enganosas, que dificultam a distinção entre fontes confiáveis e conteúdos manipulados e, ao mesmo tempo, oferecem um campo fértil para a identificação de termos e expressões que sinalizam diferentes tipos de distorção. Como sintetizado na Figura 3, essas categorias se sobrepõem e ajudam a diferenciar situações em que há apenas circulação de informação incorreta daquelas em que há uso estratégico da informação para produzir dano.

Figura 3 – "Desordem da informação".

Fonte: (IRETON; POSETTI, 2018)

De acordo com a tipologia proposta por Wardle e Derakhshan (WARDLE, 2017), a desordem informacional se estrutura em três dimensões complementares:

- Informação Incorreta: quando informações falsas são compartilhadas sem intenção de causar dano;
- Desinformação: quando há intenção deliberada de enganar ou prejudicar;
- Má-informação: quando informações verdadeiras são utilizadas com o objetivo de causar dano.

Esses três eixos foram incorporados e ampliados no manual da UNESCO (IRETON; POSETTI, 2018), intitulado *Journalism, "Fake News" & Disinformation*, que trata a desordem informacional como um fenômeno sistêmico, envolvendo dimensões políticas, tecnológicas e cognitivas.

Sob essa perspectiva, o Tribunal Superior Eleitoral (TSE) passou a adotar "desinformação" como conceito guarda-chuva para abarcar diferentes formas de manipulação e distorção de conteúdo em contextos de desordem informacional (Tribunal Superior Eleitoral, 2022). Materiais do TRE-SP explicitam que "*fake news*" é uma subcategoria específica, ligada à falsificação da forma notícia e à aparência jornalística, enquanto "desinformação" cobre um escopo mais amplo, incluindo postagens, conteúdos audiovisuais, memes e narrativas distorcidas que circulam nas redes (Tribunal Regional Eleitoral de São Paulo, 2023).

Assim, para estudos sobre circulação de informações *online* e mais especificamente, para este trabalho, o uso do termo "desinformação" mostra-se conceitualmente mais preciso e metodologicamente consistente (Tribunal Superior Eleitoral, 2022; Tribunal Regional Eleitoral de São Paulo, 2023).

A literatura especializada reforça essa ampliação conceitual. Marchiori (MARCHIORI, 2002) já identificava o paradoxo do excesso informacional, no qual o volume de dados não se converte em conhecimento. Han (HAN, 2018), por sua vez, descreve o ambiente digital como um "enxame de ruídos", no qual indivíduos isolados produzem e compartilham incessantemente fragmentos de informação sem coordenação nem filtro, gerando ruído cognitivo coletivo.

Estudos empíricos comprovam os efeitos dessa desordem. Vosoughi, Roy e Aral (VOSOUGHI; ROY; ARAL, 2018) mostram que, no *Twitter*, boatos falsos difundem-se mais longe, mais rápido e mais profundamente do que notícias verdadeiras, chegando a 1,500 pessoas cerca de seis vezes mais rápido, em grande parte associados à novidade da informação e a um perfil emocional distinto nas respostas (mais surpresa e nojo), e não ao efeito de *bots*. Cinelli et al. (CINELLI et al., 2020) analisam várias plataformas e encontram padrões de difusão semelhantes para conteúdos de fontes confiáveis e questionáveis. No *Twitter*, as estimativas de amplificação indicam que a arquitetura não discrimina a veracidade no ato da difusão, favorecendo a circulação indiscriminada de conteúdo.

Durante a pandemia, o *Covid19 Infodemic Observatory* (DOMENICO et al., 2020) verificou que cerca de 28,9% das publicações sobre COVID-19 continham informações questionáveis. Já a Organização Pan-Americana da Saúde (OPAS) (Organização Pan-Americana da Saúde (OPAS/OMS), 2020) conceituou a infodemia como uma "superabundância de informações, precisas e imprecisas, que dificulta a identificação de orientações confiáveis", destacando o papel das redes sociais como amplificadoras desse processo.

Diante desse cenário, García-Saisó et al. (GARCÍA-SAISÓ et al., 2021) enfatizam que a desordem informacional deve ser tratada como questão de saúde pública e de governança democrática, demandando estratégias de comunicação de risco, fortalecimento da checagem de fatos e políticas de alfabetização midiática.

Em síntese, a desordem informacional e, dentro dela, a desinformação, traduz a crise da racionalidade comunicacional contemporânea. Mais do que um problema semântico, trata-se de uma reconfiguração estrutural do ecossistema informativo, em que a abundância de conteúdo, a ausência de filtros e o incentivo algorítmico à polarização substituem a busca pela verdade pela simples viralização. Dessa forma, o presente trabalho se concentra especificamente no estudo da desinformação.

Modelos como o *XLM-R-Large-ClaimDetection*, utilizado neste estudo, aplicam técnicas de aprendizado profundo para identificar automaticamente frases que contêm

declarações factuais dignas de verificação. Essa abordagem se mostra promissora ao reduzir a sobrecarga humana e aumentar a escalabilidade de sistemas de monitoramento de desinformação.

2.8 *Claim Detection* e Verificação de Fatos Automatizada

O processo de verificação de fatos, ou *fact-checking*, é composto por quatro etapas principais: (1) monitoramento de informações, (2) identificação de afirmações verificáveis (*claim detection*), (3) verificação de veracidade e (4) publicação dos resultados (KONSTANTINOVSKIY et al., 2021). Entre essas etapas, a detecção automatizada de afirmações factuais tem ganhado destaque por reduzir o volume de informações analisadas manualmente e otimizar o trabalho das "agências" de checagem.

A detecção de afirmações (*claim detection*) é uma tarefa do PLN voltada à identificação, em um texto, de declarações que expressam uma posição argumentativa, isto é, sentenças que sustentam ou contestam uma determinada ideia. No caso da detecção de afirmações dependentes de contexto (*Context Dependent Claim Detection – CDCD*), proposta por Levy et al. (LEVY et al., 2014), o objetivo é reconhecer apenas as afirmações que se relacionam diretamente a um tópico específico, como uma questão controversa ou um tema de debate. Essa abordagem requer não apenas o reconhecimento da estrutura linguística de uma afirmação, mas também a compreensão de sua relevância semântica em relação ao contexto, diferenciando-a de simples definições, repetições do tópico ou informações neutras.

Afirmações detectadas podem ser tanto factuais quanto não factuais, abrangendo desde proposições verificáveis, baseadas em dados concretos, até declarações de natureza opinativa, que refletem juízos de valor, crenças ou percepções subjetivas. Essa característica torna o *claim detection* uma ferramenta importante para tarefas de mineração de argumentação, análise de debates e suporte à decisão automatizada, pois permite extrair, de grandes volumes de texto, as bases argumentativas que sustentam diferentes pontos de vista, independentemente de sua comprovação empírica.

A partir dessa perspectiva argumentativa, a tarefa de *claim detection* passou a ser também entendida como uma etapa inicial fundamental dos sistemas de checagem automática de fatos (*automated fact-checking*). Nesse contexto, o objetivo não é apenas identificar sentenças argumentativas, mas delimitar quais delas são potencialmente verificáveis, isto é, quais expressam afirmações que podem ser submetidas a um processo de comprovação de veracidade com base em evidências empíricas ou fontes confiáveis. Essa evolução conceitual é explorada por Konstantinovskiy et al. (KONSTANTINOVSKIY et al., 2021), que desenvolvem um modelo de anotação e um conjunto de dados voltados especificamente à identificação de afirmações factuais.

Os autores propõem uma abordagem mais objetiva e padronizada, evitando critérios subjetivos como "importância" ou "relevância política" da afirmação, frequentemente presentes em trabalhos anteriores. Para isso, elaboraram um esquema de sete categorias de afirmações, abrangendo desde sentenças quantitativas, predições e relações de causa e efeito até leis, regras e experiências pessoais. O modelo classifica como trechos factuais aquelas sentenças que expressam asserções sobre o mundo passíveis de verificação, distinguindo-as de opiniões, perguntas ou comentários.

Assim, o *claim detection* deixa de ser apenas uma tarefa voltada à análise argumentativa e passa a ocupar um papel central na detecção e combate à desinformação. Ao automatizar a identificação de enunciados verificáveis em discursos políticos, notícias ou redes sociais, essa técnica contribui para acelerar o processo de verificação, ampliar a cobertura dos sistemas de *fact-checking* e reduzir o tempo de exposição pública de informações falsas ou enganosas. Desse modo, o campo evolui de uma abordagem semântico-argumentativa para uma função prática e social, a de servir como ponte entre o PLN e a integridade informacional no espaço público.

Baseado na arquitetura *Transformer* de Vaswani et al. (VASWANI et al., 2017), o *XLM-R* adota um processo de pré-treinamento não supervisionado fundamentado na tarefa de *Masked Language Modeling (MLM)*, em que determinados *tokens* de uma sequência são mascarados e o modelo deve prever as palavras originais com base no contexto (CONNEAU et al., 2020). Essa abordagem permite o aprendizado de representações contextuais profundas, que capturam relações sintáticas e semânticas compartilhadas entre diferentes idiomas. Para alcançar tal desempenho, foram desenvolvidas duas configurações de arquitetura: o *XLM-R Base*, com 12 camadas e cerca de 270 milhões de parâmetros, e o *XLM-R Large*, com 24 camadas e aproximadamente 550 milhões de parâmetros, ambas empregando 16 cabeças de atenção e uma dimensão de *embedding* de 1024 unidades.

O treinamento do *XLM-R* foi realizado sobre o corpus *CC-100*, um conjunto de dados de 2,5 terabytes de textos coletados e filtrados do *CommonCrawl*, abrangendo 100 idiomas. Diferentemente de modelos anteriores, que utilizavam dados da *Wikipédia*, o *XLM-R* amplia significativamente a cobertura linguística, especialmente em línguas de poucos recursos, por meio da filtragem automatizada de textos com o *fastText* e modelos próprios de detecção de idioma. Para tokenização, foi empregado o algoritmo *SentencePiece*, que realiza a segmentação sublexical diretamente sobre texto cru, sem necessidade de regras específicas por idioma, permitindo um vocabulário unificado de 250 mil subpalavras. Esse design elimina dependências linguísticas e torna o modelo mais eficiente em cenários de *code-switching*, onde há mistura de línguas em uma mesma sentença.

Os experimentos conduzidos confirmaram a superioridade do *XLM-R* em relação a modelos como o *mBERT* e o *XLM-100*. Nos testes de inferência multilíngue (*XNLI*), o modelo alcançou ganhos médios de 14,6% em acurácia, com destaque para melhorias

expressivas em idiomas de baixo recurso, como suaíli (+15,7%) e urdu (+11,4%). Em tarefas de reconhecimento de entidades nomeadas (*NER*) e resposta a perguntas multilíngue (*MLQA*), o *XLM-R* obteve incrementos de 2,4% e 13% em *F1-score*, respectivamente. O modelo também demonstrou desempenho comparável a modelos monolíngues de ponta, como *RoBERTa* e *XLNet*, no *benchmark GLUE*, reforçando a viabilidade de um modelo unificado capaz de lidar com múltiplos idiomas sem perda significativa de desempenho individual.

Assim, o *XLM-R* representa um avanço metodológico e arquitetural significativo para o campo de PLN, provando que a escala e a diversidade linguística, quando combinadas a um treinamento profundo e bem balanceado, são suficientes para gerar representações universais e transferíveis. Essa robustez técnica torna o *XLM-R* uma base sólida para aplicações contemporâneas de detecção de afirmações e checagem automática de fatos, permitindo o *fine-tuning* de modelos especializados, como o *XLM-R-Large-ClaimDetection* utilizado no trabalho, em tarefas que exigem discernimento semântico e generalização entre diferentes idiomas e domínios textuais.

Com base no *XLM-R*, houve uma adaptação supervisionada para a tarefa de classificação de sentenças factuais e não factuais. Esse procedimento tem como objetivo aproveitar o conhecimento linguístico geral aprendido durante o pré-treinamento massivo em múltiplos idiomas e ajustá-lo para um domínio mais restrito, neste caso, o de checagem automatizada de fatos (*automated fact-checking*) (SAMI, 2024).

Tecnicamente, o *fine-tuning* envolve a reutilização dos pesos e *embeddings* do modelo *XLM-R Large*, previamente treinado com o objetivo de predição de palavras mascaradas (*Masked Language Modeling*), e a adição de uma camada de classificação no topo da rede, responsável por prever se uma sentença representa ou não uma afirmação factual. Durante o treinamento supervisionado, essa nova camada, e, em parte, as camadas internas do *Transformer*, passam por um processo de reajuste dos parâmetros com base em exemplos anotados, permitindo que o modelo aprenda padrões semânticos e discursivos associados a declarações verificáveis.

O processo de adaptação seguiu uma estratégia semi-supervisionada (*weakly-supervised*) em duas etapas. Na primeira, o modelo foi treinado com um conjunto de dados fraco, isto é, um *corpus* de mensagens do *Telegram* anotado automaticamente com o auxílio do modelo *GPT-4o*, que produziu rótulos aproximados de factualidade com base em um *prompt* de classificação. Essa etapa forneceu uma ampla base de exemplos. Em seguida, na segunda etapa, foi realizado um *fine-tuning* mais preciso utilizando o conjunto de dados manualmente anotado proveniente de comentários do *Facebook*. Essa combinação de dados fracos e fortes possibilitou ao modelo consolidar o aprendizado.

O modelo resultante foi então avaliado em dois contextos distintos. Em um conjunto de mensagens do *Telegram* anotadas por quatro codificadores humanos, atingiu uma

acurácia de 0,90, demonstrando alta consistência com o julgamento humano. No conjunto de teste com dados extraídos do *Facebook*, obteve-se uma acurácia de 0,79, o que evidencia a capacidade do modelo de se generalizar para diferentes domínios de texto. Embora o treinamento tenha sido realizado exclusivamente com dados em alemão, o modelo mantém o caráter multilíngue herdado do *XLM-R*, que foi originalmente treinado em cem idiomas.

2.9 Métricas de desempenho

2.9.1 Acurácia

A acurácia é uma métrica amplamente utilizada para avaliar o desempenho de modelos de classificação. Ela é definida como a proporção de previsões corretas em relação ao total de previsões realizadas. A fórmula para calcular a acurácia é dada por:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

onde:

- TP (*True Positives*): número de verdadeiros positivos, ou seja, casos em que o modelo previu corretamente a classe positiva.
- TN (*True Negatives*): número de verdadeiros negativos, ou seja, casos em que o modelo previu corretamente a classe negativa.
- FP (*False Positives*): número de falsos positivos, ou seja, casos em que o modelo previu incorretamente a classe positiva.
- FN (*False Negatives*): número de falsos negativos, ou seja, casos em que o modelo previu incorretamente a classe negativa.

2.9.2 Precisão

A precisão é uma métrica que avalia a qualidade das previsões positivas feitas por um modelo de classificação. Ela é definida como a proporção de verdadeiros positivos em relação ao total de previsões positivas. A fórmula para calcular a precisão é dada por:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2.5}$$

onde:

- TP (*True Positives*): número de verdadeiros positivos.
- FP (*False Positives*): número de falsos positivos.

2.9.3 *Recall*

O *recall*, ou sensibilidade, é uma métrica que avalia a capacidade de um modelo de classificação em identificar corretamente todas as instâncias positivas. É definido como a proporção de verdadeiros positivos em relação ao total de instâncias positivas reais. A fórmula para calcular o *recall* é dada por:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.6}$$

onde:

- TP (*True Positives*): número de verdadeiros positivos.
- FN (*False Negatives*): número de falsos negativos.

2.9.4 *F1-Score*

O *F1-score* é uma métrica que combina a precisão e o *recall* em uma única medida, proporcionando um equilíbrio entre essas duas métricas. Especialmente útil quando há um desequilíbrio entre as classes positivas e negativas. A fórmula para calcular o *F1-score* é dada por:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{2.7}$$

2.9.5 Índice de Jaccard

O índice de Jaccard, também conhecido como coeficiente de Jaccard, é uma métrica que avalia a similaridade entre dois conjuntos. É definido como a razão entre a interseção e a união dos conjuntos. A fórmula para calcular o índice de Jaccard é dada por:

$$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|} \tag{2.8}$$

onde:

- A e B são os conjuntos a serem comparados.

2.10 Trabalhos Relacionados

A identificação de desinformação e a análise de posicionamentos têm sido amplamente investigadas na literatura devido à sua relevância para a compreensão dos fenômenos sociais, bem como pelo impacto cultural, político e econômico associado a esses temas. Nesta seção, são apresentados os principais trabalhos relacionados a essa temática, com destaque para suas contribuições, metodologias aplicadas e os pontos de melhoria que motivaram e influenciaram o presente estudo. Inicialmente, discutimos pesquisas com

escopo geral voltadas para a classificação de desinformação, cujo aprimoramento é um dos objetivos deste trabalho. Em seguida, apresentamos o estudo inicial que serviu como base para o desenvolvimento tanto do primeiro artigo quanto desta pesquisa, corroborando para a construção de uma base sólida para estes estudos.

O primeiro trabalho em questão, (BRITO, 2024), tem como foco a análise do posicionamento dos usuários em discussões *online* sobre desinformação. A pesquisa investiga como os posicionamentos expressos em textos podem ser utilizados para identificar conteúdos potencialmente enganosos ou nocivos, com especial atenção às discussões sobre as urnas eletrônicas no Brasil, no período de fevereiro a novembro de 2022. Utilizando técnicas algorítmicas, o estudo aplica modelagem de tópicos e análise de interações nas redes sociais, com dados extraídos de postagens publicadas no *Twitter* (atualmente *X*) e conteúdo desinformativo verificado por "agências" de checagem de notícias. Tal extração textual foi auxiliada pelo processo manual na elaboração de termos-chave para obtenção de conteúdo com desinformação, sendo esta a principal lacuna que o presente estudo pretende preencher por meio da elaboração automatizada de termos-chave.

A pesquisa demonstrou a viabilidade da aplicação de técnicas de detecção de posicionamento e modelagem de tópicos para identificar desinformação e caracterizar o comportamento interacional dos usuários na propagação desse conteúdo, utilizando também técnicas de *TF-IDF* para rotulação e análise (BRITO, 2024).

O segundo trabalho, (SANTOS et al., 2023), tem como objetivo analisar os ataques ao sistema eleitoral brasileiro durante as eleições de 2022, especificamente no *Twitter*. Adotando uma abordagem interdisciplinar e o uso de ferramentas computacionais de rotulação automatizada de perfis e análise de linguagem natural, o artigo identificou os principais tipos de discursos hostis e o posicionamento político dos perfis responsáveis por esses tipos de discursos contra as urnas eletrônicas, o Tribunal Superior Eleitoral (TSE) e os magistrados do tribunal. Os dados revelaram que perfis governistas, principalmente bolsonaristas, foram os responsáveis por uma maior produção de conteúdos hostis, incluindo xingamentos às urnas eletrônicas e ataques de ódio direcionados aos ministros do TSE.

Outro projeto relevante na detecção de textos factuais utilizado como base para este trabalho é o estudo de Arslan et al. (ARSLAN et al., 2020), que introduz o conjunto de dados *ClaimBuster*, um *benchmark* para identificação de afirmações passíveis de checagem (*check-worthy factual claims*) em discursos políticos. O corpus é composto por 23,533 sentenças extraídas das transcrições de todos os debates presidenciais gerais dos Estados Unidos, anotadas manualmente em três categorias: (i) enunciados não factuais, (ii) enunciados factuais pouco relevantes e (iii) enunciados factuais relevantes para checagem. Essa estrutura de rótulos permite distinguir, dentro do subconjunto de sentenças factuais, aquelas que efetivamente merecem prioridade em processos de checagem profissional.

Os autores formalizam afirmações *check-worthy* como declarações factuais cujo

esclarecimento de veracidade é, em princípio, de interesse do público em geral, e documentam cuidadosamente o processo de anotação, incluindo diretrizes e agregação das decisões individuais dos anotadores (ARSLAN et al., 2020). O conjunto é disponibilizado em diferentes arquivos, contendo tanto as sentenças quanto os rótulos consolidados, o que viabiliza estudos sobre subjetividade na tarefa e sobre a consistência entre avaliadores humanos.

A partir desse recurso, Arslan et al. (ARSLAN et al., 2020) avaliam modelos supervisionados para detecção automática de afirmações passíveis de checagem, combinando características lexicais e semânticas das sentenças com algoritmos de aprendizado de máquina. Os resultados mostram que é possível discriminar, com desempenho competitivo, entre sentenças não factuais, factuais pouco relevantes e factuais relevantes, estabelecendo um patamar de comparação para trabalhos posteriores em *claim detection*. Assim, a principal contribuição do artigo é fornecer um conjunto de dados padronizado e amplamente reutilizável que serve como base para o desenvolvimento e a avaliação sistemática de modelos de identificação de afirmações factuais verificáveis em textos políticos.

Ademais, o estudo aplicou um método de detecção de posicionamento não supervisionado para agrupar os usuários em *clusters* polarizados, com base nas contas que retuitaram. Essa técnica foi fundamentada em estudos que sugerem que os usuários tendem a polarizar suas opiniões e formar comunidades políticas, seguindo o princípio da homofilia. A análise revelou dois grandes *clusters*: um representando usuários favoráveis à visão política dominante e outro, contrário. Esses *clusters* mostraram-se densos em interações internas, com baixa proximidade entre si, caracterizando a polarização do debate. A abordagem permitiu entender melhor as dinâmicas de propagação de discursos tóxicos e a formação de "bolhas" ideológicas dentro da rede social, reforçando a relevância da polarização como um fator na disseminação de desinformação (SANTOS et al., 2023). Esses *clusters* polarizados também foram utilizados para testar a extração de tuítes com base nos termos-chave obtidos nesta pesquisa, assim como foi feito com os termos-chave extraídos manualmente na primeira pesquisa citada. Portanto, esse artigo foi essencial para a comparação entre os métodos automáticos e manuais de extração de termos-chave que permitiu avaliar a eficácia da abordagem automatizada, oferecendo informações fundamentais sobre a viabilidade de sua aplicação em contextos de grande volume de dados.

3 Metodologia

3.1 Etapas Realizadas Neste Projeto

3.1.1 Extração e seleção de produções textuais

A primeira etapa consistiu na extração de dados para a construção de uma base de conhecimento que foi usada para análise e classificação de novos textos. Para a extração de conteúdos desinformacionais, foi utilizado *web scraping* (raspagem de dados) com a biblioteca *Python BeautifulSoup*, que manipula dados *HTML* e *XML* de sites. Com essas fontes, foram coletados dados usados como base para obtenção de termos-chave relevantes. Mais especificamente, o *web scraping* foi realizado em notícias previamente selecionadas por outro estudo (BRITO, 2024) com foco mais geral que utiliza detecção de posicionamento para identificação de conteúdo desinformativo no mesmo contexto das eleições brasileiras de 2022.

Tais conteúdos relacionados à desinformação foram coletados a partir de postagens previamente verificadas por "agências" de *fact-checking*, como G1 Fato ou Fake, Aos Fatos, UOL Confere, Estadão Verifica e Agência Lupa. O processo envolveu uma seleção criteriosa de postagens que continham informações falsas ou enganosas, já analisadas e desmentidas por essas "agências".

Além disso, a coleta foi limitada a conteúdos verificados e publicados no período de janeiro a dezembro de 2022, incluindo apenas postagens que continham os termos específicos: 'urna', 'urnas', 'eleição', 'voto impresso', 'auditável' e 'eleições'. Essas restrições foram estabelecidas para garantir que os dados retornados estivessem diretamente relacionados ao contexto das eleições brasileiras de 2022, permitindo uma análise mais focada do conteúdo desinformativo (BRITO, 2024).

Para futura comparação e classificação textual, foram utilizados dados coletados por meio da *API* v2 do *Twitter* (atual *X*). A *API* (Interface de Programação de Aplicações) é um conjunto de ferramentas que permite que desenvolvedores interajam diretamente com os serviços e dados de uma plataforma, como o *Twitter*. Por meio dela, é possível acessar informações estruturadas, como *tweets*, *retweets*, curtidas, *hashtags* e metadados, de forma programática e escalável.

Entretanto, com a transição da *API* do *Twitter* para um modelo monetizado com altos custos, a extração contínua de conteúdo tornou-se inviável. (LUPA, 2023) Ainda assim, os dados obtidos anteriormente (SANTOS et al., 2023) foram suficientes para sustentar pesquisas relacionadas, incluindo o presente estudo.

3.1.2 Pré-processamento e mapeamento de textos

O primeiro passo no processo foi a extração de sentenças relevantes por meio do cálculo do peso *Term Frequency-Inverse Document Frequency* (TF-IDF), uma métrica amplamente utilizada para avaliar a relevância de palavras e sentenças dentro de um conjunto de textos. A função desenvolvida para este fim utiliza a classe *TfidfVectorizer* da biblioteca *Scikit-learn*, que converte sentenças de uma notícia em vetores numéricos representando a importância de termos no contexto do documento analisado. Adicionalmente, a similaridade de cosseno entre as sentenças também foi calculada, permitindo identificar trechos que melhor resumem o conteúdo do texto, com base em sua relevância estatística.

Os dados coletados em fontes de "agências" de *fact-checking* foram processados utilizando técnicas de sumarização automatizada, especificamente por meio do modelo BERT (*Bidirectional Encoder Representations from Transformers*), com o objetivo de gerar resumos mais concisos e focados. O BERT é uma tecnologia baseada na arquitetura de *Transformers*, que utiliza atenção bidirecional para compreender o contexto das palavras em um texto. Essa abordagem permite capturar características semânticas tanto de palavras individuais quanto de frases inteiras, resultando em resumos que preservam o significado essencial do conteúdo original.

O processo de sumarização foi realizado utilizando uma variante específica do BERT, chamada SBERT (*Sentence-BERT*), otimizada para tarefas de similaridade e classificação de sentenças. Esse modelo foi treinado para identificar as partes mais relevantes do texto, reduzindo-o a uma forma condensada sem perder informações críticas, facilitando assim a análise de grandes volumes de dados desinformativos. Diversos modelos baseados em *Transformers* foram testados para identificar a abordagem mais eficiente em termos de desempenho e precisão. Entre os modelos avaliados estavam:

- BERT padrão (*model = Summarizer()*), que apresentou um tempo médio de processamento de 301,84 a 309,03 segundos por tarefa.
- BERT com suporte a resolução de correferências, utilizando o *CoreferenceHandler* com um grau de "ganância" (*greedyness*) ajustado para 0,4.
- DistilBERT, uma versão mais leve do BERT, configurada com camadas ocultas específicas (hidden=[-1, -2] e hidden_concat=True), que reduziu o tempo médio de execução para aproximadamente 86,87 a 96,21 segundos por tarefa.
- SBERT (*Sentence-BERT*), utilizando a versão *paraphrase-MiniLM-L6-v2*, que apresentou o melhor desempenho, com um tempo médio de processamento de 57,75 a 60,28 segundos por tarefa.

O fator determinante para a escolha do SBERT foi, principalmente, o tempo de processamento reduzido, pois os resultados dos resumos gerados não apresentaram diferenças significativas em relação aos outros modelos testados. Além disso, o SBERT demonstrou desempenho consistente tanto na extração de palavras-chave quanto em processos futuros do estudo, como a análise de conteúdo e a comparação de dados.

Combinando a sumarização e a extração de palavras-chave, foi possível obter uma visualização mais clara e organizada dos dados, reduzindo ruídos e destacando elementos fundamentais para a análise do conteúdo desinformativo, especialmente no contexto das eleições brasileiras de 2022.

Os dados para comparação foram previamente processados e analisados, abrangendo o período de julho a novembro de 2022, com foco nos debates políticos no Brasil que antecederam as eleições gerais daquele ano (SANTOS et al., 2023). Vale destacar que, assim como nos artigos em que este estudo foi baseado, os termos *tweets* e *retweets* foram mantidos e não substituídos por "*post*" e "*repost*". Essa decisão segue a terminologia amplamente conhecida e utilizada, em vez de adotar a nova nomenclatura estabelecida pela empresa atualmente denominada X.

3.1.3 Registro dos Dados Processados

O registro dos dados previamente processados foi armazenado em arquivos de documentos, devido ao volume relativamente pequeno de informações. Os textos extratos por meio do processo de *web scraping*, bem como as notícias coletadas de "agências" de *fact-checking*, foram manipulados utilizando a biblioteca *pandas* do *Python* e organizados em arquivos XLSX com tamanhos variando entre 100 KB e 200 KB cada.

Já os dados utilizados para comparações, por apresentarem maior volume, foram armazenados em arquivos CSV, com tamanhos variando entre 5500 KB e 11000 KB, dependendo do conteúdo. Esses dados incluem textos previamente processados e preparados para análise de similaridade e padrões, permitindo uma estrutura eficiente para futuras análises.

3.1.4 Obtenção dos Termos-chave

Para os modelos de classificação semântica e detecção de fatos, foi implementado um processo de segmentação adicional com expressões regulares. Essa etapa visou dividir as sentenças em fragmentos menores com base em conjunções e conectores linguísticos. Além disso, foi aplicada uma estratégia de separação específica para frases resumidas, utilizando o método de divisão *split* contido na biblioteca *regex* por conjunções. Essa abordagem permite que cada ideia expressa seja analisada isoladamente, garantindo maior precisão na classificação semântica.

As conjunções são palavras ou expressões que conectam orações ou termos dentro de uma frase, indicando relações como adição, contraste, causa, condição ou tempo. Exemplo: palavras como 'e', 'mas', 'porque' e 'enquanto' são conjunções amplamente utilizadas na língua portuguesa. No contexto do processamento de textos, essas palavras são importantes pois frequentemente sinalizam mudanças de foco ou introduzem novas ideias. Dessa forma, pode-se obter a separação dos dados factuais de não factuais necessários, filtrando dados possíveis de se conter desinformação.

Além disso, foi notado que o termo "que", em notícias provenientes de "agências" de checagem de fatos, frequentemente é sucedido por informações relevantes, sejam elas desinformação ou esclarecimentos. Com base nessa observação, o termo foi utilizado como um ponto de referência para identificar e isolar termos-chave no texto, ajudando a localizar trechos particularmente importantes para a análise e classificação semântica.

Outro processamento focado em expressões regulares foi a detecção de aspas (" ou ') em notícias, uma vez que essas normalmente remetem a citações ou falas que podem indicar pronunciamentos relacionados à desinformação ou apenas relatar declarações factuais.

O aspecto de normalização dos textos, que incluiu a remoção de espaços extras e a conversão de palavras para um formato consistente, também foi considerado, evitando problemas como redundâncias causadas por variações de capitalização, espaçamento ou outras formatações inadequadas capazes de interferir no processo de extração textual. O processo de *padding*, utilizado para adequar os textos a um tamanho padrão de sequência, desempenha um papel importante na otimização do processamento em modelos baseados em *deep learning*. Essa técnica evita problemas de incompatibilidade de tamanho entre lotes e pode reduzir drasticamente a quantidade de *tokens* desnecessários (*padding tokens*), que em casos extremos podem representar até 50 por cento dos *tokens* processados, como relatado em estudos recentes (KRELL et al., 2022).

Se tratando de abordagens mais complexas, o modelo XLM-R-Large-*ClaimDetection* foi utilizado para filtrar textos previamente processados, retirados de fontes de checagem de fatos confiáveis, com o objetivo de classificar esses textos como "factual" ou "não factual". A tarefa principal do modelo foi identificar os textos que possuem afirmações factuais, eliminando aqueles que não necessitam de verificação. É válido ressaltar que inicialmente o modelo foi treinado para dados "factuais", "não factuais" e "insignificantes" e para este projeto apenas dados "factuais" foram filtrados (RISCH et al., 2021).

O conjunto de dados utilizado para treinamento do modelo *ClaimDetection* contém 23,533 declarações extratas de todos os debates presidenciais dos EUA (ARSLAN et al., 2020). Essas declarações foram rotuladas por um processo manual e classificadas em três categorias: afirmações factuais dignas de verificação, afirmações factuais insignificantes e afirmações não factuais. Durante o estudo, foi realizada uma tentativa de treinamento do modelo utilizando esta base de dados, mas, devido a limitações de processamento,

não foi possível concluir a tarefa, nem mesmo realizando a tentativa de aprendizado por transferência e utilizando o modelo xlm-roberta-base, considerado mais leve que o modelo original.

A aplicação do modelo *ClaimDetection* baseou-se em sua capacidade de distinguir entre afirmações que são verificáveis e aquelas que não apresentam um conteúdo factual claro. A tarefa principal do modelo foi identificar textos contendo afirmações factuais que exigem confirmação ou validação, ao mesmo tempo em que elimina aqueles que não necessitam de verificação, ajudando assim a filtrar a grande quantidade de informações desnecessárias presentes nas informações de "agências" de checagem.

O processo de treinamento do modelo SVM (*Support Vector Machine*) foi realizado com o objetivo de classificar os textos presentes no mesmo conjunto de dados utilizados para o modelo XLM, separando os dados em duas categorias: "factual" e "não factual". Inicialmente, os textos foram processados e convertidos em *embeddings* utilizando o modelo XLM-Roberta, um modelo pré-treinado adequado para o processamento de textos em múltiplos idiomas. A partir desses *embeddings*, um modelo SVM com *kernel* linear foi treinado para identificar padrões nos textos, classificando-os com base em suas características semânticas extraídas. O treinamento foi realizado usando uma divisão da base de dados em conjuntos de treino e teste, sendo 80% destinados ao conjunto de treinamento e 20% ao conjunto de teste. A divisão foi estratificada, isto é, a proporção entre as classes factual e não factual foi mantida aproximadamente constante em ambos os subconjuntos. Essa escolha é importante para evitar que o modelo seja treinado ou testado em amostras artificialmente desbalanceadas, o que poderia enviesar as métricas de desempenho. Além disso, foi fixada uma semente de aleatoriedade específica (*random_state*=42), garantindo reprodutibilidade: experimentos repetidos com a mesma configuração produzem a mesma partição dos dados.

Como etapa de pré-processamento e representação, os textos foram convertidos para uma forma vetorial numérica por meio da técnica TF-IDF, utilizando-se um vetorizador de texto padrão (*TfidfVectorizer*). A frequência de cada termo em um texto (TF) é ponderada inversamente pela frequência desse termo no corpus como um todo (IDF), de modo a enfatizar palavras mais informativas e atenuar o peso de termos muito comuns. No presente estudo, foram consideradas não apenas palavras isoladas (unigramas), mas também pares consecutivos de palavras (bigramas), o que permite capturar padrões linguísticos mais ricos, como expressões típicas de linguagem factual (por exemplo, "segundo dados", "de acordo com"). Além disso, estabeleceu-se um limite máximo de características para o vocabulário e excluíram-se termos extremamente raros, buscando um equilíbrio entre riqueza representacional e controle da dimensionalidade.

Após a vetorização, foi treinado o classificador SVM com função de decisão linear. Utilizou-se um parâmetro de regularização padrão (C=1,0), que controla o compromisso

entre complexidade do modelo e capacidade de generalização para encontrar o hiperplano. Para mitigar efeitos de possível desbalanceamento entre as classes factual e não factual, adotou-se um esquema de pesos balanceados entre classes (*class_weight="balanced"*), fazendo com que erros na classe minoritária tenham maior peso na função de perda. Tal estratégia é de extrema importância para evitar que o modelo aprenda a favorecer a classe majoritária, o que poderia levar a um desempenho insatisfatório na detecção de textos "não factual".

Ademais, o PLN foi auxiliado pelo NLTK, uma biblioteca que fornece uma ampla gama de recursos para trabalhar e manipular textos, incluindo funções de tokenização, lematização, análise sintática, reconhecimento de entidades, entre outras. Após o pré-processamento com NLTK, os textos foram transformados em vetores numéricos, usados para treinar os modelos para identificação dos termos-chave presentes na base de dados. Além disso, *Redes Neurais Recorrentes (RNNs)* podem ser utilizadas para análise mais aprofundada do conteúdo textual, identificando padrões sequenciais em textos e oferecendo um método alternativo aos SVMs para tarefas de classificação. Outra abordagem selecionada para o PLN é o uso de modelos baseados em *Transformers*, como o BERT (*Bidirectional Encoder Representations from Transformers*). O BERT utiliza o treinamento bidirecional do *transformer*, permitindo que o modelo tenha um entendimento mais profundo do contexto e do fluxo da linguagem, sendo uma ferramenta promissora para tarefas voltadas à classificação de *fake news* (SANTOS, 2022).

3.1.5 Classificação Textual

A classificação dos dados foi baseada na intencionalidade por trás da mensagem emitida, destacando as diferenças entre informações deliberadamente falsas ou distorcidas, notícias que buscam intencionalmente manipular e enganar o público, propagandas, publicidade, paródia e sátira. Esses diferentes tipos de conteúdos serão analisados com base em características específicas para distinguir suas intencionalidades e seus impactos na sociedade. Este método para classificação manual segue o mesmo padrão utilizado no trabalho Posicionamento e Desinformação (BRITO, 2024), já que o objetivo final é gerar uma classificação similar ou mais eficiente do que a realizada no estudo base. Sendo assim, apenas o processo de sumarização e geração de termos-chave se diferencia do projeto original.

Além disso, as notícias filtradas do conjunto de dados de comparação por meio dos termos-chave, obtidas a partir do estudo geral de classificação de desinformação, que também apareceram no presente estudo, mantiveram as classificações previamente atribuídas, evitando retrabalho. Apenas as notícias inéditas foram submetidas a novas classificações manuais.

Importante destacar que o processo de classificação textual não sofreu modificações,

tendo como base o relatório técnico *Detecção de Posicionamento como Abordagem para Identificação de Conteúdo Desinformativo* (BRITO, 2024), pois o foco do novo estudo é verificar o impacto de detecção de fatos mais otimizadas no processo de classificação textual já realizado no artigo base. Apenas os *inputs* gerados com os novos termos-chave foram acrescentados no processo de classificação.

3.1.6 Análise da eficácia obtida por cada tecnologia

A avaliação da eficácia dos métodos empregados neste estudo foi conduzida por meio de métricas consolidadas na literatura de aprendizado de máquina e PLN, como precisão, *recall* (taxa de detecção, capaz de medir a capacidade de identificação das instâncias positivas da base de dados) (YACOUBY; AXMAN, 2020) e *F1-score*. Tais métricas possibilitam compreender de maneira abrangente a qualidade das classificações obtidas, uma vez que mensuram não apenas a capacidade de identificar corretamente as instâncias positivas, mas também a proporção de acertos entre todas as predições realizadas (precisão). O *F1-score*, por sua vez, sintetiza essas duas dimensões em uma única medida harmônica, sendo particularmente relevante em cenários nos quais há desbalanceamento entre classes, como frequentemente ocorre na detecção de desinformação. Dessa forma, a aplicação conjunta dessas métricas permite avaliar de modo robusto o desempenho dos modelos tanto de aprendizado supervisionado tradicional quanto de *aprendizado profundo*.

Além das métricas quantitativas, foi realizado testes de eficiência baseados em pesquisas anteriores, permitindo comparar os resultados obtidos com *benchmarks* já estabelecidos em processos de análise textual em larga escala. Esse tipo de comparação é fundamental para identificar se as tecnologias atualmente utilizadas ainda se mostram adequadas ao combate à disseminação de *fake news* ou se apresentam sinais de defasagem frente à evolução do fenômeno. A partir dessa análise, é possível apontar fragilidades específicas que demandam aprimoramentos, tais como baixa sensibilidade a variações linguísticas, elevado custo computacional ou dificuldades em lidar com novos padrões de manipulação textual.

Outro aspecto a ser considerado é a etapa de extração de termos-chave, que representa um elo intermediário no processo de checagem de notícias. Nesse ponto, os resultados obtidos pelas diferentes tecnologias serão confrontados com aqueles provenientes do projeto base "*Detecção de Posicionamento como Abordagem para Identificação de Conteúdo Desinformativo*" (BRITO, 2024). A comparação permitirá verificar a consistência e abrangência dos métodos, identificando quais deles são capazes de recuperar, de forma automática, um número maior de *tweets* alinhados ao escopo investigado. Métodos que se aproximarem ou superarem a eficácia do processo manual adotado no projeto de referência serão considerados eficientes. Em contrapartida, técnicas que não demonstrarem desempenho comparável serão classificadas como ineficazes, ainda que apresentem valores

satisfatórios em métricas isoladas, uma vez que a utilidade prática está diretamente vinculada à capacidade de identificar conteúdos relevantes em larga escala.

Por fim, a análise de eficácia não se limitará apenas ao desempenho preditivo, mas incluirá também a avaliação do consumo de recursos computacionais, como memória, tempo de processamento e demanda de hardware necessária para o treinamento e a inferência. Esse aspecto é decisivo quando se considera a aplicação das técnicas em cenários reais, especialmente aqueles que exigem monitoramento contínuo e em tempo real, como plataformas de checagem de fatos ou sistemas de vigilância digital. A viabilidade prática dependerá, portanto, da conjugação entre desempenho técnico e eficiência computacional, de modo que um método altamente preciso, mas inviável em termos de tempo e custo, dificilmente poderá ser adotado em larga escala.

Em síntese, a análise da eficácia de cada tecnologia deve integrar diferentes dimensões: métricas de desempenho, capacidade de extração de termos-chave, eficiência na detecção de conteúdos consistentes, consumo de recursos e adaptabilidade frente a novos contextos. Essa perspectiva multidimensional é indispensável para identificar as ferramentas mais promissoras não apenas sob a ótica da precisão técnica, mas também sob a lógica da aplicabilidade prática no combate à disseminação de desinformação em ambientes digitais.

3.1.7 Comparação da eficácia entre as tecnologias estudadas durante o projeto

A avaliação dos resultados obtidos por cada método requer uma análise criteriosa de desempenho, de modo a identificar quais tecnologias apresentam maior potencial de aplicação prática. Para isso, métricas consolidadas no campo do aprendizado de máquina, como precisão, *recall* e *F1-score*, foram utilizadas como referência fundamental. A precisão indica a capacidade do modelo de evitar falsos positivos, enquanto o *recall* mede sua sensibilidade na identificação de instâncias relevantes. Já o *F1-score*, ao combinar ambas as métricas, permite avaliar de forma equilibrada o desempenho dos classificadores em cenários nos quais tanto a exatidão quanto a abrangência são relevantes. Essas métricas, calculadas a partir das classificações realizadas nos experimentos, oferecem uma base objetiva para a comparação entre diferentes abordagens e servem como critério para a seleção das técnicas mais promissoras.

Entretanto, é importante reconhecer que o desempenho numérico não é o único fator determinante para a escolha do método mais adequado. Modelos baseados em aprendizado profundo, como as arquiteturas de *redes neurais recorrentes* ou baseadas em *Transformers*, tendem a apresentar resultados superiores em termos de precisão e capacidade de generalização, sobretudo quando aplicados a grandes volumes de dados textuais. Essa vantagem, porém, é frequentemente acompanhada por um custo computacional elevado, demandando maior capacidade de processamento, uso intensivo de memória e tempos

de execução prolongados. Tais aspectos podem limitar a aplicabilidade em contextos de larga escala ou em ambientes com restrições de recursos, como em sistemas de checagem automatizada em tempo real.

Por outro lado, métodos mais tradicionais de aprendizado de máquina supervisionado, como os classificadores baseados em SVM ou regressões lineares, embora menos sofisticados em termos de captura de dependências semânticas, apresentam vantagens quanto à eficiência e simplicidade. Seu menor custo computacional permite a implementação em sistemas mais enxutos, além de facilitar a replicabilidade em ambientes distintos sem a necessidade de infraestrutura de alto desempenho. Assim, ainda que possam oferecer resultados inferiores em alguns cenários, esses modelos se destacam pela viabilidade prática, sobretudo quando há limitação de recursos ou quando a demanda de tempo de resposta é um fator crítico.

Além das métricas de desempenho e da análise do custo computacional, outro aspecto a ser considerado é a complexidade de implementação das diferentes técnicas. O ambiente digital, caracterizado pela constante mutação nos processos de fabricação e disseminação de notícias falsas, exige modelos adaptáveis e de fácil atualização. Métodos excessivamente complexos podem não acompanhar a velocidade das mudanças no ecossistema de desinformação, tornando-se rapidamente obsoletos. Por essa razão, a comparação entre os métodos não se restringirá apenas à sua acurácia, mas também levará em conta a flexibilidade de adaptação, a robustez frente a novos padrões linguísticos e a escalabilidade para diferentes contextos.

Outro aspecto relevante a ser considerado é o tempo de processamento exigido por cada abordagem. Métodos tradicionais, como o uso de vetorização TF-IDF combinada a classificadores supervisionados, tendem a apresentar tempos de execução significativamente menores, sendo adequados para cenários em que a rapidez é fator crítico, como em análises em tempo real. Por outro lado, modelos de aprendizado profundo, especialmente aqueles baseados em arquiteturas de *Transformers*, demandam maior capacidade computacional e apresentam tempos de processamento mais elevados devido à complexidade do treinamento e da inferência. Essa diferença impacta diretamente a escalabilidade da solução, visto que aplicações em larga escala, como monitoramento contínuo de redes sociais, exigem um equilíbrio entre acurácia e velocidade de resposta. Dessa forma, a análise comparativa entre os métodos deve considerar não apenas a precisão alcançada, mas também a viabilidade temporal para sua aplicação prática em diferentes contextos.

Portanto, a análise comparativa deve integrar múltiplas dimensões, desempenho, custo computacional, tempo de processamento e complexidade de implementação, com o objetivo de identificar não apenas o modelo mais preciso, mas aquele que seja sustentável, eficiente e aplicável em cenários reais. Essa perspectiva holística é crucial para que a tecnologia desenvolvida seja capaz de responder, de maneira eficaz e contínua, ao desafio

dinâmico da detecção e mitigação da desinformação.

4 Resultados e Discussões

4.1 Resultados Obtidos

Este capítulo apresenta e discute os principais resultados empíricos obtidos ao longo do estudo. Inicialmente, são analisados os efeitos dos diferentes modelos de sumarização automática na redução de ruído e na preservação de trechos factuais dos textos. Em seguida, são avaliadas as abordagens de segmentação baseadas em conjunções e em aspas, com ênfase na capacidade de recuperar trechos relevantes em notícias e em *clusters* de *tweets*. Na sequência, discute-se o desempenho do modelo *XLM-R-Large-ClaimDetection* e do classificador SVM na identificação de afirmações factuais, comparando seus resultados com o modelo original em inglês e com o método manual de referência. Por fim, são comparadas as diferentes estratégias de classificação de desinformação em *tweets* e retuítes, destacando os compromissos entre assertividade, conservadorismo e risco de super-rotulagem.

4.1.1 Sumarização

Para o primeiro processo de sumarização automática para redução de ruído, três modelos foram avaliados: BERT-base, DistilBERT e SBERT-MiniLM. Em todos os casos, os resumos ficaram entre 25% e 27% do tamanho do texto original conforme mostrado na Figura 4, sem diferenças expressivas de compressão entre os modelos, com leve tendência do DistilBERT a produzir resumos mais concisos e do SBERT-MiniLM a manter textos um pouco maiores.

Quando se observa apenas a preservação de trechos factuais, BERT-base e SBERT-MiniLM mantiveram cerca de 25% a 26% de conteúdo classificado como factual, superando o DistilBERT, que ficou em torno de 22% conforme mostrado na Figura 5. Em termos de retenção de tópicos centrais, medida pelo índice de Jaccard entre as palavras-chave do original e do resumo, o SBERT-MiniLM apresentou o melhor resultado (0,22), seguido muito de perto pelo BERT-base (0,21), enquanto o DistilBERT apresentou desempenho inferior (0,14). Da mesma forma, quando a similaridade textual foi medida por uma métrica de similaridade (TF-IDF/cosseno), o SBERT-MiniLM obteve o maior valor (0,64), superando ligeiramente o BERT-base e o DistilBERT, ambos com 0,61. No entanto, o tempo de processamento mostrou diferenças marcantes: o BERT-base demandou em média mais de 22 segundos por texto, o DistilBERT cerca de 3 segundos e o SBERT-MiniLM pouco mais de 1 segundo, o que torna o último mais adequado a cenários de grande volume.

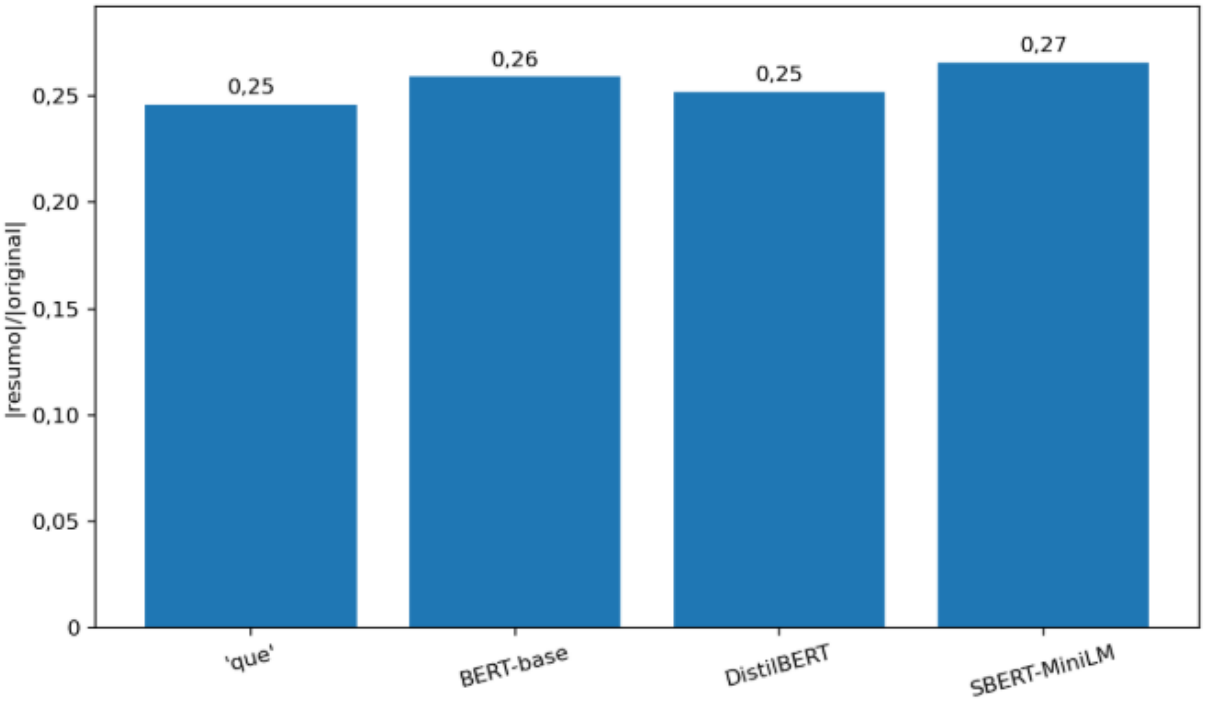

Figura 4 – *Compression ratio* médio (|resumo|/|original|) por modelo.

Fonte: elaboração própria.

4.1.2 Segmentação baseada em conjunções

Os resultados da abordagem de segmentação baseada no termo "que" demonstraram sua eficácia. Nas 29 amostras de notícias provenientes de "agências" de checagem de fatos, a segmentação baseada no termo "que" resultou em textos altamente relevantes, evidenciando ser uma técnica interessante de segmentação para textos contidos em notícias. Este método apresenta a vantagem de não exigir grande processamento, sendo uma alternativa eficiente para identificar informações relevantes em grandes volumes de texto.

Além disso, a aplicação desse método em *clusters* polarizados mostrou resultados significativos. Para o $cluster_0$, foi possível obter 2,060 *tuítes* relevantes, em comparação com os 2,056 *tuítes* obtidos utilizando termos-chave criados manualmente. Já no $cluster_1$, foram extraídos 1,089 *tuítes*, superando os 781 obtidos com termos-chave definidos manualmente. Exemplos de textos retornados por esse método incluem: "aponta que apertar 'confirma' durante a tela 'confira seu voto' anula o voto" e "usou o plenário da câmara federal para propagar informações falsas sobre a pandemia, como mostrou uma reportagem publicada pela Lupa em dezembro do ano passado". Esses resultados demonstram que a segmentação baseada em conjunções, como o termo "que", pode ser mais eficaz do que abordagens manuais tradicionais, ampliando a abrangência e precisão na identificação de conteúdos relevantes para análise de desinformação.

A aplicação do modelo XLM-R-Large-*ClaimDetection* nos dados de debates presidenciais dos EUA (ARSLAN et al., 2020), adaptados ao contexto português, apresentou

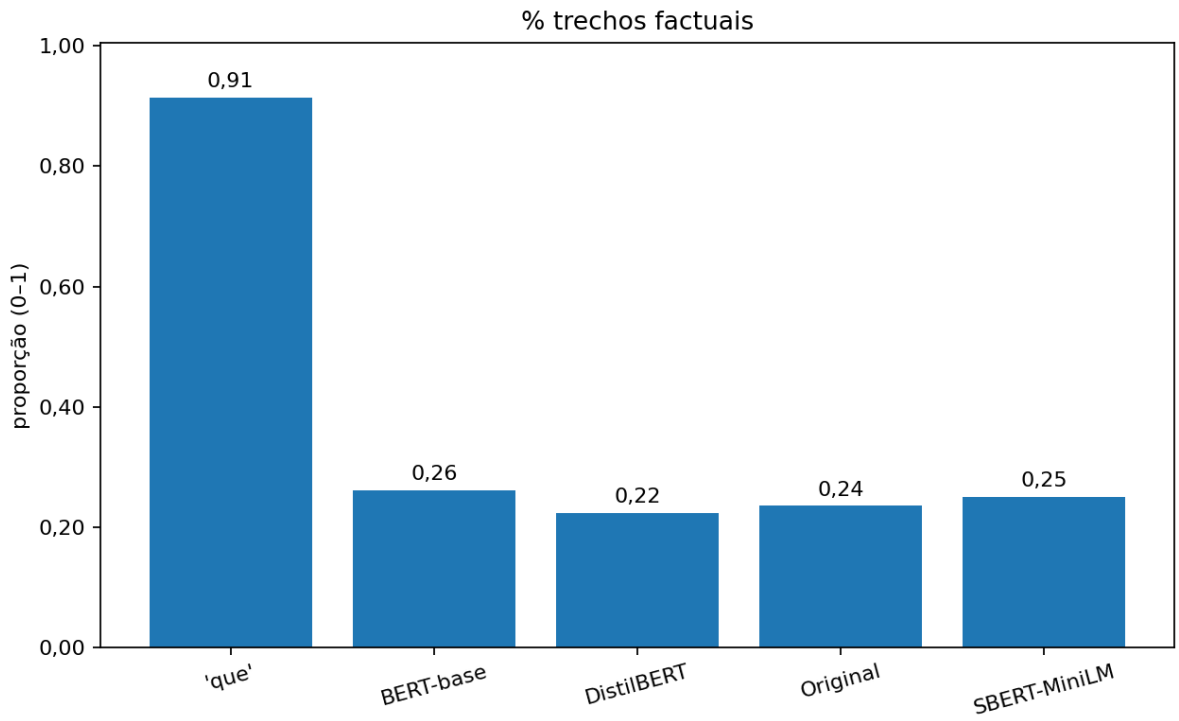

Figura 5 – Factualidade % por modelo.

Fonte: elaboração própria.

resultados promissores na tarefa de classificação de afirmações factuais. O modelo alcançou uma acurácia de 0,88, mostrando bom desempenho mesmo considerando as diferenças linguísticas entre os dados testados em português e o treinamento original em inglês.

Por meio da utilização de uma matriz de confusão foi possível verificar que o modelo classificou corretamente a maioria das afirmações factuais e não factuais, com 680 previsões corretas para sentenças factuais não relevantes e 232 para as sentenças factuais importantes. Porém, houve dificuldades de diferenciar categorias factuais não relevantes, sendo classificadas incorretamente 114 sentenças das 1034 presentes no conjunto de dados.

Esses resultados são demonstrados no relatório de classificação, que apresentou um desempenho muito elevado para as sentenças factuais não relevantes, com *precisão* de 0,99 e *recall* de 0,86, conforme dados Tabela 1. Já para as sentenças factuais relevantes, a *precisão* foi de 0,67, enquanto o *recall* atingiu um valor consideravelmente alto de 0,97, evidenciando a capacidade do modelo em identificar corretamente a maioria das afirmações importantes, ainda que tenha apresentado menor precisão nessa categoria.

Em comparação com o modelo original treinado em inglês, que obteve uma acurácia de 0,90, os resultados em português apresentaram uma leve queda. No estudo original, o modelo foi capaz de classificar afirmações factuais e não factuais com equilíbrio entre *precisão* e *recall*, alcançando valores de 0,90 em ambas as métricas, conforme apresentado na Tabela 2.

Tabela 1 – Resultados do modelo XLM-R-Large-*ClaimDetection* em português

Classe	*Precision*	*Recall*	*F1-Score*
Unimportant Factual Sentence (UFS)	0,99	0,86	0,92
Important Factual Sentence (IFS)	0,67	0,97	0,79
Acurácia Geral		0,88	

Tabela 2 – Resultados originais do modelo XLM-R-Large-*ClaimDetection* em inglês

Classe	*Precision*	*Recall*	*F1-Score*
Factual	0,88	0,92	0,90
Non-Factual	0,92	0,88	0,90
Acurácia Geral		0,90	

O baixo rendimento em encontrar sentenças factuais importantes (IFS) pode ser atribuído ao fato de o treinamento não ter sido realizado de forma ideal, devido a limitações de *hardware*. O desempenho alcançada, ainda assim, destaca a utilidade do modelo XLM-R-Large como ferramenta para filtragem e segmentação de grandes volumes de texto, facilitando a identificação de informações relevantes para análises de desinformação e processos de verificação de fatos.

Apesar de apresentar métricas menos eficientes, o modelo conseguiu extrair textos relevantes das 29 notícias analisadas conforme pode ser visto na Tabela 4. Com textos resumidos pelo BERT, foram retornados 2,391 *tuítes* no $cluster_0$ e 1,333 no $cluster_1$, já para o DistilBERT (dBERT), 1,670 e 928, respectivamente e com o SBERT, 2,040 e 1,234 como pode ser observado na Figura 6. Entre as frases relevantes extraídas, destacam-se exemplos como: “o TSE (Tribunal Superior Eleitoral) não testou a segurança das urnas para as eleições de 2022”. No entanto, o modelo também retornou frases menos relevantes, como: “no Telegram disseminam informações falsas sobre o”, “estão registrados na seção” e “circulam nas redes sociais (veja aqui)”. Esse comportamento reforça a necessidade de aprimorar a precisão na filtragem de conteúdo, evitando a extração de dados não essenciais, o que pode aumentar o tempo de processamento e reduzir a eficiência geral do sistema.

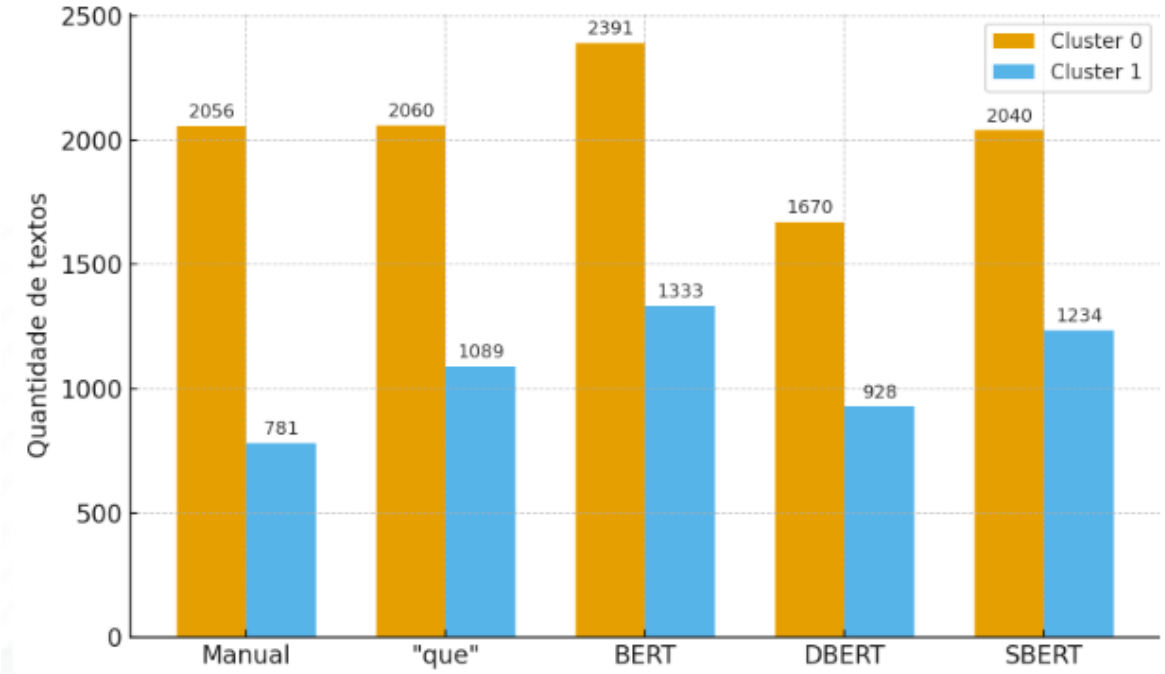

Figura 6 – Textos encontrados por método e *cluster*

Fonte: elaboração própria.

4.1.3 Segmentação baseada em aspas

Para o método envolvendo apenas aspas, houve retorno de frases para apenas 5 das 29 notícias analisadas, demonstrando sua baixa eficácia na identificação de informações relevantes. Entre os textos retornados estavam exemplos como: “Outro exemplo citado por ele é no caso de a totalização envolver, hipoteticamente, apenas duas cidades”, além de termos como “banco nacional do brasil”, “sala escura” e “vão para as nuvens”. Embora algumas frases tenham relevância contextual, o método não foi consistente o suficiente para ser amplamente aplicado, destacando a necessidade de estratégias mais eficazes para análise desse tipo de dado.

4.1.4 Classificação de afirmações factuais com XLM-R-Large-*ClaimDetection*

A aplicação do modelo XLM-R-Large-*ClaimDetection* apresentou acurácia de 0,88 em português, com bom *recall* para sentenças factuais importantes (0,97), mas menor precisão (0,67), o que indica tendência a recuperar quase tudo o que é relevante, ao custo de trazer alguns itens não essenciais.

4.1.5 Comparação com o SVM

O modelo SVM apresenta uma acurácia global moderada, em torno de 0,79. A classe 0 obteve *precision* de 0,85 e *recall* de 0,83, enquanto a classe 1, associada às sentenças

factuais/relevantes, alcançou *precision* de 0,69 e *f1-score* de 0,71, conforme Tabela 3. Esses valores indicam que o modelo é capaz de capturar uma parcela significativa das sentenças de interesse, embora ainda com ocorrência de falsos positivos e perda de algumas sentenças relevantes. Na prática, isso significa que a filtragem contribui para reduzir o volume de texto a ser analisado. Contudo, a presença de textos menos relevantes entre os resultados sugere que o modelo SVM pode beneficiar-se de ajustes adicionais ou da combinação com outras técnicas de filtragem para melhorar a qualidade do conteúdo extraído.

Tabela 3 – Resultados do SVM em português

Classe	*Precision*	*Recall*	*F1-Score*
Factual	0,6903	0,7323	0,7107
Non-Factual	0,8538	0,8263	0,8398
Acurácia geral		0,7938	

4.2 Comparação de Classificação

Pode-se notar que pelo gráfico de Distribuição de Desinformação em *tweets* presentes no *Cluster* 0 (Figura 7), no método manual de referência o C0 apresenta 33,4% de *tweets* classificados como com desinformação, 6,0% como sem desinformação e 60,5% como não rotulados. Com a aplicação dos modelos, todos os métodos de segmentação e classificação aumentaram a proporção de *tweets* classificados como com desinformação em relação ao método manual: no DistilBERT, essa proporção sobe para 47,5%, enquanto SBERT, segmentação por "que", BERT e SVM deslocam a classe "com desinformação" para a faixa de 55–62%, com a classe "sem desinformação" avançando para cerca de 9–13%. O DistilBERT foi o que apresentou o menor aumento, seguido pelo SBERT, segmentação por "que", BERT e SVM, que tiveram aumentos mais expressivos. Isso indica que esses métodos são mais agressivos na identificação de *tweets* com desinformação, o que pode ser benéfico para capturar mais conteúdo relevante, mas também pode aumentar o risco de falsos positivos.

Já para o *Cluster* 1 (Figura 8), no método manual de referência 8,4% dos *tweets* são classificados como com desinformação, 28,6% como sem desinformação e 63,0% permanecem não rotulados". Com os modelos, todos os métodos também aumentaram a proporção de *tweets* classificados como sem desinformação em comparação com o método manual, elevando essa classe para a faixa de 42–54% e a classe com desinformação para 15–17%, ao mesmo tempo em que comprimem a fração não rotulado para aproximadamente 29–43%, abaixo dos 63,0% do cenário manual. Novamente, o DistilBERT apresentou o menor aumento, seguido pela segmentação por "que", SBERT, BERT e SVM. Isso sugere que esses métodos são eficazes na identificação de *tweets* sem desinformação.

Em termos de comparação relativa ao manual, o DistilBERT tende a ficar mais próximo, sendo o que menos retira a classe "não rotulado" (C0: 43,3% e C1: 43,2%). BERT, sBERT e "que" tornam as classes rotuladas mais assertivas, principalmente *com desinformação* no C0 e *sem desinformação* no C1. Entre os resultados, o DistilBERT apresenta o desvio mais contido, enquanto BERT/sBERT/"que" oferecem maior assertividade com menor conservadorismo.

Os gráficos de retuítes, exibidos nas Figuras 9 e 10, mostram padrões semelhantes aos observados nos gráficos de *tweets*. No método manual de referência, o *Cluster* 0 (C0) apresenta 44,0% de retuítes classificados como "com desinformação", 11,2% como "sem desinformação" e 44,9% como "não rotulados". Com a aplicação dos modelos, todos os métodos de segmentação e classificação aumentaram a proporção de retuítes classificados como "com desinformação" em relação ao método manual: no DistilBERT, essa proporção sobe para 57,0%, para SBERT, segmentação por "que", BERT e SVM deslocam a classe "com desinformação" para a faixa de 57–63%, com a classe "sem desinformação" avançando para cerca de 5–13%. O DistilBERT foi o único que apresentou um decréscimo na classe "sem desinformação", onde houve apenas 2154 retuítes para "sem desinformação" em comparação aos 4493 no método manual. Os outros métodos aumentaram essa classe em relação ao manual mas de forma pouco significativa.

Já para o *Cluster* 1 (Figura 10), no método manual de referência 1,3% dos retuítes são classificados como "com desinformação", 44,5% como "sem desinformação" e 54,3% permanecem "não rotulados". Com os modelos, todos os métodos também aumentaram a proporção de retuítes classificados como "sem desinformação" em comparação com o método manual, elevando essa classe para a faixa de 51–72%, ao mesmo tempo em que comprimem a fração "não rotulado" para aproximadamente 24–54%, abaixo dos 62,6% do cenário manual. Novamente, o DistilBERT apresentou o menor aumento, seguido pela segmentação por "que", SBERT, BERT e SVM. Já a classe "com desinformação" apresentou um aumento porcentual menos significativo em relação ao manual, ficando na faixa de 1,6–2,3%, tendo o DistilBERT e SBERT bem próximos ao obtido manualmente (1,6 e 1,7% respectivamente).

O DistilBERT atua como um freio nessa tendência. Ele preserva uma parcela maior de "não rotulado" em ambos os *clusters*, aproximando a distribuição dos retuítes do que se vê no manual. Em C0, isso se traduz em um crescimento mais moderado da classe "com desinformação". Em C1, significa manter, em nível relativamente alto, a incerteza capturada pela classe "não rotulado", e, portanto, menor risco de super-rotulagem. Em resumo, entre os modelos testados, DistilBERT apresenta o desvio mais contido em relação ao padrão manual, enquanto BERT, sBERT e "que" fornecem leituras mais assertivas.

Do ponto de vista analítico, a opção entre assertividade e fidelidade orienta a escolha. Para reproduzir com maior proximidade o retrato manual e controlar falsos

positivos nos retuítes, DistilBERT é mais similar e retorna resultados mais conservadores. Para maximizar a rotulagem e obter uma visão mais clara dos polos de desinformação e informação correta, BERT, sBERT e a segmentação por "que" são mais eficazes.

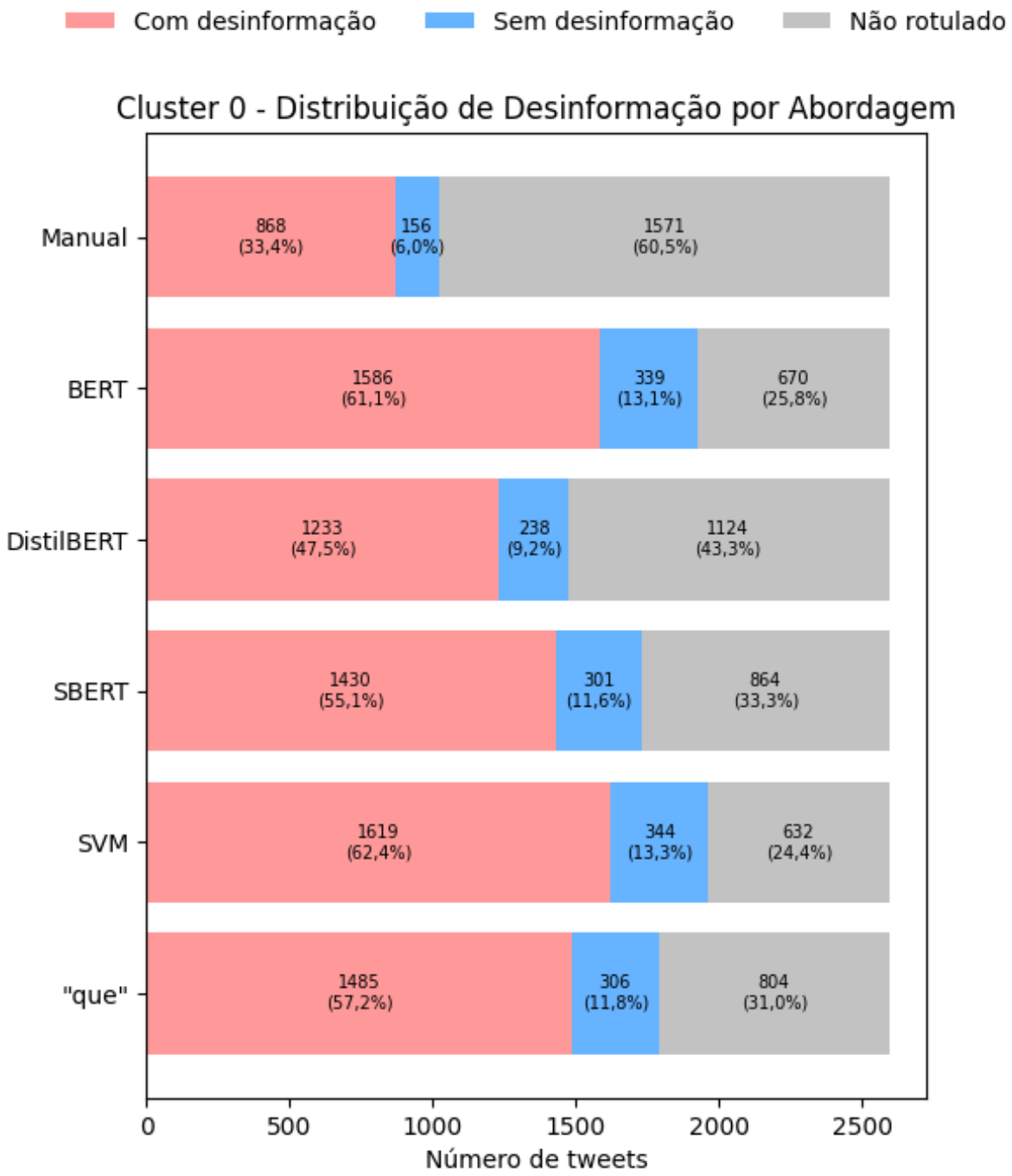

Figura 7 – Distribuição de Desinformação em *tweets* por abordagem para o *Cluster* 0 - Mês de outubro/2022

Fonte: elaboração própria.

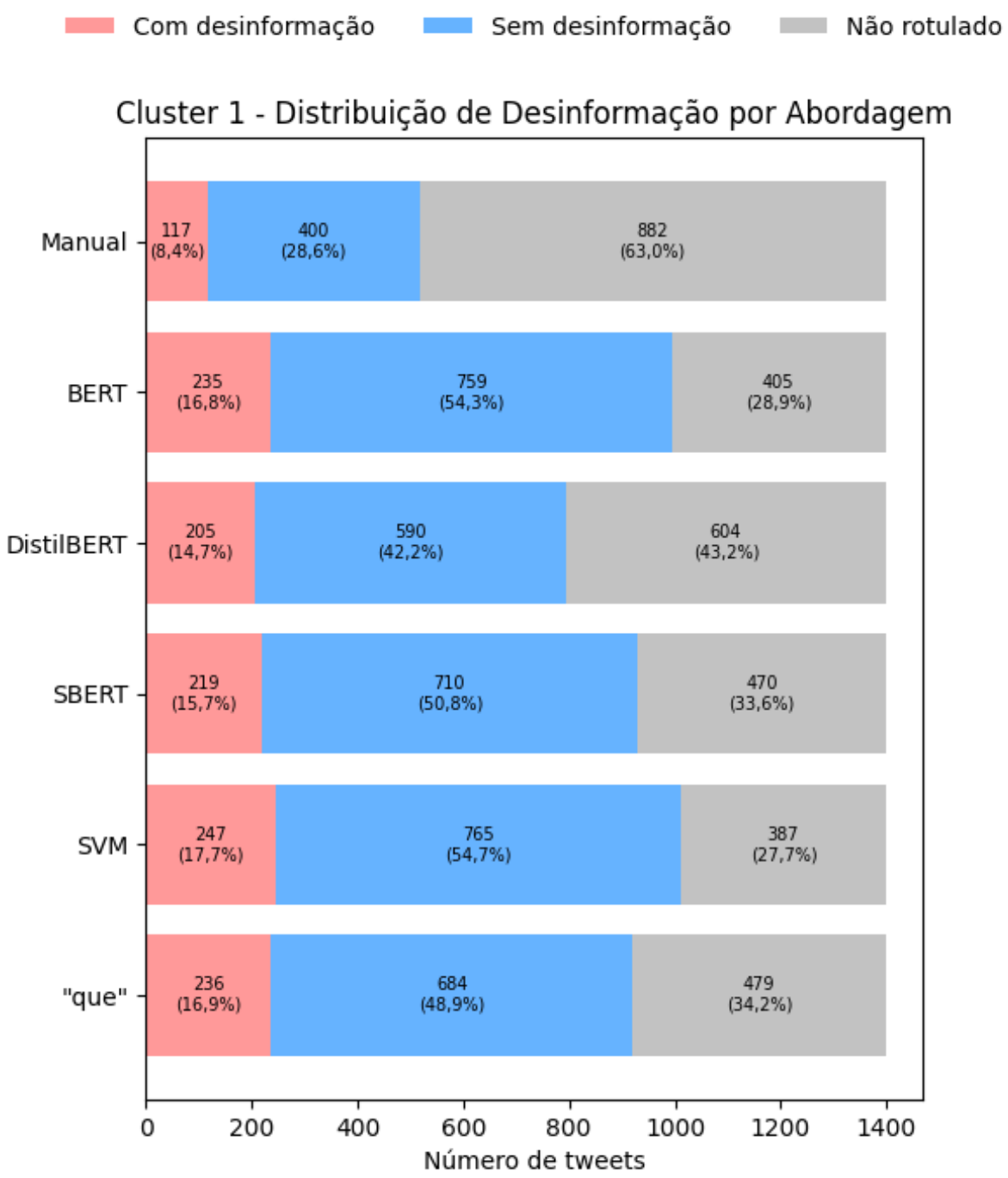

Figura 8 – Distribuição de Desinformação em *tweets* por abordagem para o *Cluster* 1 - Mês de outubro/2022

Fonte: elaboração própria.

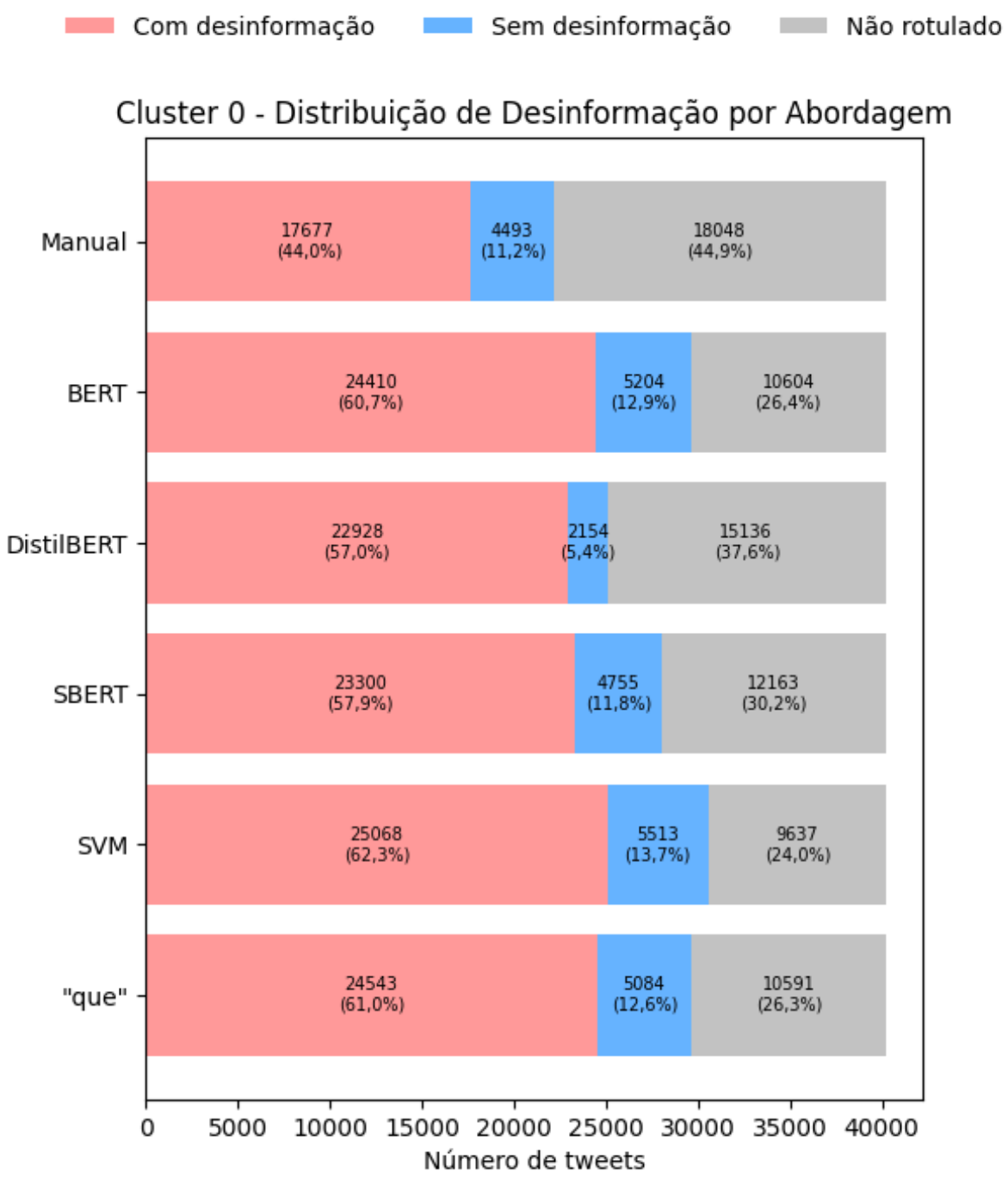

Figura 9 – Distribuição de Desinformação em retuítes por abordagem para o *Cluster* 0 - Mês de outubro/2022

Fonte: elaboração própria.

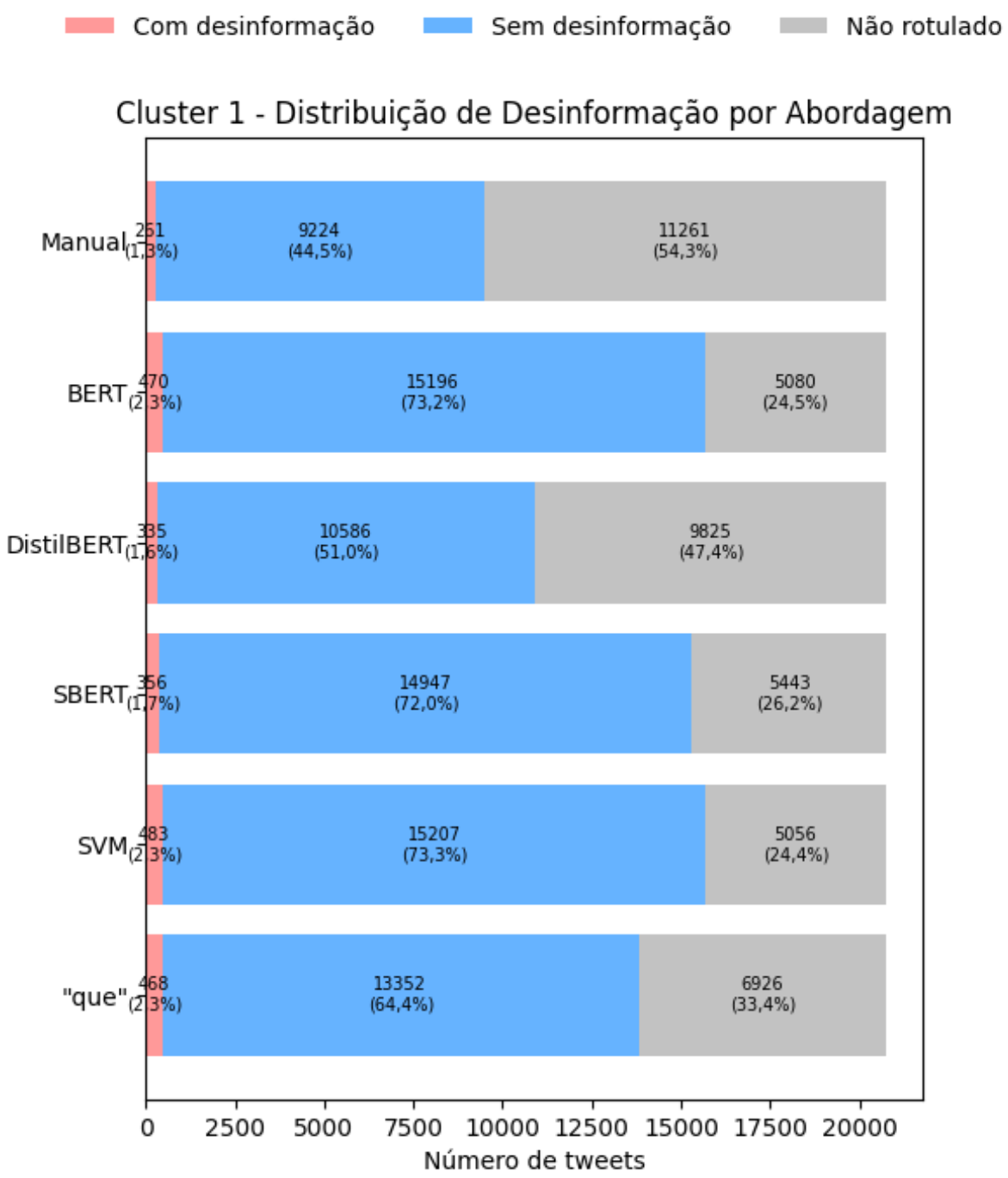

Figura 10 – Distribuição de Desinformação em retuítes por abordagem para o *Cluster* 1 - Mês de outubro/2022

Fonte: elaboração própria.

Tabela 4 – Termos-chave resultantes

Títulos das notícias	**Termos Manuais**	**Termos SVM**	**Termos DistilBERT**	**Termos SBERT**
Vulnerabilidades em urnas citadas em vídeo de 2014 já foram corrigidas	fraudar urnas, vídeo de Diego Aranha, não testou as urnas, como fiscalizar as urnas, Video viralizando Sábado, não te como TSE Fraudar as urnas	os problemas citados por ele na gravação já foram sanados pela justiça eleitoral, os problemas citados por ele na gravação já foram sanados pela justiça eleitoral, um vídeo de 2014 em	registra aleatoriamente os votos computados pelos eleitores, o tse (tribunal superior eleitoral) não testou a segurança das urnas para as eleições de 2022	o tse (tribunal superior eleitoral) não testou a segurança das urnas para as eleições de 2022
Apertar "confirma" durante tela "confira seu voto" na urna eletrônica anula voto é boato	TSE fraudar as urnas, fraudar eleições, derrota antecipada, fracasso eleitoral, Não votar nas urnas eletrônicas, muito mais votos registrados	estamos chegando na semana final da campanha eleitoral do primeiro turno, ao contrário, foi possível ver	apertar "confirma" durante a tela "confira seu voto" anula o voto, anulação é falsa	a tela em questão não tem qualquer relação com anulação
Correntes no WhatsApp e no Telegram mentem sobre o que eleitor pode ou não fazer no domingo	Apertar confira seu voto, perderá voto, votos não computadorizados, votos serão anulados, teclar ok ou confirma	mesários mal-intencionados tentariam boicotar o pleito, deixando de entregar comprovantes de votação, de coletar as assinaturas no livro de registro	no telegram disseminam informações falsas sobre o	aos fatosf: compartilhe correntes de mensagens no whatsapp
Acusação de fraude eleitoral domina correntes de WhatsApp em grupos monitorados	FRAUD3 NA ELE1ÇÃ0,Bolsonaro não pode deixar haver, Bolsonaro não deixa ter segundo turno, Fraude na eleição	esta reportagem foi feita numa colaboração entre agência pública, aos fatos	esta reportagem foi feita numa colaboração entre agência pública, aos fatos	esta reportagem foi feita numa colaboração entre agência pública, aos fatos

5 Considerações Finais

Este trabalho mostrou que é possível reduzir ruído e ampliar cobertura na coleta de conteúdo relevante para desinformação combinando etapas simples de PLN com modelos modernos baseados em *Transformers* e regex. Na sumarização, embora *BERT-base*, *DistilBERT* e *SBERT-MiniLM* tenham mantido proporções semelhantes do texto original, o *SBERT* destacou-se pelo melhor equilíbrio entre retenção semântica e custo computacional, tornando-o o candidato mais indicado para cenários de alto volume. Já o *DistilBERT* apareceu como um modelo mais conservador, preservando incerteza e reduzindo risco de super-rotulagem, ao passo que o *BERT-base* priorizou a retenção de trechos factuais completos.

A principal contribuição metodológica foi a segmentação guiada por conjunções, mais especificamente pelo termo "que". Essa estratégia, de baixo custo e fácil implementação, recuperou conjuntos maiores e mais pertinentes de trechos quando comparada a termos manuais, além de superar abordagens baseadas apenas em aspas, que se mostraram pouco consistentes. Em dados de referência com dois *clusters* polarizados, a segmentação por "que" e os modelos *BERT*/*sBERT* comprimiram a classe "não rotulado" e reforçaram o contraste entre *clusters* (C0 mais "com desinformação"; C1 mais "sem desinformação"), enquanto o *DistilBERT* manteve distribuição um pouco mais próxima do retrato manual.

Na detecção de afirmações factuais, o *XLM-R-Large-ClaimDetection* alcançou desempenho satisfatório mesmo fora do idioma de treino, com alta sensibilidade para sentenças relevantes (*recall* elevado) e menor precisão adequada à etapa de triagem automática. Em paralelo, o modelo *SVM* exibiu uma acurácia moderada; sua utilidade prática se destaca sobretudo pelo baixo custo computacional, o que o torna adequado como linha de base em cenários com recursos limitados.

Foram utilizadas bases de dados públicas para fins de análise, quantificação, comparação e classificação, com orientação da Profª Drª Denise Hideko Goya. A disponibilização da rede institucional e do acesso a artigos voltados à comunidade acadêmica contribuiu para a elaboração do *pipeline* proposto e possibilitou testar, em ambiente controlado, cenários de monitoramento em larga escala.

Em síntese, o *pipeline* proposto, sumarização leve (preferencialmente com *SBERT* para melhor escalabilidade), segmentação linguística simples (com "que") e um classificador sensível para triagem (*XLM-R*), mostrou ganhos relevantes de cobertura e eficiência sem depender de infraestrutura mais complexa.

5.1 Limitações

Apesar dos resultados positivos, algumas limitações importantes permaneceram. Em primeiro lugar, há restrições de hardware para realização de um *fine-tuning* ideal dos modelos de maior porte, em especial do *XLM-R-Large-ClaimDetection*. Essas limitações de capacidade computacional restringiram os experimentos à utilização de modelos pré-treinados, sem ajuste fino mais aprofundado aos dados específicos em português.

Em segundo lugar, o estudo depende de um modelo pré-treinado majoritariamente em inglês para tarefas aplicadas a textos em português. Essa diferença linguística pode impactar a eficácia do modelo, especialmente em nuances culturais e contextuais presentes na desinformação política local apesar de ser utilizado um modelo multilíngue.

5.2 Trabalhos Futuros

Como perspectiva para trabalhos futuros, recomenda-se, em primeiro lugar, calibrar *thresholds* com validação cruzada por *cluster*, de forma a adaptar os pontos de corte às características específicas de grupos de mensagens com perfis distintos. Em segundo lugar, sugere-se empilhar filtros híbridos (regras + *embeddings*) após o *XLM-R*, combinando a interpretabilidade de regras linguísticas com o poder de generalização de representações densas. Em terceiro lugar, é relevante testar esquemas de empacotamento eficiente de sequências e técnicas de *knowledge distillation* para reduzir custo computacional, fortalecendo a aplicabilidade do *pipeline* em cenários de monitoramento contínuo e em tempo quase real.

Além disso, seria relevante realizar uma comparação sistemática entre a Abordagem Linguística implementada neste estudo e modelos baseados em Abordagem de Redes descritos na literatura. A primeira foca na análise do conteúdo de mensagens enganosas, identificando padrões linguísticos que indicam engano, como o uso de pronomes, conjunções e palavras associadas a emoções negativas. Já a Abordagem de Redes utiliza informações de rede, por exemplo, metadados de mensagens ou consultas estruturadas em redes de conhecimento, para calcular medidas agregadas de engano (CONROY; RUBIN; CHEN, 2015). Ambas as abordagens tendem a incorporar técnicas de aprendizado de máquina para treinar classificadores especializados, e uma comparação sistemática poderia esclarecer em quais contextos cada uma delas apresenta vantagens relativas.

Adicionalmente, um ponto relevante consiste em realizar o *fine-tuning* do modelo *XLM-R-Large-ClaimDetection* para o português e comparar seu desempenho tanto com a abordagem linguística proposta neste trabalho quanto com modelos baseados em redes. Por conta das limitações de hardware enfrentadas durante este estudo, o *fine-tuning* do *XLM-R-Large-ClaimDetection* não foi possível, o que restringiu a análise ao uso do modelo

pré-treinado em inglês. Portanto, realizar esse ajuste fino para o português e avaliar seu desempenho em comparação com outras abordagens configura um passo importante para aprofundar a compreensão sobre a eficácia dos modelos de detecção de desinformação em diferentes idiomas e contextos.

Por fim, recomenda-se explorar processos de sumarização com a utilização de LLMs (*Large Language Models*) para avaliar se esses modelos conseguem capturar melhor o contexto e nuances do texto original, potencialmente melhorando a retenção semântica em comparação com os modelos baseados em *Transformers* utilizados neste estudo.

Referências

ARSLAN, F. et al. A benchmark dataset of check-worthy factual claims. *AAAI*, v. 53, 2020. Citado 4 vezes nas páginas 21, 22, 26 e 34.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, v. 5, p. 135–146, 2017. Disponível em: <https://transacl.org/ojs/index.php/tacl/article/view/999>. Citado na página 8.

BRITO, R. M. P. Detecção de posicionamento como abordagem para identificação de conteúdo desinformativo. *Relatório Técnico. XVII Simpósio de Iniciação Científica da UFABC*, 2024. Citado 4 vezes nas páginas 21, 23, 28 e 29.

CINELLI, M. et al. The covid-19 social media infodemic. *Scientific Reports*, v. 10, n. 1, p. 16598, 2020. Disponível em: <https://doi.org/10.1038/s41598-020-73510-5>. Citado 2 vezes nas páginas 3 e 15.

CONNEAU, A. et al. Unsupervised cross-lingual representation learning at scale. *ACL*, 2020. Disponível em: <https://aclanthology.org/2020.acl-main.747.pdf>. Citado na página 17.

CONROY, N.; RUBIN, V.; CHEN, Y. Automatic deception detection: Methods for finding fake news. *ResearchGate*, 2015. Disponível em: <https://www.researchgate.net/publication/281818865_Automatic_Deception_Detection_Methods_for_Finding_Fake_News>. Citado 2 vezes nas páginas 3 e 46.

COSTA, M. Ângelo A.; MARTINS, B. Uma comparação sistemática de diferentes abordagens para a sumarização automática extrativa de textos em português. *Linguamática*, v. 7, n. 1, p. 23–40, 2015. Citado na página 9.

DOMENICO, M. D. et al. *Covid19 Infodemic Observatory*. 2020. <http://dx.doi.org/10.17605/OSF.IO/N6UPX>. Disponível em: OSF (Open Science Framework). Citado na página 15.

FRANZESE, A. *Gen Z and Millennials now more likely to communicate with each other digitally than in person*. 2017. Disponível em: <https://www.prnewswire.com/news-releases/gen-z-and-millennials-now-more-likely-to-communicate-with-each-other-digitally-than-in-person-300537770.html>. Citado na página 4.

GARCÍA-SAISÓ, S. et al. Infodemia en tiempos de covid-19. *Revista Panamericana de Salud Pública*, Organización Panamericana de la Salud, v. 45, p. e89, 2021. Disponível em: <https://doi.org/10.26633/RPSP.2021.89>. Citado 2 vezes nas páginas 3 e 15.

GUPTA, V.; LEHAL, G. S. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, Academy Publisher, v. 2, n. 3, p. 258–268, 2010. Citado na página 9.

HACOHEN-KERNER, Y.; MILLER, D.; YIGAL, Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE*, v. 15, n. 5, p. e0232525, 2020. Citado na página 8.

HAN, B.-C. *No enxame: perspectivas do digital*. Petrópolis, RJ: Vozes, 2018. ISBN 9788532658517. Citado na página 15.

IRETON, C.; POSETTI, J. (Ed.). *Journalism, 'Fake News' & Disinformation: Handbook for Journalism Education and Training*. Paris: UNESCO Publishing, 2018. ISBN 978-92-3-100281-6. Disponível em: <https://unesdoc.unesco.org/ark:/48223/pf0000265552>. Citado 2 vezes nas páginas 3 e 14.

ITU. Almost three-quarters of the population are online. *ITU Statistics*, 2025. Disponível em: <https://www.itu.int/itu-d/reports/statistics/2025/10/15/ff25-internet-use/>. Citado na página 2.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition, online draft. ed. [s.n.], 2025. Online manuscript, 3rd edition draft, released August 24, 2025. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>. Citado na página 7.

KONSTANTINOVSKIY, L. et al. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *ACL Anthology*, 2021. Disponível em: <https://aclanthology.org/2021.acl-long.14/>. Citado 4 vezes nas páginas 3, 5, 6 e 16.

KRELL, M. M. et al. Efficient sequence packing without cross-contamination: Accelerating large language models without impacting performance. *arXiv*, 2022. Disponível em: <https://arxiv.org/abs/2107.02027>. Citado na página 26.

LEVY, R. et al. Context dependent claim detection. *ACM*, 2014. Citado na página 16.

LUPA, A. *Acesso pago a dados do Twitter pode prejudicar projetos que monitoram desinformação*. 2023. Acesso em: 18 nov. 2025. Disponível em: <https://lupa.uol.com.br/jornalismo/2023/02/03/api-twitter-projetos>. Citado na página 23.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999. Citado na página 8.

MARCHIORI, P. Z. A ciência e a gestão da informação: compatibilidades no espaço profissional. *Ciência da Informação*, Brasília, v. 31, n. 2, p. 72–79, 2002. Disponível em: <https://revista.ibict.br/ciinf/article/view/962>. Citado na página 15.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into text. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411. Disponível em: <https://aclanthology.org/W04-3252/>. Citado na página 12.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [s.n.], 2013. Disponível em: <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality>. Citado na página 8.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. Citado na página 9.

NENKOVA, A.; MCKEOWN, K. Automatic summarization. *Foundations and Trends in Information Retrieval*, Now Publishers Inc., v. 5, n. 2–3, p. 103–233, 2011. Citado na página 9.

Organização Pan-Americana da Saúde (OPAS/OMS). *Entenda a infodemia e a desinformação na luta contra a COVID-19*. Brasília, 2020. Disponível em: <https://iris.paho.org/handle/10665.2/52054>. Citado 2 vezes nas páginas 3 e 15.

OSHIKAWA, R.; QIAN, J.; WANG, W. Y. A survey on natural language processing for fake news detection. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020. p. 6086–6093. Citado na página 5.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://nlp.stanford.edu/pubs/glove.pdf>. Citado na página 8.

PENNYCOOK, G.; RAND, D. The psychology of fake news. *Trends in Cognitive Sciences*, v. 25, n. 5, p. 399–400, 2021. Citado na página 2.

RAPOZA, K. Can 'fake news' impact the stock market? *Forbes*, 2017. Citado na página 5.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3982–3992. Disponível em: <https://aclanthology.org/D19-1410/>. Citado na página 12.

RISCH, J. et al. Overview of the germeval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In: *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*. Duesseldorf, Germany: Association for Computational Linguistics, 2021. Disponível em: <https://aclanthology.org/2021.germeval-1.1>. Citado na página 26.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. *Communications of the ACM*, v. 18, n. 11, p. 613–620, 1975. Citado na página 8.

SAMI. *Model Card for Model ID*. 2024. HuggingFace. Disponível em: <https://huggingface.co/Sami92/XLM-R-Large-ClaimDetection>. Citado na página 18.

SANTOS, P. D. et al. Democracia sob ataque: polarização política e produção de conteúdos hostis no twitter nas eleições de 2022. *Revista Debates*, 2023. Disponível em: <https://seer.ufrgs.br/index.php/debates/article/view/129776>. Citado 4 vezes nas páginas 21, 22, 23 e 25.

SANTOS, R. *Processamento de Linguagem Natural: Recursos, Ferramentas e Aplicações para a Língua Portuguesa*. 2022. Dissertação de Mestrado. Disponível em: <https://sucupira-legado.capes.gov.br/sucupira/public/consultas/coleta/trabalhoConc

lusao/viewTrabalhoConclusao.jsf?popup=true&id_trabalho=11601526>. Citado na página 28.

SILVA, P. V. D. d. Pandemia e infodemia nas mídias: análise da desordem informacional no twitter. *AtoZ: novas práticas em informação e conhecimento*, v. 9, n. 2, p. 148–159, 2020. Citado 2 vezes nas páginas 2 e 13.

SOUZA, R. B. R. "fake news", pós-verdade e sociedade do capital: o irracionalismo como motor da desinformação jornalística. *FAMECOS*, v. 26, n. 3, 2019. Citado na página 2.

Tribunal Regional Eleitoral de São Paulo. *Fake news x desinformação: entenda qual é a diferença entre os termos*. 2023. Disponível em: <https://www.tre-sp.jus.br/comunicacao/noticias/2023/Agosto/fake-news-x-desinformacao-entenda-qual-e-a-diferenca-entre-os-termos>. Citado 3 vezes nas páginas 3, 14 e 15.

Tribunal Superior Eleitoral. *Programa Permanente de Enfrentamento à Desinformação no âmbito da Justiça Eleitoral: Plano Estratégico — Eleições 2022*. Brasília: [s.n.], 2022. Disponível em: <https://www.justicaeleitoral.jus.br/desinformacao/arquivos/programa-permanente-de-enfrentamento-a-desinformacao-novo.pdf>. Citado 2 vezes nas páginas 14 e 15.

TURNER, R. E. An introduction to transformers. *arXiv*, 2024. Disponível em: <https://arxiv.org/pdf/2304.10557>. Citado na página 12.

UMAIR, M.; SULTANA, T.; LEE, Y.-K. Pre-trained language models for keyphrase prediction: A review. *ICT Express*, v. 10, n. 4, p. 871–890, 2024. Citado na página 13.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2017. Citado 4 vezes nas páginas 10, 11, 12 e 17.

VOSOUGHI, S.; ROY, D.; ARAL, S. The spread of true and false news online. *Science*, v. 359, n. 6380, p. 1146–1151, 2018. Disponível em: <https://doi.org/10.1126/science.aap9559>. Citado 2 vezes nas páginas 3 e 15.

WARDLE, H. D. C. Os três tipos de desordem informacional: Desinformação (dis-information), informação falsa (mis-information) e informação maliciosa (mal-information). *PUC*, 2017. Disponível em: <https://periodicos.pucminas.br/SapereAude/article/view/32503>. Citado 2 vezes nas páginas 3 e 14.

YACOUBY, R.; AXMAN, D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. *ACL Anthology*, 2020. Disponível em: <https://aclanthology.org/2020.eval4nlp-1.9/>. Citado na página 29.

ZHOU, X.; ZAFARANI, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, v. 53, 2020. Citado na página 13.