Comparação de Técnicas de Classificação para Identificação de *Fake News* em Português

Matheus Ferreira Pereira[1], **Edimar Manica**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS)
Rua Nelsi Ribas Fritsch, 1111 – Ibirubá – RS – Brasil

matheusferreirap67@gmail.com, edimar.manica@ibiruba.ifrs.edu.br

Abstract. *The advance of the Internet has resulted in an increase in the dissemination of information, but the quality of this information cannot be guaranteed, leading to the problem of Fake News. These are responsible for different types of damage to society. To solve this problem, there are different automatized approaches. This work compared different techniques used to identify Fake News, seeking to help understand what is more effective for the task in the Portuguese language. Nine research questions were raised and experiments were performed on a database with 7,200 real news in order to answer them. The best result was achieved by the BERTimbau model, a pre-trained BERT model in Portuguese, reaching F1-Measure of 99%.*

Resumo. *O avanço da Internet resultou em um aumento na disseminação de informações, mas não é possível garantir a qualidade dessas informações, levando ao problema de Fake News. As mesmas são responsáveis por diferentes tipos de danos a sociedade. Para resolver esse problema, existem distintas aproximações automatizadas. Este trabalho comparou diferentes técnicas usadas para a identificação de Fake News, buscando ajudar a entender o que é mais eficaz para a tarefa no idioma Português. Foram levantadas 9 questões de pesquisa e realizados experimentos em uma base de dados com 7.200 notícias reais para respondê-las. O melhor resultado foi atingido pelo modelo BERTimbau, um modelo BERT pré-treinado em português, alcançando Medida-F1 de 99%.*

1. Introdução

A constante expansão da Internet transformou a forma como as pessoas se comunicam. Plataformas de mídia social, como Twitter[1] e Facebook[2], facilitam a distribuição de informação entre usuários ao redor do mundo, mas a qualidade das informações distribuídas não necessariamente se equiparam com os meios tradicionais (ZHANG; GHORBANI, 2020).

A desinformação e a propaganda não são fenômenos recentes, existem pelo menos desde a Roma Antiga. Também é evidenciado que ao longo do tempo, as mesmas passaram por transformações, se tornando cada vez mais comuns. Um grande avanço de desinformação emerge junto do surgimento de mídias sociais, pois essas servem como enormes ferramentas de propagação. Essas mudanças levaram a popularização de um

[1]Disponível em <https://twitter.com/home>. Último acesso em 22/06/2022.
[2]Disponível em <https://www.facebook.com/>. Último acesso em 22/06/2022.

termo para essas informações falsas, *Fake News*, traduzido livremente do inglês, como Notícias Falsas (POSETTI; MATTHEWS, 2018).

Esse termo pode ser abrangente e também é constantemente aplicado de formas diferentes, gerando debate sobre sua definição, mas que sempre apresenta certas características, sendo essas: a) a capacidade de apropriar a aparência e o sentimento de uma notícia verdadeira; b) a forma como os artigos são escritos e até mesmo a aparência visual dos sites buscam imitar notícias verdadeiras; c) as imagens possuírem fontes com a finalidade de parecer mais como uma notícia legítima, imitando algum tipo de credibilidade (TANDOC; LIM; LING, 2018).

Estudos demonstram que notícias falsas são capazes de afetar a população em geral, recentemente evidenciadas em situações políticas (RECUERO; GRUZD, 2019) e de saúde pública (JúNIOR et al., 2020). Além de terem até mesmo a capacidade de influenciar eleições ao disseminar falsas informações sobre os candidatos envolvidos (ALLCOTT; GENTZKOW, 2017).

Segundo Allcott e Gentzkow (2017) e Júnior et al. (2020), *Fake News* possuem uma capacidade de espalhar desinformação em diferentes escalas, causando distintos tipos de danos à sociedade. No entanto, identificar que uma notícia é falsa não é uma tarefa trivial.

Um estudo realizado por (BOND; DEPAULO, 2006) demonstrou que a capacidade das pessoas, sem preparo preliminar ou ajuda em tempo real, para julgar mentiras em relação a verdades, atinge uma média de apenas 54% de julgamento correto. Nesse estudo, foram classificadas corretamente 47% das mentiras como "enganosas" e 61% das verdades como "não enganosas".

Se forem levados em consideração os impactos negativos que *Fake News* podem causar para a população conforme levantado anteriormente. Também for considerado que *Fake News* tem a intenção de enganar as pessoas e essas, naturalmente, possuem uma dificuldade em discernir o que é verdade. É possível identificar que existe a necessidade de investigações sobre soluções automatizadas para averiguar a veracidade de uma notícia (SHU et al., 2017).

Atualmente, existe uma série de técnicas diferentes que podem ser utilizadas para o processo de detecção automática de *Fake News*. Porém, não existe uma solução já estabelecida como a definitiva. Nesse sentido, este artigo realiza um conjunto de experimentos com o intuito de comparar a eficácia de algumas das principais técnicas de identificação de *Fake News* presentes na literatura. O estudo tem como foco o idioma português. Os experimentos foram realizados em um *corpus* de notícias em português, contendo 7.200 notícias, sendo 3.600 falsas e 3.600 verdadeiras. O objetivo dos experimentos era avaliar algumas das técnicas, parâmetros e algoritmos mais eficazes para identificação de *Fake News*. Para alcançar esse objetivo, foram definidas as seguintes questões de pesquisa:

1. "a técnica é mais eficaz removendo a acentuação das palavras?"
2. "a técnica é mais eficaz ao remover os números?"
3. "a técnica é mais eficaz ao remover as *stop words*"
4. "a técnica é mais eficaz ao utilizar *Stemming*
5. "qual é a quantidade de n-gramas mais eficaz"

6. “qual é a quantidade de *features* mais eficaz"
7. “é mais eficaz utilizar apenas a frequência do termo para ponderar as palavras?"
8. “qual o modelo de classificação que possui a melhor eficácia entre SVM e Regressão Logística"
9. “qual o modelo de classificação que possui a melhor eficácia entre BERTimbau e SVM?"

O restante do trabalho está organizado da seguinte forma. A Seção 2 define os conceitos necessários para compreensão deste trabalho. A Seção 3 apresenta e compara os trabalhos relacionados. A Seção 4 descreve a metodologia utilizada. A Seção 5 apresenta e discute os resultados obtidos. Por fim, a Seção 6 concluí o trabalho e sugere trabalhos futuros.

2. Fundamentação Teórica

Esta seção apresenta os conceitos necessários para a compreensão deste trabalho. Este trabalho se encontra dentro do campo de Aprendizado de Máquina. Essa técnica é compreendida como “a capacidade da máquina de adquirir seu próprio conhecimento ao extrair padrões de dados sem tratamento” (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma outra técnica relevante para este trabalho é a de Processamento de Linguagem Natural - *Natural Language Processing* (NLP). Essa técnica utiliza Aprendizado de Máquina para processar e interpretar a língua humana (VASILAKES; ZHOU; ZHANG, 2020). Bases de dados específicas para NLP, como a utilizada neste trabalho, geralmente são chamadas de *corpus*. Segundo Percy et al. (1996), um *corpus* é “uma coletânea de porções de linguagem que são selecionadas e organizadas de acordo com critérios linguísticos explícitos, a fim de serem usadas como uma amostra da linguagem”.

Será necessário utilizar técnicas de mineração de dados para o desenvolvimento do trabalho. Mineração de dados é a descoberta de estruturas interessantes, inesperadas ou preciosas em grande bancos de dados. Para auxiliar a extração desse conhecimento útil, é utilizada uma combinação de estatística com ideias, ferramentas e métodos da Ciência da Computação (HAND, 2007).

A subseção 2.1 especifica o modelo de processo de mineração de dados adotado. A subseção 2.2 descreve diferentes formas para a representação de dados. A subseção 2.3 define a tarefa de classificação e explica os modelos de classificação de dados utilizados no trabalho.

2.1. CRISP-DM

Esta seção apresenta o modelo de processo de mineração de dados adotado para a elaboração deste trabalho, CRISP-DM (*CRoss Industry Standard Process for Data Mining* - Processo Padrão Inter-Indústrias para Mineração de Dados).

Esse modelo tem por objetivo inicial, catalogar e guiar os passos mais comuns em projetos de mineração de dados, tendo se tornado “o padrão de fato para desenvolver projetos de mineração de dados e descoberta de conhecimento” (MARBáN et al., 2009). Apesar das mudanças e o avanço da área, CRISP-DM continua uma referência por mais

de 20 anos, sendo ainda relevante na área de ciência de dados como uma das metodologias mais utilizadas (MARTINEZ-PLUMED et al., 2021).

O modelo descreve diferentes passos que precisam ser realizados e fornece conselhos para os mesmos. Segundo Wirth (2000), na sua versão genérica o ciclo de vida de um projeto seguindo o modelo CRISP-DM é dividido em seis fases, conforme a Figura 2, que são:

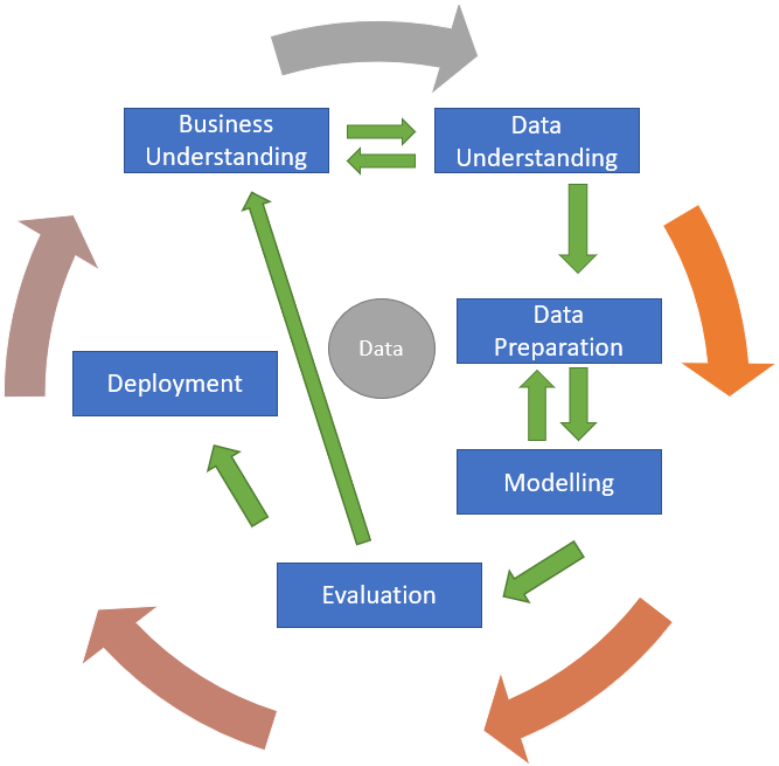

Figura 1. Fases genéricas do ciclo de vida do CRISP-DM
Fonte: (WIRTH, 2000)

- *Business Understanding* (Compreensão do Negócio) - A fase inicial que foca em entender os objetivos e requisitos do projeto a partir de uma perspectiva de negócio;
- *Data Understanding* (Compreensão dos Dados) - Consiste em coletar dados, identificar suas características, averiguar suas qualidades e explorá-los;
- *Data Preparation* (Preparação dos Dados) - Todas as atividades necessárias para a criação da base de dados final, incluindo, selecionar, limpar, construir, integrar e formatar dados;
- *Modeling* (Modelagem) - Nessa fase, várias técnicas de aprendizado de máquina são selecionadas e aplicadas em modelos diferentes, também são criados modelos de design de testes. Por fim, um modelo final é treinado, até que seus parâmetros sejam capazes de atingir valores otimizados;
- *Evaluation* (Avaliação) - Avaliação e revisão dos resultados, também determina os próximos passos para o projeto;

- *Deployment* (Entrega) - Pode representar tanto um simples relatório final, quanto a implementação contínua do processo, dependendo do projeto em que está inserido.

A sequência das fases não é restrita, o objetivo das flechas é apenas indicar quais são as fases mais importantes, e as frequentes dependências e interações entre elas. As flechas exteriores representam o ciclo da mineração de dados, uma vez que esse não precisa ser finalizado quando a solução é entregue, já que novas questões de negócio e dados podem surgir a partir dos anteriores (WIRTH, 2000).

2.2. Representação de dados

Esta subseção descreve de forma sucinta as representações de dados que são utilizadas neste trabalho, são elas *Bag of Words* e *Tem Frequency Inverse Document Frequency*. A seguir cada representação é explicada.

Bag of Words (BOW) - É um modelo simplificado que transforma textos em vetores de palavras. Cada coluna do vetor representa uma palavra, o valor em cada linha representa a quantidade de vezes que a palavra aparece numa sentença. Essa aproximação foca apenas na ocorrência das palavras, ignorando o contexto (GOLDBERG, 2017).

Term Frequency Inverse Document Frequency (TF-IDF) - É uma medida estatístico que representa, como o nome sugere, a frequência do termo em relação ao inverso da frequência nos documentos. Em outras palavras, a medida leva em consideração o número de aparições de uma palavra em um documento, sendo equilibrado pela frequência total da palavra em todo o *corpus*. Buscando então demonstrar a relevância de uma palavra em dado *corpus* de documentos (RAMOS, 2003). O peso se torna maior quando um termo ocorre várias vezes em um acervo pequeno de documentos. O peso se torna menor quando um termo ocorre poucas vezes em um único documento ou ocorre muitas vezes em muitos documentos. Essa medida pode ser incorporada a representação BoW para atribuir pesos para as palavras.

2.3. Modelos de classificação

Esta subseção apresenta a tarefa de classificação, e explica os modelos de classificação que foram utilizados neste trabalho.

A mineração de dados pode ser classificada em diferentes tarefas, sendo as principais: Descrição, Classificação, Estimação, Predição, Agrupamento e Associação. Este trabalho se enquadra na tarefa de classificação supervisionada.

Uma das tarefas mais comuns, a Classificação, tem por objetivo identificar se determinada amostra pertence ou não a uma classe. Nesta tarefa, é analisado um conjunto de dados fornecido, esses dados são rotulados com qual classe cada amostra pertence, isto serve para que um modelo seja capaz de notar padrões da classe e "aprender" como classificar uma nova amostra (Aprendizado Supervisionado) (LAROSE, 2005).

Neste trabalho, foram utilizados três modelos de classificação: Regressão Logística, Máquina de Vetores de Suporte e BERT. A seguir, cada um desses modelos é explicado.

Regressão Logística - *Logistic Regression* (LR), é um método de Aprendizado de Máquina supervisionado para classificação. Esse método pode utilizar diferentes tipos de

dados característicos, permitindo dizer quais dados e combinações de dados são úteis. Isso ocorre através de uma curva a qual representa a relação das variáveis. Essa curva vai de 0 a 1, então classificando apenas como falso e verdadeiro em relação a ela (GOODFELLOW; BENGIO; COURVILLE, 2016).

Máquina de Vetores de Suporte - *Support Vector Machine* (SVM) - é um modelo linear de aprendizado de máquina supervisionado. SVM constrói um hiperplano em espaços dimensionais maiores para usá-los para classificação, regressão ou diferentes tarefas. SVM move o conjunto de dados para dimensões maiores através de funções *kernel*. Isso permite que seja possível visualizar melhor a linha que separa as classes de dados. Essa linha é estabelecida a partir dos pontos mais próximos entre cada classe, esses chamados por Vetores de Suporte. A distância entre a linha e os Vetores de Suporte é chamado de margem. O objetivo é maximizar essa margem para melhorar a classificação, uma vez que quanto maior é a margem, menor é o erro de generalização do classificador e sobreajuste (BURGES, 1998).

Bidirectional Encoder Representations from Transformers (BERT) é um algoritmo de aprendizado profundo de NLP. É um modelo pré treinado que utilizou mais de 3,3 bilhões de palavras. Ele pode ser afinado para tarefas específicas. O intuito do BERT é ajudar a entender a linguagem ambígua e o contexto presente. Ele é um modelo bidirecional, ou seja, é capaz de ler textos e levar em consideração o contexto da esquerda para a direita e da direita para a esquerda de uma única vez. O modelo BERT possui sua própria forma de representar dados e criar seu vocabulário. A sua forma de representação permite com que o mesmo entenda o significado contextual de cada sentença (ROGERS; KOVALEVA; RUMSHISKY, 2020). Um exemplo é que o modelo é capaz de diferenciar a seguinte frase: “Eu quero comprar um Jaguar para andar nas ruas", ao diferenciar o significado de Jaguar sendo como a marca de carro e não do animal. Um modelo de BERT foi pré-treinado na língua portuguesa, esse é chamado por BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020).

3. Trabalhos Relacionados

Esta seção descreve e compara os trabalhos relevantes relacionados ao tema. Esses trabalhos foram selecionados por meio de buscas no Google Scholar[3] e Mendeley[4] a partir das seguintes palavras-chaves: “*Fake News*”, “*Fake News Detection*” e “*Fake News Automatic Detection*”. Os resultados foram filtrados a partir do idioma e relevância. Os idiomas considerados foram português e inglês. A relevância considerada foi a estabelecida pelo algoritmo da plataforma utilizada. Os primeiros resultados de cada plataforma foram analisados por meio do título e resumo para ver se contemplavam o escopo do trabalho, até que um total de cinco fossem selecionados.

Em (MONTEIRO et al., 2018) e (SILVA et al., 2020) encontram-se duas publicações diferentes que introduzem o *corpus* utilizado neste trabalho. Os autores reconheceram tanto os problemas causados pelas *Fake News*, quanto o progresso das tecnologias para a detecção automática das mesmas. Também, foi identificada a ausência até então de um *corpus* da língua portuguesa, dificultando o progresso no desenvolvimento, avaliação

[3]Disponível em: <https://scholar.google.com.br/>. Último acesso em: 28/07/2022.

[4]Disponível em: <https://www.mendeley.com/>. Último acesso em: 28/07/2022.

e comparação de diferentes métodos para a detecção automática de *Fake News* em português. Para a finalidade de resolver esse problema, foi criado um *corpus*, denominado, Fake.Br Corpus. Mais detalhes sobre o *corpus* é encontrado na Seção 4.1 deste trabalho.

Na experimentação realizada por Monteiro et al. (2018), os textos foram normalizados (em número de palavras) ao truncar de acordo com a notícia alinhada. Foi utilizado SVM como método de classificação. Como técnicas foram utilizadas as formas de representação BOW e POS, bem como características linguísticas que pudessem servir de indicadores de conteúdo falso. Para avaliação, foi usada validação cruzada de 5 *folds* e as métricas precisão, revocação, acurácia e Medida-F1. Como resultados, BOW atingiu 88% de Medida-F1. Por outro lado, combinando com indicadores linguísticos obteve-se 89%. A partir desses resultados, os autores concluíram que os mesmos não ajudaram a melhorar consideravelmente a eficácia. Fake.Br foi o primeiro *corpus* para detecção de *Fake News* em português.

O trabalho realizado por Monteiro et al. (2018) foi continuado em (SILVA et al., 2020), onde novas experimentações foram feitas, toda amostra foi representada com três técnicas: BoW utilizando TF-IDF para ajustar o peso dos *tokens* de cada documento, Word2Vec (MIKOLOV et al., 2013) e FastText (BOJANOWSKI et al., 2017). Ajustes e normalização como a conversão de todos os textos para minúsculo foram realizados. Experimentos foram realizados com as notícias em seu tamanho completo e também truncadas. Foram utilizados indicadores linguísticos. Como métodos foram utilizados SVM, Regressão Logística, Árvores de Decisão, Floresta Aleatória, Bagging e AdaBoost. Para medidores de performance foram utilizados: Taxa de bloqueio de notícias legítimas, Taxa de *Fake News* detectadas, Taxa de precisão de *Fake News* e Medida-F1. Notícias com tamanho completo obtiveram melhor performance do que notícias com tamanho truncado, fazendo o autor concluir que existe um possível viés no tamanho, uma vez que notícias verdadeiras tendem a ser maior que notícias falsas. Então, os resultados a serem considerados foram os que utilizaram texto truncado. A Regressão Logística com BoW atingiu 93.7% de Medida-F1. Ao combinar os resultados da Regressão Logística com BoW e indicadores linguísticos foi obtido o melhor resultado de 96.5% de Medida-F1.

Outro trabalho relacionado a identificação de *Fake News* foi realizado por Jwa et al. (2019). Esse teve por objetivo a construção de BAKE, um modelo de detecção automático de *Fake News* em inglês. O modelo é baseado em BERT e busca aprimorá-lo ao mitigar o problema de desequilíbrio de dados. Seu foco principal foi analisar a relação entre o título e o corpo de um artigo para a identificação automática de *Fake News*. Inicialmente, o modelo BERT foi pré-treinado utilizando uma base de dados de notícias genéricas, CNN-Daily Mail, criando o exBAKE. Após, a base de dados FNC-1, específica para *Fake News*, foi usada para refinamento. O conjunto de treinamento foi composto por pares de títulos e seus textos, incluindo o apropriado rótulo de classe para cada par. Também foram adicionados pares sem rótulos de classe. Um total de 2,587 pares do FNC-1 foram utilizados. Foi utilizada a Medida F1 como métrica de avaliação. Testes foram feitos comparando com outros modelos do estado da arte. No total exBAKE foi capaz de atingir o estado da arte em F1. Os resultados sugerem que BERT é o melhor modelo para esse tipo de tarefa devido a sua natureza de contextualização profunda. Também sugerem que incorporar conhecimento extra de grandes base de dados de notícias é útil para a tarefa de identificação. Por fim, o trabalho foi capaz de proporcionar uma ferramenta de

identificação de Fake News automatizada em tempo real, que segundo o autor, não existia previamente.

Outro trabalho de referência para a identificação de *Fake News* foi realizado por Wang (2017). Nesse trabalho foi criado um *corpus*, manualmente rotulado, de *fake news* em inglês. O *corpus* foi criado para que servisse de base e referência para o problema de detecção automática, constituído por 12.836 pequenas declarações. Essas declarações possuem diversos contextos. Também combina metadados com texto. O trabalho também realizou experimentos buscando comparar diferentes métodos e técnicas. Para os experimentos, foram utilizados os seguintes modelos de classificação: Regressão Logística, SVM, Bi-LSTMs e CNNs. A base de dados foi dividida em três partes, uma para treinamento (10.269), outra para validação (1.284) e outra para teste (1.283). Foi utilizada uma representação de dados Word2Vec com 300 dimensões pré-treinadas na base de dados do Google News, encontrada em (MIKOLOV et al., 2013). Também foram refinados todos os hiperparâmetros na base de dados de validação. Como métrica de avaliação foi utilizada a precisão. Os autores foram capazes de entregar um *corpus* de grande porte funcional e curado. Foi possível identificar uma melhora nos modelos SVM e LR, segundo o autor, Bi-LSTM não teve boa performance por causa do sobreajuste. Os modelos CNNs tiveram a melhor performance, resultando em uma precisão de 27% em teste.

Tabela 1. Comparação dos trabalhos relacionados

Referência	[Monteiro et al. 2018]	[Silva et al. 2020]	[Wang 2017]	[Jwa et al. 2019]	Este trabalho
Corpus utilizado	Fake.Br	Fake.Br	LIAR	FNC-1 CNN-DailyMail	Fake.Br
Idioma do corpus	Português	Português	Inglês	Inglês	Português
Tamanho	7.200 notícias	7.200 notícias	12.836 pequenas declarações	2.587 notícias	7.200 notícias
Representação de dados	BoW, POS	BoW (TF-IDF) Word2Vec FastText	Word2Vec	BERT	BoW (TF-IDF) BERT
Métodos de classificação	SVM	LR, SVM DT, RF Bagging AdaBoost	LR SVM Bi-LSTMs CNNs	BERT BAKE exBAKE	LR SVM BERT
Métricas de avaliação	Acurácia Precisão, Revocação Medida-F1	LBR FCR FPR Medida-F1	Precisão	Medida-F1	Precisão Revocação Medida-F1
Melhores resultados	BoW	LR + BoW	CNNs	exBAKE	BERTimbau

A Tabela 1 apresenta uma análise comparativa entre os trabalhos relacionados e o presente trabalho. Observa-se na tabela que existem diversas técnicas que podem ser utilizadas para o processo de detecção de *Fake News* e o aperfeiçoamento do mesmo.

Todos os melhores resultados foram distintos, mesmo para trabalhos que utilizaram as mesmas técnicas. Por outro lado, as medidas de eficácia continuam sendo quase sempre as mesmas aplicadas em modelos de classificação.

Por fim, para a realização desse trabalho de conclusão de curso, foi utilizado o *corpus* apresentado em (MONTEIRO et al., 2018) e (SILVA et al., 2020). Foram utilizadas as técnicas SVM e Regressão Logística que aparecem em diversos trabalhos. De (JWA et al., 2019) foi utilizada a aproximação do modelo BERT. Foi utilizada a técnica BoW utilizando TF-IDF para balancear os pesos dos *tokens* conforme (SILVA et al., 2020). Este trabalho utiliza BERT em português que não está presente em nenhum dos trabalhos relacionados.

4. Metodologia

Essa seção apresenta a metodologia utilizada no trabalho. A Figura 2 (a) ilustra a aplicação do modelo criado neste trabalho, onde uma notícia é inserida no detector automático de Fake News, a mesma é classificada como verdadeira ou falsa, a partir do modelo criado. A Figura 2 (b) descreve de forma genérica as etapas principais para obter o modelo final.

O modelo adotado foi o CRISP-DM para elaborar as diferentes etapas da metodologia. Primeiramente foi necessário selecionar uma base de dados rotulada em português. Após, foram selecionadas e avaliadas diferentes técnicas de pré processamento conforme presente na subseção 5.1. Então, foram selecionados e avaliados diferentes parâmetros conforme presente na subseção 5.2. Após, foram selecionados e avaliados diferentes algoritmos conforme descrito nas subseções 5.3 e 5.4. Por fim, ocorre a criação do modelo final, que serve como base para um detector automático de *Fake News*. No detector é inserida uma notícia, que é processada, por fim, a notícia é retornada como verdadeira ou falsa.

Para facilitar a clareza do trabalho, são denominadas técnicas de pré-processamento: a remoção de números, acentos, *stop words* e o *Stemming* e serão consideradas apenas parte do pré-processamento. E como parâmetros serão considerados apenas os parâmetros presentes no algoritmo TF-IDF. Isso se torna algo necessário ressaltar já que existe uma sobreposição comum entre os elementos, uma vez que, por exemplo, *stop words* podem ser parâmetros dentro de um modelo.

Para facilitar a compreensão, neste trabalho são denominadas técnicas de processamento: a remoção de números, acentos, *stop words* e o uso de *Stemming*. Por outro lado, são denominados parâmetros a quantidade de n-gramas e o número de *features* e a inclusão do IDF. Essa diferenciação se faz necessária uma vez que em algumas ferramentas as técnicas de pré-processamento podem ser parâmetros dos algoritmos de classificação.

A subseção 4.1 descreve as características da base de dados utilizada. A subseção 4.2 especifica as métricas utilizadas no trabalho.

4.1. Base de Dados

Para a base de dados, foi utilizado o *corpus* Fake.br-Corpus (MONTEIRO et al., 2018). Esse é composto de notícias verdadeiras e falsas em português brasileiro. Foi feito um alinhamento para cada notícia falsa, onde foi coletada uma verdadeira correspondente, que se não necessariamente a nega, é relacionada ao tema.

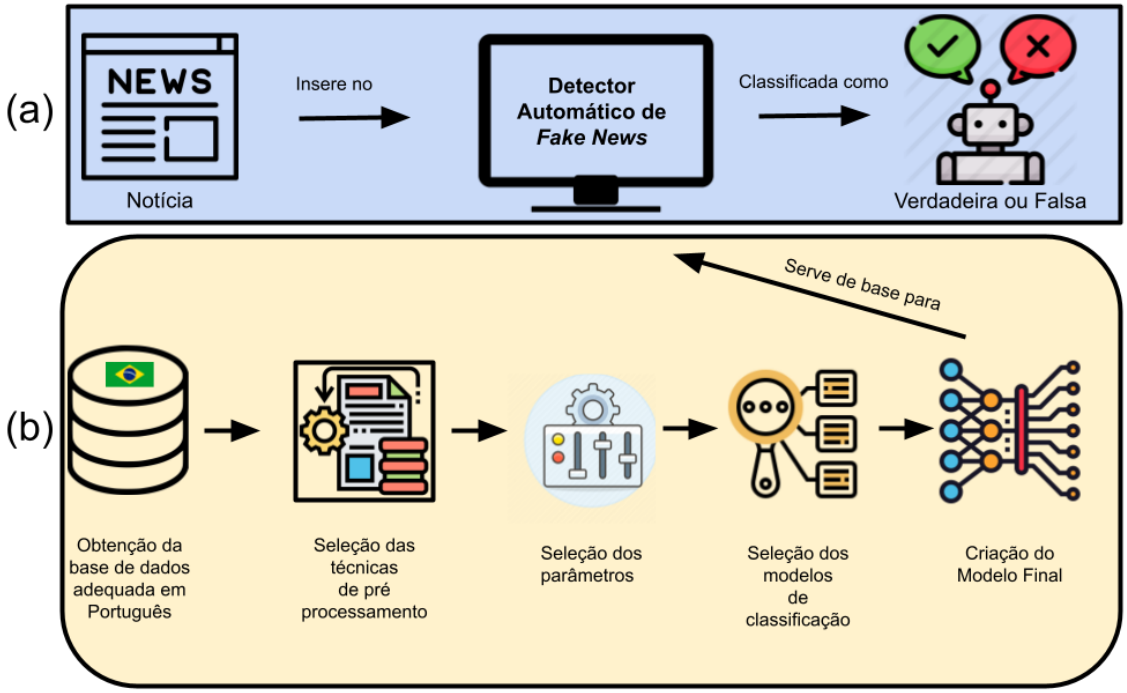

Figura 2. Visão Geral do Trabalho (Fonte: Autor)

Outras medidas também foram adotadas: a) notícias estão em formato de texto simples, para serem mais apropriadas para NLP; b) as notícias precisam ter tamanhos similares (normalmente no número de palavras) para evitar viés; c) normalização do tamanho da notícia ocorre quando necessário e d) as notícias foram coletadas em um período específico de tempo para manter um único estilo de escrita.

No total foram coletadas 7.200 notícias, sendo metade falsas e metade verdadeiras, filtrando as notícias que são meia verdades[5]. Todas as notícias foram separadas em 6 categorias: política (58%), TV e celebridades (21,4%), sociedade e notícias diárias (17,7%), ciência e tecnologia (1,5%), economia (0,7%) e religião (0,7%). Também foram salvos todos os metadados, contendo o nome do autor, o link, a categoria, a data, informações estatísticas e características linguísticas.

As notícias também são separadas em categorias e possuem metadados, mas essas informações não foram utilizadas neste trabalho. Para a realização dos experimentos, foram usadas as notícias com tamanho truncado mas sem pré-processamento, essas foram colocadas em um documento CSV contendo duas colunas, uma para o texto e outra para a classe. A classe possui dois valores possíveis: verdadeira ou falsa.

4.2. Métricas

Para avaliar a eficácia das técnicas de processamento, dos diferentes parâmetros e algoritmos, foram adotadas métricas tradicionais para tarefas de classificação. Essas métricas são calculadas a partir dos conceitos de Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo (MARIANO et al., 2020). No contexto de *Fake News*:

[5]"Afirmação que omite parte dos fatos ou das informações, principalmente quando é feita propositadamente com o objetivo de enganar alguém" (MEU..., 2022)

- **Verdadeiro Positivo** (VP) é uma notícia falsa que foi identificada corretamente como falsa;
- **Falso Negativo** (FN) é uma notícia falsa que foi identificada como verdadeira.
- **Falso Positivo** (FP) é uma notícia verdadeira que foi identificada como falsa;
- **Falso Negativo** (FN) é uma notícia falsa que foi identificada como verdadeira.

Os conceitos de Verdadeiro Positivo, Falso Negativo, Falso Positivo e Falso Negativo, são utilizados então para formular as métricas. Abaixo segue uma descrição e a fórmula das três métricas utilizadas neste trabalho:

Revocação (R) representa a proporção de *Fake News* que foram corretamente identificadas, sendo definida pela seguinte equação:

$$R = \frac{VP}{VP + FN} \tag{1}$$

Precisão (P) representa a proporção das notícias que foram classificadas como falsas que realmente eram falsas, sendo definida pela seguinte equação:

$$P = \frac{VP}{VP + FP} \tag{2}$$

Medida-F1 é a média harmônica entre Revocação e Precisão, sendo definida pela seguinte equação:

$$F1 = 2 * \frac{P * R}{P + R} \tag{3}$$

5. Resultados e discussão

Esta seção apresenta os diferentes experimentos que foram realizados com o objetivo de avaliar a eficácia das técnicas de preprocessamento, parâmetros e dos modelos de classificação para identificar *Fake News* e discute seus resultados. A Subseção 5.1 apresenta os experimentos relacionados ao pré-processamento de dados, A Subseção 5.2 contém os experimentos relacionados aos parâmetros. Por fim, a Subseção 5.3 descreve os experimentos comparando os modelos de classificação.

5.1. Experimentos de Análise do Pré-processamento

O objetivo desses experimentos foi verificar se a técnica se torna mais eficaz ao aplicar diferentes técnicas de pré-processamento. A metodologia adotada foi testar as diferentes hipóteses, usando como *baseline* o resultado mais eficaz dos experimentos anteriores. Todos os experimentos dessa etapa utilizaram o modelo SVM com a representação BoW utilizando pesos do TF-IDF e Validação Cruzada de 5 camadas.

O **Experimento 1** tem como objetivo verificar a seguinte questão de pesquisa: "`a técnica é mais eficaz removendo a acentuação das palavras?`". Nesse experimento, foi utilizado como *baseline* o texto puro sem modificações, comparando com o texto sem acentuação. Os resultados são apresentados na Figura 3, onde observa-se que não houve diferença uma vez que os dois resultados obtiveram 88% de F1. Dessa forma, remover a acentuação não torna a técnica mais eficaz e nos próximos experimentos o *baseline* continua sendo o texto puro. A partir desse resultado,

pode ser levantado uma hipótese de que erros de digitação e palavras com significados ambíguos não possuem um impacto considerável no contexto de classificar uma notícia falsa.

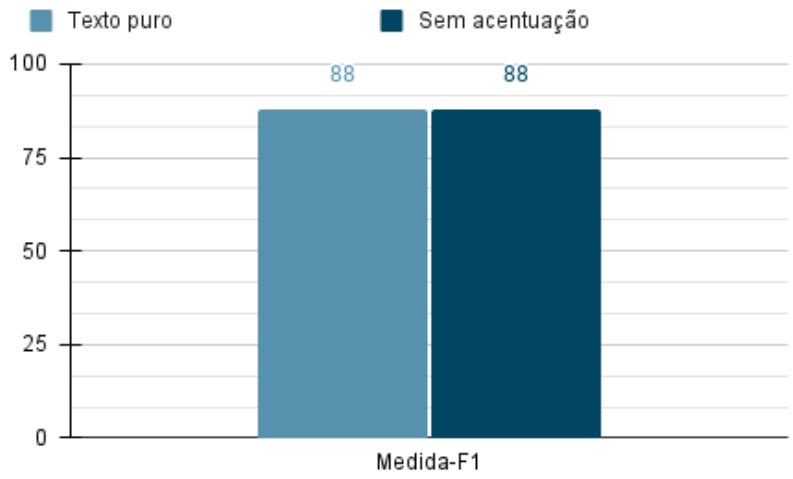

Figura 3. Experimento 1 - Análise da remoção da acentuação

Para o **Experimento 2**, a questão de pesquisa foi a seguinte: "`a técnica é mais eficaz ao remover os números?`". O *baseline* foi o texto puro sem modificações que agora foi comparado com o texto sem os números. Os resultados são apresentados na Figura 4. Novamente, observa-se que não houve uma diferença uma vez que ambos obtiveram 88% de F1. Dessa forma, remover a acentuação não torna a técnica mais eficaz, mantendo o texto puro novamente como *baseline*. A partir do resultado, podem ser inferidas duas diferentes hipóteses. A primeira é de que a remoção de números não possui uma diferença, uma vez que números não afetam nessa classificação. A segunda é de que, enquanto a remoção de certos números, como valores aleatórios, pode melhorar a eficácia. A remoção de números importantes, como partidos políticos, tem o efeito inverso, diminuindo a eficácia.

O **Experimento 3** busca responder a seguinte questão de pesquisa: "`a técnica é mais eficaz ao remover as stop words`". *Stop words* são palavras comuns que podem ter pouco ou até nenhum significado importante. É comum removê-las em diversas tarefas de NLP. Listas de *stop words* geralmente são compostas por artigos, preposições, pronomes e conjunções (HVITFELDT; SILGE, 2021). A lista de *stop words* utilizada foi a versão em português presente na biblioteca NLTK (NLTK, 2022). O *baseline* foi comparado com o texto sem as *stop word*s. Os resultados são apresentados na Figura 5. É possível notar um decréscimo na Medida F1 ao remover as *stop words*, na qual o valor diminui para 87%. Portanto é possível notar que a remoção de *stop words* foi prejudicial para a eficácia, então foi mantido o texto puro como *baseline*.

O **Experimento 4** visa responder a seguinte questão de pesquisa: "`a técnica é mais eficaz ao utilizar Stemming`". O *Stemming*, ou em português, Stemização, é a redução de uma palavra a sua raiz gramatical. Sua utilidade se encontra para relacionar palavras similares que usam da mesma raiz (HVITFELDT; SILGE, 2021). Um exemplo seria "Casinha", "Casa"e "Casebre"sendo reduzidos a "Cas". O *baseline* foi

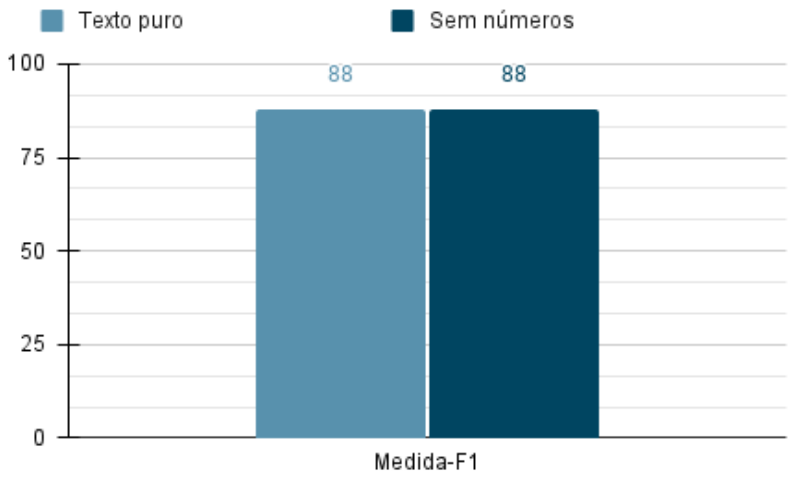

Figura 4. Experimento 2 - Análise da remoção dos números

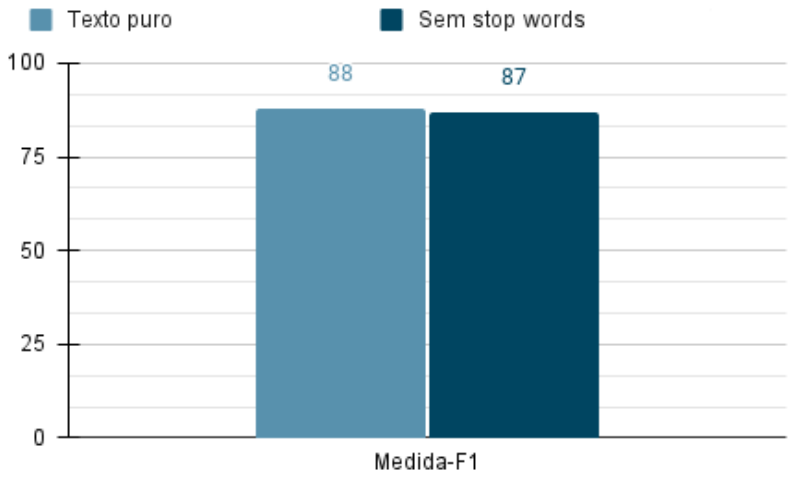

Figura 5. Experimento 3 - Análise da remoção das *stop words*

comparado com o texto aplicando *Stemming*. Os resultados são apresentados na Figura 6. Ambos os resultados foram de 88%, não havendo diferença. Logo, a utilização de *Stemming* não torna a técnica mais eficaz e nos próximos experimentos o *baseline* continua sendo o texto puro.

5.2. Análise dos parâmetros

Os experimentos nessa etapa buscam testar quais parâmetros do TF-IDF são mais eficazes para o problema. Em questão foram testados três parâmetros: quantidade de n-gramas, quantidade de *features* e a desativação do IDF. Primeiramente ocorreram testes utilizando diferentes variações de n-gramas. N-gramas representam uma sequência de n itens dado um texto ou fala. Os itens podem ser fonemas, sílabas, letras ou palavras, dependendo da aplicação, geralmente sendo palavras. N-gramas servem para capturar ordem de palavras que do contrário seria perdido (HVITFELDT; SILGE, 2021). Alguns exemplos de n-gramas podem ser:

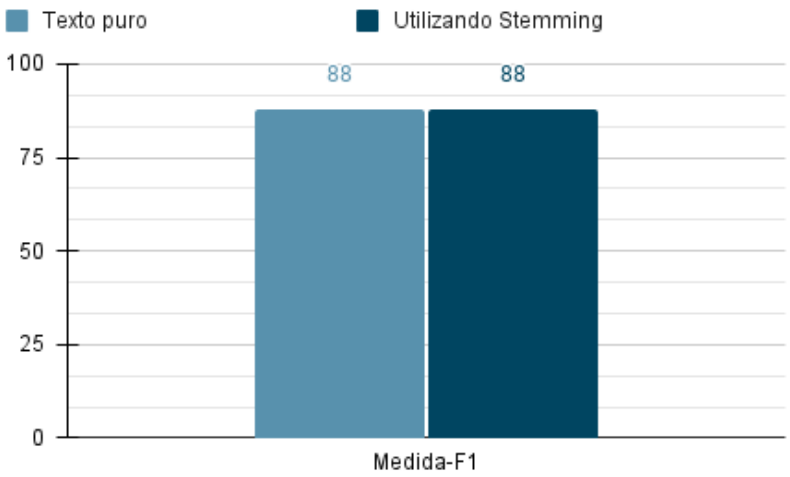

Figura 6. Experimento 4 - Análise do uso de *Stemming*

- **Unigrama**, apenas uma palavra, por exemplo: “Olá", “dia", “meu".
- **Bigrama**, duas palavras, por exemplo: “Ar fresco", “Robin Hood", “Árvore alta".
- **Trigrama**, três palavras, por exemplo: “Eu e você", “A pequena sereia", “era uma vez".

Para os experimentos, foram utilizadas as seguintes siglas:

- (1,1): Representa apenas unigramas;
- (1,2): Representa apenas unigramas e bigramas;
- (1,3): Representa unigramas, bigramas e trigramas;
- (2,2): Representa apenas bigramas;
- (2,3): Representa apenas bigramas e trigramas;
- (3,3): Representa apenas trigramas.

Ocorreram então testes utilizando todas as opções de n-gramas listadas acima. Após, o melhor foi escolhido para ser utilizado testando diferentes valores de *features*, valores incluindo de 100 até 150.000. Por fim, foi testado desativar o IDF. Todos esses experimentos utilizaram BoW com TF-IDF e SVM com Validação Cruzada de 5 camadas.

O **Experimento 5** busca responder a seguinte questão de pesquisa: “`qual é a quantidade de n-gramas mais eficaz`". Os resultados são apresentados na Figura 7. Para isso foram comparadas todas as opções de n-gramas listadas acima. Todas as opções foram testadas com 50 mil *features*. É possível notar que (1,2) e (1,3) apresentaram os melhores resultados, alcançando uma Medida-F1 de 90%. Enquanto (3,3) apresentaram os piores resultados, alcançando apenas 83% na Medida-F1. A diferença entre as quantidades de n-gramas ocorre na precisão, possuindo revocações similares em torno de 89% a 90%. Uma vez que os dois melhores possuem os mesmos resultados, foram necessários mais experimentos para verificar qual dos dois se sobressai.

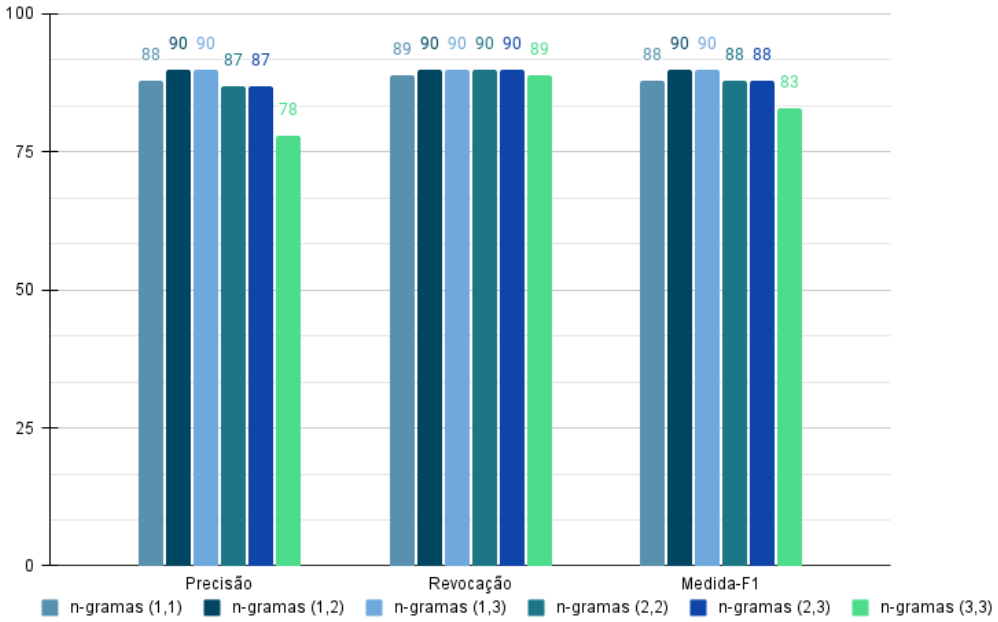

Figura 7. Experimento 5 - Comparação de diferentes quantidades de n-gramas

O **Experimento 6** busca concluir a resposta para a questão de pesquisa anterior. No *Experimento 6*, as duas opções de n-gramas mais eficazes no *Experimento 5* foram comparadas, respectivamente, com 10 mil *features* e 100 mil *features*. A Figura 8 apresenta os resultados. É possível notar que enquanto ambos apresentam os mesmos resultados com apenas 10 mil *features*, com o aumento de *features* para 100 mil, (1,3) passa a atingir 91% de precisão. Vale ressaltar que ambos apresentaram Revocação e Medida-F1 igual. Isso faz com que o uso de unigramas, bigramas e trigramas em conjunto seja a técnica mais eficaz e por isso é usada nos próximos experimentos.

Para o **Experimento 7**, é levantada a seguinte questão de pesquisa: "`qual é a quantidade de features mais eficaz`". O mesmo foi testado com a melhor quantidade de n-gramas encontrada, sendo essa (1,3). Foi testado 100, 1000, 10000, e aumentando em 10 mil, até alcançar 150 mil. A figura 9 apresenta os resultados. Para facilitar a visualização da Figura 9, foram mantidos apenas os valores em que a precisão ou a medida F1 aumenta ou diminuí. É possível observar que com o aumento de *features*, também ocorre o aumento da eficácia da técnica. Porém, esse aumento para de ocorrer após 70 mil *features*, quando atinge o ápice de 91% de precisão. Nota-se que a precisão passa a diminuir voltando para 90% quando atinge o valor de 150 mil *features*.

O **Experimento 8** representa a questão de pesquisa: "`é mais eficaz usar apenas a frequência do termo para ponderar as palavras?`". Para esse experimento foram utilizados os melhores valores até então, 70 mil *features* e (1,3) n-gramas. A Figura 10 apresenta os resultados. Nota-se que a presença dos pesos do IDF faz com que seja mais eficaz, em vez de usar apenas a frequência do termo para ponderar as palavras, uma vez que sua utilização representa ganho de 4 pontos percentuais de F1. Ressalta-se que a perda ao não utilizar os pesos do IDF foi a mesma na precisão e revocação, representando 2 pontos percentuais.

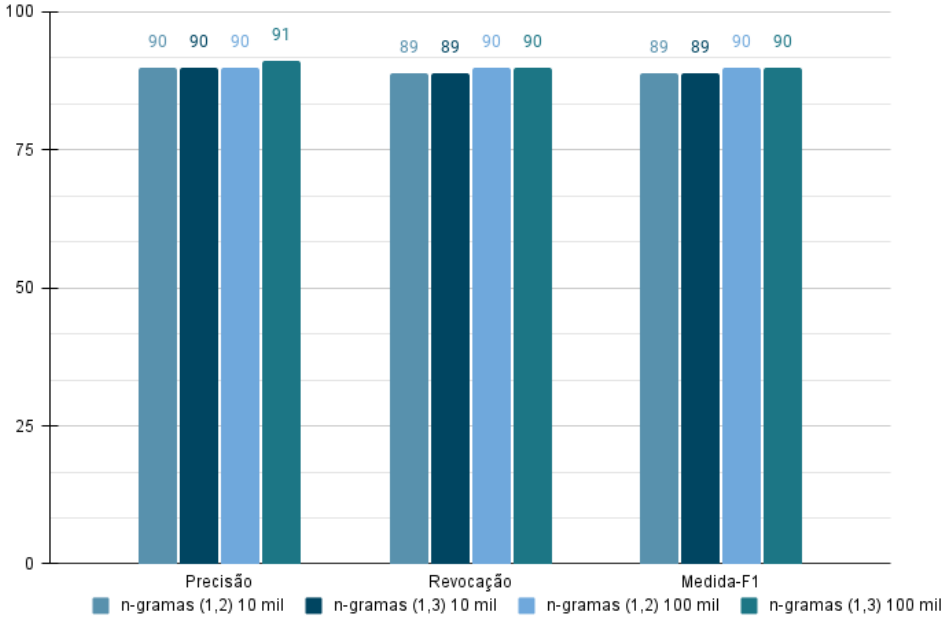

Figura 8. Experimento 6 - Comparando (1,2) e (1,3) com 10 mil e 100 mil *features*

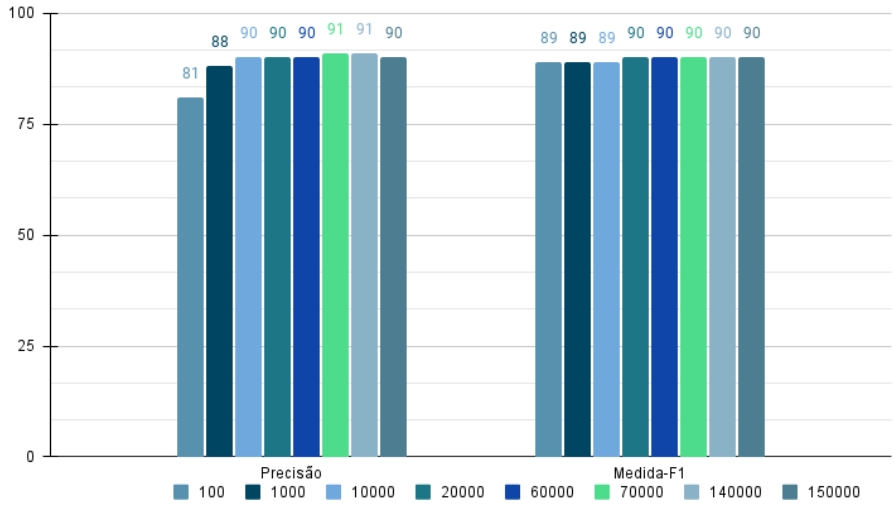

Figura 9. Experimento 7 - Comparando as diferentes quantidades de features e sua eficácia

5.3. Análise dos modelos de classificação

Os experimentos dessa etapa buscam comparar e avaliar a eficácia dos diferentes modelos de classificação. A Subsubseção 5.3.1 contém a comparação entre SVM e Regressão Logística. A Subsubseção 5.3.2 contém a comparação entre entre SVM e BERT.

5.3.1. SVM x Regressão Logística

Nessa subseção está o **Experimento 9** que busca responder a seguinte questão de pesquisa: "`qual o modelo de classificação que possui`

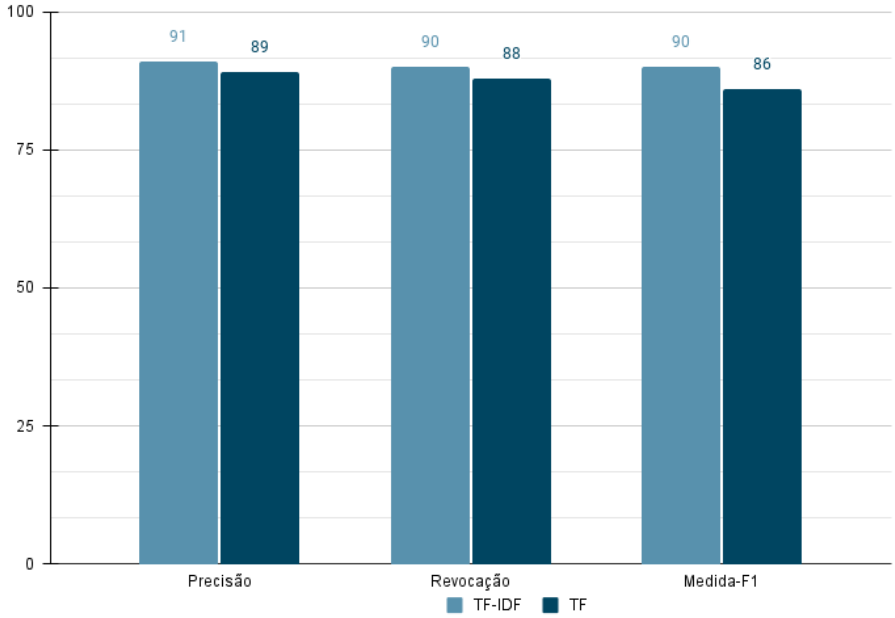

Figura 10. Experimento 8 - Desativando os pesos do IDF

`a melhor eficácia entre SVM e Regressão Logística?`". Para esse teste, foi utilizado BoW com TF-IDF, utilizando 70 mil *features* e (1,3) n-gramas. Para a Regressão Logística, foram usados os parâmetros padrão. Para a avaliação foi utilizado Validação Cruzada de 5 camadas. A Figura 11 apresenta os resultados, onde nota-se que SVM teve uma melhor eficácia, atingindo 90% na Medida-F1 em comparação aos 89% da Regressão Logística. Vale ressaltar que a Regressão Logística obteve a mesma revocação que SVM (90%). Porém, sua precisão foi de 87% onde foi 4 pontos percentuais menores que a de SVM.

5.3.2. BERTimbau x SVM

Nessa subsubseção está contido o **Experimento 10**, esse que visa responder a seguinte questão de pesquisa: "`qual o modelo de classificação que possui a melhor eficácia entre BERTimbau e SVM?`".

O melhor algoritmo até então, SVM, seguiu da mesma forma conforme apresentado no Experimento 9. Para o modelo BERT foi utilizado o modelo pré-treinado em português BERTimbau. O conjunto de dados foi separado em três partes: 80% dos dados foram utilizados para treinar, 10% utilizados para testar e 10% para validar. Foram utilizadas 5 épocas durante o treinamento e 5 épocas durante a validação. Épocas representam uma passagem inteira pelo conjunto de dados estabelecido. O tamanho de lote utilizado foi de 2 para caber na memória do Google Colab. O tamanho de lote representa o número de amostras processadas antes de atualizar o modelo. Para a avaliação foi utilizado Validação Cruzada de 5 camadas. As três métricas do modelo BERTimbau atingiram 99%, conforme mostrado na Figura 12. Ressalta-se que sua Medida F1 acabou sendo 8 pontos percentuais maior que a F1 do SVM.

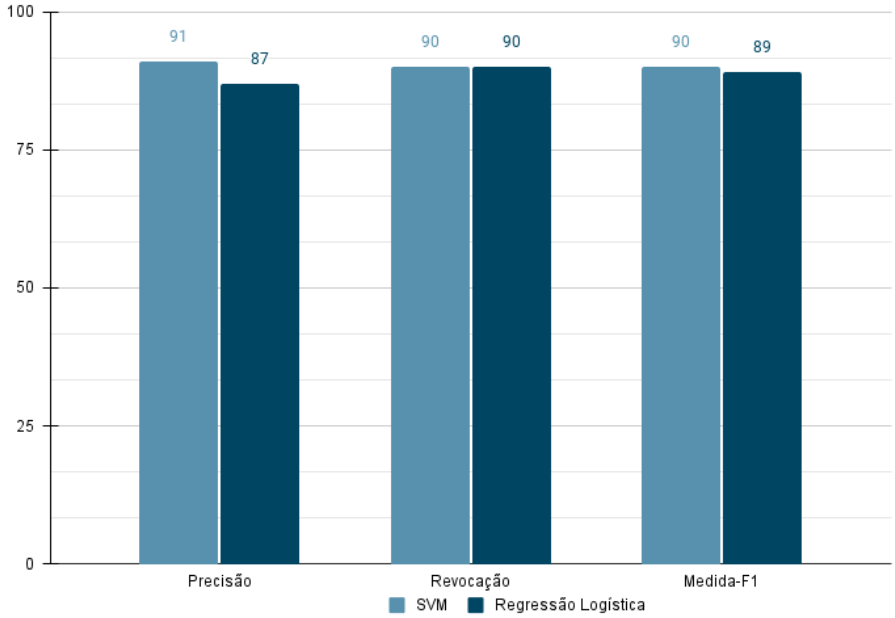

Figura 11. Experimento 9 - Comparando SVM e Regressão Logística

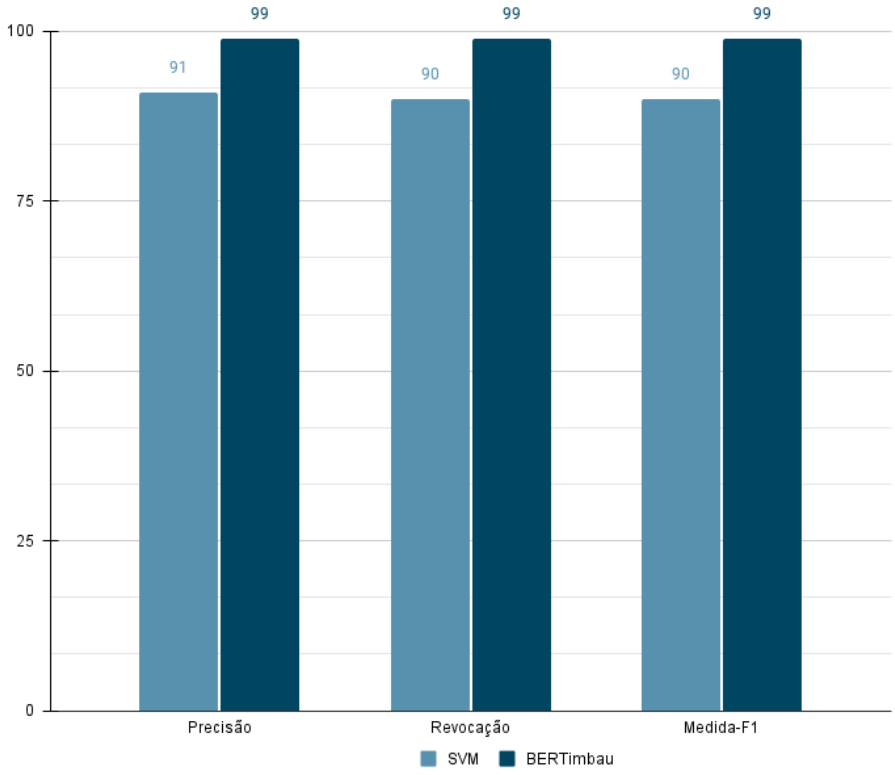

Figura 12. Experimento 10 - Comparando BERTimbau e SVM

6. Conclusão

Neste trabalho foram feitos experimentos para comparar técnicas de pré-processamento, parâmetros e modelos de classificação a fim de identificar *Fake News*

no idioma português. Para isso foram elaboradas 9 questões de pesquisa. Foram feitos experimentos para responder cada questão de pesquisa. Para a metodologia foi utilizado o modelo CRISP-DM. Para a experimentação foi utilizado um *corpus* de *Fake News* em português com 7.200 notícias. Para a avaliação foi escolhido três métricas e validação cruzada de 5 camadas.

A conclusão foi que não é eficaz remover a acentuação, números, *stop words* ou aplicar *Stemming*. O melhor número de parâmetros para o TF-IDF foram 70 mil *features* utilizando unigramas, bigramas e trigramas (1,3). Em relação aos modelos de classificação, o mais eficaz foi de BERTimbau, com uma medida F1 de 99%, representando um ganho de 10% em relação ao SVM e um ganho de 11% em relação a Regressão Logística.

O melhor modelo de classificação para a língua portuguesa neste trabalho foi BERT, assim como (JWA et al., 2019) constata para a língua inglesa. É preciso destacar que o BERT, diferentemente dos outros modelos, ajuda a entender linguagem ambígua, usando os textos de ambas as direções para estabelecer um contexto de uma palavra. Isso permite com que ele retorne diferentes vetores para uma mesma palavra, contanto que o texto em sua volta varie. Esse fator faz com que sua capacidade de analisar linguagem natural seja muito maior. A seguir estão listadas sugestões para trabalhos futuros:

- Experimentar com os parâmetros do SVM, buscando afinar mais o modelo
- Experimentar com os parâmetros da Regressão Logística, buscando afinar mais o modelo
- Experimentar com os parâmetros de BERT, buscando afinar mais o modelo
- Levar em consideração o tempo que leva para processar modelos e parâmetros
- Testar o modelo final em um outro *corpus* de Fake News, a fim de validar mais os resultados
- Elaborar uma aplicação para a utilização do modelo final

Referências

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, v. 31, p. 31–47, 2017. ISSN 08953309.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, v. 5, p. 135–146, 2017.

BOND, C. F.; DEPAULO, B. M. Accuracy of deception judgments. *Personality and Social Psychology Review*, v. 10, p. 214–234, 2006. ISSN 10888683.

BURGES, C. J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, Kluwer Academic Publishers, v. 2, p. 121–167, 1998. ISSN 13845810.

GOLDBERG, Y. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, v. 10, p. 1–309, 2017. ISSN 19474040.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. 1st. ed. [S.l.]: MIT press, 2016.

HAND, D. J. Principles of data mining. In: . [S.l.: s.n.], 2007. v. 30, p. 621–622. ISSN 01145916.

HVITFELDT, E.; SILGE, J. *Supervised Machine Learning for Text Analysis in R*. 1st. ed. [S.l.: s.n.], 2021.

JWA, H. et al. exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert). *Applied Sciences (Switzerland)*, v. 9, p. 4062–4071, 2019. ISSN 20763417.

JúNIOR, J. H. de S. et al. Da desinformação ao caos: uma análise das fake news frente à pandemia do coronavírus (covid-19) no Brasil. *Cadernos de Prospecção*, v. 13, p. 331, 2020. ISSN 2317-0026.

LAROSE, D. T. *Discovering Knowledge in Data: An Introduction to Data Mining*. 1st. ed. [S.l.: s.n.], 2005.

MARBáN, O. et al. Toward data mining engineering: A software engineering approach. *Information Systems*, v. 34, p. 87–107, 2009. ISSN 03064379.

MARIANO, D. C. et al. *Data Mining*. 1ª. ed. Porto Alegre, RS: Grupo A, 2020.

MARTINEZ-PLUMED, F. et al. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, v. 33, p. 3048–3061, 2021. ISSN 15582191.

MEU Dicionário. 2022. Disponível em: <https://www.meudicionario.org>. Último acesso em: 2022-08-08.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: *1st International Conference on Learning Representations, ICLR 2013, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. [S.l.: s.n.], 2013.

MONTEIRO, R. A. et al. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: *Computational Processing of the Portuguese Language*. [S.l.]: Springer International Publishing, 2018. p. 324–334. ISBN 978-3-319-99722-3.

NLTK. 2022. Disponível em: <https://www.nltk.org/>. Último acesso em: 2022-12-12.

PERCY, C. E. et al. *Synchronic corpus linguistics : papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)*. [S.l.]: Rodopi, 1996. 289 p. p. ISBN 9042000198 (bound)9̊042000279 (paper).

POSETTI, J.; MATTHEWS, A. A short guide to the history of 'fake news' and disinformation. *ICFJ (International Center for Journalists)*, v. 1, p. 1–20, 2018. ISSN 0890-0523.

RAMOS, J. Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, v. 242, 2003.

RECUERO, R.; GRUZD, A. Cascatas de fake news políticas: um estudo de caso no twitter. *Galáxia (São Paulo)*, 2019. ISSN 1519-311X.

ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, v. 8, p. 842–866, 2020. ISSN 2307387X.

SHU, K. et al. Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, v. 19, p. 22–36, 2017. ISSN 1931-0145.

SILVA, R. M. et al. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, v. 146, p. 113199, 2020. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417420300257>.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: Pretrained bert models for brazilian portuguese. In: . [S.l.: s.n.], 2020. v. 12319 LNAI, p. 403–417. ISSN 16113349.

TANDOC, E. C.; LIM, Z. W.; LING, R. Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, v. 6, p. 1–153, 2018. ISSN 2167082X.

VASILAKES, J.; ZHOU, S.; ZHANG, R. *Natural language processing*. [S.l.]: Elsevier, 2020. 123-148 p. ISBN 9780128202739.

WANG, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In: . [S.l.: s.n.], 2017. v. 2, p. 422–427.

WIRTH, R. Crisp-dm : Towards a standard process model for data mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, p. 43–56, 2000.

ZHANG, X.; GHORBANI, A. A. An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management*, Elsevier Ltd, v. 57, p. 102025, 3 2020. ISSN 03064573.