Detecção Automática de Desinformação Relacionada à Covid-19 no Brasil

João M. M. Couto[1], Isadora Salles[1], Breno Pimenta[1], Samuel Assis[1], Leandro Araújo[1], Julio C. S. Reis[2], Fabrício Benevenuto[1]

[1]Depto. Ciência da Computação – Universidade Federal de Minas Gerais (UFMG) – Brasil

[2]Departamento de Informática – Universidade Federal de Viçosa (UFV) – Brasil

{joaocouto,isadorasalles,brenopimenta,samuelassis,leandroaraujo}@dcc.ufmg.br

jreis@ufv.br,fabricio@dcc.ufmg.br

Abstract. *The spread of fake news impacts crucial elements of democratic governance. Various efforts attempt to identify fake news instances through information captured after their propagation in the web. We propose a detection methodology at early stages of propagation by offering an analysis that comprehends the training of thousands of prediction models utilizing a wide range of hyperparameters and textual features extracted from suspicious news instances. In this process, we develop a novel database of Covid-19 fake news instances propagated in Brazil. Our results reveal the most relevant attribute sets and the predictive power of different supervised classifiers for this problem in Brazil.*

Resumo. *A disseminação de notícias falsas tem impacto em diversas áreas cruciais da governança democrática. Muitas abordagens de identificação destas noticiais tomam como base a exploração de informações capturadas depois de sua propagação nas redes. Propomos uma metodologia de detecção em estágio inicial de propagação. Efetuamos uma análise exploratória que compreende o treinamento de milhares de modelos utilizando conjuntos diversos de parâmetros e atributos textuais extraídos de notícias suspeitas. Neste processo, desenvolvemos uma base inédita de notícias falsas propagadas no Brasil relativas à Covid-19. Resultados revelam os conjuntos de atributos mais relevantes e o poder de classificadores supervisionados para este problema no Brasil.*

1. Introdução

Plataformas digitais alteraram fundamentalmente a maneira que notícias são produzidas e disseminadas, trazendo com si uma nova gama de oportunidades e desafios em igual proporção. Nessa conjuntura, uma nova era de desinformação desafia até mesmo as mais fortes democracias do mundo. Vários agentes desse ecossistema, muitas vezes financiados por entidades políticas, utilizam plataformas digitais para veicular grandes campanhas de desinformação. Essas campanhas buscam manipular a opinião pública acerca de tópicos específicos, construindo narrativas e criando polarização na sociedade [Ribeiro et al. 2019]. Com o surgimento da Covid-19, o fenômeno da desinformação ganhou uma nova dimensão [Ferrara 2020]. Ao incluir o questionamento de medidas sanitárias e promover a adoção de medicamentos sem eficácia comprovada, seus efeitos escapam ao campo das ideias, causando perda de vidas em larga escala. De fato, cerca

de dois meses após a divulgação do primeiro caso de Covid-19 no mundo, a Organização Mundial de Saúde (OMS) publicou um relatório declarando que a pandemia foi acompanhada por uma "infodemia" [Massarani et al. 2021].

Dada a dimensão do problema e visando ampliar a limitada escalabilidade dos processos de verificação de fatos por agências jornalísticas reconhecidas, vários esforços têm surgido no sentido de automatizar a identificação de notícias falsas a partir de padrões recorrentes [Reis et al. 2019b, Vargas et al. 2022]. Um importante fator frequentemente negligenciado neste tipo de trabalho é a existência de uma forte correlação entre o desempenho dos classificadores, os conjuntos de atributos utilizados e a natureza da desinformação, sobretudo a sua temática. Por exemplo, diferentes modelos oferecem separação entre notícias verdadeiras e falsas utilizando explicações amplamente diversificadas [Reis et al. 2019a]. Em particular, isso significa que classificar desinformação no contexto da pandemia do Covid-19 requer um estudo dos atributos disponíveis na literatura para averiguar quais combinações destes oferecem explicações mais robustas considerando esta temática.

Neste trabalho, apresentamos uma investigação do poder discriminatório de atributos textuais propostos para identificação de desinformação relacionada à Covid-19 no Brasil. Em outras palavras, analisamos o impacto causado por diferentes conjuntos de atributos utilizados em modelos de detecção de desinformação, na busca por uma modelagem capaz de detectar desinformação no contexto da Covid-19 no Brasil que estabeleça um compromisso entre custo computacional, e.g. para computação dos atributos, e desempenho desses modelos.

Para isso, construímos uma base de dados de notícias brasileiras de baixa credibilidade relacionadas ao contexto da Covid-19, todas extraídas de portais com pelo menos uma notícia falsa desmentida por uma agência de checagem de fatos. Além disso, construímos uma coleção de notícias verdadeiras sobre o mesmo tema, extraídas de portais de notícias de alta credibilidade. No total, nossa base de dados possui 290 notícias de baixa credibilidade e 1181 notícias de alta credibilidade. Nela computamos uma série de atributos extraídos a partir do texto das notícias e investigamos o poder preditivo de diferentes classificadores supervisionados.

Através da abordagem proposta, os melhores classificadores gerados utilizam um número reduzido de atributos em comparação com o total de atributos computados. Simultaneamente, utilizando a abordagem de seleção de atributos e hiperparâmetros proposta, observamos um ganho de desempenho, metrificado em AUC (*Area Under the Curve*), em todos os tipos de classificadores testados. Esses resultados evidenciam o sucesso na redução do custo computacional dos modelos e na identificação de notícias de baixa credibilidade relacionadas à Covid-19 no Brasil. Demonstramos que é possível reduzir o custo computacional gasto no cálculo dos atributos sem perda no desempenho da classificação, viabilizando a aplicação de métodos como esse em cenários reais.

O restante deste artigo está organizado da seguinte forma: a Seção 2 oferece um sumário de trabalhos relacionados. Uma descrição do processo de criação da base de dados utilizada, bem como o processo de extração de atributos e modelagem da detecção automática de desinformação incluindo detalhes da configuração experimental é apresentada, respectivamente, nas Seções 3, 4, 5 e 6. A Seção 7 apresenta uma análise do

desempenho dos modelos criados. Finalmente, a Seção 8 conclui o artigo.

2. Trabalhos Relacionados

Existem vários esforços dedicados a investigar o fenômeno da desinformação [Vosoughi et al. 2018, Lazer et al. 2018, Shu et al. 2017]. De maneira muito simplificada, esses trabalhos podem ser divididos em dois grupos principais. O primeiro *(i)* está focado em prover melhor compreensão do fenômeno. Nesse contexto, o trabalho apresentado em [Vosoughi et al. 2018] mostra evidências de que um conteúdo que possui desinformação (i.e., notícia falsa) tende a se espalhar mais rapidamente em comparação aos demais. Por outro lado, no trabalho de Lazer et al. os autores ressaltam que o entendimento do fenômeno da desinformação é um desafio complexo, e que deve ser abordado de maneira interdisciplinar [Lazer et al. 2018].

Já o segundo *(ii)* grupo de trabalhos existentes compreende aqueles focados em prover uma solução para resolução e/ou mitigação do problema. Particularmente, esses esforços discutem diferentes abordagens para detecção de desinformação [Reis et al. 2019b, Shu et al. 2017, Volkova et al. 2017, Conroy et al. 2015]. Parte destes trabalhos utilizam padrões típicos de um conteúdo com desinformação como atributos para treinamento de um classificador. Por exemplo, no trabalho de Pérez-Rosas et al. foi conduzido um conjunto de experimentos de aprendizado para construir um detector de desinformação usando atributos linguísticos (e.g., atributos de legibilidade, sintaxe e psicolinguísticos) [Pérez-Rosas et al. 2017]. Similarmente, no trabalho apresentado em [Volkova et al. 2017] foram construídos modelos linguísticos para classificar um conteúdo, especificamente notícias, como suspeitas ou verdadeiras. Por outro lado, o estudo [Shu et al. 2017] oferece uma revisão da literatura existente contemplando abordagens para detecção de desinformação, de uma perspectiva de mineração de dados, incluindo técnicas para extração de atributos e modelagem. Finalmente, no trabalho apresentado em [Reis et al. 2019b], foi criada uma abordagem para detecção de desinformação (i.e., notícias falsas) baseada em um modelo híbrido a partir de atributos extraídos do conteúdo, reputação da fonte e ambiente (i.e., redes sociais).

De forma geral, os trabalhos de ambos os grupos (i.e., *(i)* e *(ii)*) necessitam do uso de uma base de dados rotulada. Logo, existem também esforços focados em prover bases de dados públicas contendo fatos rotulados como verdadeiros ou falsos. A maioria das bases de dados públicas é composta por conteúdos escritos em Inglês [Reis et al. 2020]. No contexto do Brasil, o número de bases de dados públicas é bastante limitado. A maioria dos trabalhos disponíveis agregam a criação da base com o uso de técnicas de aprendizado para avaliar o desempenho de uma detecção automática no repositório criado. Um exemplo disso se dá na forma do trabalho de Monteiro et al. [Monteiro et al. 2018]. No trabalho de Charles et al. os autores apresentam um repositório de dados flexível com notícias falsas e notícias confiáveis escritas no idioma Português [Charles et al. 2022]. Essas duas bases de dados são mais gerais, não possuindo foco em algum tema específico. Mais relacionado ao contexto do nosso trabalho, no estudo de Martins et al. foi apresentado o COVID-19.BR, uma base de dados manualmente rotulada contendo mensagens do *WhatsApp* sobre o Covid-19 escritas no idioma Português [Martins et al. 2021]. No entanto, diferentemente dos esforços anteriores, estamos focados em desinformação disseminada por veículos de baixa credibilidade. Tal investigação é importante uma vez que nos propomos a identificar este conteúdo precocemente, a partir de atributos textuais, antes mesmo

que ele tenha sido disseminado, por exemplo, em uma plataforma digital, o que pode viabilizar o seu uso na prática. Além disso, o padrão de um conteúdo publicado por uma mídia (e.g., portal de notícias) possui características particulares que tendem a ser diferentes de uma informação postada em uma plataforma menos formal como *WhatsApp*. Estes aspectos certamente possuem impacto no desempenho de um modelo baseado em atributos textuais, o que ressalta a importância da realização deste estudo.

Nas seções a seguir são apresentados detalhes referentes à metodologia adotada neste trabalho, incluindo uma breve descrição dos processos de criação da base de dados (Seção 3) e extração de atributos (Seção 4), bem como informações pertinentes à proposta de geração dos modelos que incluem uma descrição das abordagens exploradas (Seção 5).

3. Base de Dados

Na exploração do fenômeno da desinformação em diferentes áreas do conhecimento, incluindo Ciência da Computação, é primordial que esforços sejam feitos utilizando um conjunto de dados rotulado por anotadores credenciados, em particular, jornalistas ou agências de verificação de fatos reconhecidas por sua experiência de domínio [Reis et al. 2020].

Neste trabalho, a base de dados explorada foi construída a partir de esforços anteriores [Couto et al. 2022], onde coletamos e agrupamos *websites* (ou domínios) de alta e baixa credibilidade. Neste contexto, é válido ressaltar que um domínio é categorizado como "baixa credibilidade" se ele já tiver publicado e/ou publica notícias que, comprovadamente, gerem desinformação. A comprovação de que uma notícia gera desinformação se dá pela existência de um veredito fornecido por uma agência de checagem de fatos brasileira reconhecida internacionalmente[1]. Em particular, as verificações de fatos utilizadas nesse processo comprobatório englobam todos os vereditos publicados no Brasil entre 2013 e 2021 [Couto et al. 2021]. Por fim, o conteúdo publicado por um domínio de baixa credibilidade é, por extensão, considerado de baixa credibilidade, que no contexto da criação desta base de dados é referenciada como "desinformação".

Assim, implementamos coletores de dados para rastrear as publicações veiculadas por esses domínios de alta e baixa credibilidade, permitindo a construção de um conjunto de dados que inclui o título e conteúdo textual de notícias em ambos os grupos de credibilidade. Por fim, filtramos as publicações obtidas de forma a incluir apenas notícias associadas à Covid-19 e para isso retiramos apenas notícias que não apresentam nenhuma palavra-chave associada com a temática[2]. A base de dados final é composta de 290 instâncias de notícias de baixa credibilidade (i.e., desinformação) e 1181 instâncias de notícias de alta credibilidade publicadas do inicio da pandemia até maio de 2022. Esse repositório foi utilizado como base para o processo de computação dos atributos apresentados a seguir.

4. Extração de Atributos

Os atributos utilizados neste trabalho foram coletados a partir das notícias compiladas em nossa base de dados. Especificamente, sua extração se deu nos títulos das notícias.

[1] `https://www.poynter.org/ifcn/`

[2] O conjunto de palavras-chave utilizadas neste processo foi extraído do "Relatório COVID-19 no WhatsApp", disponível em `http://www.monitor-de-whatsapp.dcc.ufmg.br/reports/pdfs/2020_02_a_2021_06_report_tematico_covid_completo.pdf` bem como no Apêndice A.

Trabalhos anteriores mostraram que o título (ou *headline*) de uma notícia tem como objetivo atrair a atenção do leitor apresentando o conteúdo de forma rápida e resumida, podendo determinar como as pessoas leem as notícias [Reis et al. 2015]. Além disso, neste mesmo trabalho os autores atestam que uma parcela significativa dos leitores compartilham informações apenas a partir da leitura de uma manchete. De forma geral, acreditamos que atributos extraídos a partir de conteúdo textual podem revelar características peculiares da desinformação que podem ser úteis para identificá-la. No total, foram implementados 103 atributos que foram agrupados em seis categorias, que são descritas a seguir. Neste contexto, é válido ressaltar que esses atributos foram implementados com base em esforços anteriores do nosso grupo de pesquisa [Reis et al. 2019b, Reis et al. 2019a].

Nota. Foram realizados experimentos considerando atributos extraídos do título e texto. No entanto, percebeu-se que o volume de atributos aumenta gerando ruídos que prejudicam o desempenho dos modelos. Assim, optamos por explorar somente atributos extraídos da manchete da notícia.

1. *Atributos de sintaxe (SINT)* englobam características em nível de sentença incluindo, por exemplo, métricas de qualidade e da legibilidade de um texto. No total foram implementados 7 atributos desta categoria (e.g., medidas de legibilidade e qualidade de texto). Esses atributos foram amplamente utilizados em esforços anteriores para tarefas de detecção de notícias falsas [Conroy et al. 2015].
2. *Atributos lexicais (LEXI)* consistem em atributos em nível de caracteres de palavras. Neste trabalho foram implementados 16 atributos desta categoria, incluindo aspectos como a contabilização de palavras únicas, pontuação, palavras escritas com letras em maiúsculo, tamanho médio das palavras, presença de *hashtags*, número de palavras na maior e na menor frase.
3. *Atributos gramaticais (GRAM)* capturam informações relativas à regras de uso da língua. Implementamos 10 atributos desta categoria incluindo a frequência e porcentagem de classes morfológicas das palavras, como verbos ser e estar, verbos auxiliares, conjunções, pronomes e preposições.
4. *Atributos semânticos (SEMA)* envolvem aspectos de significado do texto. Foram extraídos 3 atributos semânticas que consistem em indicadores de toxicidade, ameaça e insulto. Para isso foi utilizada a Perspective API[3]. Esta API faz uso de modelos de aprendizado de máquina para medir esses indicadores dado um texto de entrada. Há outros indicadores fornecidos nessa API, no entanto, os três mencionados foram os que apresentaram maior relevância dado a tarefa de distinção de desinformação e o uso para a língua portuguesa. Esse recurso tem sido bastante explorado em trabalhos relacionados.
5. *Atributos de subjetividade (SUBJ)*, em alto nível, mensuram o sentimento associado a um texto. A extração desses atributos foi realizada com o uso da API *Cloud Natural Language*[4]. Essa ferramenta disponibilizada pela Google [White and Rege 2020], aplica processamento de linguagem natural (PLN) fornecendo modelos pré-treinados e que podem ser utilizados como serviço pela plataforma. Esse serviço fornece duas métricas, pontuação e magnitude de sentimentos. A pontuação corresponde ao viés emocional do texto, enquanto a magnitude

[3] `https://www.perspectiveapi.com`
[4] `https://cloud.google.com/natural-language/docs`

indica a força da emoção (positiva ou negativa). Além dos indicadores, foi realizada uma classificação combinando as duas métricas, totalizando 3 atributos de subjetividade. A classificação de sentimentos é definida em quatro classes: *Clearly positive*, onde apresenta alta pontuação e alta magnitude ; *Clearly negative*, associado com baixa pontuação e alta magnitude ; *Neutral*,média pontuação e baixa magnitude ; *Mixed*, indicativa de pontuação média e magnitude alta.

6. *Atributos psicolinguísticos (PSIC)* consistem em atributos em nível de categoria das palavras. Para isso, foi utilizada como base a versão 2007 do *Linguistic Inquiry and Word Count* (LIWC) para extração e análise da distribuição de 64 atributos psicolinguísticos [Tausczik and Pennebaker 2010], que incluem categorias de palavras como saúde, religião, dinheiro, trabalho, tempo, sexualidade, certeza e sentimento. Desde a sua concepção, o LIWC tem sido amplamente utilizado para uma série de tarefas diferentes, incluindo detecção de notícias falsas em plataformas digitais [Reis et al. 2019b].

Na próxima seção apresentamos como os atributos implementados foram explorados para geração dos modelos.

5. Proposta para Geração dos Modelos

Avaliar com exatidão o potencial discriminativo dos atributos apresentados na Seção 4 demandaria a enumeração de todos os possíveis subconjuntos seguido da geração de um modelo treinado utilizando cada um deles. Naturalmente, essa busca exaustiva no espaço de soluções é computacionalmente inviável. Neste contexto, propusemos uma estratégia em que os atributos selecionados para o treinamento de modelos são escolhidos aleatoriamente e aplicados numa gama de diferentes tipos de modelagens e parâmetros, possibilitando uma amostragem não enviesada, em termos de seleção de atributos, do espaço de soluções. Aqui, é importante ressaltar que esforços recentes ratificam o potencial deste tipo de metodologia para permitir a composição de modelagens competitivas que buscam mapear padrões no intuito de diferenciar desinformação (i.e., notícias falsas) das demais [Reis et al. 2019a]. Em particular, essa estratégia pode ser divida em três etapas principais.

1) Combinações de atributos. Aqui, todos os possíveis conjuntos de até X atributos são enumerados, onde X varia progressivamente de 1 a 15, gerando milhares de combinações possíveis. Ou seja, 103 modelos com apenas 1 atributo (uma vez que este é o número total de atributos explorados), 5050 modelos com combinações de 2 atributos, e assim por diante. Ao final desta etapa, extraímos uma amostra de 12.500 conjuntos de atributos (modelos) selecionados para cada tamanho, totalizando 187.500 combinações a serem exploradas nas etapas seguintes. Vale ressaltar que o número reduzido de possibilidades para modelos com menos atributos foi compensado ao amostrarmos a mesma quantidade de combinações para os diferentes tamanhos, independente do total de combinações enumeradas originalmente.

2) Ajuste de hiperparâmetros por modelagem/tamanho. A segunda etapa compreendeu a determinação, para cada dupla de tipo de modelagem e tamanho de modelo (número de atributos presentes), o conjunto de hiperparâmetros que leva ao melhor desempenho. Para tal, uma técnica de otimização de hiperparâmetros foi proposta. Primeiro as abordagens de aprendizado de máquina a serem testadas foram elencadas, são elas: Florestas

de Decisão Aleatória (RF, do termo em inglês *Random Forest*, XGBoost (XGB, do termo em inglês *eXtreme Gradient Boosting*), Árvores de Decisão (DT, do termo em inglês *Decision Tree* e Máquina de Vetores de Suporte (SVM, do termo em inglês *Support Vector Machines*. Essas abordagens serão detalhadas nas seções a seguir. Para cada uma delas, identificamos o conjunto dos principais hiperparâmetros e uma faixa de valores a ser explorada no entorno do valor padrão de cada um deles. Depois, amostramos aleatoriamente 50 conjuntos de atributos de cada tamanho, ou seja $1, 2, ..., 15$, totalizando 750 combinações. Por fim, enumeramos todas as possíveis combinações de hiperparâmetros dentro das faixas de valores de cada modelagem previamente definidos e, utilizando cada uma delas, treinamos 750 modelos correspondentes às 50 combinações de atributos dos 15 tamanhos. Assim, se uma modelagem tem 3 hiperparâmetros principais e cada um deles 5 valores em sua faixa de valores, foram treinados $5^3 \times 50$ modelos por tamanho, para um total de 93.750 modelos por tipo de modelagem. Ao final desse processo, para cada grupo de modelos com o mesmo número de atributos, foi calculada a média de desempenho obtida nos grupos em termos de AUC (*Area under the ROC Curve*) considerando-se cada uma das combinações de hiperparâmetros. Então, foi escolhido, para cada tamanho de modelo e tipo de modelagem, o conjunto de hiperparâmetros com o melhor desempenho médio nos 50 modelos treinados.

3) Treinamento e avaliação de modelos. A última etapa consistiu na realização do treinamento e avaliação dos modelos compostos pelas 187.500 combinações de atributos da primeira etapa a menos das 750 utilizadas para ajuste de hiperparâmetros na segunda. Para tal, um modelo de cada tipo (RF, XGB, DT, SVM) foi instanciado para cada uma dessas combinações, utilizando o conjunto de hiperparâmetros correspondente ao tipo de modelagem e tamanho de cada combinação de atributos encontrada na etapa anterior. Ao final desse processo, foi efetuada uma ordenação dos modelos de cada tipo a partir da AUC obtida em cada um deles.

Nota. É válido ressaltar que o número de combinações de atributos (i.e., 15), as abordagens e métricas exploradas foram selecionadas com base em esforços anteriores [Reis et al. 2019a]. Em particular o número máximo de atributos explorados nos modelos foi definido com base em experimentos preliminares. Conforme discutido nas seções a seguir, os resultados apresentados reforçam que este limiar é suficiente para gerar modelos com desempenho satisfatório. Para modelos com mais de 15 atributos não observamos melhora significativa em termos de desempenho, por outro lado, computar atributos é um processo que em alguns casos, pode ser custoso.

6. Configuração Experimental das Abordagens Exploradas

Nesta seção apresentamos uma breve descrição da configuração experimental das abordagens exploradas que são importantes para fins de reprodutibilidade do trabalho. Relacionamos ainda os melhores hiperparâmetros (coluna "Melhor") obtidos em cada uma das abordagens exploradas, que são base para os resultados apresentados na Seção 7.

Florestas de Decisão Aleatória (RF). As Florestas de Decisão Aleatória (ou *Random Forest*) [Breiman 2001] são uma modelagem baseada na combinação de modelos mais fracos para o aprendizado de tarefas de classificação. Nela, um conjunto de árvores de decisão são treinadas e uma instância é classificada de acordo com a classe atribuída pela maioria delas. Para avaliar o potencial de modelos RF na identificação de desinformação

relacionada à Covid-19, identificamos os seguintes hiperparâmetros como os mais fortemente correlacionados com o desempenho deste tipo de modelagem: "`criterion`", "`n_estimators`", "`min_smaples_split`". A Tabela 1 apresenta os valores padrão[5] e explorados para cada um dos hiperparâmetros mencionados bem como o utilizado no modelo de melhor desempenho cujo resultado será discutido na seção a seguir.

	Hiperparâmetros RF		
	Padrão	**Explorados**	**Melhor**
criterion	gini	gini, entropy, log_loss	gini
n_estimators	100	25 / 100 / 150	150
min_samples_split	2	2 / 4 / 6	4

Tabela 1. Hiperparâmetros padrão, explorados e melhor configuração obtida usando RF.

XGBoost (XGB). O *eXtreme Gradient Boosting* ou XGBoost [Chen and Guestrin 2016], é um tipo de modelagem de combinação de árvores de decisão que agrupa a classificação de vários modelos simples para gerar uma classificação unificada no qual novos modelos são introduzidos e treinados especificamente para acertar a classificação das instâncias do dado erroneamente classificado pelas árvores anteriores. Neste tipo de modelagem, exploramos diferentes valores para os seguintes hiperparâmetros: "`max_depth`", "`learning_rate`" e "`eval_metric`". A Tabela 2 apresenta uma visão geral dos hiperparâmetros explorados.

	Hiperparâmetros XGB		
	Padrão	**Explorados**	**Melhor**
max_depth	6	3 / 6 / 9	3
learning_rate (eta)	0,3	0,25 / 0,3 / 0,35	0,35
eval_metric	logloss	error, logloss, auc	logloss

Tabela 2. Hiperparâmetros padrão, explorados e melhor configuração obtida usando XGB.

Árvores de Decisão (DT). Árvores de decisões (ou *Decision Tree*) [Breiman et al. 2017] foram utilizadas devido ao seu rápido tempo de treinamento. Isso possibilitou uma alta frequência de testes dos métodos implementados antes que fossem aplicados a modelagens de treinamento mais complexo e demorado. Avaliamos o potencial das árvores de decisão na tarefa de identificar desinformação relacionada à Covid-19 explorando combinações de valores dos seguintes hiperparâmetros: "`criterion`", "`min_samples_leaf`", "`min_samples_split`", "`max_features`". A Tabela 3 sumariza os esforços desta investigação.

Máquina de Vetores de Suporte (SVM). Máquina de Vetores de Suporte ou *Support Vector Machines* [Joachims 1998], é um tipo de modelagem supervisionada que busca identificar o hiperplano em um espaço N-dimensional (onde N é o número de atributos utilizados no modelo) que maximize a distância até instâncias das duas classes que buscamos classificar. Neste esforço, avaliamos o potencial deste tipo de modelagem através da

[5]Os valores padrão apresentados nas Tabelas da Seção 6 foram extraídos do pacote scikit-learn disponível para a linguagem Python.

	Hiperparâmetros DT		
	Padrão	**Explorados**	**Melhor**
criterion	gini	gini, entropy, log_loss	gini
min_samples_leaf	1	1 / 2 / 3	2
min_samples_split	2	2 / 4 / 6	4
max_features	None	sqrt, log2, None	None

Tabela 3. Hiperparâmetros padrão, explorados e melhor configuração obtida usando DT.

metodologia de exploração do espaço de solução proposta. Aqui, identificamos os hiperparâmetros "`kernel`", "`degree`", "`gamma`"e "`shrinking`"como os mais fortemente correlacionados com o desempenho do modelo nesta tarefa. Na Tabela 4 observamos os valores padrão, explorados e escolhidos (para o melhor modelo) de cada um desses parâmetros.

	Hiperparâmetros SVM		
	Padrão	**Explorados**	**Melhor**
kernel	rbf	linear, poly, rbf	poly
degree	3	2-5	4
gamma	scale	auto, scale	scale
shrinking	True	True, False	True

Tabela 4. Hiperparâmetros padrão, explorados e melhor configuração obtida usando SVM.

Na próxima seção são discutidos os resultados obtidos a partir das abordagens exploradas incluindo uma breve discussão das melhores configurações de hiperparâmetros obtidas para cada uma delas, que foram já relacionadas nesta seção por motivos de limitação de espaço.

7. Resultados

Nesta seção apresentamos os resultados experimentais do nosso estudo. Para cada tipo de modelagem explorada, comparamos o melhor classificador obtido com um modelo equivalente treinado com todos os atributos calculados. A Tabela 5 apresenta o desempenho de classificação, metrificado em AUC, obtido a partir de cada uma das abordagens descritas anteriormente considerando dois contextos diferentes: (*i*) utilizando o subconjunto de atributos com melhor desempenho na classificação (primeira coluna) e (*ii*) utilizando todos os 103 atributos calculados (segunda coluna). Para este tipo de modelagem, o melhor modelo encontrado (i.e., RF) utiliza apenas 12 dos atributos explorados neste trabalho, possibilitando a classificação de nossa base de dados com 0,790 de AUC. Este desempenho é significativamente melhor do que o obtido utilizando todos os atributos calculados: 0,753 de AUC.

De forma geral, observamos que apesar dos modelos encontrados pela metodologia proposta utilizarem um número muito menor de atributos, eles são igualmente capazes de classificar instâncias de desinformação de Coronavírus quando comparados com modelos treinados utilizando todos os atributos disponíveis. Este resultado, observado em

	Melhor (AUC)	Todos (AUC)	Atributos melhor modelo	#Atributos
RF	**0,790**	0,753	ameaça (SEMA), insulto (SEMA), #palavras letra maiúscula (LEXI), categ. lar (PSIC), índice Coleman-Liau (Sintaxe), #letras (LEXI), categ. funcional (PSIC), categ. família (PSIC), categ. afeto (PSIC), categ. visão (PSIC), categ. relatividade (PSIC), sentimentos confusos (SUBJ)	12
XGB	0,762	0,759	insulto (SEMA), #palavras letra maiúscula (LEXI), tamanho médio das frases (Lexicical), %preposições (GRAM), categ. pronomes impessoais (PSIC), categ. conquista (PSIC), sentimentos confusos (SUBJ)	7
DT	0,714	0,637	toxicidade (SEMA), ameaça (SEMA), #palavras letras maiúscula (LEXI), % normalizações linguísticas (GRAM), categ. funcional (PSIC), categ. primeira pessoa (PSIC), categ. conjunções (PSIC), categ. causa (PSIC)	8
SVM	0,726	0,613	#palavras letras maiúsculo (LEXI), #letras (LEXI), #conjunções (GRAM), #preposições (GRAM), categ. afeto (PSIC), categ. positividade (PSIC), categ. escuta (PSIC)	7

Tabela 5. Resultados experimentais.

todos os tipos de modelagens, oferece evidência do poder da metodologia proposta. Aqui é importante relembrar que o melhor classificador é definido como a combinação do subconjunto de atributos e valores de hiperparâmetros que resultou na maior AUC para cada tipo de modelo.

Na exploração das combinações de hiperparâmetros, através das Tabelas 1 e 2 observamos que no caso dos melhores modelos encontrados de tipo RF e XGB, apenas um dos três hiperparâmetros explorados manteve seu valor padrão ("`criterion`" no caso de RF e "`eval_metric`" no caso de XGB). Já no caso dos melhores modelos de tipo DT e SVM, isso foi o caso para apenas dois dos quatro hiperparâmetros explorados ("`gini`" e "`max_features`" em modelos DT, "`gamma`" e "`shrinking`" em modelos SVM). Estes resultados indicam um ganho de desempenho possibilitado pela aplicação da metodologia proposta. Além disso, por meio da Tabela 5 podemos notar que todas as modelagens obtiveram um melhor desempenho (em termos de AUC) quando treinadas utilizando o melhor subconjunto de atributos em comparação com os modelos treinados com todos os atributos disponíveis (i.e., 103). O ganho observado foi maior nos modelos de tipo SVM e DT, todavia um desempenho de classificação parecido nos dois contextos é suficiente evidência para a capacidade da metodologia proposta.

Nossos resultados reforçam que através da amostragem do espaço de atributos proposta obtemos modelos com igual ou melhor capacidade de classificação utilizando um número muito menor de atributos, consequentemente reduzindo o custo computacional necessário para a classificação de uma nova instância. Eles demonstram ainda que estes modelos identificam padrões capazes de mapear com sucesso a natureza do conteúdo no ecossistema de desinformação associado com a Covid-19 no Brasil.

8. Conclusão

Neste trabalho nós apresentamos diversas contribuições relevantes para o nosso cenário de estudo. Primeiramente, nós apresentamos uma base de dados rotulada para notícias falsas e notícias verdadeiras acerca do COVID-19 no Brasil. Em seguida, pesquisamos um grande número de trabalhos relacionados e implementamos variados tipos de atributos extraídos de conteúdo presentes nesses estudos, com o intuito de detectar notícias falsas.

Por fim, foi proposta uma modelagem para avaliar o potencial discriminativo dos atributos. Foi analisado o desempenho em termos de AUC de cada um dos tipos de modelagem implementados em dois contextos: (i) utilizando todos os atributos; (ii) utilizando apenas o melhor subconjunto de atributos selecionado com a nossa abordagem. Observou-se um descolamento entre o desempenho obtido nos dois contextos, para todos os tipos de modelagem. Essa observação fortemente suporta a motivação da metodologia: reduzir o custo computacional despendido para o cálculo de atributos sem perda de desempenho classificatório. Utilizando apenas atributos disponíveis tão cedo quanto a publicação de uma nova notícia e reduzindo significantemente o custo computacional do cálculo de atributos ao utilizar uma quantidade significantemente menor deles, obtivemos um desempenho em classificação superior àquela obtida utilizando todos os atributos disponíveis. Como contribuição final, pretendemos compartilhar a base de dados construída neste trabalho com a comunidade científica, abrindo a possibilidade de melhorar os resultados aqui alcançados. Como trabalhos futuros, pretendemos investigar a abordagem proposta em novos conjuntos de dados, de diferentes contextos, bem como avaliar aspectos de variabilidade e erros dos modelos propostos e realizar uma análise individual de informatividade dos atributos utilizados.

Agradecimentos. Este trabalho foi parcialmente financiado pelo MPMG, projeto Capacidades Analíticas, CNPQ, FAPEMIG e FAPESP.

Referências

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Charles, A. C., Ruback, L., and Oliveira, J. (2022). Fakepedia corpus: A flexible fake news corpus in portuguese. In *Proc. of the Int'l Conference on Computational Processing of the Portuguese Language*, pages 37–45.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Conroy, N. K., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proc. of the Association for Information Science and Technology*, pages 1–4.

Couto, J. M. M., Pimenta, B., de Araújo, I. M., Assis, S., Reis, J. C., da Silva, A. P. C., Almeida, J. M., and Benevenuto, F. (2021). Central de fatos: Um repositório de checagens de fatos. In *Proc. of the Dataset Showcase Workshop (DSW)*, pages 128–137.

Couto, J. M. M., Reis, J. C., Cunha, Í., Araújo, L., and Benevenuto, F. (2022). Caracterizando websites de baixa credibilidade no brasil. In *Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 503–516. SBC.

Ferrara, E. (2020). What types of covid-19 conspiracies are populated by twitter bots? *arXiv preprint arXiv:2004.09531*.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Martins, A. D. F., Cabral, L., Mourao, P. J. C., Monteiro, J. M., and Machado, J. (2021). Detection of misinformation about covid-19 in brazilian portuguese whatsapp messages using deep learning. In *Proc. of the Brazilian Symposium on Databases (SBBD)*, pages 85–96.

Massarani, L. M., Leal, T., Waltz, I., and Medeiros, A. (2021). Infodemia, desinformação e vacinas: a circulação de conteúdos em redes sociais antes e depois da covid-19. *Liinc em Revista*, 17(1):e5689.

Monteiro, R. A., Santos, R. L., Pardo, T. A., Almeida, T. A. d., Ruiz, E. E., and Vale, O. A. (2018). Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Proc. of the Int'l Conference on Computational Processing of the Portuguese Language*, pages 324–334.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019a). Explainable machine learning for fake news detection. In *Proc. of the ACM Conference on Web Science*, pages 17–26.

Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019b). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

Reis, J. C., Melo, P., Garimella, K., Almeida, J. M., Eckles, D., and Benevenuto, F. (2020). A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proc. of the Int'l AAAI Conference on Weblogs and Social Media*, pages 903–908.

Reis, J. C. S., de Souza, F., Vaz de Melo, P., Prates, R., Kwak, H., and An, J. (2015). Breaking the news: First impressions matter on online news. In *Proc. of the Int'l AAAI Conference on Web and Social Media*, pages 357–366.

Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., Oana Goga, K. P. G., and Redmiles, E. M. (2019). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proc. of the ACM Conference on Fairness, Accountability, and Transparency*.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Vargas, F., DÁlessandro, J., Rabinovich, Z., Benevenuto, F., and Pardo, T. A. (2022). Rhetorical structure approach for online deception detection: A survey. In *Proc. of the Int'l Conference on Language Resources and Evaluation*, pages 357–366.

Volkova, S., Shaffer, K., Jang, J. Y., and Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 647–653.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

White, T. E. and Rege, M. (2020). Sentiment analysis on google cloud platform. *Issues in Information Systems*, 21(2):221–228.

Apêndice A - Palavras-Chave Covid-19

"covid", "COVID-19", "Corona", "coronavirus", "virus chinês", "pandemia", "contágio", "FiqueEmCasa", "NãoFiqueEmCasa", "Fique em casa", "lavar as mãos", "uso de máscaras", "usar máscaras", "isolamento social", "alcool em gel", "CPI da COVID", "CPI da pandemia", "cloroquina", "hidroxicloroquina", "ivermectina", "tratamento precoce", "kit covid", "vacina" "vacinação", "coronavac", "astrazeneca", "butantan", "imunidade de rebanho", "Anvisa", "Agência Nacional de Vigilância Sanitária", "pfizer", "Ministro da Saúde", "Mandetta" "Pazuello", "arma biológica", "Moderna", "BionTech", "Fiocruz", "Janssen", "lockdown", "quarentena", "variante", "Covaxin"