Detecção Automática de Desinformação em Diferentes Cenários: Eleições nos Estados Unidos e no Brasil

Julio C. S. Reis[1], **Fabrício Benevenuto**[2]

[1]Universidade Federal de Viçosa (UFV) – Brasil
[2]Universidade Federal de Minas Gerais (UFMG) – Brasil

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

Abstract. *In this work, we present an investigation of the potential of features for misinformation detection considering different scenarios (i.e., presidential elections in the United States and Brazil). In order to do this, we collect data from these two events and compute features explored in the previous efforts in both datasets. We then propose a methodology for unbiased model generation using the XGB classifier, whose performance of built models was measured in terms of AUC. Last, we conduct an experiment based on Pareto-Efficiency that allowed us to identify features that can be useful for generating models with high performance to identify misinformation disseminated in different scenarios.*

Resumo. *Neste trabalho apresentamos uma investigação do potencial de atributos para detecção de desinformação considerando diferentes cenários (i.e., eleições presidenciais nos Estados Unidos e no Brasil). Para isso, reunimos dados destes dois eventos e computamos atributos explorados em trabalhos anteriores em ambos os repositórios. Depois, propomos uma metodologia para geração imparcial de modelos usando o classificador XGB, cujo desempenho dos modelos gerados foi mensurado em termos de AUC. Por fim, conduzimos um experimento baseado na Fronteira de Pareto que nos permitiu identificar atributos que podem ser úteis para a geração de modelos com alto desempenho para identificação de desinformação disseminada em diferentes cenários.*

1. Introdução

Redes sociais online e aplicativos de mensagem instantânea, impactaram consideravelmente a forma como os usuários interagem, se conectam, e se comunicam no ambiente online, remodelando vários dos ecossistemas de informação existentes. Em especial, essas plataformas digitais mudaram drasticamente a maneira como notícias são produzidas, disseminadas e consumidas em nossa sociedade. Estudos revelam que 62% dos adultos americanos usam essas plataformas como fonte primária para consumo deste tipo de conteúdo (i.e., notícias) [Mitchell 2016]. No Brasil, segundo um levantamento realizado pelo Instituto *Reuters*, esse percentual chega à 66% [Report 2018].

Essas mudanças, no entanto, contribuíram para que essas plataformas se tornassem espaços bastante vulneráveis, favorecendo a disseminação de campanhas de desinformação, que além de enganarem e/ou confundirem as pessoas em contextos sensíveis como saúde e política, reduzem a credibilidade dos meios de comunicação nesses ambientes [Li et al. 2016]. Embora a disseminação da desinformação (e/ou notícias

falsas) não seja um problema novo[1] e existam vários esforços com objetivo de entender esse fenômeno [Lazer et al. 2018, Vosoughi et al. 2018], não é surpreendente o fato de que a maioria destes estudos estejam dedicados à proposição de abordagens automatizadas para detecção de desinformação [Conroy et al. 2015, Ruchansky et al. 2017, Reis et al. 2019b]. Apesar da inegável importância dos esforços existentes nesta direção, eles são, em sua maioria, trabalhos simultâneos que propõem atributos e/ou soluções complementares. De maneira resumida, o problema é reduzido à uma tarefa de classificação binária, na qual o conteúdo disseminado é rotulado como "Desinformação" (ou notícia falsa) e técnicas baseadas em aprendizado de máquina supervisionado são então aplicadas com objetivo de distingui-lo dos demais baseadas em modelos treinados em repositórios de dados oriundos de cenários específicos. Logo, pouco se sabe sobre o potencial desses atributos propostos na literatura para identificação de desinformação em cenários distintos. Diante deste contexto, o objetivo deste trabalho é investigar se existe um conjunto de atributos que pode ser útil para a geração de modelos com alto desempenho que sejam capazes detectar desinformação (ou notícia falsa) disseminada em plataformas digitais (e.g., Facebook e WhatsApp) considerando diferentes cenários (i.e., eleições presidenciais em diferentes países: Estados Unidos e Brasil).

Para isso, primeiramente, selecionamos repositórios de dados dos diferentes cenários de interesse (i.e., eleições nos Estados Unidos e no Brasil). Depois, a partir de nossos esforços anteriores [Reis et al. 2019b, Reis and Benevenuto 2021], onde conduzimos uma breve revisão sistemática da literatura com o objetivo de identificar os principais atributos que foram propostos para detecção de desinformação em plataformas digitais, extraímos, de ambos os repositórios, 163 atributos (em comum) com potencial para identificação de um conteúdo contendo desinformação. Considerando que esses atributos podem ter uma variedade de interações não lineares complexas, utilizamos um algoritmo de aprendizado de máquina (i.e., XGB [Chen and Guestrin 2016]) com flexibilidade significativa, rápido e eficaz para diversas tarefas, incluindo classificação. Em seguida, propomos uma metodologia para geração imparcial de modelos em termos de seleção de atributos que os compõem. Em outras palavras, a abordagem proposta contempla a seleção aleatória de atributos para a composição de cada um dos modelos, o que viabiliza uma investigação do potencial de cada um desses atributos na tarefa de distinguir um conteúdo contendo desinformação dos demais. Depois disso, para cada conjunto de atributos gerado a partir da metodologia proposta, construímos e avaliamos o desempenho dos modelos nos diferentes cenários de interesse. No total, foram gerados 247.941 modelos. Por fim, exploramos um conceito oriundo da Economia (i.e., Fronteira de Pareto) para conduzir uma investigação minuciosa do potencial dos atributos explorados para detecção de desinformação disseminada em plataformas digitais considerando dois eventos importantes: as eleições presidenciais nos Estados Unidos e no Brasil em 2016 e 2018, respectivamente.

De forma geral, nossos resultados revelam que há um conjunto de atributos com potencial para identificação de desinformação disseminada nos diferentes cenários analisados. Em outras palavras, apresentamos evidências de que existem atributos extraídos do ambiente (e.g., medidas de disseminação do conteúdo na Web), do conteúdo (e.g., informações relacionadas à mensagens de endosso associadas) e da fonte (e.g., viés

[1] `http://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535`

político) da informação, que podem ser úteis para geração de modelos com alto desempenho capazes de detectar desinformação disseminada em eleições de diferentes países. Esperamos que os resultados obtidos forneçam insumos para a construção futura de abordagens agnósticas e robustas para a identificação de desinformação considerando diferentes cenários.

O restante do trabalho está organizado como segue. Na Seção 2 apresentamos uma breve descrição dos trabalhos relacionados. Depois, na Seção 3, descrevemos a metodologia experimental adotada neste trabalho. Os resultados são apresentados e discutidos na Seção 4. Por fim, na Seção 5 concluímos este trabalho.

2. Trabalhos Relacionados

"Desinformação" pode ser definida como "*uma notícia ou mensagem publicada e propagada pela mídia, contendo informações falsas, independentemente dos meios e motivos que a embasam*" [Sharma et al. 2019][2]. Recentemente, o estudo deste fenômeno em plataformas digitais tem atraído a atenção de pesquisadores de diversas áreas, incluindo jornalismo, ciências políticas e computação. Neste contexto, existe uma parcela significativa desses esforços que estão focados em entender características deste fenômeno em diferentes contextos como saúde [Ferrara 2020, Cinelli et al. 2020] e política [Bessi and Ferrara 2016], e considerando ainda, plataformas digitais distintas como Twitter [Bovet and Makse 2019], WhatsApp [Resende et al. 2019b, Resende et al. 2019a] e YouTube [Lemos et al. 2021]. Em [Ferrara 2020], por exemplo, o autor investiga como contas automatizadas (i.e., *bots*) contribuíram para o espalhamento de rumores e conspirações no Twitter, durante a pandemia de COVID-19. Por outro lado, no trabalho conduzido por Resende et al. [Resende et al. 2019b], os autores coletaram dados de grupos públicos (politicamente orientados) do WhatsApp, e apresentam evidências e/ou características da desinformação disseminada neste ambiente durante dois importantes eventos no Brasil: a greve nacional de caminhoneiros e a campanha presidencial brasileira de 2018. De maneira geral, esses esforços destacam o potencial da desinformação disseminada em plataformas digitais para manipulação da opinião pública em diversos contextos, o que ressalta a importância de que este fenômeno seja investigado de maneira multidisciplinar [Lazer et al. 2018], abrangendo, inclusive, a proposição de soluções que sejam efetivas para conter e/ou mitigar o problema.

Neste contexto, uma forma eficaz de detectar a desinformação disseminada em plataformas digitais é a checagem direta de fatos, normalmente realizada por jornalistas especialistas associados à agências de checagem de fatos como "Snopes.com", "PolitiFact" e "FactCheck.org"[3], nos Estados Unidos, e "Aos fatos", "Lupa", e "Boatos.org"[4], no Brasil. No entanto, esta tarefa pode se tornar um processo demorado, uma vez que requer uma análise detalhada para embasamento do veredito. Consequentemente, a checagem de fatos tradicional não consegue acompanhar o enorme volume de informações gerado atualmente no ambiente online [Ciampaglia et al. 2015]. Tal fato tem motivado o surgimento de vários esforços que propõem a verificação automática de fatos [Atanasova et al. 2019] incluindo abordagens para detecção automática de desinformação disseminada em diferentes plataformas, contextos e cenários [Conroy et al. 2015, Volkova et al. 2017,

[2]É válido ressaltar que esta é a definição do termo "Desinformação" adotada neste trabalho.

[3]`www.snopes.com`, `www.politifact.com/`, `www.factcheck.org/`

[4]`aosfatos.org`, `piaui.folha.uol.com.br/lupa/`, `www.boatos.org`

Reis and Benevenuto 2021, Bang et al. 2021]. Grande parte desses esforços, de forma geral, exploram técnicas baseadas em inteligência artificial, tais como aprendizado de máquina supervisionado [Conroy et al. 2015, Volkova et al. 2017, Reis et al. 2019b], ativo [Bhattacharjee et al. 2017], profundo [Wang et al. 2018, Kumar et al. 2020], dentre outros, para a proposição de modelos e/ou abordagens computacionais que sejam efetivas na contenção ou mitigação dos impactos ocasionados pelo problema. Resumidamente, esses trabalhos exploram características da informação disseminada (e.g., fonte, padrões textuais do conteúdo, medidas de propagação em plataformas digitais) para treinar classificadores que sejam capazes de identificar desinformação propagada em cenários específicos. Assim, é difícil avaliar o potencial desses atributos propostos para identificação da desinformação disseminada considerando diferentes cenários, como por exemplo, eleições em diferentes países (i.e., Estados Unidos e Brasil). Esta é a lacuna de pesquisa que este trabalho busca preencher.

3. Metodologia Experimental

Nesta seção são apresentados detalhes relativos à metodologia experimental adotada neste trabalho. Primeiramente, introduzimos os repositórios de dados selecionados de diferentes cenários (i.e., eleições nos Estados Unidos e no Brasil). Em seguida, apresentamos uma breve descrição dos atributos para detecção automática de desinformação disseminada em plataformas digitais explorados neste trabalho. Por fim, a estratégia para geração dos modelos, incluindo informações do algoritmo e métrica utilizados, bem como detalhes relativos ao experimento baseado na Fronteira de Pareto, são apresentados.

3.1. Repositórios de Dados

Durante esta etapa, nós conduzimos um breve levantamento de repositórios de dados disponíveis que nos permitissem investigar o potencial de (um mesmo conjunto de) atributos para detecção de desinformação em diferentes cenários. Neste contexto, focamos na obtenção de dados oriundos de eventos políticos que foram notoriamente impactados pela disseminação de campanhas de desinformação [Bessi and Ferrara 2016, Bovet and Makse 2019, Resende et al. 2019b], a saber: eleições presidenciais nos Estados Unidos e no Brasil nos anos de 2016 e 2018, respectivamente. Para definição dos repositórios de dados explorados neste trabalho foram considerados os seguintes critérios: (i) disponibilidade do conteúdo textual e informações relativas à fonte e propagação da informação em plataformas digitais, o que nos permite computar grande parte dos atributos propostos na literatura para identificação de desinformação e; (ii) representatividade dos dados em termos de volume de instâncias, que acreditamos que deva ser coerente com mundo real, onde a proporção de desinformação publicada pode ser considerada desbalanceada em comparação ao volume de informações produzido diariamente nestes ambientes, o que torna a tarefa ainda mais desafiadora. A Tabela 1 apresenta uma visão geral dos repositórios de dados selecionados que contemplam os cenários de interesse deste trabalho, seguido de uma breve descrição de características gerais de ambos.

Tabela 1. Visão geral dos repositórios de dados selecionados.

Cenário	#Instâncias	#Desinformação (%)	Período
Eleições nos Estados Unidos	2.018	302 ($\approx 15\%$)	2016
Eleições no Brasil	4.524	135 ($\approx 3\%$)	2018

Eleições nos Estados Unidos. Para contemplação do cenário eleitoral americano, exploramos o repositório de dados *BuzzFace*, disponibilizado em [Santia and Williams 2018]. De maneira resumida, ele contém 2.282 artigos de notícias rotulados por jornalistas do *BuzzFeed* relacionados às eleições presidenciais americanas de 2016 [Silverman et al. 2016]. Este repositório de dados do *BuzzFace* foi enriquecido com mais de $1,6$ milhões de comentários associados às notícias, além de dados de compartilhamentos e reações dos usuários no Facebook. As informações neste repositório são classificadas em 4 categorias[5]: "Verdadeiras", que representam 73% de todo conteúdo, "Majoritariamente Falsas" (4%), "Mistura entre Informações Verdadeiras e Falsas" (11%), e for fim, "Não é Fato" (12%). Para fins de simplicidade, agrupamos em uma única classe, que chamamos de "Desinformação", as instâncias classificadas como "Majoritariamente Falsas"e "Mistura entre Informações Verdadeiras e Falsas", e descartamos todo o conteúdo categorizado como "Não é Fato".

Eleições no Brasil. O repositório de dados construído em nossos esforços anteriores [Resende et al. 2019b, Reis et al. 2020] foi utilizado como representante do cenário eleitoral brasileiro. Em suma, ele é composto por mensagens disseminadas no WhatsApp durante o período eleitoral brasileiro de 2018, contendo 4.524 mensagens distintas[6] compartilhadas nesta plataforma entre agosto e outubro de 2018. Essas mensagens foram disseminadas em 414 grupos únicos e foram compartilhadas por 17.465 usuários distintos [Reis and Benevenuto 2021]. Neste contexto, é válido destacar que as 135 instâncias classificadas como "Desinformação"(cerca de 3% mensagens) foram rotuladas a partir de uma abordagem automatizada baseada em buscas no Google que é descrita com mais detalhes em [Reis et al. 2020]. O restante das mensagens (97%), foi classificado como informação "Não Verificada", uma vez que para este grupo, baseado na abordagem apresentada em [Resende et al. 2019b, Reis et al. 2020, Reis and Benevenuto 2021], a veracidade de seu conteúdo não foi necessariamente verificada por especialistas (e.g., agências de checagem de fatos). Por fim, o repositório de dados explorado neste trabalho (i.e., contendo mensagens rotuladas como "Desinformação") está disponível publicamente em: `http://doi.org/10.5281/zenodo.3779157`.

3.2. Extração de Atributos

Conforme discutido em trabalhos anteriores [Reis et al. 2019b], atributos para identificação de desinformação podem ser divididos em 3 grupos principais: (1) atributos extraídos do conteúdo, como por exemplo o tamanho e sentimento associado ao texto, (2) atributos extraídos da fonte (e.g., viés político do editor/produtor da informação), e por fim (3) atributos extraídos de ambiente, que de forma geral capturam aspectos da disseminação do conteúdo em plataformas digitais (e.g., número de compartilhamentos no WhatsApp ou no Facebook, etc) e na Web como um todo. Para este trabalho foram explorados atributos para detecção automática de desinformação implementados em nossos esforços anteriores [Reis et al. 2019b, Reis et al. 2019a, Reis and Benevenuto 2021], que apresentam uma descrição mais detalhada do processo de computação dos referidos atributos. A Tabela 2 apresenta um sumário dos 163 atributos implementados em ambos os repositórios de dados de interesse (i.e., eleições nos Estados Unidos e no Brasil) nos

[5]Traduzidas de forma livre pelos autores deste trabalho.

[6]Optamos por filtrar apenas mensagens que divulgam informações por meio de imagens uma vez que esforços anteriores mostraram que as imagens são o tipo de conteúdo de mídia mais frequente do WhatsApp, bem como uma importante fonte de desinformação [Resende et al. 2019b].

Tabela 2. Visão geral dos atributos implementados.

Extraído do(a)...	Grupo de Atributos	Descrição Geral (Exemplos)	Total
Conteúdo	Propriedades da Imagem (IMAG)	Número de rostos em uma imagem associada ao conteúdo, rótulos, cores, objetos, etc	8
	Atributos Sintáticos (SINT)	Atributos em nível de sentença, indicadores de qualidade do texto (ex.: métricas de legibilidade), etc	31
	Atributos Lexicais (LEXI)	Atributos em nível de caracteres e palavras, incluindo número de palavras, pronomes, verbos, pontuações, etc	47
	Atributos Psicolinguísticos (PSIC)	Sinais adicionais de linguagem persuasiva, como raiva, tristeza, etc. e indicadores de linguagem tendenciosa	26
	Estrutura Semântica (SEMA)	Rótulos, informações contextuais, medição de toxicidade do texto	8
	Subjetividade (SUBJ)	Medidas de subjetividade e análise de sentimentos	4
Fonte	Editor (EDIT)	Informações relacionadas ao editor/produtor do conteúdo	3
	Viés (VIES)	Alinhamento político do conteúdo (ex.: direita, esquerda, centro)	3
Ambiente (Plataforma Digital e Web)	Engajamento Interno (Plataforma Digital) (ENGA)	Métricas de engajamento do conteúdo na plataforma (e.g., número de compartilhamentos), em diferentes janelas de tempo	15
	Propagação Externa (Web) (PROP)	Informações relativas ao espalhamento de uma imagem associada à informação fora da plataforma digital (i.e., na Web)	5
	Padrões Temporais (TEMP)	A taxa na qual compartilhamentos são feitos internamente na plataforma para diferentes janelas de tempo (em segundos)	13

3 grupos relacionados[7]. Neste contexto, é válido ressaltar que não foram considerados, neste trabalho, atributos que são específicos e/ou dependentes de determinada plataforma, como por exemplo, informações sobre a credibilidade da fonte que são extraídas especificamente do repositório de dados das eleições americanas que disponibiliza o domínio responsável pela publicação da informação e, da mesma forma, dos grupos de WhatsApp onde as informações foram divulgadas, que são atributos extraídos especificamente do repositório de dados das eleições brasileiras.

3.3. Estratégia para Geração dos Modelos

Inicialmente, é preciso mencionar que, no contexto deste trabalho, a tarefa de detecção de desinformação consiste em: dado um conteúdo não rotulado $c \in \mathcal{C}$, um modelo para detecção de desinformação atribui uma pontuação $S(c) \in [0, 1]$ que indica a probabilidade de que c contenha desinformação. Neste cenário, um limiar τ pode ser definido de forma que uma função de previsão F seja capaz de distinguir um conteúdo com "Desinformação" dos demais.

Neste contexto, uma abordagem exata para avaliar o impacto real (ou potencial) de cada um dos 163 atributos implementados para detecção de desinformação, exigiria uma enumeração exaustiva de todas as possíveis combinações desses atributos, que seriam utilizadas, posteriormente, para geração dos modelos preditivos. Obviamente, avaliar todas essas combinações possíveis (i.e., subconjuntos de atributos) é computacionalmente inviável. Diante disso, nós amostramos o que chamamos de espaço de atributos $a \in \mathcal{A}$, selecionando aleatoriamente cada atributo a para compor um modelo m. Especificamente, o processo foi iniciado com a enumeração de todos as possíveis combinações de 1 e 2 atributos (i.e., 163 e 13.203, respectivamente). Em seguida, para cada combinação de 2

[7]É importante destacar que atributos extraídos do conteúdo não foram computados considerando apenas a informação (i.e., texto da notícia) isoladamente, mas também a manchete e a imagem associada à uma ela, bem como qualquer mensagem de endosso que tenha sido publicada em conjunto. Assim, especialmente para informações embutidas em imagens e vídeos – caso bastante comum no repositório de dados das eleições brasileiras, por exemplo – aplicamos técnicas de processamento de imagem (e.g., reconhecimento óptico de caracteres (OCR)) fornecido pelo *Google Vision API* disponível em `https://cloud.google.com/vision`, para extração do texto associado (i.e., etapa de pré-processamento), o que nos permitiu computar atributos de conteúdo.

atributos, incluímos aleatoriamente (e de maneira uniforme) neste subconjunto um novo atributo a. Este processo foi repetido até a obtenção de conjuntos de 20 atributos, que foram utilizados para a geração do espaço de modelos $\mathcal{M}$, composto por 247.941 modelos m. É importante destacar que, cuidadosamente, em cada execução do processo, garantimos que cada atributo seja incluído o mesmo número de vezes e ainda, que nenhum atributo apareça mais de uma vez em um mesmo subconjunto de atributos $a...a_n$, mantendo constante o número de modelos gerados (i.e., 13.203), independentemente do número de atributos que compõem um determinado conjunto (i.e., 2 até 20 atributos).

Algoritmo de Classificação. Os atributos implementados para detecção da desinformação podem ter uma variedade de interações não lineares complexas. Assim, capturar essas interações requer um algoritmo de classificação com flexibilidade significativa. Por este motivo, utilizamos um algoritmo de aprendizado de máquina que explora a estrutura de gradiente *boosting* em seu núcleo, ou seja, máquinas de gradiente *boosting*. Em alto nível, a ideia principal das máquinas de gradiente *boosting* é combinar vários modelos simples em um mais robusto. Mais especificamente, uma sequência de modelos é treinada iterativamente de forma que os erros dos modelos já treinados sejam considerados durante o treinamento de um novo modelo a ser incluído na sequência. Em outras palavras, a cada iteração, os erros são calculados e um modelo é ajustado com base nos erros identificados durante o processo de treinamento antecessor. Por fim, a contribuição de cada modelo intermediário para o modelo final é encontrada a partir da minimização do erro global do modelo definitivo. De fato, ajustar esses modelos é computacionalmente desafiador, então usamos uma implementação de alto desempenho de máquinas de gradiente *boosting*, chamada XGBoost (do inglês *eXtreme Gradient Boosting*), ou simplesmente XGB [Chen and Guestrin 2016].

Métrica de Avaliação. Para avaliação do desempenho dos modelos gerados, utilizamos a métrica área sob a curva ROC (em inglês, *Area Under the ROC Curve*) ou AUC [Baeza-Yates and Ribeiro-Neto 1999], que leva em consideração a relação sensibilidade-especificidade do modelo. Basicamente, a AUC serve como uma estimativa do equilíbrio desejado entre encontrar todos os positivos verdadeiros – "Desinformação" (i.e., taxa de falsos positivos igual a 0 ou taxa de verdadeiros positivos igual a 1) e reduzir o número de avaliações em conteúdo sem desinformação (i.e., taxa de falso negativos também próxima de 0). Logo, quanto maior a AUC, melhor o resultado, no sentido de que se atinge um melhor compromisso entre as duas medidas (verdadeiros positivos e falsos negativos) ao longo dos vários pontos de corte testados, evidenciando a capacidade do modelo proposto para classificar o conteúdo desejado. Nesse sentido, é válido destacar que AUC é robusta ao desbalanceamento de classe uma vez que considera todos os limites de classificação possíveis.

Por fim, para cada modelo, realizamos uma validação cruzada de 5 partições (i.e., treino e teste). Em outras palavras, o repositório de dados foi dividido em 5 partições, das quais 4 são usadas como dados de treinamento, e o restante (i.e., 1 partição) é usada para teste. O processo é então repetido 5 vezes, alternando, em cada execução, a partição usada para teste, produzindo, assim, 5 resultados. Logo, os valores de AUC reportados neste trabalho consistem no valor médio calculado a partir dos 5 resultados obtidos, o que representa, para cada modelo, uma estimativa do seu desempenho na tarefa de interesse.

3.4. Fronteira de Pareto

Depois, com o intuito de investigar o potencial dos atributos apresentados na Seção 3.2 para identificação de desinformação disseminada nas eleições presidenciais nos Estados Unidos e no Brasil, nós propomos um experimento baseado na eficiência de Pareto [Yoo and Harman 2007, Lin et al. 2019].

De forma geral, a eficiência (ou otimalidade) de Pareto estabelece uma relação de compromisso na alocação de recursos, na qual é impossível realocá-los de forma que todos os recursos sejam melhorados de forma conjunta[8]. Em outras palavras, "quando alguma ação pode ser feita para melhorar a situação de pelo menos uma pessoa sem prejudicar outra, então, ela deve ser feita". Esta ação (ou relação de compromisso) pode ser chamada de melhoria de Pareto [Ribeiro et al. 2014]. Na otimização multiobjetivo, esse conceito pode ser aplicado para determinação do que chamamos de Fronteira Pareto, isto é, o conjunto de pontos factíveis que não são dominados por nenhum outro ponto.

Assim, neste trabalho, este conceito foi explorado para investigar se existe um conjunto de atributos que sejam úteis para identificação de desinformação disseminada em diferentes cenários. Neste caso, o conjunto de atributos mais eficiente é aquele que não pode melhorar ainda mais um objetivo (por exemplo, modelos com alto desempenho em termos de AUC em um determinado repositório de dados), sem prejudicar o outro objetivo (ou seja, modelos com alto desempenho gerados a partir de outros repositórios de dados). Em suma, isso nos permite identificar o potencial de (conjuntos) atributos para geração de modelos com desempenho satisfatório na tarefa de detectar desinformação considerando dados oriundos de eleições em diferentes países: Estados Unidos e Brasil, conforme resultados apresentados e discutidos na próxima seção.

4. Resultados

Conforme mostrado na Figura 1, cada conjunto de atributos possível (i.e., modelo m) está associado à um ponto em um diagrama de dispersão bidimensional (i.e., espaço de modelos M. Neste contexto, cada ponto é representado como [x,y], onde cada coordenada x, y corresponde ao desempenho dos modelos em termos de AUC em cada um dos cenários analisados, ou seja, eleições nos Estados Unidos e no Brasil. Em outras palavras, os modelos foram treinados usando um conjunto de atributos [a_1,a_2..a_n] considerando os diferentes repositórios de dados apresentados na Seção 3.1, e partir disso, o desempenho deles foi avaliado em termos de AUC, conforme explicitado na Seção 3.3.

De forma geral, os 247.941 pontos no gráfico correspondem ao total de conjuntos de atributos e, consequentemente, todos os modelos $m \in \mathcal{M}$ gerados a partir da estratégia também apresentada na Seção 3.3. Os pontos azuis conectados pela linha vermelha indicam a escolha ótima em termos do conjunto de atributos (i.e., modelos m), formando um limite de Pareto sob o espaço de escolha restante (abaixo e à esquerda, pontos na cor cinza). Esses pontos não são dominados por nenhum outro ponto do diagrama de dispersão [Zames et al. 1981, Palda 2011], ou seja, eles representam casos para os quais nenhuma melhoria (de Pareto) é possível, sendo, portanto, conjuntos de atributos para identificação de desinformação mais prováveis de serem simultaneamente úteis para construir modelos com alto desempenho considerando dados oriundos dos diferentes

[8] https://blog.enacom.com.br/2019/01/14/fronteira-pareto

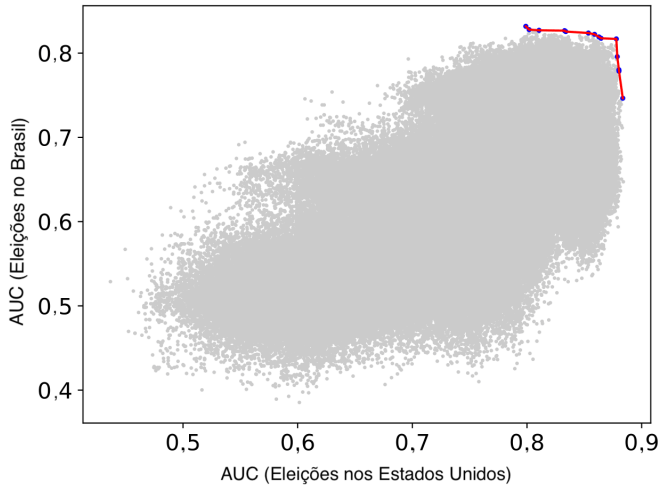

Figura 1. Eficiência de Pareto.

cenários analisados (i.e., eleições americanas e brasileiras). Assim, para obter uma compreensão profunda desses atributos que podem ser úteis para identificar desinformação nos diferentes cenários analisados, focamos nos 14 modelos, destacados em azul (i.e., gerados a partir de subconjuntos de atributos) que compõem a Fronteira de Pareto, destacada pela linha vermelha. Primeiramente, observamos que a AUC destes modelos está no intervalo de $[0,80 - 0,88]$ para o repositório de dados das eleições dos Estados Unidos e $[0,75 - 0,83]$ para o repositório de dados das eleições brasileiras. Além disso, dentre os modelos que compõem a Fronteira de Pareto, notamos que 28% deles (i.e., 4 modelos) obtiveram desempenho inferior a 0,8 em termos de AUC. De maneira geral, os modelos gerados apresentam um bom compromisso (*trade-off*) entre o desempenho em ambos os repositórios de dados.

Em seguida, para entendermos melhor a relação entre os atributos implementados e o desempenho dos modelos gerados – que compõem a Fronteira de Pareto – em cada um dos repositórios de dados analisados, calculamos a relação de prevalência de cada um deles. Os resultados são apresentados na Tabela 3. Observamos que os atributos extraídos do ambiente estão presentes em todos os modelos que constituem a fronteira (e.g., o número de sites/domínios que publicaram a imagem associada à notícia na Web (100%), o volume deles que são incomuns (100%) e padrões temporais (44%)). Além disso, alguns atributos extraídos do conteúdo também são muito frequentes nesses modelos (e.g., identificador de texto de endosso associado (55%) e contagem de palavras maiúsculas (55%)). Por último, atributos extraídos da fonte (e.g., informações sobre o viés político do editor/produtor da informação (44%)) e outros atributos extraídos do conteúdo, incluindo propriedades das imagens associadas (e.g., contagem de objetos de imagem (44%)) bem como informações referentes à estrutura semântica do conteúdo (44%)), são bastante prevalentes neste grupo de modelos. Em suma, concluímos que existem atributos que podem ser úteis para a geração de modelos com alto desempenho capazes de identificar a desinformação disseminada em diferentes cenários.

5. Conclusão

Neste trabalho apresentamos uma investigação do potencial de atributos para detecção da desinformação disseminada em diferentes cenários: eleições nos Estados Unidos e no

Tabela 3. Top-10 atributos nos modelos que compõem a Fronteira de Pareto.

Atributo	Grupo	(%)
cont_disseminacao_urls_web	Ambiente (PROP	100
cont_disseminacao_urls_dominios_incomuns	Ambiente (PROP)	100
idt_mensagem_endosso	Conteúdo (SINT)	55
cont_palavras_maiusculo	Conteúdo (LEXI)	55
taxa_prop_432000_segundos	Ambiente (TEMP)	44
objetos_imagem	Conteúdo (SEMA)	44
palavras_uso_pronomes	Conteúdo (LEXI)	44
cont_objetos_imagem	Conteúdo (IMAG)	44
palavras_foco_passado	Conteúdo (PSYC)	44
vies_politico	Fonte (VIES)	44

Brasil. Para isso, reunimos dados oriundos destes dois eventos e baseados em esforços anteriores, implementamos 163 atributos propostos na literatura com potencial para distinguir um conteúdo contendo "Desinformação" dos demais. Em seguida, propomos uma metologia que combina um algoritmo de classificação eficiente (i.e., XGB) para geração de modelos não enviesados em termos de atributos que os compõem, e mensuramos o desempenho deles em termos de AUC. Por fim, conduzimos um experimento baseado na Fronteira de Pareto que nos permitiu realizar uma análise aprofundada dos atributos que podem ser úteis para a geração de modelos que sejam efetivos na realização da referida tarefa.

De forma geral, nossos resultados apresentam evidências que existem atributos com potencial discriminativo satisfatório para geração de modelos com alto desempenho para detecção de desinformação considerando diferentes cenários, como por exemplo, eleições em diferentes países. Especificamente, mostramos que atributos extraídos do ambiente (e.g., medidas de propagação do conteúdo na Web) são os mais discriminativos considerando ambos os cenários, seguidos dos atributos extraídos do conteúdo (e.g., mensagens de endosso associadas, uso de palavras em formato maiúsculo, etc) e da fonte (e.g., viés ou alinhamento político), respectivamente. Acreditamos que nossos resultados forneçam insumos importantes para a proposição futura de ferramentas que sejam minimamente agnósticas e robustas para a identificação de desinformação disseminada em diferentes cenários.

Agradecimentos. Este trabalho foi parcialmente financiado pelo Ministério Público de Minas Gerais (MPMG), Projeto Capacidades Analíticas, CNPq, FAPEMIG e FAPESP.

Referências

Atanasova, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Karadzhov, G., Mihaylova, T., Mohtarami, M., and Glass, J. (2019). Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM press New York.

Bang, Y., Ishii, E., Cahyawijaya, S., Ji, Z., and Fung, P. (2021). Model generalization on covid-19 fake news detection. In *Proc. of the Int'l Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*.

Bessi, A. and Ferrara, E. (2016). Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11).

Bhattacharjee, S. D., Talukder, A., and Balantrapu, B. V. (2017). Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *Proc. of the IEEE Int'l Conf. on Big Data (Big Data)*.

Bovet, A. and Makse, H. A. (2019). Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proc. of the Int'l ACM Conf. on Knowledge Discovery and Data Mining (KDD)*.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. (2015). Computational fact checking from knowledge networks. *Plos One*, 10(6).

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. (2020). The covid-19 social media infodemic. *Scientific reports*, 10(1):1–10.

Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proc. of the Annual Meeting of the Association for Information Science and Technology (ASIS&T)*.

Ferrara, E. (2020). What types of covid-19 conspiracies are populated by twitter bots? *First Monday*.

Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., and Akbar, M. (2020). Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2).

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.

Lemos, A. L. M., Bitencourt, E. C., and dos Santos, J. G. B. (2021). Fake news as fake politics: the digital materialities of youtube misinformation videos about brazilian oil spill catastrophe. *Media, Culture & Society*, 43(5):886–905.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2016). A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, 17(2):1–16.

Lin, X., Chen, H., Pei, C., Sun, F., Xiao, X., Sun, H., Zhang, Y., Ou, W., and Jiang, P. (2019). A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proc. of the Int'l ACM Conf. on Recommender Systems (RecSys)*.

Mitchell, A. (2016). Key findings on the traits and habits of the modern news consumer. http://www.pewresearch.org/fact-tank/2016/07/07/modern-news-consumer/.

Palda, K. F. (2011). *Pareto's Republic and the new Science of Peace*. Filip Palda.

Reis, J. C. and Benevenuto, F. (2021). Supervised learning for misinformation detection in whatsapp. In *Proc. of the Brazilian Symp. on Multimedia and the Web (WebMedia)*.

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 Conf. on Web Science*.

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 Web and Social Media (ICWSM)*.

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

Report, D. N. (2018). Statistic of the week: How brazilian voters get their news. https://reutersinstitute.politics.ox.ac.uk/risj-review/statistic-week-how-brazilian-voters-get-their-news.

Resende, G., Melo, P., Reis, J. C. S., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019a). Analyzing textual (mis)information shared in whatsapp groups. In *Proc. of the Int'l ACM Conf. on Web Science*.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., and Benevenuto, F. (2019b). (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *Proc. of the ACM Web Conference (WWW)*.

Ribeiro, M. T., Ziviani, N., Moura, E. S. D., Hata, I., Lacerda, A., and Veloso, A. (2014). Multiobjective pareto-efficient approaches for recommender systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1–20.

Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In *Proc. of the Int'l ACM Conf. on Inform. and Knowledge Manag. (CIKM)*.

Santia, G. and Williams, J. (2018). Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proc. of the Int'l AAAI Conf. on Web. and Soc. Med. (ICWSM)*.

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., and Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42.

Silverman, C., Strapagiel, L., Shaban, H., Hall, E., , and Singer-Vine, J. (2016). Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis.

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 (ACL)*.

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

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proc. of the Int'l ACM Conf. on Knowledge Discovery and Data Mining (KDD)*.

Yoo, S. and Harman, M. (2007). Pareto efficient multi-objective test case selection. In *Proc. of the Int'l Symp. on Software Testing and Analysis (ISSTA)*.

Zames, G., Ajlouni, N., Ajlouni, N., Ajlouni, N., Holland, J., Hills, W., and Goldberg, D. (1981). Genetic algorithms in search, optimization and machine learning. *Information Technology Journal*, 3(1):301–302.