AUTOMATA: Um Ambiente para Combate Automático de *Fake News* em Redes Sociais Virtuais

Uma Experiência no Contexto da Pandemia de COVID-19

Augusto José M. da Fonseca
CEFET-RJ
augusto.fonseca@eic.cefet-rj.br

Carlos Henrique da S. Moreira
UniSãoJosé-RJ
kmoreira620@gmail.com

Gabriel Resende Machado
PUC-Rio
gresende@inf.puc-rio.br

Paulo Márcio Souza Freire
IME-RJ
paulomsfreire@ime.eb.br

Ronaldo Ribeiro Goldschmidt
IME-RJ
ronaldo.rgold@ime.eb.br

Abstract

The propagation of fake news on social media has been increasing significantly in the last years. Despite the existence of applications that aim at suppressing the proliferation of fake news written in Portuguese, it was noticed the proposed solutions present a passive behavior regarding two aspects: (i) they are limited to identifying potential fake news only from content introduced by their users, as well as (ii) the absence of any procedures to combat disinformation. Therefore, this article presents *AUTOMATA*, an automated tool for combating fake news written in Portuguese. *AUTOMATA* periodically monitors posts made on social media and relies on Artificial Intelligence to detect suspicious fake news. After the detection process, *AUTOMATA* adopts a two-pronged approach to autonomously mitigate the widespread of this content, either by emitting posts on social media for warning about potential fake news or by sending the detected content to be curated by fact-checking agencies. This article also reports an experience in partnership with Ministério da Saúde for the application of *AUTOMATA* in the context of the COVID-19 pandemic in Brazil.

Keywords: Fake News, Disinformation, Artificial Intelligence, Classification, Fake News Detection.

1 Introdução

O problema de combater *fake news* (i.e., notícias falsas divulgadas intencionalmente [6]) não é recente [2]. Contudo, sua complexidade vem aumentando em função do crescimento do volume e da velocidade de divulgação dessas notícias nos meios digitais, em especial, nas Redes Sociais Virtuais (RSVs), tais como *Twitter*, *Facebook*, entre outras [1].

Como exemplo do poder de influência negativa das *fake news*, pode ser citado o caso da pandemia de *COVID-19*, que matou mais de seis milhões de pessoas pelo mundo[1], e que foi objeto de inúmeras *fake news* divulgadas em RSVs [5].

Em busca de mitigar os efeitos nocivos das *fake news*, ferramentas computacionais que possam auxiliar no combate a esse tipo de notícia têm sido desenvolvidas, sendo poucas delas voltadas para notícias escritas em Língua Portuguesa[2,3,4,5,6,7]. Até onde foi possível observar, tais ferramentas apresentam uma postura passiva, sob dois aspectos: (i) se limitam a identificar possíveis *fake news* apenas a partir de notícias a elas apresentadas pelos seus usuários; (ii) não atuam para tentar mitigar a propagação dessas notícias.

Diante do exposto, o presente artigo tem como objetivo apresentar o *AUTOMATA*, um ambiente computacional de combate *automático* a *fake news* escritas em Língua Portuguesa e divulgadas em RSVs. Tal ambiente monitora de forma autônoma e periódica as RSVs, em busca de postagens relacionadas a notícias sobre assuntos de interesse previamente indicados pelo usuário do *AUTOMATA*. Em seguida, o ambiente utiliza técnicas de Inteligência Artificial (IA) para classificar tais notícias como possíveis *fake news*. Por fim, no caso de notícias classificadas como suspeitas, o *AUTOMATA* as remete automaticamente para análise por Agências de Checagem de Fatos (ACFs), como também emite alertas de possíveis *fake news* junto às RSVs visando, desta forma, mitigar o espalhamento da desinformação. A fim de demonstrar a viabilidade prática do *AUTOMATA*, foi implementado um protótipo funcional do ambiente e realizada uma experiência em parceria com o Ministério da Saúde de aplicação deste protótipo no contexto da pandemia da COVID-19 no Brasil. O protótipo desenvolvido monitora notícias postadas no *Twitter*, utiliza a abordagem de detecção de *fake news* baseada em *Crowd Signals* proposta em [3] e gera uma interface com notícias suspeitas para análise pela ACF *Boatos.Org*.

[1] www.who.int/emergencies/diseases/novel-coronavirus-2019.
[2] https://nilc-fakenews.herokuapp.com/about.
[3] https://sites.google.com/view/detector-de-fake-news/.
[4] http://www.fakepedia.org/.
[5] https://dl.acm.org/doi/abs/10.1145/3323503.3361698.
[6] https://dl.acm.org/doi/abs/10.1145/3323503.3360648.
[7] https://dl.acm.org/doi/abs/10.1145/3470482.3479467.

In: XXI Workshop de Ferramentas e Aplicações (WFA 2022), Curitiba, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2022.

ISSN 2596-1683

2 *AUTOMATA*

O *AUTOMATA* tem como base a arquitetura macro-funcional conceitual ilustrada na Figura 1. Os módulos de *Monitoramento*, *Detecção*, *Intervenção*, *Scraping* e *FeedBack ACF* compõem o eixo principal de processamento do ambiente e são executados automaticamente de forma encadeada compondo um ciclo de processamento. O *AUTOMATA* possui uma *engine* responsável por controlar cada ciclo de processamento e realizar auto-recuperação em caso de falhas. Além disso, por meio do parâmetro *frequency*, é possível configurar a *engine* para executar os ciclos de processamento em intervalos de tempo pré-definidos. A configuração deste e de outros parâmetros necessários ao funcionamento do *AUTOMATA* é feita por meio do *Painel Administrativo*, um componente do ambiente que permite acompanhar os ciclos de processamento e as notícias analisadas. Os próximos parágrafos detalham os módulos do *AUTOMATA* e o *Painel Administrativo*. O vídeo disponível em *https://youtu.be/Go7tLl-WQl8* aprofunda esse detalhamento, abrangendo a descrição das entradas/saídas, parâmetros, interfaces e relatórios da ferramenta.

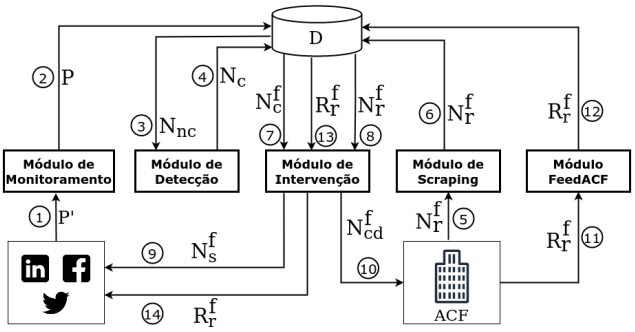

Figura 1. Arquitetura Macro-Funcional do *AUTOMATA*.

Módulo de *Monitoramento:* o módulo de *Monitoramento* é responsável por selecionar um conjunto P de postagens realizadas nas RSVs que envolvam uma ou mais palavras-chave previamente definidas em função do domínio de interesse do usuário do *AUTOMATA* para combater *Fake News* (*e.g.* política, economia, saúde). Formalmente, este módulo recebe como entrada um conjunto de postagens P' obtidas nas RSVs (fluxo de dados 1), onde cada $p \in P'$ possui o atributo $p.W$ que contém o conjunto de palavras do texto postado em p. Em seguida, o módulo utiliza o conjunto de palavras-chave W para gerar o conjunto $P = \{p \in P' | p.W \cap W \neq \emptyset\}$ que é, então, armazenado no repositório D (fluxo de dados 2). A fim de evitar duplicação de notícias[8], o módulo de *Monitoramento* verifica se cada p enviada em P é similar a alguma notícia n *recente* (i.e., publicada nos últimos $window\_size_m$ dias na RSV, onde $window\_size_m$ é um parâmetro previamente configurado no ambiente) pertencente a um histórico de notícias N armazenado em D. Para tanto, o módulo aplica uma variação do algoritmo de distância de edição aos textos em $p.W$ e $n.W$ para cálculo de similaridade entre os textos. Caso o valor de similaridade retornado pelo algoritmo seja superior ao parâmetro s do *AUTOMATA*, o módulo registra em D a associação de p a n.[9] Caso contrário, o módulo insere uma nova notícia em N com os dados de p. Por fim, é importante mencionar que a execução do módulo de *Monitoramento* em cada ciclo de processamento ocorre durante um intervalo de tempo configurado por meio do parâmetro *stream_time*, expresso em minutos.

Módulo de *Detecção:* o módulo de *Detecção* recupera, a partir de D (fluxo de dados 3), o conjunto de notícias não classificadas N_{nc} a fim de avaliar se cada $n_{nc} \in N_{nc}$ pode ou não ser *fake news*. Para tanto, o referido módulo implementa a abordagem de detecção de *fake news* baseada em *crowd signals implícitos* originalmente proposta em [3], premiada em [4] e estendida em [2]. Em resumo, tal abordagem utiliza modelos de IA para classificar notícias como possíveis *fake news*, levando em conta, para isso, a reputação baseada no comportamento prévio (*signals*) dos usuários da RSV (membros do *crowd*) na divulgação tanto de notícias anteriores quanto da notícia a ser classificada. Os *crowd signals* utilizados pela abordagem em questão são considerados *implícitos* pois são inferidos pelos modelos de *IA*. Dispensam, portanto, a participação humana para opinar explicitamente sobre a classificação da notícia como demandado pelas soluções do estado da arte baseadas em *crowd signals*. Ao final de sua execução, este módulo gera o conjunto N_c (fluxo de dados 4), onde cada $n_c \in N_c$ é enriquecida com dois atributos: $n_c.y$ e $n_c.q$. O primeiro contém a opinião do *AUTOMATA* sobre a classe de n_c (i.e., $n_c.y \in \{fake, not\text{-}fake\}$) e o segundo indica a probabilidade estimada pelo ambiente de que n_c realmente pertença a $n_c.y$.

Módulo de *Scraping*: antes de iniciar o módulo de *Intervenção*, o *AUTOMATA* realiza, por meio do módulo de *Scraping*, uma busca por notícias já checadas e rotuladas como *fake news* por ACFs. Esta busca é necessária pois permitirá ao módulo de *Intervenção* evitar reenviar para as ACFs notícias já checadas e rotuladas por elas como sendo *fake news*. Em termos formais, o módulo de *Scraping* obtém o conjunto de notícias $N_r^f = \{n | n.r = fake\}$ junto às ACFs (fluxo de dados 5) e o armazena em D (fluxo de dados 6). Convém notar que o atributo $n.r$ de uma notícia n indica o rótulo atribuído a n, onde $n.r \in \{fake, not\text{-}fake\}$.

Módulo de *Intervenção*: o módulo de *Intervenção* atua automaticamente junto às RSVs e ACFs, informando sobre as possíveis *fake news* identificadas pelo módulo de *Detecção*. Para tanto, a partir do conjunto N_c armazenado em D, recupera (via fluxo de dados 7) o subconjunto $N_c^f = \{n_c \in N_c \mid n_c.y = fake \wedge n_c.q \geq q_{min} \wedge n_c.dt \geq f_{dt}(window\_size_i)\}$, onde $n_c.dt$

[8] Cabe mencionar que, para o *AUTOMATA*, os conceitos de *Postagem* e *Notícia* correspondem, respectivamente, a um texto publicado por um usuário da RSV e a um texto que divulga um acontecimento socialmente relevante.

[9] Vale ressaltar que postagens distintas podem ser associadas a uma mesma notícia.

é a data de divulgação da notícia n_c e $f_{dt}(window\_size_i)$ uma função que retorna a data referente a $window\_size_i$ dias anteriores a data corrente. $window\_size_i$ e q_{min} são parâmetros do ambiente previamente configurados que correspondem, respectivamente, ao comprimento de janela de observação (expresso em número de dias) e a um limiar mínimo de probabilidade de uma notícia ser *fake*. Em resumo, N_c^f corresponde ao conjunto de notícias suspeitas de serem falsas segundo a opinião do *AUTOMATA* que tenham sido publicadas recentemente (i.e., nos últimos $window\_size_i$ dias) nas RSVs. Em seguida, N_c^f é ordenado de forma decrescente em relação aos valores de $n_c.ns$ e $n_c.q$, sendo $n_c.ns$ o número de compartilhamentos de n_c nas RSVs. Neste momento, o módulo de *Intervenção* seleciona os *num_records* (parâmetro do sistema) primeiros registros para criar o conjunto N_{cd}^f de notícias candidatas à intervenção. Dessa forma, são priorizadas para intervenção as notícias mais recentes, suspeitas de serem *fake news* e com maior propagação nas RSVs (i.e., aquelas com possível maior potencial de causar desinformação nas RSVs). Em seguida, o módulo de *Intervenção* recupera do repositório D (via fluxo de dados 8) o conjunto N_r^f de notícias rotuladas como *fake news* pelas ACFs anteriormente coletadas pelo módulo de *Scraping*. Então, para cada notícia candidata $n_{cd}^f \in N_{cd}^f$ é aplicado o mesmo algoritmo de similaridade utilizado no módulo de *Monitoramento* para avaliar se o texto de n_{cd}^f é similar ao texto de pelo menos uma das notícias em N_r^f, empregando o mesmo limiar s usado no módulo de *Monitoramento*. Caso a notícia seja avaliada como similar, esta é incluída no conjunto de notícias similares N_s^f e removida de N_{cd}^f. Tal remoção visa evitar o envio de notícias já rotuladas pelas ACFs. Ao final desse processamento são, então, realizadas automaticamente duas ações de intervenção: *(i)* para cada notícia $n_s^f \in N_s^f$ é realizada uma postagem nas RSVs (fluxo de dados 9) alertando que o texto em n_s^f é uma possível *fake news* e; *(ii)* as notícias remanescentes em N_{cd}^f são enviadas para as ACFs (fluxo de dados 10). O grande espaço de busca gerado pelas postagens em RSVs e as possíveis limitações operacionais dos usuários do *AUTOMATA* poderiam tornar inviável o uso da ferramenta. No entanto, o módulo de *Intervenção* foi projetado para adaptar-se à capacidade operacional tanto dos usuários do *AUTOMATA* quanto das ACFs parceiras por meio da configuração dos diversos parâmetros indicados na descrição deste módulo.

Módulo de *FeedBack ACF*: o módulo de *FeedBackACF* processa de forma automática o conjunto R_r^f, retorno de uma ACF (fluxo de dados 11) sobre o conjunto de notícias suspeitas N_{cd}^f encaminhado a ela, atualizando o repositório D (fluxo de dados 12). Formalmente, $R_r^f = \{n_{cd}^f \in N_{cd}^f | n_{cd}^f.r = fake\}$. É importante ressaltar que, sempre que R_r^f é atualizado em D, o módulo de *Intervenção* automaticamente recupera essas notícias (fluxo de dados 13) e, para cada $n_r^f \in R_r^f$, realiza uma postagem nas RSVs (fluxo de dados 14), alertando que o texto de n_r^f é uma *fake news* confirmada pela ACF.

Módulo de *Painel Administrativo*: o *Painel Administrativo* do *AUTOMATA* é um componente desacoplado da ferramenta e desenvolvido na forma de uma aplicação *web*. Os usuários devidamente credenciados podem acessá-lo e visualizar em detalhe ou de forma consolidada todas as notícias processadas pelo *AUTOMATA*. A Figura 2 apresenta um exemplo de relatório fornecido pelo *Painel Administrativo*. Como recurso adicional, o *Painel Administrativo* conta com uma funcionalidade de curadoria que permite que a análise de uma notícia possa ser realizada pelos próprios usuários do *AUTOMATA* especialistas no domínio de interesse, desobrigando a necessidade de atuação junto a uma ACF. Em resumo, usuários podem, sob sua própria conta e risco, rotular notícias suspeitas e disparar os alertas nas RSVs por meio desta funcionalidade. É importante mencionar que todas as alterações realizadas no repositório D por meio do *Painel Administrativo*, assim como suas respectivas autorias, permanecem armazenadas para fins de auditoria, caso necessário. Por último, como comentado no início desta seção, todos os parâmetros dos módulos do *AUTOMATA* são configuráveis por meio de interface no *Painel Administrativo*.

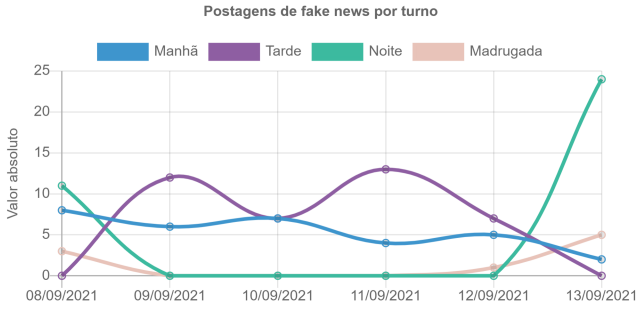

Figura 2. Exemplo de Relatório do *Painel Administrativo*.

3 Experiência no Contexto da COVID-19

A fim de avaliar a viabilidade da aplicação do *AUTOMATA* no combate automático a *fake news*, foi realizada uma experiência em parceria com o Ministério da Saúde no contexto da Pandemia da COVID-19 no Brasil. Foi implementado um protótipo[10] baseado na arquitetura macro-funcional como descrita na Seção 2 tendo o *Twitter* e a *Boatos.org* como RSV e ACF, respectivamente.

A experiência ocorreu do dia 08/09/2021 ao dia 15/09/2021. A *engine* foi configurada para executar 1 ciclo de processamento a cada intervalo de 6 horas (*frequency=21.600 seg*). Tal intervalo teve por estratégia permitir monitorar a RSV

[10]Foram utilizadas as tecnologias Python, PHP e PostgreSQL. O código fonte e a documentação do protótipo podem ser obtidos em *https://github.com/projeto-confia/automata*. O *AUTOMATA* encontra-se em processo de registro como software livre e sua licença específica será definida ao final desse processo. O *hardware* utilizado possui: 1 processador Intel Xeon Silver 4208, 32 GiB DIMM DDR4 3200MHz e 2 discos SATA 2TB 7.200RPM 6Gbps.

nos diferentes períodos do dia (manhã, tarde, noite e madrugada). O módulo de monitoramento foi configurado para realizar a coleta durante 20 minutos (*stream_time=1.200 seg*) em cada ciclo. O conjunto de palavras-chave utilizado foi W = {"COVID-19", "CORONAVÍRUS", "VACINA", "PANDEMIA", "PFIZER", "CORONAVAC", "JOHNSON", "OXFORD", "ASTRAZENECA", "SPUTNIK"}. As configurações adotadas para os demais parâmetros foram definidas empiricamente, e são apresentadas na Tabela 1.

Tabela 1. Parâmetros definidos para o *AUTOMATA*.

Configurações	
limiar sim. $s = 0.7$	q_{min} = 0.9
$window\_size_m = 30$	$num\_records$ = 4
$window\_size_i$ = 7	

O total de postagens coletadas durante toda a experiência foi de 29.002 postagens. O volume médio diário de postagens coletadas foi de aproximadamente 9.069 postagens sendo aproximadamente 765 durante o período da madrugada, 1.207 durante o período da manhã, 2.635 durante o período da tarde e 4.462 durante o período da noite. Tal volume de postagens coletadas correspondeu a 14.274 notícias das quais 0,8% (115 notícias) foram classificadas como *fake news* e 0,3% (42 notícias) foram selecionadas para Intervenção. Vale ressaltar que a abordagem de detecção de *fake news* utilizada no *AUTOMATA* alcançou, aproximadamente, 92% de acurácia e 2% falsos positivos, conforme descrito em [2]. A partir desse resumo estatístico, é possível projetar para um monitoramento de 24h por dia que o volume de postagens coletadas poderia chegar a mais de 160.000 unidades diárias. Tal projeção poderia aumentar ainda mais no caso de inclusão de mais palavras-chave.

Diante das estatísticas obtidas, é possível perceber não somente a adequação do *AUTOMATA* para atuar no combate a *fake news* de forma automática como também a necessidade de uma ferramenta desse tipo para lidar com o alto volume de dados envolvidos neste combate. Por exemplo, só o volume diário de postagens demandaria um custo consideravelmente elevado de recursos humanos para realizar de forma não automatizada as tarefas de monitoramento, detecção e intervenção, sem contar a possibilidade de falhas humanas no processo. Além disso, por meio dos diversos parâmetros que podem ser configurados no *AUTOMATA* é possível ajustar seu ciclo de processamento, adequando-o ao hardware disponível, ao número de ACFs apoiadoras e à estratégia de comunicação na RSV.

4 Considerações Finais

A disseminação de *fake news* vem aumentando consideravelmente em volume e velocidade nas RSVs. Atualmente, já existem ferramentas computacionais para o combate automático de *fake news*, sendo que poucas delas são voltadas para analisar conteúdos em Português. Até onde foi possível observar, essas ferramentas apresentam uma postura passiva, sob dois aspectos: (i) se limitam a identificar possíveis *fake news* apenas a partir de notícias a elas apresentadas pelos seus usuários, e (ii) não atuam de forma ostensiva para mitigar a propagação dessas notícias.

Diante desse contexto, o presente trabalho apresentou as seguintes contribuições: (1) o *AUTOMATA*, uma ferramenta para combate automático a *fake news* escritas em Português nas RSVs[11]. Baseado em IA, o *AUTOMATA* - diferentemente das ferramentas existentes - é pró-ativo, atuando periodicamente tanto no monitoramento de notícias postadas em RSVs, como também na intervenção por meio de duas vertentes: pela emissão de alertas de possíveis *fake news* nas RSVs, e pela interação com ACFs para curadoria das notícias detectadas como suspeitas; (2) o relato de uma experiência em parceria com o Ministério da Saúde de aplicação do protótipo no contexto da Pandemia da COVID-19 no Brasil.

Como trabalhos futuros, podem ser destacados: (1) o monitoramento de outras RSVs e a interação com outras ACFs; (2) a avaliação de outros modelos de IA no módulo de *Detecção*; (3) análise de conteúdo multimídia associado às notícias (imagens, vídeos e áudios); (4) priorização de palavras-chave no monitoramento nas RSVs; (5) personalização de alertas em função do perfil dos usuários das RSVs; (6) diversificação dos canais de envio de alertas, como *Whatsapp* e *SMS*; e (7) avaliação do *AUTOMATA* em outros idiomas.

Agradecimentos

Este trabalho recebeu apoio das seguintes instituições: MCTI, MS, CNPq (401662/2020-9), Boatos.Org, IME e AGITEC.

Referências

[1] N. J. Conroy, V. L. RFubin, and Y. Chen. 2015. Automatic Deception Detection: Methods for Finding Fake News. *Assoc. Information Science and Technology* 52. https://doi.org/10.1002/pra2.2015.145052010082

[2] P. Freire et al. 2021. Fake news detection based on explicit and implicit signals of a hybrid crowd. *Expert Systems with Applications* 183 (2021). https://doi.org/10.1016/j.eswa.2021.115414

[3] P. Freire and R. Goldschmidt. 2019. Fake News Detection on Social Media via Implicit Crowd Signals. In *Proc. of the 25th WebMedia* (Rio de Janeiro, Brazil). ACM, New York, NY, USA, 521–524. https://doi.org/10.1145/3323503.3360626

[4] P. Freire and R. Goldschmidt. 2020. Combatendo Fake News nas Redes Sociais via Crowd Signals Implícitos. In *Anais do XVI ENIAC*. SBC, Porto Alegre, RS, 424–435. https://doi.org/10.5753/eniac.2019.9303

[5] Y. Mejova and K. Kalimeri. 2020. Advertisers Jump on Coronavirus Bandwagon: Politics, News, and Business. *ArXiv* abs/2003.00923 (2020).

[6] K. Shu, A. Silva, S. Wang, J. Tang, and H. Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1, 22–36. https://doi.org/10.1145/3137597.3137600

[11] Vale ressaltar que o *AUTOMATA* é adaptável para outros idiomas, e pode ser configurado de acordo com a capacidade de operação dos seus usuários.