Analyzing the Potential of Feature Groups for Misinformation Detection in WhatsApp

Patrick De Angeli
Department of Informatics
Universidade Federal de Viçosa (UFV)
Viçosa, Minas Gerais, Brazil
patrick.angeli@ufv.br

Julio C. S. Reis
Department of Informatics
Universidade Federal de Viçosa (UFV)
Viçosa, Minas Gerais, Brazil
jreis@ufv.br

ABSTRACT

The new mass media (e.g., social networks, and instant messaging applications) have drastically changed how users generate, propagate, and consume information. Given this scenario, an emerging problem is the abuse of these platforms for the propagation of disinformation that, consequently, affects the credibility of the news ecosystem in these environments. The dissemination of misinformation on these platforms has become a worldwide phenomenon, and scalable strategies to contain or mitigate the problem are scarce. Thus, using automated approaches to detect disinformation on digital media can help journalists and fact-check teams identify content that needs to be verified. In this context, in this work, we investigate the potential of feature groups for automatic misinformation detection shared on digital platforms. Our results reveal that using a set of features from some groups, can be useful to build models with satisfactory performance, which can make the application of the model viable in a practical scenario.

Keywords: Fake News, Misinformation Detection, WhatsApp, Features

1 Introdução

As plataformas digitais, incluindo redes sociais e aplicativos de mensagem instantânea, estão presentes em mais de 70% das telas iniciais dos *smartphones* dos usuários, e, neste contexto, o WhatsApp, pode ser considerado o aplicativo de comunicação mais utilizado no Brasil[1]. Particularmente, 66% dos brasileiros utilizam esta plataforma para consumo de diversos tipos de informação, incluindo notícias [9], sobre assuntos distintos. No entanto, um problema impulsionado pelo crescente uso de plataformas digitais, e especificamente o WhatsApp, para o consumo de notícias, está relacionado à utilização desses ambientes para disseminação de campanhas de desinformação com o objetivo de enganar a população em determinado contexto e/ou manipular a opinião pública.

Recentemente, um estudo conduzido pela Fiocruz[2], mostrou que no início da pandemia da COVID-19 70% da desinformação relacionada ao coronavírus foi disseminada dentro da referida plataforma. Explorando dados de 2018, estudos revelaram a utilização do WhatsApp para a propagação de desinformação, por grupos politicamente orientados, considerando dois importantes eventos no Brasil naquele ano: as eleições presidenciais e a greve dos caminhoneiros[10]. Ademais, editorial do *The Guardian*[3] revelou que cerca de 42% das mensagens propagadas durante as eleições brasileiras foram verificadas por veículos de imprensa. Assim, é possível que um volume significativo de mensagens contendo desinformação tenha circulado nessa plataforma durante este período sem que tenha sido realizada uma checagem dos fatos associada.

Dentro deste contexto, um dos maiores desafios relacionados ao combate da desinformação em plataformas digitais, está relacionado à proposição de soluções que sejam úteis para mitigação do problema. Isso é importante pois, de forma geral, essas plataformas, incluindo o WhatsApp, não conseguem identificar facilmente um conteúdo contendo desinformação frente ao enorme volume de informações geradas diariamente. Além disso, o custo de verificar uma informação pode ser alto o que torna, impraticável, em alguns casos, a realização de checagens de fato em tempo real. Por exemplo, existem vereditos que podem demorar dias para serem construídos, uma vez que requerem uma análise e/ou investigação detalhada para suportá-los [13]. Especificamente sobre o WhatsApp, pode-se destacar ainda a dificuldade de acesso às mensagens, devido a criptografia de ponta a ponta, bem como o volume informações propagadas nos diferentes grupos públicos tornam o problema ainda mais desafiador.

Observando nuances do fenômeno da desinformação, e considerando os inúmeros prejuízos e/ou impactos ocasionados pela propagação deste tipo de conteúdo envolvendo assuntos como saúde e política, surgiram vários esforços que, além de fornecerem um entendimento do fenômeno,

[1] https://www.mobiletime.com.br/pesquisas/uso-de-apps-no-brasil-junho-de-2022/

[2] https://portal.fiocruz.br/noticia/pesquisa-revela-dados-sobre-fake-news-relacionadas-covid-19

[3] https://www.theguardian.com/world/2019/oct/30/whatsapp-fake-news-brazil-election-favoured-jair-bolsonaro-analysis-suggests

In: II Concurso de Trabalhos de Iniciação Científica (CTIC 2022), Curitiba, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2022.

ISSN 2596-1683

propõem abordagens para conter o problema [3, 6, 8, 14]. De forma geral, pesquisas sobre esse assunto buscam identificar padrões e atributos típicos deste tipo de conteúdo, testando-os em conjunto, não investigando por exemplo, se é possível a obtenção de um desempenho satisfatório na tarefa de identificação deste tipo de conteúdo, usando grupos específicos de atributos, e consequentemente um conjunto menor deles, o que poderia, por exemplo, viabilizar a aplicação de uma ferramenta automática em um cenário real prático, já que computar atributos pode ser computacionalmente custoso em diversos cenários.

Exposto o contexto, neste trabalho investigamos o desempenho de modelos gerados a partir de grupos de atributos propostos e implementados em trabalhos anteriores [6, 8]. Para isso, exploramos uma base de dados do WhatsApp, com informações da última eleição presidencial brasileira, ou seja 2018, que notavelmente foi marcada pela disseminação de desinformação.[4]. Em seguida, avaliamos o desempenho de uma abordagem clássica de aprendizado de máquina (i.e., *XGBoost*) que se mostrou bastante promissora neste contexto considerando esforços anteriores [6, 8]. De forma geral, os resultados obtidos mostram um grau útil de poder discriminativo dos grupos de atributos investigados para a tarefa proposta, que consiste em identificar uma mensagem contendo desinformação, podendo ser útil para apoiar checadores de fatos no processo de identificação de um conteúdo que necessite ser checado. Além disso, acreditamos que ao explorar grupos de atributos, e consequentemente um conjunto menor deles, diminuímos o custo associado ao processo de computação dessas informações, o que pode ser considerado um passo importante no sentido de investigar a viabilidade de implantação dessas abordagens automatizadas em ambientes reais.

O trabalho está organizado do seguinte modo. Na Seção 2, são brevemente discutidos os trabalhos relacionados. Em seguida na Seção 3, é apresentada a metodologia experimental que foi utilizada no trabalho, o conjunto de dados utilizados e dos atributos implementados para detecção de desinformação. Os resultados são apresentados e discutidos na Seção 4. Por fim, a Seção 5 conclui este trabalho e apresenta direções para trabalhos futuros.

2 Trabalhos Relacionados

Existem vários trabalhos com foco em propor abordagens para detecção de desinformação. Em suma, esses esforços implementam vários atributos extraídos da informação propagada, e propõem e/ou avaliam o desempenho de diferentes abordagens baseadas em aprendizado de máquina na realização da referida tarefa [8, 12].

Por exemplo, em trabalhos anteriores [7, 8], os autores exploraram cerca de 200 atributos para detecção de desinformação disseminada durante as eleições americanas de 2016, e usando um conjunto de dados disponível na literatura [11], avaliaram aspectos de desempenho e explicabilidade de abordagens baseadas em aprendizado de máquina. Já em [6], Reis et al. utilizaram uma base de dados das eleições brasileiras em 2018, na qual foram explorados aproximadamente 181 atributos para detecção de desinformação disseminada no WhatsApp. De forma geral, embora os resultados apresentados sejam promissores, existe um alto custo associado à computação de todos os atributos investigados, o que pode ser um fator impeditivo para implantação de abordagens desta natureza em cenários práticos. Além disso, em alguns casos, a computação de alguns atributos pode ser difícil, considerando por exemplo, a dependência de acesso a serviços externos para captura de determinadas informações (e.g., propagação de um conteúdo na Web).

Assim, de forma complementar aos esforços relacionados, usamos dados coletados e atributos implementados anteriormente [4, 6] para investigar o desempenho de abordagens automáticas na tarefa de identificar desinformação. No entanto, diferente dos trabalhos anterior, aqui, estamos interessados em investigar se é possível gerar modelos com desempenho satisfatório e/ou equivalente explorando diferentes grupos de atributos, e consequentemente um número reduzido deles.

3 Metodologia

Na presente seção, apresentamos detalhes relacionados à metodologia proposta para a realização do trabalho. Primeiramente, é importante mencionar que todos os experimentos foram realizados utilizando a ferramenta *Google Colab*[5]. A escolha desta ferramenta se deu em razão de sua aplicação de fácil e gratuito acesso, dado que a quantidade de dados colocada para ser treinada é relativamente pequena e pode ser executada diretamente no navegador (*browser*). A Figura 1 apresenta uma visão geral da metodologia proposta, que será detalhada a seguir.

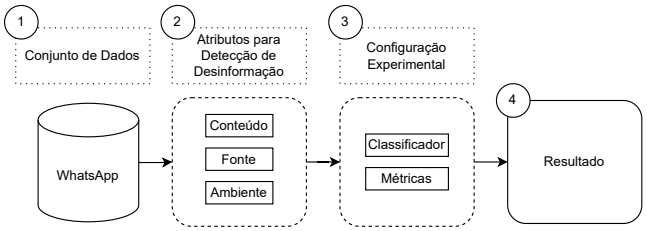

Figura 1. Visão geral da metodologia.

1) Conjunto de Dados. O aplicativo Whatsapp tem se tornado objeto de estudos no atual momento, dada a sua capacidade amplificada de disseminar informações de forma rápida. O grande volume de informações que circula nesta plataforma, atrelado à dificuldade em capturar/coletar as mensagens (criptografia de ponta-a-ponta), dificulta, potencialmente, a realização de pesquisas no ambiente. Tal fato

[4]https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html

[5]https://colab.research.google.com

Tabela 1. Visão geral dos atributos para detecção de desinformação explorados.

Grupo	Breve Descrição dos Atributos	Total
Conteúdo	Propriedades textuais (e.g., quantidade de palavras, palavras por sentença), atributos relacionados à sintaxe, semântica, e visuais como imagens compartilhadas (face, cor, objetos), etc.	31
Fonte	Informações relativas ao grupo (ou grupos) em que a mensagem disseminada, etc.	10
Ambiente	Contagem de compartilhamentos, se foi compartilhado em grupos não políticos, informações sobre o ambiente que foi disseminado, etc.	9

emerge a problemática residente na desafiadora tarefa de reconhecer quem está disseminando ou quem foi o autor de um conteúdo desinformativo. Contudo, para investigar as abordagens automatizadas de aprendizado de máquinas, treinando-as para gozar de capacidade para classificar desinformação, exploramos um conjunto de dados construído que compreende mensagens disseminadas pelo Whatsapp durante o período eleitoral brasileiro presidencial do ano de 2018.

Especificamente, neste trabalho, utilizamos a base de dados construída em [6, 10], que contém 4.524 mensagens rotuladas, disseminadas em 414 grupos únicos e compartilhados por 17.465 usuários únicos do WhatsApp. É importante mencionar que, baseado em uma abordagem que explora o rótulo dado por agências de checagem de fatos brasileiras, cerca de $\approx 3\%$ (i.e., 135) dessas mensagens foram rotuladas como desinformação.

2) Atributos para Detecção de Desinformação. Trabalhos anteriores mostraram que os atributos que podem ser utilizados para detectar informação falsa elencam-se em 3 grupos: *i)* atributos textuais de conteúdo (e.g., propriedades textuais e visuais associadas a um conteúdo) *ii)* atributos extraídos da fonte de informação (e.g., editor, de quem publicou a mensagem) *iii)* atributos extraídos do ambiente, que envolvem dinâmicas de propagação em plataformas digitais e na Web como um todo. A Tabela 1 apresenta uma visão geral dos atributos explorados neste trabalho. No total foram explorados 50 atributos distribuídos entre os diferentes grupos (i.e., conteúdo, fonte e ambiente). É válido ressaltar que esses atributos foram selecionados com base em seu poder discriminativo para realização da tarefa reportado em esforços anteriores [6]. Maiores detalhes sobre o processo de implementação dos atributos podem ser obtidos em [5].

3) Configuração Experimental. Neste trabalho exploramos um classificador bastante robusto para tarefas de classificação (incluindo detecção de desinformação [8], que não demanda muita capacidade computacional, chamado XGBoost [2] (ou simplesmente, XGB).

Para verificação do desempenho da abordagem, foram adotadas métricas amplamente utilizadas para tarefas de Aprendizado de Máquina e Recuperação de Informação, sendo elas: MacroF1, que nos permite avaliar de forma adequada a performance das abordagens em um cenários desbalanceado, e *area under the ROC curve* (AUC), uma métrica para classificação binária frequentemente usada como medida de qualidade do desempenho dos modelos [1].

De forma geral o classificador "aprendeu" um modelo a partir de um conjunto de dados previamente rotulado e, em seguida, valemo-nos deste modelo para distinguir instâncias contendo "desinformação" das demais. Os experimentos foram realizados com validação cruzada de 5 partições. Os resultados obtidos durante esta etapa do trabalho serão apresentadas na próxima seção.

4 Resultados

A Tabela 2 apresenta os resultados experimentais obtidos neste trabalho, utilizando o classificador XGB e os atributos explorados e descritos na seção anterior (ver Tabela 1), considerando todas as combinações possíveis considerando os grupos analisados (i.e., Conteúdo, Fonte e Ambiente). Podemos observar que os melhores resultados foram obtidos explorando atributos dos grupos de Conteúdo + Fonte, com 0,82 e 0,95 em termos de AUC e Macro-F1, respectivamente. Especificamente em termos de AUC, atributos extraídos da fonte, de forma isolada ou combinada (e.g., + atributos extraídos do Ambiente), também mostraram resultados bastante promissores.

Em seguida, para o melhor resultado identificado anteriormente, foi realizada uma investigação inicial do potencial prático dessas abordagens automáticas para detecção de desinformação na base de dados do WhatsApp. Para isso inspecionou-se a curva de ROC do classificador XGB utilizando atributos extraídos do Conteúdo + Fonte, conforme mostrado na Figura 2. De forma geral, podemos notar que é possível escolher um limite para classificar corretamente a maioria das mensagens que contém desinformação (Taxa de

Tabela 2. Resultados experimentais usando o classificador XGB.

Conjuntos de Atributos	AUC	Macro-F1
Conteúdo	0,68 (± 0,31)	0,80 (± 0,31)
Fonte	0,80 (± 0,19)	0,84 (± 0,19)
Ambiente	0,79 (± 0,39)	0,86 (± 0,39)
Conteúdo + Fonte	0,82 (± 0,18)	0,95 (± 0,18)
Conteúdo + Ambiente	0,77 (± 0,22)	0,84 (± 0,22)
Fonte + Ambiente	0,79 (± 0,20)	0,86 (± 0,20)
Conteúdo + Fonte + Ambiente	0,78 (± 0,21)	0,84 (± 0,21)

Verdadeiros Positivos ≈ 1), enquanto classifica-se erroneamente cerca de 60% do conteúdo restante. Esses resultados são próximos aos resultados apresentados em esforços anteriores [6], o que fornece evidências de que podemos gerar modelos com desempenho satisfatório utilizando grupos de atributos. Conjecturamos que essa diminuição do número de atributos, e consequentemente do custo computacional para computá-los, pode ser importante para viabilizar a implantação de abordagens automatizadas para detecção de desinformação em cenários práticos.

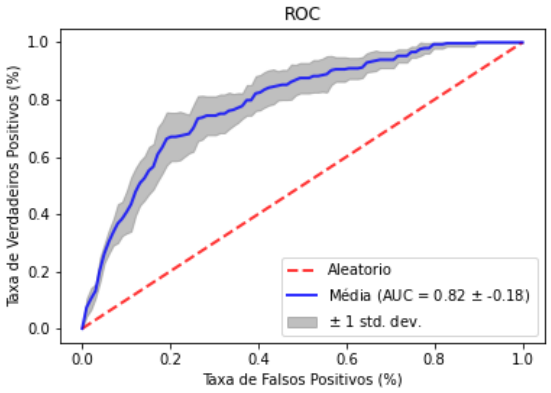

Figura 2. Curva ROC para o melhor resultado obtido (Conteúdo + Fonte) com o uso do classificador XGB.

5 Conclusão

Neste trabalho, avaliamos o desempenho de uma abordagem supervisionada (i.e., XGB) na tarefa de identificar desinformação disseminada em grupos públicos do WhatsApp. Especificamente, a partir de trabalhos anteriores, investigamos o potencial de grupos de atributos para geração de modelos que sejam efetivos na realização da tarefa. Em outras palavras, ao analisar grupos de atributos, paralelamente analisamos se é possível diminuir o custo computacional associado (em termos do número de atributos computados e explorados pelo modelo) ao mesmo tempo que o desempenho da abordagem não é prejudicada. Acreditamos que isso possa ser um passo importante no sentido de avaliar o potencial de aplicação prático dessas abordagens para conter a desinformação disseminada em plataformas digitais.

De forma geral, nossos resultados mostram que é possível gerar bons modelos com um número reduzido de (grupos de) atributos, usando o XGB. Os melhores resultados de classificação podem detectar corretamente quase todas as mensagens contendo desinformação, enquanto classificam incorretamente aproximadamente 60%, o que acreditamos já seja suficiente para detectar e apoiar a tarefa de checagem de fatos no processo de identificação do conteúdo a ser verificado. Outro aspecto importante está relacionado ao fato de que, computar atributos a partir do conteúdo e da fonte é menos restritivo em comparação a computação de atributos extraídos do ambiente que dependem de informações relacionadas à propagação do conteúdo.

Por fim, espera-se que esse trabalho seja capaz de fornecer insumos para a proposição de soluções práticas para identificação de desinformação não somente relacionada ao âmbito eleitoral, mas, futuramente, a saúde e outros contextos, e em outras plataformas (e.g., Telegram). Como trabalhos futuros, pretendemos conduzir analises individuais de informatividade dos atributos implementados e estratégias alternativas de seleção de atributos bem como investigar o potencial de técnicas de aprendizado de maquinas mais sofisticadas (e.g., aprendizado profundo) na tarefa de detectar desinformação.

Agradecimentos. Este trabalho foi parcialmente financiado pela FAPEMIG e CNPq (Edital UFV PIBIC/CNPq 2021-2022).

Referências

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern information retrieval.* Vol. 463. ACM press New York.
[2] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proc. of the Int'l ACM Conf. on Knowledge Discovery and Data Mining (KDD).*
[3] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. In *Proc. of the Annual Meeting of the Association for Informa. Science and Tech. (ASIS&T).*
[4] Philipe Melo, Fabrício Benevenuto, Daniel Kansaon, Vitor Mafra, and Kaio Sá. 2021. Monitor de WhatsApp: Um Sistema para Checagem de Fatos no Combate à Desinformação. In *Proc. of the Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)* (Minas Gerais). 79–82.
[5] Julio C. S. Reis et al. 2020. Towards automatic fake news detection in digital platforms: properties, limitations, and applications. (2020).
[6] Julio C. S. Reis and Fabrício Benevenuto. 2021. Supervised Learning for Misinformation Detection in WhatsApp. In *Proceedings of the Brazilian Symposium on Multimedia and the Web.* 245–252.
[7] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Explainable machine learning for fake news detection. In *Proc. of the ACM Conference on Web Science.* 17–26.
[8] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems* 34, 2 (2019), 76–81.
[9] Digital News Report. 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.
[10] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *Proc. of the ACM Web Conference (WWW).*
[11] Giovanni C Santia and Jake Ryland Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proc. of the Int'l AAAI Conference on Web and Social Media.*
[12] Renato M Silva, Roney LS Santos, Tiago A Almeida, and Thiago AS Pardo. 2020. Towards automatically filtering fake news in Portuguese. *Expert Systems with Applications* 146 (2020), 113199.
[13] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proc. of the ACL Workshop on Language Technologies and Computational Social Science.* 18–22.
[14] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 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).*