Métodos de Detecção de Fake News: Uma Comparação entre as Abordagens de Crowd Signals e Ensembles

Uriel Merola[1], **Paulo M. S. Freire**[2],
Ronaldo R. Goldschmidt[2], **Jorge Soares**[1]

[1]CEFET/RJ - Rio de Janeiro – Brasil

[2]Instituto Militar de Engenharia (IME) - Rio de Janeiro – Brasil

uriel.merola@gmail.com, {paulomsfreire,ronaldo.rgold}@ime.eb.br,
jorge.soares@cefet-rj.br

Abstract. *The rise of fake news dissemination is due to the easy generation and consumption of information provided by digital media. To identify them, the approach based on hybrid crowd signals (HCS) combines signals (opinions about their truthfulness) collected from users or machine learning classifiers (ML). Although promising, the HCS approach employs a naive method (Naive Bayes) to combine the signals and infer which news articles are false. Thus, this study questions whether the use of Ensemble methods to combine opinions provided by ML classifiers used in HCS can enhance the resulting classification models. Preliminary experiments with the datasets used in HCS reveal evidence supporting the hypothesis.*

Resumo. *A crescente disseminação de fake news deve-se à facilidade de criação e consumo de informações nos meios digitais. Para identificá-las, a abordagem baseada em crowd signals híbridos (HCS) combina sinais (opiniões sobre sua veracidade) coletados de usuários ou de classificadores de aprendizado de máquina (AM). Embora promissora, a abordagem HCS emprega um método ingênuo (Naive Bayes) para combinar os sinais e inferir quais notícias são falsas. Assim, o presente trabalho questiona se o uso de métodos Ensemble para conjugar opiniões fornecidas pelos classificadores de AM usados na HCS pode aprimorar os modelos de classificação resultantes. Experimentos preliminares com os datasets usados na HCS revelam indícios de validade da hipótese.*

1. Introdução

Os meios digitais de divulgação de notícias (MDDN) (ex. jornais on-line e redes sociais), devido ao seu baixo custo, vêm facilitando o consumo de notícias [Freire and Goldschmidt, 2019]. Contudo, alguns MDDN têm ampliado a proliferação de uma categoria particular de notícia falsa cuja divulgação acontece de forma intencional: as *Fake News* [Freire and Goldschmidt, 2019].

Dentre as abordagens mais promissoras de detecção de *Fake News*, estão as baseadas em *Crowd Signals*. Essa abordagem utiliza informações sobre a capacidade histórica dos usuários dos MDDN em acertar ou errar, ao opinar sobre a classificação das notícias passadas como *fake* ou não, a fim de aferir a reputação desses usuários na identificação de notícias falsas [Tschiatschek et al., 2018]. Nessa abordagem, as opiniões (*signals*)

dos usuários dos MDDN (*crowds*) sobre novas notícias a serem classificadas são conjugadas, considerando a reputação desses usuários para concluir acerca da classificação dessas notícias. Apesar de possuir um interessante potencial pelo caráter colaborativo, essa abordagem é fortemente dependente da vontade do usuário explicitar sua opinião referente a cada nova notícia a ser classificada.

Contornando a dificuldade de obtenção das opiniões explícitas dos usuários, a abordagem proposta por Souza Freire et al. [2021], denominada *Hybrid Crowd Signals* (HCS), considera que qualquer ação de divulgação (publicação ou compartilhamento) de notícia é um sinal implícito (*opinião implícita*) dado pelo usuário de que ele acredita que tal notícia é verdadeira, independentemente dessa ação ser maliciosa ou não. A abordagem HCS possui dois métodos: HCS-I e HCS-F. No HCS-I a classificação das notícias é realizada com base nas opiniões implícitas dos usuários. No HCS-F, complementar ao HCS-I, a classificação é composta, além das opiniões implícitas dos usuários, por opiniões explícitas de máquinas, ou seja, classificadores construídos a partir de algoritmos de aprendizado de máquina (AM).

Apesar dos resultados promissores do método HCS-F, o mesmo aplica inferência bayesiana ingênua como forma de combinar as classificações fornecidas pelas máquinas sobre a notícia a ser detectada (i.e., um método ingênuo). Diante disso, este estudo levanta a seguinte hipótese: *utilizar Ensembles para combinar as classificações fornecidas pelas máquinas sobre a notícia a ser analisada pode viabilizar a construção de métodos de detecção de Fake News mais robustos que o HCS-F. Ensembles* são comitês de modelos de aprendizado de máquina (classificadores) que, por serem utilizados em conjunto para diminuir a variância e, consequentemente, aumentar a acurácia das predições [Zhang and Ma, 2012], justificam a hipótese levantada.

No início de um caminho de busca por evidências de validade da hipótese acima, o presente trabalho teve como objetivo realizar uma primeira fase de experimentos visando comparar tanto *Ensembles* que combinam as opiniões explícitas das máquinas do HCS-F quanto *Ensembles* compostos por modelos gerados por algoritmos de AM (diferentes das máquinas do HCS-F), com o método HCS-I (método que utiliza somente as opiniões implícitas dos usuários). Realizados nos mesmos cinco *datasets* em que o método HCS-I foi originalmente avaliado, os experimentos revelaram alguns *Ensembles* com resultados superiores ao HCS-I.

2. Metodologia dos Experimentos e Resultados

De forma a comparar os *Ensembles* desenvolvidos neste trabalho e o método HCS-I (*baseline*), foram utilizados nos experimentos os mesmos *datasets*[1] empregados em Souza Freire et al. [2021]. A Tabela 1 apresenta um resumo estatístico desses conjuntos de dados. Cada um deles possui notícias com as seguintes informações: *id* (identificador da notícia), *rótulo* (sendo *rótulo* $\in \{real, fake\}$), e *classificações das máquinas*[2]. As máquinas utilizadas em cada *dataset* também encontram-se indicadas na Tabela 1.

Para a definição formal dos *Ensembles*, seja um conjunto T de algoritmos de classificação em que $T \neq \emptyset$. Seja, ainda, um *Ensemble* $E(C, I, M)$ voltado à tarefa de

[1]Dataset Link: https://github.com/Uriel-Merola/EHCS/tree/main/Datasets

[2]No contexto do método HCS-F, uma máquina é um classificador de notícias como *real* ou *fake*.

classificação, no qual C ($C \neq \emptyset$ e $\forall c \in C$, c é um modelo de classificação gerado a partir algum t, tal que $t \in T$) é o conjunto de classificadores da camada de base de E, I é a estratégia de integração das saídas de C (i.e., $I \in \{\text{Combinação, Votação}\}$), e M é o método que implementa tal estratégia. Detalhes sobre treinamento, avaliação e aplicação de *Ensembles* podem ser obtidos em [Zhang and Ma, 2012].

Datasets	Notícias *Fake*	Notícias Não *Fake*	Máquinas utilizadas em Souza Freire et al. [2021]
Gossip	5000	5000	*Random Forest*, *XGBoost* e *SVM*
PolitiFact	300	300	*Random Forest*, *XGBoost* e *SVM*
Gossip2	1200	1200	*Random Forest*, *XGBoost* e *SVM*
FakeNewsSet	300	300	*Random Forest*, *XGBoost* e *SVM*, *DMText* e *FNE*
FakeBr	3600	3600	*DMText* e *FNE*

Tabela 1. Síntese dos *Datasets* utilizados nos experimentos deste trabalho.

Para os experimentos deste trabalho, $T = \{$*DMText*, *FNE* , *XGBoost*, *Random Forest (RF)*, *Support Vector Machines (SVM)*, *K-Nearest Neighbors (KNN)*, *Decision Tree (DT)*, *Naive Bayes (NB)*, *Multilayer Perceptron (MLP)*, *Logistic Regression (LR)*$\}$[3]. As escolhas por esses algoritmos remetem a opções clássicas na área de *ML*, com boa diversidade em seus vieses de busca, ou por terem um histórico bem sucedido na detecção de *Fake News* (como no caso do *DMText* [Moraes et al., 2019] e do *FNE* [de Souza et al., 2020]). A Figura 1 descreve graficamente a metodologia adotada nos experimentos. Ademais, as próximas subseções apresentam o detalhamento das suas cinco etapas.

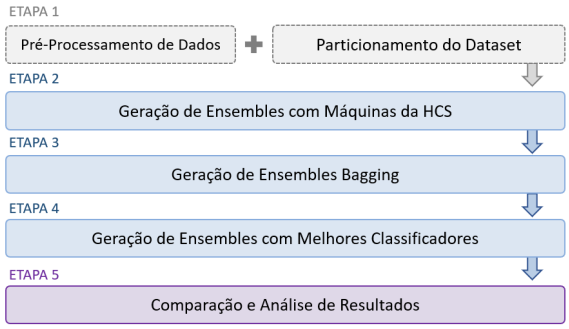

Figura 1. Etapas da metodologia de experimentação adotada

2.1. Pré-Processamento e Particionamento do Dataset

Na Etapa 1, o pré-processamento envolveu a conversão dos valores alfanuméricos dos atributos categóricos em valores numéricos (i.e. *fake* $\rightarrow$ *0* e *real* $\rightarrow$ *1*). Tal conversão visou garantir a compatibilidade dos valores dos atributos com as demandas de algoritmos de T em relação aos tipos de dados. Passaram por essa conversão os valores dos atributos *rótulo* e *classificações das máquinas*.

Ainda na Etapa 1, cada *dataset* foi particionado em 50% dos dados para teste e 50% dos dados para treinamento dos modelos de classificação. A fim de viabilizar uma comparação justa dos *Ensembles* deste trabalho com o método HCS-I, a divisão de cada *dataset* foi a mesma realizada em Souza Freire et al. [2021].

[3]Code Link: https://github.com/Uriel-Merola/EHCS/tree/main/Ensembles

2.2. Geração de Ensembles com Máquinas do HCS-F

A Etapa 2 teve como objetivo construir *Ensembles* $E_j(C, I, M)$, nos quais C fosse formado por modelos de classificação gerados a partir das mesmas máquinas utilizadas pelo método HCS-F (i.e., $\forall c \in C, \exists t \in T'$, tal que c foi gerado por t, sendo $T' = \{$*DMText*, *FNE*, *SVM*, *RF*, *XGBoost*$\}$) e *I=Combinação*. As especificações dos *Ensembles* gerados estão indicadas na Tabela 2.

Ensembles	E_1	E_2	E_3	E_4	E_5	E_6	E_7
Especificações	$M =$ *KNN*	$M =$ *DT*	$M =$ *NB*	$M =$ *SVM*	$M =$ *MLP*	$M =$ *LR*	$M =$ *RF*

Tabela 2. Especificações dos *Ensembles* construídos com máquinas do HCS-F

Uma vez definidos os *Ensembles* a serem implementados, foi executado o processo de validação cruzada com 10 conjuntos de cada *Ensemble* aplicado ao conjunto de testes. Para tanto, foram adotados os valores *default* dos hiperparâmetros dos algoritmos de C. A Tabela 3 apresenta os resultados obtidos pelos *Ensembles* implementados.

Modelos	FakeNewsSet	Gossip	Gossip2	PolitiFact	FakeBr
HCS-I	**0.9179± 0.0397**	**0,9389± 0.0094**	**0,9078±0.0238**	**0,9013± 0.0484**	**0,9987± 0.0028**
E_1	0.7225±0.0600	0.5355±0.0900	0.5003± 0.0294	0.4905±0.0144	0.9337± 0.0095
E_2	0.9152±0.0628	0.9255±0.1423	0.7915±0.1799	0.6804±0.1078	0.9337± 0.0095
E_3	0.8898±0.0635	0.9265±0.1420	0.7932±0.1826	0.6995±0.1227	0.9337± 0.0095
E_4	0.9119±0.0611	0.9255±0.1423	0.7915±0.1799	0.6960±0.0957	0.9337± 0.0095
E_5	0.9086±0.0576	0.9259±0.1424	0.7915±0.1799	0.6680±0.1199	0.9337± 0.0095
E_6	0.8991±0.0563	0.9259±0.1424	0.7907±0.1752	0.6712±0.1254	0.9337± 0.0095
E_7	0.9119±0.05913	0.9259±0.1425	0.7907±0.1811	0.6804±0.1078	0.9337± 0.0095

Tabela 3. Acurácia média $\pm$ desvio-padrão dos *Ensembles* com máquinas da HCS-F no processo de validação cruzada com 10 conjuntos em cada *dataset*

2.3. Geração de Ensembles Bagging

A fim de variar os tipos de *Ensemble* criados na Etapa 2, a Etapa 3 consistiu em gerar *Ensembles* do tipo *bagging* a partir dos algoritmos do conjunto $T'' \subset T$, sendo $T'' = \{$*KNN*, *DT*, *SVM*, *MLP*, *NB*, *LR*, *RF*$\}$[4]. Um *Ensemble* do tipo *bagging* é um modelo de classificação construído a partir da aplicação de um algoritmo de classificação A em N instâncias do *dataset* D em análise, geradas por meio da técnica de *bootstrap* (ou seja, amostragem aleatória com reposição). Desta forma, a camada de base do *bagging* é formada por N modelos de classificação $c_{A,k}$, $k = 1, ..., N$, construídos a partir do mesmo algoritmo A aplicado às N instâncias de D. Também nesta etapa, foram adotados os valores *default* dos hiperparâmetros dos algoritmos de classificação em T''. As especificações dos *Ensembles* do tipo *bagging* gerados estão indicadas na Tabela 4. É importante ressaltar que foi escolhido $N = 10$ e que, por definição, nos modelos do tipo *bagging*, *I=Votação* e *M* é o método que implementa a eleição por votação majoritária simples.

[4]Tal escolha visou o uso de algoritmos de classificação tradicionais da área de *ML*.

Ensembles	Especificações
E_8	$C = \{c_{KNN,1}, c_{KNN,2}, ..., c_{KNN,10}\}$
E_9	$C = \{c_{DT,1}, c_{DT,2}, ..., c_{DT,10}\}$
E_{10}	$C = \{c_{NB,1}, c_{NB,2}, ..., c_{NB,10}\}$
E_{11}	$C = \{c_{SVM,1}, c_{SVM,2}, ..., c_{SVM,10}\}$
E_{12}	$C = \{c_{MLP,1}, c_{MLP,2}, ..., c_{MLP,10}\}$
E_{13}	$C = \{c_{LR,1}, c_{LR,2}, ..., c_{LR,10}\}$
E_{14}	$C = \{c_{RF,1}, c_{RF,2}, ..., c_{RF,10}\}$

Tabela 4. Especificações dos *Ensembles bagging* implementados

De forma análoga à Etapa 2, foi executado o processo de validação cruzada com 10 conjuntos para cada Ensemble da Etapa 3 aplicado ao conjunto de testes. A Tabela 5 apresenta os resultados obtidos pelos *Ensembles bagging* implementados.

Modelos	FakeNewsSet	Gossip	Gossip2	PolitiFact	FakeBr
HCS-I	0.9179± 0.0397	**0,9389± 0.0094**	**0,9078±0.0238**	**0,9013± 0.0484**	**0,9987± 0.0028**
E_8	0.9086±0.0609	0.9291±0.1429	0.8000±0.1801	0.6930±0.0992	0.9337± 0.0095
E_9	**0.9183±0.0565**	0.9267±0.1429	0.8008±0.1789	0.6835±0.1111	0.9337± 0.0095
E_{10}	0.8960±0.0617	0.9265±0.1420	0.7966±0.1856	0.6995±0.1227	0.9337± 0.0095
E_{11}	**0.9183±0.0565**	0.9261±0.1426	0.8008±0.1789	0.6991±0.0989	0.9337± 0.0095
E_{12}	0.9023±0.0558	0.9267±0.1428	0.7932±0.1791	0.6837±0.1168	0.9337± 0.0095
E_{13}	0.9024±0.0592	0.9265±0.1427	0.7915±0.1762	0.6806±0.1171	0.9337± 0.0095
E_{14}	0.9151±0.0600	0.9265±0.1428	0.7907±0.1811	0.6867±0.1094	0.9337± 0.0095

Tabela 5. Acurácia média ± desvio-padrão dos *Ensembles* bagging no processo de validação cruzada com 10 conjuntos em cada *dataset*

2.4. Geração de Ensembles com Melhores Classificadores

Na Etapa 4, os melhores modelos de classificação gerados para cada *dataset* nas etapas anteriores foram utilizados na formação de novos *Ensembles*, em que *I=Votação* e M o método de votação majoritária simples. A intenção desta escolha foi procurar tirar proveito do conhecimento sobre os *datasets* adquirido pelos referidos modelos. As especificações dos *Ensembles* construídos para cada *dataset* e seus respectivos desempenhos estão indicadas na Tabela 6.

Ensembles	Especificações	Dataset	Acurácia/Desvio Padrão	HCS-I
E_{15}	$C = \{E_2, E_9, E_{11}\}$	FakeNewsSet	**0.9183± 0.0566**	0.9179± 0.0397
E_{16}	$C = \{E_8, E_9, E_{13}\}$	Gossip	0.9263± 0.1427	**0,9389± 0.0094**
E_{17}	$C = \{E_8, E_9, E_{11}\}$	Gossip2	0.8008± 0.1790	**0,9078±0.0238**
E_{18}	$C = \{E_9, E_{11}, E_{13}\}$	PolitiFact	0.6837± 0.1062	**0,9013± 0.0484**
E_{19}	$C = \{E_1, E_3, E_4\}$	FakeBr	0.9337± 0.0096	**0,9987± 0.0028**

Tabela 6. Especificações e resultados dos *Ensembles* com melhores classificadores - validação cruzada com 10 conjuntos em cada *dataset*

2.5. Comparação e Análise de Resultados

Na Etapa 5, foi realizada a comparação dos resultados gerados nos experimentos. Em geral, os *Ensembles* criados com máquinas do HCS-F (Tabela 2) apresentaram desempenhos inferiores aos do método HCS-I em todos os *datasets* (Tabela 3), indicando sinais de robustez do método que congrega opiniões dos usuários dos MDDN.

Os modelos *Ensemble* do tipo *bagging* (Tabela 4), em sua maioria, apresentaram um desempenho (Tabela 5) superior em relação aos modelos *Ensemble* construídos com as máquinas utilizadas pelo HCS-F. Tais resultados sinalizam para uma superioridade da técnica *bagging* na formação de bons *Ensembles* para detecção de *Fake News*. Por outro lado, quando comparados ao método HCS-I, os *Ensembles bagging* também apresentaram resultados inferiores, com exceção dos *Ensembles* E_9 (baseado no *DT*) e E_{11} (baseado no *SVM*), que obtiveram acurácia média levemente superior no *dataset FakeNewsSet*.

De forma análoga ao comentado acima, os *Ensembles* formados a partir dos melhores classificadores obtidos nos experimentos anteriores também só conseguiram superar o método HCS-I no *dataset FakeNewsSet* (Tabela 6). Entretanto, excetuando-se os resultados do *Ensemble* E_{18} no *dataset PolitiFact*, os demais *Ensembles* com melhores classificadores apresentaram resultados relativamente próximos aos do HCS-I, indicando que o uso de *Ensembles* pode ser uma alternativa interessante para conjugar os resultados das máquinas do HCS-F.

3. Considerações Finais

O presente trabalho teve como objetivo realizar experimentos visando avaliar se o uso de *Ensembles*, como forma de integrar os pareceres de modelos de classificação, poderia superar o método HCS-I que conjuga opiniões de usuários dos MDDN para identificação de *Fake News*. Este trabalho é o primeiro passo de uma iniciativa de pesquisa que busca melhorar o desempenho do método HCS-F, derivado do HCS-I, que combina opiniões de usuários e de máquinas (modelos de classificação) a fim de detectar *Fake News*.

Realizados nos mesmos *datasets* utilizados na avaliação do HCS-I, os experimentos produziram resultados em que, apesar de poucos *Ensembles* terem superado o HCS-I, a maioria apresentou resultados relativamente próximos aos do HCS-I. Diante disso, além da utilização de outras métricas de avaliação, as seguintes iniciativas encontram-se em andamento: (i) otimização dos valores dos hiperparâmetros dos algoritmos de classificação; (ii) busca pela melhoria do desempenho dos *Ensembles* ao fornecer ao método que implementa a estratégia de integração, além das saídas dos modelos de classificação, também as entradas recebidas por tais modelos; e (iii) incorporação dos *Ensembles* no HCS-F.

Referências

de Souza, M. P., da Silva, F. R. M., Freire, P. M. S., and Goldschmidt, R. R. (2020). A linguistic-based method that combines polarity, emotion and grammatical characteristics to detect fake news in portuguese. WebMedia.

Freire, P. and Goldschmidt, R. (2019). Uma introdução ao combate automático às fake news em redes sociais virtuais. SBBD.

Moraes, M. P., de Oliveira Sampaio, J., and Charles, A. C. (2019). Data mining applied in fake news classification through textual patterns. WebMedia.

Souza Freire, P. M., Matias da Silva, F. R., and Goldschmidt, R. R. (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.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. International WWW.

Zhang, C. and Ma, Y. (2012). *Ensemble Machine Learning: Methods and Applications*. Springer.