Detecção de Fake News via Sinais Implícitos de Crowds: Uma Abordagem para Mitigar o Cold-Start no Cálculo da Reputação

Viviane Antonia Corrêa Thomé
vthome@ime.eb.br
Instituto Militar de Engenharia
Rio de Janeiro, Brasil

Paulo Márcio Souza Freire
paulo.freire@dde.faetec.rj.gov.br
Fundação de Apoio à Escola Técnica
Rio de Janeiro, Brasil

Ronaldo Ribeiro Goldschmidt
ronaldo.rgold@ime.eb.br
Instituto Militar de Engenharia
Rio de Janeiro, Brasil

ABSTRACT

The evolution of Digital Media for News Distribution has changed how people share content. Anyone can freely share information, including fake news (i.e., false information shared intentionally). In this scenario, fake news detection approaches have been proposed, with notable attention given to the one that uses crowd signals. Such approaches explore the collective sense by combining opinions (i.e., signals) of a high number of users (i.e., crowd), considering the reputations of these users regarding their capacity to identify fake news. Although promising, the crowd signals approaches have a significant limitation when many users in the crowds have not interacted with prior news. This lack of information leads to a cold-start problem when calculating those users' reputations. The present work raises the following hypothesis: the performance of the crowd signals-based detection models can be improved if they mitigate the cold-start problem by inferring the users' reputation based on the behavior those users would present in the face of prior news. This hypothesis is grounded on the fact that people tend to share news that seems familiar to their beliefs. In a search to validate the raised hypothesis, SCS, a crowd signals-based fake news detection method, is proposed. SCS considers similarities among texts (e.g., title and content) of past news to infer the reputation of users with the cold-start problem. BERT, a known Large Language Model (LLM) was used to provide embeddings that represent the texts. Preliminary experimental results demonstrate the effectiveness of the proposed method when compared to the crowd signals-based SOTA in detecting fake news.

KEYWORDS

Crowd Signals, Detecção de Fake News, LLM

1 INTRODUÇÃO

Os Meios Digitais de Divulgação de Notícias (MDDN), como redes sociais e aplicativos de mensagens, estão cada vez mais integrados na vida das pessoas. Qualquer pessoa com conexão à Internet pode compartilhar, em tempo real e a um baixo custo, suas próprias experiências e conteúdos por meio da divulgação de mensagens, de vídeos ou de imagens [2, 8, 15].

Entretanto, a evolução dos MDDN tem apresentado um efeito colateral indesejado: pessoas sendo expostas a todo tipo de informação, incluindo notícias falsas divulgadas intencionalmente, as chamadas *fake news* [1, 10, 11]. Exemplos de *fake news* são as informações falsas relacionadas às fortes enchentes que atingiram mais de 400 cidades no Rio Grande do Sul[1]. Tais informações prejudicaram os esforços de assistência à população afetada.

Diante dos efeitos negativos da disseminação de *fake news* nos MDDN, pesquisas científicas têm sido propostas com o objetivo de detectar automaticamente esse tipo nocivo de notícia [1, 6, 8]. Neste contexto, explorar métodos computacionais que analisam a reputação dos usuários em relação à sua capacidade de identificar *fake news* tem se apresentado como uma opção interessante [3, 13, 14]. Esses métodos utilizam o conceito de *crowd signals* combinando as opiniões (i.e., *signals*) de um grande número de usuários (i.e., *crowd*) para a classificação de notícias como *fake* (f) ou *not fake* ($\overline{f}$). Para tanto, cada opinião é ponderada de acordo com a respectiva reputação do usuário em opinar (i.e., acertos e erros em opiniões anteriores). Inclusive, a opinião de cada usuário participante do *crowd* pode ser obtida de duas formas. Na forma explícita, o usuário opina se uma notícia é f ou $\overline{f}$, por meio de uma funcionalidade raramente disponível no meio digital onde a notícia foi divulgada. Na forma implícita, a opinião do usuário é inferida com base em seu comportamento em relação à notícia divulgada no meio digital.

Embora os métodos baseados em *crowd signals* tenham demonstrado resultados promissores na detecção de *fake news*, uma parcela significativa dos usuários participantes dos *crowds* apresenta problema de *cold-start* no cálculo de suas reputações. Isso ocorre, pois os métodos baseados em *crowd signals* dependem do histórico de opiniões dos usuários sobre notícias passadas para o cálculo de suas reputações. Essas informações históricas nem sempre estão disponíveis.

Assim sendo, este estudo levanta a seguinte hipótese: *os desempenhos dos métodos de detecção de* fake news *utilizando crowd signals podem alcançar melhores resultados se eles mitigarem o problema de cold-start por meio da inferência das reputações dos usuários baseada no comportamento que esses usuários teriam diante de notícias passadas*. Na direção de se obter resultados experimentais que forneçam evidências de validade da hipótese levantada, o presente artigo propõe o *SCS*, método de detecção de *fake news* baseado em crowd signals que busca mitigar o problema de *cold-start* no cálculo das reputações dos usuários. Adaptado a partir do método *HCS-I* [13], o SCS considera similaridades entre textos associados às notícias (e.g. título e conteúdo das postagens) a fim de realizar as inferências sobre o comportamento dos usuários mediante notícias divulgadas no passado. A fim de calcular as similaridades entre notícias, o método proposto utiliza o BERT, um robusto modelo de linguagem de grande porte (do inglês, *LLM - Large Language Models*) para prover os *embeddings* (i.e. representações vetoriais)

[1] https://netlab.eco.ufrj.br/post/enchentes-norio-grande-do-sul-uma-análise-da-desinforma%C3%A7%C3%A3o-multiplataforma-sobre-o-desastre-climáti

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2024). Juiz de Fora, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2966-2753

dos textos associados às notícias. A inspiração para formulação do SCS se apoia no trabalho de Schwarz e Jalbert [9], onde os autores descrevem que as pessoas tendem a compartilhar notícias que lhes pareçam familiares, dada a similaridade de opiniões. Para tanto, as notícias precisam ser compatíveis com suas crenças preexistentes.

Os experimentos preliminares com o SCS foram realizados sobre um *dataset* com notícias escritas em língua portuguesa (pt-br). Os resultados obtidos pelo método proposto apresentaram desempenhos superiores aos do método original HCS-I, provendo, portanto, os primeiros indícios de validade da hipótese formulada.

Este artigo está organizado como segue: a Seção 2 apresenta os trabalhos do estado da arte no combate automático às *Fake News* mais fortemente ligados à pesquisa ora descrita; a Seção 3 recorda resumidamente o método HCS-I, utilizado como base para o desenvolvimento do SCS; a Seção 4 apresenta o método SCS propriamente dito; os resultados dos experimentos realizados são expostos e debatidos na Seção 5. Por fim, na Seção 6 estão as considerações finais do estudo realizado, destacando as contribuições da pesquisa e as iniciativas de trabalhos futuros.

2 RELATED WORK

Este estudo considera trabalhos que propuseram métodos de detecção de *fake news* baseados em *crowd signals*. Nesses estudos, as notícias são classificadas como f ou $\overline{f}$, por meio da combinação das opiniões (i.e, *signals*) de usuários (i.e, *crowd*), onde essas opiniões são ponderadas com base na reputação do usuário. A reputação de cada usuário é obtida a partir da sua capacidade histórica em opinar (i.e., acertos e erros em opiniões anteriores).

O método *Detective* [14] tem por objetivo mitigar a disseminação de *fake news*, visando interromper a sua propagação. Para tanto, o *Detective*, por meio de uma funcionalidade experimental presente na rede social *Facebook*, coletava as opiniões explícitas dos usuários sobre uma dada notícia. Com o objetivo de classificar a notícia como f ou $\overline{f}$, uma inferência *bayesiana* ponderava essas opiniões com base nas respectivas reputações dos usuários membros do *crowd*.

O trabalho HCS [13], diferente de *Detective*, apresenta uma abordagem que utiliza as opiniões implícitas dos usuários membros do *crowd*, na tarefa de detecção de *fake news*. Na HCS há dois métodos disponíveis, o *HCS-I (implicit)* e *HCS-F (full)*. O *HCS-I* coleta as opiniões implícitas dos usuários, a partir da divulgação da notícia por esses usuários. O *HCS-F*, além das opiniões implícitas dos usuários divulgadores, considera a opinião fornecida por outros métodos de detecção de *fake news* existentes na literatura. Ambos, assim como o *Detective*, utilizam a inferência *bayesiana* para ponderação das opiniões com base nas reputações dos membros do *crowd*. Os métodos da HCS demonstraram desempenho similar ao *Detective*, apesar de não dependerem de uma funcionalidade no meio digital para coletar a opinião explícita do usuário e, nem tão pouco, da boa vontade do usuário em opinar.

O TA-TCS [3], assim como o *HCS-I*, considera as opiniões implícitas dos usuários divulgadores e pondera essas opiniões com base na reputação desses usuários. Entretando, com o objetivo de obter uma detecção antecipada, utiliza a propagação temporal das notícias. O TA-TCS se diferencia da HCS por incluir uma nova etapa de particionamento dos intervalos de divulgação da notícia. O TA-TCS demonstrou assertividade similar ao *HCS-I*, além de obter a detecção antecipada.

Embora esses métodos baseados em *crowd signals* tenham demonstrado resultados promissores na detecção de *fake news*, observa-se que parte dos usuários não possuem dados históricos de divulgação de notícias para os cálculos de suas reputações. Essa carência caracteriza o problema de *cold-start* no cálculo de suas reputações. Isso ocorre porque os métodos baseados em *crowd signals* dependem do histórico de opiniões dos usuários sobre notícias passadas para o cálculo de suas reputações. A Seção 3 apresenta um resumo do funcionamento do método HCS-I utilizado como base neste estudo.

3 REVISÃO DO MÉTODO HCS-I

O HCS-I é um método baseado em crowd signals que busca detectar se uma dada notícia é f ou $\overline{f}$, combinando as opiniões implícitas dos usuários divulgadores dessa notícia. O HCS-I considera que o fato de um usuário divulgar uma notícia é um sinal implícito de que, na opinião desse usuário, a notícia é $\overline{f}$ [13]. Tal princípio se inspira na citação do filósofo Habermas [7], segundo a qual toda ação comunicativa traz consigo uma inevitável pretensão à verdade. É importante destacar que cada opinião implícita é ponderada utilizando a reputação do respectivo usuário. Essa reputação é calculada com base nos acertos e erros que esse usuário obteve ao opinar, de forma implícita, sobre notícias do passado e que, portanto, sabe-se quais são f e quais são $\overline{f}$. Com base nessa ponderação, o HCS-I pode ser beneficiado, inclusive, na ocorrência de opiniões erradas e/ou maliciosas. Para detectar *fake news* em um MDDN, o método HCS-I executa as três etapas representadas na Figura 1 para cada notícia $n^D \in N^D$, onde N^D representa o conjunto de notícias a serem analisadas. A seguir apresentam-se os detalhes de cada etapa.

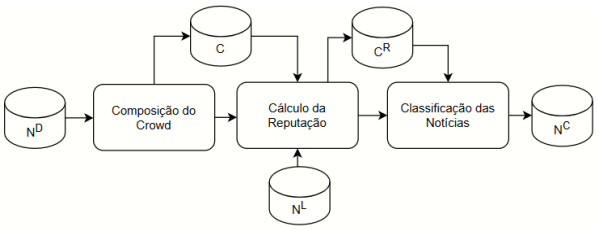

Figura 1: Visão Geral das Etapas do Método HCS-I.

Composição do Crowd: o objetivo é construir o *Crowd* C, identificando no conjunto de usuários, do meio digital, aqueles que divulgaram a notícia n^D a ser analisada. Dessa forma, ao final desta etapa, cada $c_i \in C$ é um usuário divulgador de notícia n^D.

Cálculo da Reputação: responsável por medir a reputação de cada membro $c_i \in C$. A reputação de cada c_i é expressa pela probabilidade de acertar e, consequentemente, de errar as opiniões já fornecidas por c_i sobre as notícias f ou $\overline{f}$ divulgadas no passado. Assim, a aferição da reputação dos membros de C depende da disponibilidade de um conjunto de notícias N^L, já rotuladas como f ou $\overline{f}$. Para isto, cada notícia $n^L \in N^L$ possui seu respectivo rótulo $Y^*(n^L)$, cujo valor $y^*(n^L) \in \{f, \overline{f}\}$. O cálculo da reputação de cada membro c_i é realizado com base na sua opinião sobre as notícias rotuladas existentes em N^L. Dessa forma, é necessário calcular para cada membro c_i sua probabilidade de acertar (e errar) sua opinião sobre notícias *fake* e acertar (e errar) sua opinião sobre notícias *not fake*. Para tanto, nesta etapa, o método HCS-I constrói, para cada c_i, uma

matriz de opinião O_{c_i} como representada na Figura 2 (a). Cada componente n_{rs} de O_{c_i} ($r, s \in \{f, \overline{f}\}$) indica a quantidade de notícias sinalizadas como r por c_i, dado que os rótulos reais dessas notícias são s. Se c_i é um usuário divulgador da notícia n^D, HCS-I utiliza as notícias divulgadas por c_i, armazenadas em N^L, para preencher a primeira linha de O_{c_i}. Inicialmente $n_{\overline{f}\overline{f}} = 0$ e $n_{\overline{f}f} = 0$. Uma vez que c_i tenha decidido divulgar uma notícia n^L, esse comportamento é assumido como um sinal implícito de que c_i considera n^L como *not fake*. Isto é, $Y_{c_i}(n^L) = \overline{f}$. Assim, para cada $n^L \in N^L$ divulgado por c_i, se $Y^*(n^L) = \overline{f}$, então $n_{\overline{f}\overline{f}} = n_{\overline{f}\overline{f}} + 1$, senão $n_{\overline{f}f} = n_{\overline{f}f} + 1$.

Para preencher a segunda linha de O_{c_i} seria necessário recuperar todas as notícias que c_i visualizou, mas optou por não divulgá-las, por considerá-las *fake*. Como tal informação não se encontra disponível, o método HCS-I busca inferir a capacidade de c_i em identificar notícias *fake* por meio de dois critérios a seguir:

(1) Deve-se preservar a capacidade de c_i em acertar ou errar na sinalização. Desta forma, $n_{ff}/(n_{f\overline{f}} + n_{ff})$ deve ser equivalente a $n_{\overline{f}\overline{f}}/(n_{\overline{f}\overline{f}} + n_{\overline{f}f})$.
(2) O método HCS-I deve calcular o número de exemplos sinalizados nas duas classes, preservando a proporcionalidade da primeira linha de O_{c_i}. Assim, $n_{f\overline{f}}$ deve ser dado por $(n_{\overline{f}f}/n_f) \times n_{\overline{f}}$, onde n_f e $n_{\overline{f}}$ representam, respectivamente, o total de notícias *fake* e *not fake* em N^L.

A Figura 2 (b) apresenta exemplo de preenchimento parcial da matriz na primeira linha. A Figura 2 (c) apresenta o preenchimento completo da segunda linha. Considerando $n_f = 30$ e $n_{\overline{f}} = 60$, a primeira linha representa a situação em que um dado usuário divulgou 15 notícias, sendo 3 f e 12 $\overline{f}$. A primeira linha é preenchida conforme os rótulos reais das notícias. A segunda linha é preenchida de forma proporcional de acordo com a capacidade apurada sobre o usuário de acertar e de errar. Portanto, para completar a matriz na primeira coluna da segunda linha, tem-se $n_{f\overline{f}} = (n_{\overline{f}f}/n_f) \times n_{\overline{f}}$ ou $n_{f\overline{f}} = 3/30 \times 60 = 6$. Para a segunda coluna da segunda linha tem-se $n_{ff} = (n_{\overline{f}\overline{f}} \times n_{f\overline{f}})/n_{\overline{f}f}$ ou $n_{ff} = (12 \times 6)/3 = 24$.

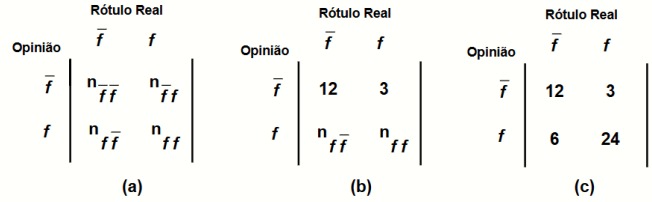

Figura 2: Matriz de Opinião

Com base na versão completa de O_{c_i}, o método HCS-I pode realizar as inferências das probabilidades $\theta_{c_i,f}$ e $\theta_{c_i,\overline{f}}$ para cada usuário c_i membro do *Crowd*, onde: $\theta_{c_i,f} = n_{ff}/(n_{\overline{f}f} + n_{ff})$ e $\theta_{c_i,\overline{f}} = n_{\overline{f}\overline{f}}/(n_{\overline{f}\overline{f}} + n_{f\overline{f}})$. Os elementos $\theta_{c_i,f}$ e $\theta_{c_i,\overline{f}}$ indicam a probabilidade de c_i em sinalizar uma notícia n^D como $\overline{f}$, dado que a notícia é de fato $\overline{f}$ ($P(Y_{c_i}(n) = \overline{f} | Y^*(n) = \overline{f})$). Ou, da mesma forma, que seja sinalizada como f, dado que a notícia é de fato f ($P(Y_{c_i}(n) = f | Y^*(n) = f)$). Esse processo de inferência resulta na reputação de cada usuário c_i, apresentada na matriz de reputação R_{c_i} abaixo. Com essas probabilidades calculadas, o HCS-I é capaz de armazenar em C^R cada $c_i \in C$ com a sua respectiva reputação R_{c_i} (i.e. $C^R = \{{c_i}^R / {c_i}^R = (c_i, R_{c_i})$ e $c_i \in C\}$). Importante destacar que quando c_i não possui histórico de divulgação de notícias (i.e., c_i não divulgou qualquer notícia em N^L), $\theta_{c_i,\overline{f}} = \theta_{c_i,f} = 50\%$.

$$R_{c_i} = \begin{bmatrix} \theta_{c_i,\overline{f}} & 1 - \theta_{c_i,f} \\ 1 - \theta_{c_i,\overline{f}} & \theta_{c_i,f} \end{bmatrix}$$

Classificação das Notícias: Nesta etapa, o HCS-I utiliza as reputações R_{c_i} de todos os membros do *crowd* retornados por C^R para calcular a probabilidade de n^D ser f ou $\overline{f}$. Seguindo uma abordagem *bayesiana*, o método utiliza as Equações 1 e 2, onde ω e, respectivamente, 1 - ω, representam a probabilidade da notícia ser f ou $\overline{f}$. A classe correspondente à maior dentre as duas probabilidades é, portanto, o resultado gerado pelo HCS-I.

$$P(Y^*(n) = f) = \omega . \prod_{c_i \in C^R} (1 - \theta_{c_i,f}) \quad (1)$$

$$P(Y^*(n) = \overline{f}) = (1 - \omega) . \prod_{c_i \in C^R} \theta_{c_i,\overline{f}} \quad (2)$$

4 MÉTODO PROPOSTO SCS

O método SCS é uma adaptação do método HCS-I. Ele busca fazer inferências sobre os comportamentos dos usuários participantes de C (membros do crowd) que não divulgaram notícias no passado (i.e., não divulgaram qualquer notícia em N^L). Por meio de similaridades entre as notícias a serem analisadas ($n^D \in N^D$) e as notícias do passado cujos rótulos já são conhecidos ($n^L \in N^L$), procura-se refinar o cálculo das reputações dos usuários $c \in C$. Para tanto, um LLM deve ser utilizado para gerar as representações vetoriais dos textos vinculados às notícias (exs: títulos, conteúdos, etc).

O método SCS segue a mesma dinâmica de execução do método HCS-I, com as mesmas etapas: *Composição do Crowd*, *Cálculo da Reputação* e *Classificação da Notícia*. No entanto, diferentemente do HCS-I, o SCS não atribui 50% à probabilidade dos usuários sem histórico de divulgação opinarem que uma notícia seja f. Em vez disso, o SCS executa quatro passos adicionais ao final da Etapa *Cálculo da Reputação* do HCS-I, para tornar possíveis as inferências de opiniões desses usuários. Tais passos estão ilustrados na Figura 3. Cada passo encontra-se descrito a seguir.

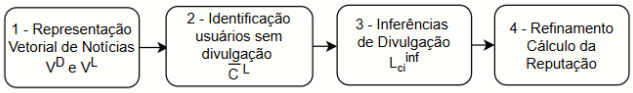

Figura 3: Passos Adicionais à Etapa *Cálculo da Reputação*.

Representação Vetorial de Notícias: Nesta etapa, um modelo de LLM pré-treinado, escolhido pelo analista responsável pelos experimentos, faz a extração das características semânticas de cada notícia e a representa vetorialmente por meio dessas características. Dessa forma, cada notícia a ser analisada $n^D \in N^D$ e cada $n^L \in N^L$, que tenha sido divulgada por um membro participante do *crowd* C, passará a contar com uma representação vetorial. Ao final, o conjunto V^D deverá conter as representações vetoriais das notícias a serem analisadas e o conjunto V^L os vetores que representam notícias previamente rotuladas.

Identificação Usuários sem Divulgação: o objetivo desta etapa é identificar cada c_i que não possui histórico de divulgação de notícias entre as notícias em N^L. Neste momento, C é dividido em

dois subconjuntos: C^L e $\overline{C^L}$ contendo, respectivamente, membros do *crowd* que possuem e que não possuem histórico de divulgação de notícias, entre as notícias em N^L.

Inferências de Divulgação: para cada $c_i \in \overline{C^L}$, seleciona-se cada notícia a ser analisada $n^D \in N^D$ que tenha sido divulgada por c_i e a sua representação vetorial $v^D \in V^D$. Para cada v^D, é feito um cálculo de similaridade, comparando v^D com a representação vetorial v^L de cada $n^L \in N^L$ (i.e., cada notícia divulgada no passado). Este cálculo deve ser implementado por meio de uma função de similaridade $f_S(\cdot,\cdot)$, gerando uma lista ordenada decrescente dos valores $f_S(v^D, v^L)$. Um critério de corte é aplicado sobre esta lista para selecionar as *Top-k* notícias cujas representações vetoriais v^L sejam mais similares à representação vetorial v^D da notícia n^D. Tal lista resultante do corte é armazenada em $L^{inf}_{c_i}$ e representa as inferências sobre notícias do passado que sejam mais similares à notícia divulgada no presente por c_i e que deverá ser analisada pelo método de detecção de *fake news*. Tal conjunto será considerado pelo SCS de forma a mitigar o problema do *cold-start*.

Refinamento Cálculo da Reputação: para cada membro do *crowd* $c_i \in C$, o preenchimento da 1ª linha de O_{c_i} é realizado da seguinte forma: (i) se $c_i \in C^L$, o cálculo considera cada $n^L \in N^L$ divulgado por c_i; (ii) se, por outro lado, $c_i \in \overline{C^L}$, o cálculo utiliza cada inferência de notícia $n^L \in L^{inf}_{c_i}$. Em ambos os casos, esta etapa utiliza as mesmas equações para o preenchimento das matrizes O_{c_i} e R_{c_i}, conforme apresentado na Seção 3.

Para exemplificar, considere o cenário com 30 notícias f, 60 $\overline{f}$ e um usuário $c_x \in C$ que divulgou a notícia n^D e não tenha divulgações prévias. Considere também que, durante o passo *Inferência de Divulgação*, tenham sido identificadas 10 notícias n^L similares a n^D. Ao analisar cada notícia similar, foi constatado que dentre as *Top-k* (supondo $k = 5$), 4 são f e 1 é $\overline{f}$. Para preencher a 1ª linha de O_{c_x} é preciso considerar as opiniões implícitas inferidas: $n_{\overline{ff}} = 1$ e $n_{\overline{f}f} = 4$. Para a 2ª linha deve-se proceder com o cálculo proporcional: $n_{f\overline{f}} = 8$ e $n_{ff} = 2$. Importante notar que, neste mesmo exemplo, as componentes da matriz O_{c_x} receberiam valor zero.

5 EXPERIMENTOS E RESULTADOS

Para a análise do método proposto foram realizados experimentos preliminares com o *dataset FakeNewsSet* [4]. O *FakeNewsSet* contém textos de notícias escritas em português e extraídas do Twitter™. É composto por 600 notícias, nas quais 300 são *fake* e 300 são *not fake*. Possui 16.024 usuários e 27.059 divulgações.

Nos testes realizados foi aplicada a técnica de validação cruzada com 10 *folds*. Em cada iteração, 90% das notícias foram alocadas no conjunto que contém o histórico das notícias N^L (para cálculo da reputação) e, os outros 10% das notícias foram alocadas no conjunto N^D (para detecção). Para a extração das representações vetoriais dos textos vinculados às notícias (título e conteúdo) foi utilizado o LLM pré-treinado BERT, um dos modelos do estado da arte em aplicações envolvendo linguagem natural [5, 12].

No passo *Inferências de Divulgação*, foi utilizado o parâmetro k do *Top-k* para indicar a quantidade de notícias similares inferidas a serem consideradas nas análises. Nos experimentos, *Top-k* foi configurado com $k = 5$, um valor ímpar fixado a fim de impedir a ocorrência de empates nos cálculos das probabilidades entre f e $\overline{f}$.

Tabela 1: Resultados dos experimentos.

Método	Acurácia ($\mu \pm \sigma$)	Precisão ($\mu \pm \sigma$)	Recall ($\mu \pm \sigma$)	F1 ($\mu \pm \sigma$)
HCS-I	0.957±0.024	0.932±0.049	0.984±0.022	0.957±0.026
SCS_{text}	**0.980±0.018**	**0.975±0.024**	0.983±0.024	**0.979±0.026**
SCS_{title}	**0.980±0.016**	**0.996±0.011**	0.962±0.032	**0.979±0.019**

Para avaliação das similaridades foram utilizados textos relacionados às notícias como os conteúdos (*text*) e os títulos (*title*).

A Tabela 1 mostra os resultados dos experimentos comparando os desempenhos obtidos pelos métodos HCS-I e SCS, nos 10 *folds* da validação cruzada. As métricas escolhidas para a avaliação dos experimentos foram a Acurácia, Precisão, Recall e F-1. O SCS demonstrou melhores resultados que o HCS-I em quase todas as métricas, com exceção do *Recall*. Tal diferença é explicada pelo fato do novo método ter conseguido aumentar a sua capacidade em identificar notícias $\overline{f}$ com maior precisão, impactando a métrica *Recall*, visto que há um *trade-off* entre estas duas métricas.

Para análise estatística dos resultados, foi aplicado o teste *Wilcoxon Signed Ranks* para a métrica Acurácia. A hipótese nula (H_0) considera que não existe diferença entre as médias dos métodos SCS e HCS-I na classificação das notícias. A Tabela 2 evidencia que a H_0 foi rejeitada para as versões *text* e *title* do SCS, onde p-value < 0.05. Tais fatos são as primeiras evidências de superioridade do SCS em relação ao HCS-I, em sintonia com a hipótese levantada no início do trabalho de que, ao ajustar um método de detecção de fake news baseado em crowd signals para mitigar o problema do *cold-start*, tal método poderia melhorar os resultados produzidos por sua versão original. Tal ajuste foi feito buscando melhorar o cálculo da reputação dos usuários sem histórico de divulgação de notícias no passado. Para tanto, buscou-se inferir qual seria o comportamento desses diante de notícias divulgadas no passado. Após o ajuste, o percentual de *cold-start* caiu de 57, 67% para 0, 0%.

Tabela 2: Resultados do Teste de hipótese *Wilcoxon SR*.

	SCS	HCS-I	P-Value	H_0
SCS_{text}	0.980	0.957	**0.017**	**Rejeitada**
SCS_{title}	0.980	0.957	**0.027**	**Rejeitada**

6 CONSIDERAÇÕES FINAIS

Este trabalho teve como principal contribuição obter os primeiros resultados experimentais que apontam para a validade da seguinte hipótese levantada: os desempenhos de métodos de detecção de *fake news* via *crowd signals* podem ser melhorados se tais métodos mitigarem o problema de *cold-start* por meio da inferência das reputações dos usuários baseada no comportamento que esses usuários teriam diante de notícias passadas. Tais resultados foram obtidos em um experimento preliminar realizado em um *dataset* com notícias escritas em pt-br, com o método SCS, outra contribuição desta pesquisa. Como trabalhos futuros, considera-se a realização de novos experimentos do método SCS com outros modelos LLM e com outros *datasets*, com número maior de notícias e, também, escritos em inglês. Adicionalmente, espera-se avaliar o uso de opiniões explícitas de máquinas (i.e. modelos de aprendizado de máquina) incorporados aos *crowds* e de cálculos de reputação de usuários por temática de notícias, visando melhorar o desempenho do SCS.

REFERÊNCIAS

[1] Majed Alkhamees, Saleh Alsaleem, Muhammad Al-Qurishi, Majed Al-Rubaian, and Amir Hussain. 2021. User trustworthiness in online social networks: A systematic review. *Applied Soft Computing* 103 (2021), 107–159.

[2] David Camacho, M Victoria Luzón, and Erik Cambria. 2021. New research methods algorithms in social network analysis. *Future Generation Computer Systems* 114 (2021), 290–293. https://doi.org/10.1016/j.future.2020.08.006

[3] Argus Antonio Barbosa Cavalcante, Paulo Márcio Souza Freire, Ronaldo Ribeiro Goldschmidt, and Claudia Marcela Justel. 2024. Early detection of fake news on virtual social networks: A time-aware approach based on crowd signals. *Expert Systems with Applications* 247 (2024), 123350. https://doi.org/10.1016/j.eswa.2024.123350

[4] Flávio Roberto Matias da Silva, Paulo Márcio Souza Freire, Marcelo Pereira de Souza, Gustavo de A. B. Plenamente, and Ronaldo Ribeiro Goldschmidt. 2020. FakeNewsSetGen: a Process to Build Datasets that Support Comparison Among Fake News Detection Methods. In *Proceedings of the Brazilian Symposium on Multimedia and the Web* (São Luís, Brazil) *(WebMedia '20)*. Association for Computing Machinery, New York, NY, USA, 241–248. https://doi.org/10.1145/3428658.3430965

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[6] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The Future of False Information Detection on Social Media: New Perspectives and Trends. *ACM Comput. Surv.* 53, 4, Article 68 (jul 2020), 36 pages. https://doi.org/10.1145/3393880

[7] Jürgen Habermas. 1982. Teoría de la Acción Comunicativa: Complementos y Estudios Previos. Madri: Cátedra. *Teorema: International Journal of Philosophy* (1982).

[8] Rohit Kumar Kaliyar and Navya Singh. 2019. Misinformation Detection on Online Social Media-A Survey. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 1–6. https://doi.org/10.1109/ICCCNT45670.2019.8944587

[9] Norbert Schwarz and Madeline Jalbert. 2019. *When (Fake) News Feels True: Intuitions of Truth and the Acceptance and Correction of Misinformation.*

[10] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. DEFEND: Explainable Fake News Detection. (2019), 395–405. https://doi.org/10.1145/3292500.3330935

[11] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[12] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

[13] Paulo Márcio Souza Freire, Flávio Roberto Matias da Silva, and Ronaldo Ribeiro Goldschmidt. 2021. Fake news detection based on explicit and implicit signals of a hybrid crowd: An approach inspired in meta-learning. *Expert Systems with Applications* 183 (2021), 115414. https://doi.org/10.1016/j.eswa.2021.115414

[14] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. (2018), 517–524. https://doi.org/10.1145/3184558.3188722

[15] Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. 2017. Rumor Gauge: Predicting the Veracity of Rumors on Twitter. *ACM Trans. Knowl. Discov. Data* 11, 4, Article 50 (jul 2017), 36 pages. https://doi.org/10.1145/3070644