Combating *Fake News* on Social Media by Implicit Crowd Signals

Paulo Márcio Souza Freire[1], **Ronaldo Ribeiro Goldschmidt**[1]

[1]Instituto Militar de Engenharia (IME)
Rio de Janeiro – RJ – Brasil

Abstract. *Fake News dissemination is an acknowledged problem on social media. One of the main approaches to automatically detect this type of news is based on reputation, especially the one use crowd signals. Although promising, this approach depend on information that is not always available: the explicit opinion of the users about the news concerning whether they are fake or not. To overcome this drawback, this article proposes a implicit crowd signal-based method that does not demand the users' explicit opinion. Experiments provided evidence that the proposed method can detect Fake News without demanding the explicit opinion of the users and without compromising the results obtained by the state-of-the-art crowd signal-based method.*

Resumo. *A disseminação do Fake News é um problema conhecido nas redes sociais. Uma das principais abordagens para detectar, automaticamente, este tipo de notícia é baseada na reputação, em especial a que utiliza Crowd Signals. Embora promissora, esta abordagem depende de informações nem sempre disponíveis: a opinião explícita dos usuários sobre as notícias serem fake ou não. Para superar esta desvantagem, este artigo propõe um método, baseado em Crowd Signals Implícitos, que não exige a opinião explícita dos usuários. Experimentos forneceram evidências de que o método proposto pode detectar Fake News sem exigir a opinião explícita dos usuários e sem comprometer os resultados obtidos pelo estado da arte dos métodos baseados em Crowd Signals.*

1. Introdução

Com o surgimento de redes sociais virtuais de fácil acesso e baixo custo, ou simplesmente redes sociais, a disseminação de notícias não está mais restrita às mídias tradicionais, tais como: rádio, televisão, revistas e jornais impressos. Assim, a cada dia, as pessoas vêm aumentando o consumo de notícias on-line [Vosoughi et al. 2017].

Apesar dos benefícios advindos desta acessibilidade, a rede social permite que qualquer um, independentemente de sua reputação, divulgue (publique/propague) notícias com intenso poder de espalhamento. Desta forma, as redes sociais amplificaram um problema antigo: a disseminação de *Fake News*. O termo *Fake News* corresponde às notícias falsas publicadas de forma intencional [Shu et al. 2017a].

A recente proliferação de *Fake News*, nas redes sociais, tem sido uma fonte de preocupação generalizada. Essa apreensão deve-se ao poder da influência das *Fake News* na sociedade. Um exemplo é a análise feita pelo *Buzzfeed*[1] onde, a partir das 20 principais notícias falsas sobre as eleições americanas de 2016, criadas por sites fraudulentos,

[1]Buzzfeed - https://https://www.buzzfeed.com

foram geradas quase 1,5 milhões de atividades de engajamento de usuários no *Facebook* [Farajtabar et al. 2017].

Nos últimos anos, tanto a academia quanto a indústria estudam como combater *Fake News* nas redes sociais. Este combate apresenta-se como não trivial, tanto pelo volume de publicações quanto pela velocidade das suas respectivas propagações. Assim, o emprego de abordagens computacionais, devido à sua maior velocidade de atuação, vem se destacando neste combate [Ruchansky et al. 2017].

Diversos estudos propuseram abordagens para combater *Fake News* em redes sociais. Visando facilitar uma classificação destas diferentes abordagens, este artigo propõe e aplica um modelo comparativo. A partir desta classificação, é possível identificar a existência de abordagens, baseadas na reputação do usuário, que apresentam duas características que as tornam alternativas interessantes. A primeira é a não obrigatoriedade na utilização do conteúdo da notícia, pois a atual similaridade entre as notícias *fake* e não *fake* dificulta a sua distinção [Liu and BrookWu 2018]. A segunda é a não necessidade do uso dos dados relativos ao perfil do usuário na rede social, haja vista a dificuldade em se obter tais informações, atualmente consideradas sigilosas [Shu et al. 2017b].

Dentre as soluções fundamentadas na reputação que não utilizam dados do conteúdo da notícia e nem do perfil do usuário, se destacam aquelas que usam a abordagem baseada em *Crowd Signals* [Tschiatschek et al. 2018] [Sharma et al. 2019]. Nesta abordagem, os usuários da rede social devem fornecer suas opiniões explícitas, informando se as notícias acessadas são *fake* ou não. Uma opinião explícita é um rótulo sinalizado pelo usuário, sobre uma determinada notícia, através de uma funcionalidade específica da rede social. Esta abordagem classifica uma nova notícia a como *fake* ou não, combinando a opinião explícita dos usuários sobre a com as suas respectivas reputações. Tal reputação é obtida a partir da capacidade do usuário em acertar ou errar o rótulo correto das notícias anteriormente recebidas por ele. Inclusive, *Crowd Signals* envolve trabalho colaborativo que já produziu resultados robustos em outras áreas, como segurança na *Web* [Chia and Knapskog 2012], avaliação de sites de *phishing* [Moore and Clayton 2008] e checagem de fatos [Kim et al. 2018] [Sethi 2017].

Embora promissora, a abordagem baseada em *Crowd Signals* depende da opinião explícita dos usuários sobre as notícias. Esta dependência tem uma desvantagem significativa: a opinião explícita dos usuários sobre as notícias nem sempre está disponível. Duas razões principais podem causar essa privação de opinião do usuário. A primeira razão é que a maioria das redes sociais não fornece uma funcionalidade para coletar a opinião do usuário sobre as notícias. A segunda e mais importante é que, mesmo quando esta funcionalidade está disponível, o usuário não pode ser forçado a indicar sua opinião sobre cada notícia. Na verdade, essa abordagem depende da boa vontade dos usuários em opinar sobre as notícias que chegam até eles através da rede social.

Portanto, nesta pesquisa, nossa pergunta é: *Dada uma rede social, é possível detectar Fake News via Crowd Signals Implícitos, não exigindo a opinião explícita dos usuários sobre as notícias, assim como não comprometendo os resultados da classificação obtidos pelo estado da arte dos métodos baseados em Crowd Signals*?

A fim de responder positivamente à questão supracitada, neste artigo, propomos um método que considera *Crowd Signals Implícitos* para detectar *Fake News* nas redes

sociais. Portanto, o método proposto infere as opiniões dos usuários a partir de seu comportamento de divulgação. Testes com dois *datasets* forneceram evidências experimentais de que o método proposto pode detectar *Fake News* de maneira tão eficiente quanto pelo estado da arte dos métodos baseados em *Crowd Signals* que precisam das opiniões explícitas dos usuários.

O restante do artigo está organizado da seguinte forma. A seção 2 apresenta e aplica um modelo comparativo entre abordagens relacionadas ao combate automatico às *Fake News* nas redes sociais, dando destaque à abordagem baseada em *Crowd Signals*. A seção 3, por sua vez, descreve o método proposto. Posteriormente, detalhes sobre os experimentos e os resultados obtidos são apresentados na Seção 4. Finalmente, a seção 5 conclui o artigo, destacando as principais contribuições do trabalho e indicando perspectivas para futuras investigações.

2. Trabalhos Relacionados

Para classificar os trabalhos vinculados ao combate automático às *Fake News* nas redes sociais é proposto e, em seguida, aplicado um modelo comparativo que viabilize uma distinção entre abordagens computacionais.

2.1. Proposta de Modelo Comparativo

O combate às *Fake News* em redes sociais, por meio de abordagens computacionais, possui uma variedade de aspectos que podem ser considerados. Com o objetivo de facilitar a comparação e a consequente classificação das referidas abordagens, tais aspectos são categorizados na Figura 1. As próximas subseções detalham cada um destes aspectos.

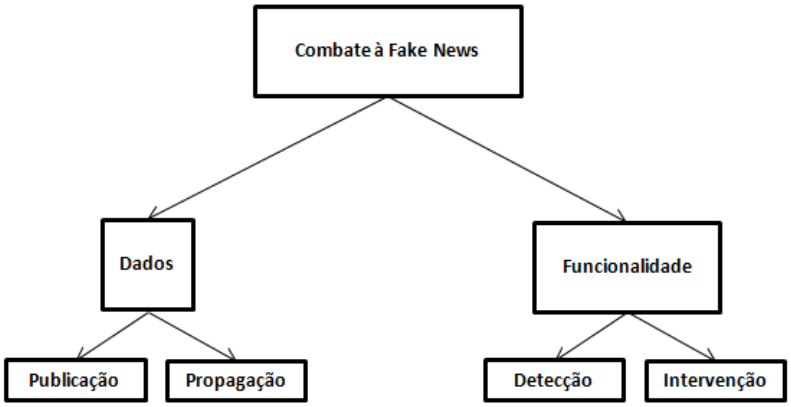


Figura 1. Aspectos considerados em Abordagens

2.1.1. Dados

Aspecto relacionado aos dados que podem ser utilizados pelas abordagens computacionais de combate às *Fake News*. Este aspecto subdivide-se em dados obtidos a partir da *Publicação* da notícia, como também aqueles associados com a sua *Propagação*.

Os dados de *Publicação* representam as informações inerentes ao surgimento da notícia na rede social. Estes dados podem ser classificados em *Notícia, Usuário, Assunto*

e Temporalidade. No que diz respeito à Notícia, a abordagem pode ser capaz de analisar dados oriundos da publicação a partir de diferentes tipos de *Mídia* (*Texto, Áudio e Imagem*). Independente da *Mídia*, a análise do *Conteúdo* pode ser realizada de forma *Léxica, Sintática, Semântica e Legibilidade*. Com relação ao *Usuário* publicador, a abordagem pode identificar diferentes *Tipos*, tais como: humano, *bot* ou *cyborg*. Pode-se analisar também dados referentes ao *Perfil* do usuário na rede social, tais como: identificação e idade. Outro aspecto relevante está relacionado à *Reputação* do publicador, que pode estar vinculada à sua capacidade em identificar ou publicar *Fake News*. A abordagem pode também utilizar o *Assunto* abordado no momento da publicação. Assim, é possível tratar Especificidades, tais como: relacionamento entre assuntos ou assuntos controversos. Outro aspecto leva em consideração a *Relevância* do assunto publicado, haja vista que assuntos em voga motivam a criação de *Fake News*. A variação das características de uma notícia com o passar do tempo, torna a *Temporalidade* mais um relevante recurso para a identificação de *Fake News*.

Os dados de *Propagação* representam as informações obtidas após a publicação, consequentemente, aquelas inerentes ao espalhamento da notícia na rede social (ex: curtida/*like*, comentário/*reply* ou compartilhar/*retweet*). Portanto, estes dados podem ser classificados em *Contribuição, Usuário, Assunto, Temporalidade e Rede*. No que diz respeito à *Contribuição*, *Usuário*, *Assunto* e *Temporalidade* a abordagem pode ser capaz de analisar os dados oriundos da propagação, a partir dos mesmos aspectos anteriormente citados na *Publicação*. Ademais, as informações relacionadas à *Rede* criada, a partir da propagação da notícia, possibilitam não só a identificação de uma *Fake News* como uma possível atuação contra a mesma.

2.1.2. Funcionalidade

Além dos dados coletados, as abordagens automáticas de combate às *Fake News* podem, basicamente, possuir duas funcionalidades: *Detecção e Intervenção*. O objetivo da detecção é identificar uma notícia divulgada na rede social como sendo intencionalmente falsa. Após a sua detecção, a intervenção é a atuação que será empregada contra a *Fake News* visando seu bloqueio ou sua mitigação.

A *Detecção* automática de *Fake News* pode ser interpretada como um problema de classificação binária onde dada uma rede social $\mathcal{G}$, uma notícia a e um conjunto de postagens (publicações/propagações) $\mathcal{P}$, relacionadas com a, são espalhadas através da $\mathcal{G}$ por um conjunto de usuários U em um intervalo de tempo t. Assim, o referido classificador binário $\mathcal{F}$ deve aprender, a partir dos dados, a prever se a é uma *fake news* ou não, como formalmente indicado na Equação 1.

$$\mathcal{F}(\mathcal{G}, a, \mathcal{P}, U, t) = \begin{cases} 1, & \text{se } a \text{ é uma } \textit{fake news}; \\ 0, & \text{caso contrário.} \end{cases} \tag{1}$$

A *Intervenção* automática pode se subdividir em proativa ou reativa. A abordagem reativa busca combater os efeitos da *Fake News* a partir do momento da sua detecção. De outra forma, a abordagem proativa pode atuar antes mesmo da referida detecção, tendo então um comportamento preventivo.

2.2. Revisão dos Trabalhos Relacionados

Diversos estudos propuseram abordagens para combater *Fake News* em redes sociais. Visando facilitar uma classificação destas diferentes abordagens, este artigo propõe e aplica o modelo comparativo tratado na SubSeção 2.1.1, conforme mostram as Tabelas 1 e 2. Cabe ressaltar que, nestas duas Tabelas, as células não preenchidas indicam a não utilização do respectivo aspecto no trabalho correspondente.

Tabela 1. Comparação entre abordagens - Dados de Publicação

Id	Dados							
	Publicação							
	Notícia		Usuário			Assunto		Temporalidade
	Mídia (Texto, Áudio e Imagem)	Conteúdo (Léxica, Sintática, Semântica e Legibilidade)	Tipo (Humano, Bot e Cyborg)	Perfil	Reputação	Especificidades	Relevância	
[Pérez-Rosas et al. 2018]	Texto	Léxica, Sintática, Semântica e Legibilidade						
[Janze and Risius 2017]	Texto e Imagem	Léxica						
[Wang 2017]	Texto	Léxica e Semântica		X		Relaciona Assuntos		
[Farajtabar et al. 2017]								
[Ruchansky et al. 2017]	Texto	Léxica e Semântica		X	X			X
[Woloszyn and Nejdl 2018]					X	Assuntos controversos		
[Zhang et al. 2018]	Texto	Semântica						
[Bhatt et al. 2018]	Texto	Semântica						
[Tschiatschek et al. 2018]					X			X
[Wu and Liu 2018]					X			
[Nasim et al. 2018]			Bot	X				X
[Buntain and Golbeck 2017]	Texto	Léxica e Semântica		X				X
[Gilda 2017]	Texto	Léxica e Semântica						
[Shu et al. 2017b]	Texto	Léxica e Semântica		X	X			
[Rubin et al. 2015]	Texto	Semântica						
[Liu and BrookWu 2018]				X				X
[Qian et al. 2018]	Texto	Semântica						
[Shu et al. 2019]	Texto	Léxica e Semântica		X				X

Tabela 2. Comparação entre abordagens - Dados de Propagação

Id	Dados								
	Propagação								
	Contribuição		Usuário			Assunto		Temporalidade	Rede
	Mídia (Texto, Áudio e Imagem)	Conteúdo (Léxica, Sintática, Semântica e Legibilidade)	Tipo (Humano, Bot e Cyborg)	Perfil	Reputação	Especificidades	Relevância		
[Pérez-Rosas et al. 2018]									
[Janze and Risius 2017]	Texto	Léxica							X
[Wang 2017]									
[Farajtabar et al. 2017]									X
[Ruchansky et al. 2017]	Texto	Léxica e Semântica		X	X			X	X
[Woloszyn and Nejdl 2018]									
[Zhang et al. 2018]									
[Bhatt et al. 2018]									
[Tschiatschek et al. 2018]					X			X	X
[Wu and Liu 2018]					X				X
[Nasim et al. 2018]			Bot	X				X	X
[Buntain and Golbeck 2017]	Texto	Léxica e Semântica		X				X	X
[Gilda 2017]									
[Shu et al. 2017b]				X	X				X
[Rubin et al. 2015]									
[Liu and BrookWu 2018]				X				X	X
[Qian et al. 2018]	Texto	Semântica							
[Shu et al. 2019]				X				X	X

A partir da classificação apresentada pelas Tabelas 1 e 2, é possível identificar a existência de abordagens, baseadas na reputação do usuário, que apresentam duas características que as tornam alternativas interessantes. A primeira é não ser obrigatória a utilização do conteúdo da notícia, pois a atual similaridade entre as notícias *fake* e não *fake* dificulta esta forma de detecção [Liu and BrookWu 2018]. A segunda é a não necessidade do uso dos dados relativos ao perfil do usuário na rede social, haja vista a dificuldade em se obter tais informações sigilosas [Shu et al. 2017b].

Dentre as abordagens computacionais que utilizam a reputação dos usuários sem a utilização dos dados do conteúdo da notícia e do perfil do usuário, uma das mais promissoras é a abordagem cuja reputação é obtida a partir da opinião do usuário sobre a notícia [Tschiatschek et al. 2018]. Em resumo, ela usa a reputação dos usuários em acertar ou errar o rótulo correto (classe) das notícias analisadas por eles. Esta relação de acerto e erro é usada como *Crowd Signals* para detectar se a notícia é ou não *fake*.

2.3. Revisão dos Trabalhos Relacionados Baseados em *Crowd Signals*

Até onde pudemos constatar, o [Tschiatschek et al. 2018] é o principal estudo que seguiu a abordagem baseada em *Crowd Signals* para detectar *Fake News*. Ele propôs um método baseado em *Crowd Signals* chamado *Detective*. Em essência, este método utiliza um classificador bayesiano binário, cujos fundamentos são descritos abaixo.

Dada uma rede social $\mathcal{G}$, a entrada do *Detective* contém os seguintes elementos: um intervalo de tempo t (chamado época, por exemplo, um dia), um conjunto de usuários U de $\mathcal{G}$, um *dataset* D com notícias previamente rotuladas e uma notícia específica a ser analisada a. D contém notícias com dois tipos de rótulos: o real (correto) e o sinalizado pelo usuário (opinião). O atributo que contém o rótulo real para qualquer notícia x é representado pela variável $Y^*(x)$ e seu valor dado por $y^*(x)$, pertencente a $\{f, \overline{f}\}$, no qual $y^*(x) = f$ (resp. $y^*(x) = \overline{f}$) significa que a notícia x é fake (resp. não fake). Denotado por uma variável $Y_u(x)$, o rótulo sinalizado é aquele atribuído por um usuário u para uma notícia x. Seu valor $y_u(x)$ pertence a $\{f, \overline{f}\}$ onde $y_u(x) = f$ (resp. $y_u(x) = \overline{f}$) significa que u sinalizou x como *fake* (resp. não *fake*). É importante notar que, diferente de outras notícias, $y^*(a)$ é desconhecido e deve ser previsto pelo *Detective*.

Inicialmente, *Detective* aplica as funções $\pi^t(a)$ e $\psi^t(a)$ ao D. Enquanto a primeira função retorna o conjunto de usuários que viram a notícia *a* até o final da época *t*, a outra retorna o conjunto completo de usuários que sinalizaram *a* como *fake* no final de *t*.

O *Detective* pode assumir que não há abstinência na sinalização e, para cada usuário $u \in \pi^t(a)$, calcula $\theta_{u,\overline{f}}$ e $\theta_{u,f}$, considerando as notícias sinalizadas por u antes de t. Assim, $\theta_{u,\overline{f}}$ (resp. $\theta_{u,f}$) é a probabilidade de *u* sinalizar uma notícia *x* como não *fake* (resp. *fake*), dado que *x* é realmente não *fake* (resp. fake). Em ambos os casos, o cálculo da probabilidade é limitado ao conjunto de notícias revisadas por u antes de t. Portanto, para cada usuário u, o *Detective* sumariza o histórico de atividades de sinalização de *u* pela correspondente matriz $\mathcal{M}_u$, genericamente definida como segue.

$$\begin{vmatrix} \theta_{u,\overline{f}} & 1 - \theta_{u,f} \\ 1 - \theta_{u,\overline{f}} & \theta_{u,f} \end{vmatrix}$$

onde:

- $\theta_{u,\overline{f}} = P(Y_u(x) = \overline{f} \mid Y^*(x) = \overline{f})$
- 1 - $\theta_{u,\overline{f}} = P(Y_u(x) = f \mid Y^*(x) = \overline{f})$
- $\theta_{u,f} = P(Y_u(x) = f \mid Y^*(x) = f)$
- 1 - $\theta_{u,f} = P(Y_u(x) = \overline{f} \mid Y^*(x) = f)$

Por fim, seguindo uma abordagem bayesiana, o *Detective* usa as Equações 2 e 3 para calcular as probabilidades de *a* ser *fake* e não *fake*, respectivamente. Sendo ω (resp. $1 - \omega$) a probabilidade a priori de que qualquer notícia seja *fake* (resp. não *fake*). Ambas as Equações consideram a capacidade dos usuários de acertar, assim como, de errar suas opiniões de acordo com seu voto. Portanto, o *Detective* pode se beneficiar quando os usuários acertarem ou errarem, mesmo que eles mostrem incapacidade [Freeman 2017] ou má intenção ao avaliar as notícias. A classe correspondente à maior probabilidade é a opinião do *Detective* sobre a e, portanto, sua saída.

$$P(Y^*(a) = f) = \omega \cdot \prod_{u \in \psi^t(a)} \theta_{u,f} \cdot \prod_{u \in \pi^t(a) \setminus \psi^t(a)} (1 - \theta_{u,f}) \tag{2}$$

$$P(Y^*(a) = \overline{f}) = (1 - \omega) \cdot \prod_{u \in \psi^t(a)} (1 - \theta_{u,\overline{f}}) \cdot \prod_{u \in \pi^t(a) \setminus \psi^t(a)} \theta_{u,\overline{f}} \tag{3}$$

Apesar dos resultados promissores do *Detective*, relatados em [Tschiatschek et al. 2018], este método exige a opinião explícita dos usuários sobre as notícias, o que é um requisito limitante, uma vez que essas informações nem sempre estão disponíveis.

3. Método Proposto

Chamado *ICS* (Implict Crowd Signals), o método proposto neste artigo é uma adaptação do *Detective*, o método desenvolvido pelo trabalho que representa o estado da arte das abordagens baseadas em *Crowd Signals*, descrito na subseção 2.3. Ambos são classificadores bayesianos para a detecção de *Fake News* nas redes sociais e usam a opinião dos usuários sobre notícias passadas. No entanto, o *ICS* difere do *Detective* na forma como interpreta o *Crowd Signals*. Enquanto o segundo exige a opinião explícita dos usuários sobre os rótulos (*fake* ou não *fake*) das notícias, nossa abordagem infere opiniões a partir dos padrões de comportamento dos usuários ao divulgar notícias. Portanto, o *ICS* não precisa de todos os dados do conjunto de entrada D do *Detective*, sendo substituído por um conjunto de dados D', que possui quase os mesmos dados de D, exceto pelos rótulos das notícias sinalizados pelos usuários (ou seja, D' não contém as opiniões dos usuários sobre as notícias).

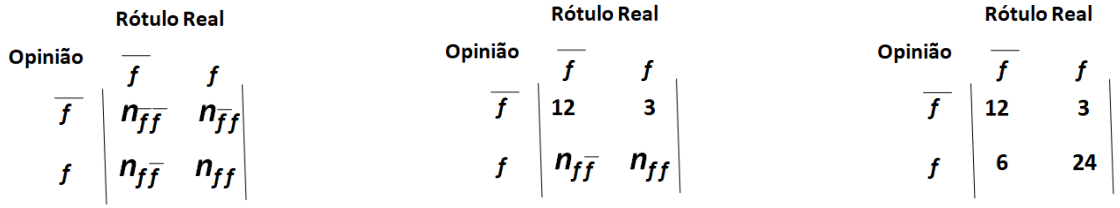
Inicialmente, no estágio de treinamento, para cada usuário u, *ICS* deve inferir as probabilidades $\theta_{u,f}$ e $\theta_{u,\overline{f}}$. Para isso, utiliza uma matriz de opinião $\mathcal{O}_u$ genericamente definida na Figura 2a. Cada componente n_{rs} em $\mathcal{O}_u$ ($r, s \in \{f, \overline{f}\}$) indica a quantidade de notícias sinalizadas como r por u, dado que os rótulos reais dessas notícias são s. O *ICS* também usa n_f e $n_{\overline{f}}$ para preencher $\mathcal{O}_u$. Eles representam o total de notícias *fake* e não *fake* em D', respectivamente.

O método proposto considera as notícias anteriormente divulgadas por u (registradas em D') para preencher a primeira linha de $\mathcal{O}_u$. Uma vez que u decidiu divulgá-las, o *ICS* assume este ato como um sinal implícito de que u considera estas notícias como não *fake*. Observe que os rótulos reais dessas notícias são conhecidos e estão disponíveis em D'. A Figura 2b apresenta um exemplo parcial em que u divulga 15 notícias, sendo 3 *fake* e 12 não *fake*.

Para preencher a segunda linha de $\mathcal{O}_u$, também seria necessário recuperar a quantidade de notícias que u visualizou, mas decidiu não propagar considerando-as *fake*. No entanto, esta informação não está disponível. Esta indisponibilidade acontece, pois não é suficiente saber quais notícias chegaram a u, haja vista que não há garantia de que u tenha visto e decidido não propagá-las. Assim, o *ICS* deve inferir tal informação com base na capacidade de u em identificar as notícias não *fake*. Para este fim, usa dois critérios:

- Primeiramente, ele deve preservar a capacidade do usuário de acertar ou errar suas suposições, isto é, $n_{ff}/(n_{f\overline{f}}+n_{ff})$ deve ser aproximadamente igual a $n_{\overline{f}\overline{f}}/(n_{\overline{f}\overline{f}}+n_{\overline{f}f})$, um valor conhecido.
- Em seguida, ele deve comutar o número relativo de exemplos sinalizados nas duas classes, preservando a proporcionalidade apresentada na primeira linha de $\mathcal{O}_u$, ou seja, $n_{f\overline{f}}$ deve ser dado por $(n_{\overline{f}f}/n_f) \times n_{\overline{f}}$.

A Figura 2c ilustra esta inferência para o exemplo apresentado na Figura 2b, considerando $n_f = 30$ e $n_{\overline{f}} = 60$.



(a) Definição Genérica **(b)** Exemplo Parcial **(c)** Exemplo Completo

Figura 2. Matriz de Opiniâo $\mathcal{O}_u$

Com base na versão completa de $\mathcal{O}_u$, o *ICS* infere as probabilidades $\theta_{u,f}$ e $\theta_{u,\overline{f}}$ conforme indicado abaixo:

- $\theta_{u,f} = n_{ff}/(n_{\overline{f}f} + n_{ff})$
- $\theta_{u,\overline{f}} = n_{\overline{f}\overline{f}}/(n_{\overline{f}\overline{f}} + n_{f\overline{f}})$

Similar ao *Detective*, o cálculo da probabilidade é limitado ao conjunto de notícias revisadas por u antes de uma determinada época t. Portanto, com estas probabilidades, o *ICS* é capaz de representar a atividade de sinalização implícita de *u*, observada pela correspondente matriz $\mathcal{M}_u$.

Por fim, como *Detective*, dada uma notícia a a ser analisada, o *ICS* usa a regra bayesiana para concluir se a é *fake* ou não. No entanto, diferente do *Detective*, o *ICS* não sabe quais usuários consideraram a como *fake*, uma vez que D' não contém tais informações. Portanto, as funções $\pi^t(a)$ e $\psi^t(a)$ não podem ser aplicadas a D'. Desta forma, as Equações 2 e 3 tiveram que ser adaptadas como indicado nas Equações 4 e 5

para lidar com essa lacuna de informação. O primeiro ajuste é que o *ICS* considera uma nova função $\chi^t(a)$ que recupera de D' os usuários que divulgaram a. Neste ponto, o *ICS* assume que a decisão do u divulgar a é um sinal implícito de que a não é *fake*. Assim, $\chi^t(a)$ recupera os usuários que não consideram a como *fake*. O segundo ajuste é que os fatores $\prod_{u \in \psi^t(a)} \theta_{u,f}$ e $\prod_{u \in \psi^t(a)} (1 - \theta_{u,\overline{f}})$ das Equações 2 e 3 foram substituídos por 1 nas Equações 4 e 5. Estas substituições refletem a visão otimista do *ICS* em relação aos usuários desconhecidos que podem considerar a como *fake*. De acordo com essa visão, o *ICS* assume que esses usuários têm uma probabilidade máxima: para acertar notícias opinando como *fake* quando a é, na verdade, *fake* ($\prod_{u \in \psi^t(a)} \theta_{u,f} = 1$); errar notícias opinando como *fake* quando a é, na verdade, não *fake* ($\prod_{u \in \psi^t(a)} (1 - \theta_{u,\overline{f}}) = 1$). Em ambas as Equações, as substituições levam às maiores probabilidades posteriores.

$$P(Y^*(a) = f) = \omega \cdot \prod_{u \in \chi^t(a)} (1 - \theta_{u,f}) \tag{4}$$

$$P(Y^*(a) = \overline{f}) = (1 - \omega) \cdot \prod_{u \in \chi^t(a)} \theta_{u,\overline{f}} \tag{5}$$

4. Experimentos e Resultados

Apesar da relevância do problema de detecção *Fake News* nas redes sociais, os *datasets* que contêm dados reais, neste cenário, ainda estão raramente disponíveis para download. Como conseqüência, a maioria das pesquisas relacionadas à detecção de *Fake News* adaptou *datasets* originalmente criados para investigar outros problemas em redes sociais, como divulgação de *Rumor* [2] [Ruchansky et al. 2017]. Esses datasets adaptados, geralmente, não contêm informações importantes para a detecção de *Fake News*, como rótulos *fake* / não *fake*. Além disso, a maioria desses *datasets* (adaptados ou originalmente criados para detecção de *Fake News*) não descrevem a propagação das notícias nas redes sociais, como uma mesma notícia divulgada por vários usuários e várias notícias divulgadas por um mesmo usuário. Assim, não há um consenso sobre os *datasets* de referência para este problema [Shu et al. 2017a]. Desta forma, escolhemos os dois *datasets* pertencentes ao repositório FakeNewsNet [Shu et al. 2017a], cujos dados das notícias foram obtidos a partir do Twitter. Tanto o primeiro *dataset* quanto o segundo possuem o nome do site utilizado como fonte de consulta para rotular as notícias, respectivamente o *BuzzFeed e o PolitiFact* [3]. Nossa escolha foi guiada por três razões principais. Primeiro, esses *datasets* foram criados para o específico propósito de detecção de *Fake News* e contêm, para cada notícia, seu rótulo real, ou seja, a indicação de que a notícia é *fake* ou não. Em segundo lugar, eles descrevem a propagação das notícias nas redes sociais. Por fim, eles foram usados e disponibilizados por publicações recentes e relevantes [Shu et al. 2017a] [Shu et al. 2017b] [Shu et al. 2019] [Sharma et al. 2019]. A Tabela 3 fornece uma visão estatística geral dos *datasets* escolhidos.

Assim, cada *dataset* foi dividido em dois subconjuntos de notícias não sobrepostos: treinamento e teste, com uma proporção de dados de, respectivamente, 70% e 30%.

[2] Diferente de *Fake News*, um *Rumor* é uma informação não verificada (verdadeira ou falsa) [Vosoughi et al. 2017].

[3] https://www.politifact.com/

Tabela 3. *Datasets* usados nos Experimentos

Dataset	Não *Fake News*	*Fake News*	Usuários	Média de usuários por notícia
BuzzFeed	91	91	15257	125,16
PolitiFact	120	120	23865	136,63

Foram avaliados e comparados os dois métodos de detecção de *Fake News*: o *ICS* e o *Detective*. Como apresentado anteriormente, enquanto o segundo exige as opiniões explícitas dos usuários sobre as notícias, o primeiro infere as opiniões dos usuários com base no comportamento histórico de divulgações destes usuários.

Para o método *Detective*, a metodologia experimental foi semelhante à metodologia ótima seguida por [Tschiatschek et al. 2018]. As probabilidades θ foram aleatoriamente designadas aos usuários, criando três grupos de usuários: *bom* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0.9$), *indiferente* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0.5$) e *spammer* ($\theta_{u,\overline{f}} = \theta_{u,f} = 0,1$). Assim como [Tschiatschek et al. 2018], foi assumido que nenhum usuário se absteve de dar sua opinião sobre as notícias a serem analisadas. Embora não esteja claramente indicado em [Tschiatschek et al. 2018], também foi assumido que cada usuário deveria sinalizar aleatoriamente uma notícia de acordo com a probabilidade atribuída ao seu grupo. Por exemplo: dada uma notícia a para ser analisada por u, um *bom* usuário. De acordo com a configuração definida para *bons* usuários, u deve acertar ou errar para a com probabilidades de 90% e 10%, respectivamente. Além disso, foi usado o método da roleta para decidir se cada usuário deve acertar ou errar o rótulo real de uma notícia. Embora pouco realista, esta metodologia, ao ser executada, faz com que o *Detective* maximize a precisão das suas classificações.

Conforme descrito na seção 3, o método proposto *ICS* não requer configuração manual. Assim, as probabilidades θ foram calculadas, automaticamente, a partir do comportamento histórico de divulgações dos usuários armazenados em D' (isto é, o *dataset* de treinamento). Para comparar os métodos de detecção de *Fake News*, foi utilizado o *holdout* como critério de avaliação e a acurácia como métrica de desempenho. A Tabela 4 resume os resultados dos experimentos. A principal constatação, a partir desses resultados, é que as diferenças de precisão entre os dois métodos, em ambos os conjuntos de dados, ocorrem na segunda casa decimal. Estes valores aproximados indicam que o *ICS* produziu resultados comparáveis aos produzidos pelo *Detective*, ressaltando que, diferentemente do *Detective*, o *ICS* foi submetido à uma metodologia experimental mais realista, assim como, dispensa as opiniões explícitas dos usuários. Em resumo, os resultados obtidos fornecem evidências experimentais de que *Crowd Signals* implícitos podem ser usados para detectar *Fake News* em redes sociais, em vez da opinião explícita.

Tabela 4. Acurácia dos métodos de detecção de *Fake News*

Método	BuzzFeed	PolitiFact
Detective	**0.9835**	**0.9791**
ICS	0.9333	0.9402

5. Conclusão

Cada vez mais pessoas estão consumindo notícias das redes sociais, ao invés dos canais tradicionais. Tal tendência amplificou a disseminação de *Fake News*, isto é, as notícias intencionalmente falsas. Este tipo de notícia pode ter significativos impactos sociais negativos, por exemplo, a manipulação da opinião em larga escala. Uma das principais abordagens para detectar, automaticamente, as *Fake News* é baseada em *Crowd Signals*, ou seja, opiniões manifestadas por usuários da rede social sobre uma notícia ser *fake* ou não. Embora promissora, esta abordagem tem uma desvantagem importante: depende da, nem sempre disponível, opinião explícita dos usuários sobre as notícias. Para superar esta dificuldade, o presente artigo propôs um método baseado em *Crowd Signals Implícitos* (*ICS*) que não exige a opinião explícita dos usuários para detectar as *Fake News*. De fato, o método proposto infere as opiniões dos usuários a partir do seu comportamento histórico de divulgações. Testes forneceram evidências experimentais de que o método proposto é comparável ao estado da arte dos métodos baseados em *Crowd Signals* explícitos. Nossas iniciativas para trabalhos futuros incluem experimentos com outros *datasets*, assim como, visando melhorar os resultados obtidos pelo *ICS*, a investigação de outras formas de inferir a opinião implícita dos usuários.

Referências

Bhatt, G., Sharma, A., Sharma, S., Nagpal, A., Raman, B., and Mittal, A. (2018). Combining neural, statistical and external features for fake news stance identification. In Comp. Proc. of the The Web Con. Int WWW Con Steering Committee.

Buntain, C. and Golbeck, J. (2017). Automatically identifying fake news in popular twitter threads. In IEEE Int Con on Smart Cloud.

Chia, P. H. and Knapskog, S. J. (2012). Re-evaluating the wisdom of crowds in assessing web security. In The 15th Int Con on Financial Cryptography and Data Security, Berlin, Heidelberg. Springer-Verlag.

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., and Zha, H. (2017). Fake news mitigation via point process based intervention. In Int Con on Machine Learning.

Freeman, D. M. (2017). Can you spot the fakes?: On the limitations of user feedback in online social networks. In Int Con on WWW.

Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. In IEEE Student Con on Research and Development.

Janze, C. and Risius, M. (2017). Automatic detection of fake news on social media platforms. In PACIS.

Kim, J., Tabibian, B., Oh, A., Schölkopf, B., and Gomez-Rodriguez, M. (2018). Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In ACM Int Con on Web Search and Data Mining.

Liu, Y. and BrookWu, Y. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In AAAI Con on Artificial Intelligence.

Moore, T. and Clayton, R. (2008). Financial cryptography and data security. chapter Evaluating the Wisdom of Crowds in Assessing Phishing Websites.

Nasim, M., Nguyen, A., Lothian, N., Cope, R., and Mitchell, L. (2018). Real-time detection of content polluters in partially observable twitter networks. In Comp. Proc. of the The Web Con.

Pérez-Rosas, V., B. Kleinberg, A. L., and Mihalcea, R. (2018). Automatic detection of fake news. In Int Con on Comput. Linguistics.

Qian, F., Gong, C., Sharma, K., and Liu, Y. (2018). Neural user response generator: Fake news detection with collective user intelligence. In Int Joint Con on Artificial Intelligence.

Rubin, V. L., Conroy, N. J., and Chen, Y. (2015). Towards news verification: Deception detection methods for news discourse.

Ruchansky, N., Seo, S., and Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. In ACM on Con on Inform. and Knowledge Management.

Sethi, R. J. (2017). Crowdsourcing the verification of fake news and alternative facts. In ACM Con on Hypertext and Social Media.

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 Trans. Intell. Syst. Technol.

Shu, K., Mahudeswaran, D., and Liu, H. (2019). Fakenewstracker:a tool for fake news collection,detection, and visualization. Com. Mat. Or. The.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017a). Fake news detection on social media: A data mining perspective. SIGKDD Ex. New.

Shu, K., Wang, S., and Liu, H. (2017b). Exploiting tri-relationship for fake news detection. In arXiv.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In Comp. Proc. of the The Web Con.

Vosoughi, S., Mohsenvand, M. N., and Roy, D. (2017). Rumor gauge: Predicting the veracity of rumors on twitter. ACM Trans. Know. Disc. Data.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In The Assoc. for Comput. Linguistics.

Woloszyn, V. and Nejdl, W. (2018). Distrustrank: Spotting false news domains. In ACM Con on Web Science.

Wu, L. and Liu, H. (2018). Tracing fake-news footprints: Characterizing social media messages by how they propagate. In ACM Int Con on Web Search and Data Mining.

Zhang, Q., Yilmaz, E., and Liang, S. (2018). Ranking-based method for news stance detection. In Comp. Proc. of the The Web Con.