Rotulação de perfis de usuários propagadores de desinformação através de análise de personalidade em redes sociais

Palavras-Chave: *fake news*, Aprendizado de máquina, Processamento de linguagem Natural

Autores/as:

Wagner Alexandre Martins Junior (Aluno) UNICAMP-FT

Profa. Dra. Ana Estela Antunes da Silva (orientadora) UNICAMP-FT

Prof. Dr. Guilherme Palermo Coelho (coorientador) UNICAMP-FT

INTRODUÇÃO

Apesar do alto uso e da democratização do acesso à informação, muitas vezes a qualidade do conteúdo gerado em mídias sociais é questionável, sendo comum o surgimento e compartilhamento de notícias falsas (*fake news*) ou sem fundamentos. Segundo pesquisa feita pela Câmara dos Deputados e Senado Federal (2019), 83% dos entrevistados já identificaram algum tipo de notícia falsa, sendo que a grande maioria afirma que esse tipo de informação ganha mais visibilidade que notícias legítimas, demonstrando alto potencial de desinformação e influência nas ações da população em eventos como eleições e referendos.

Tendo isso em vista, surgem vários esforços para detecção de *fake news* utilizando diferentes abordagens que vão desde a rotulação manual de notícias falsas, como realizado pelas agências de verificação Fato Ou Fake e Agência Lupa, até a utilização de algoritmos de Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (AM) para a rotulação automática de fake news (SHU *et al.,*2020; MONTEIRO *et al.,*2018). Uma das técnicas utilizadas para construção destes algoritmos é a Supervisão Social Fraca (do Inglês, Social Weak Supervision) ou SSF, que permite a identificação precoce de fake news dando maior foco às características dos usuários em geral, sejam eles autores ou receptores de conteúdo, e suas conexões, que ao conteúdo das notícias como feito pelas agências de verificação, ou seja, essa abordagem coloca em foco o escopo social e psicológico dos usuários para a avaliação de conteúdo.

Utilizando a abordagem de Supervisão Social Fraca, SHU *et al.* (2018) identificaram uma relação forte entre publicação de *fake news* e tipos de usuários em redes sociais. A pesquisa rotulou usuários do *Twitter* propagadores de notícias falsas de acordo com características explícitas dos perfis como verificação do perfil, data de cadastro na rede social e número de postagens, em conjunto com características implícitas do perfil, como gênero, idade e personalidade, apontando uma relação entre tipo de usuário e propagação de notícias falsas ainda não explorada em algoritmos de rotulação automática de notícias.

Em estudo similar, os trabalhos de SHU et al., 2018; DA SILVA et al., 2018; RAMOS, 2018; HARTMAN et al., 2017 propõem as metodologias para a análise de personalidade utilizando o método BIG 5, que mensura cinco dimensões fundamentais do comportamento humano, que são: Liberalidade (ou Abertura à experiência), Meticulosidade, Extroversão, Agradabilidade e Instabilidade (ou Neuroticismo). Além disso, também detalham a utilização de vetores de palavras (*word embeddings*) para rotulação de personalidades, sendo os trabalhos de DA SILVA et al., 2018 e HARTMAN et al., 2017 direcionados à língua portuguesa.

Tendo em vista a importância do tema e o fato de que ele ainda é pouco explorado, principalmente em trabalhos direcionados à detecção de *fake news* em língua portuguesa, este trabalho traz a proposta de investigar a relação entre fake news e características de perfis de usuários através de métodos de SSF, PLN e AM, para posterior utilização em um modelo de detecção automática de notícias falsas em língua portuguesa.

OBJETIVO

O objetivo geral deste projeto foi apresentar uma relação entre características de perfis de usuários do Twitter, obtida por meio de análise de características explícitas e análise de personalidade, e publicação de fake news.

METODOLOGIA

Como não há um *dataset* acessível com rotulação de usuários do *Twitter* como propagadores de *fake news* e frases de autoria desses usuários, este trabalho precisou montar sua própria base de dados com essas informações, sendo que as informações coletadas por essa base serão usadas como entrada para os algoritmos de classificação de *fake news* de acordo com a personalidade. A coleta de dados de usuários é realizada através de scrapping, atividade de coleta automática de informação a partir de algoritmos.

A metodologia utilizada para a construção da base de dados dos usuários segue categorização de usuários semelhante à feita por SHU *et al.* (2018), onde a cada usuário é atribuído um nível de confiança de acordo com a quantidade de *fake news* que publicou (*tweet*), compartilhou (*retweet*) ou gostou (*like*), sendo que usuários com grau de confiança menor seriam aqueles com maior número de eventos relacionados a notícias falsas, entendendo-se eventos como *tweet*, *retweet* e *like* de *fake news*. Após a coleta de informação dos usuários, foi utilizado o dataset B5-POST (DA SILVA et al., 2018), que contém rotulação de frases de usuários do *Facebook* de acordo com as cinco personalidades presentes no método BIG 5, para o treinamento de algoritmos de AM e PLN, com o objetivo de verificar a existência de relação entre propagação de notícias falsas e perfis de personalidade. Para a tarefa de classificação, foi escolhido o classificador SVM, porque a literatura indicou este classificador como o de melhor desempenho para a tarefa em questão (SHU et al., 2018). Para cada personalidade, treinou-se um classificador SVM separadamente. Nas seções a seguir as etapas de extração, rotulação de usuários e verificação de relação de personalidade são mais bem definidas.

1 - Diagrama de *Scrapping* de usuários

O diagrama da Figura 1 ilustra cada etapa do processo de extração de características dos usuários da rede social Twitter que espalham *fake news*.

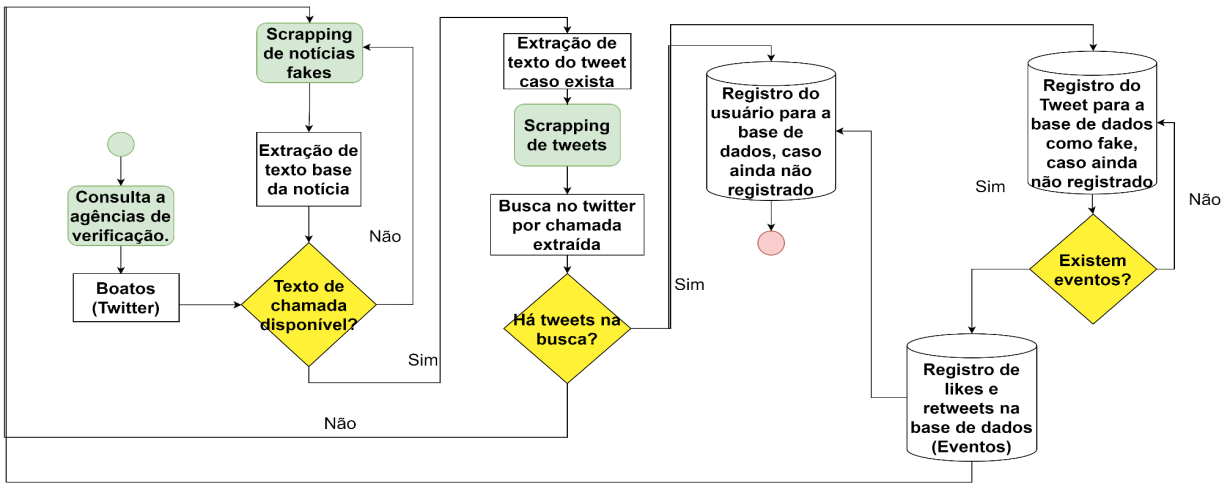

Figura 1: Processo de *scrapping* de usuários e notícias falsas

2 - Coleta de notícias falsas

Para a realização da coleta de notícias falsas, foi utilizada a agência de rotulação Boatos, com busca específica para notícias falsas espalhadas pelo *Twitter*. Em cada análise de notícias falsas feita pela agência, há disponibilização do texto publicado no *Twitter* ou explicação do conteúdo da notícia falsa. Caso o texto do *tweet* esteja disponível na análise, ele é extraído e rotulado como notícia falsa.

3 - Scrapping de usuários

A partir da coleta do corpo do tweet, descrita no passo anterior, é feita a busca na rede social por tal *tweet*. A busca é configurada para retornar exatamente o mesmo texto, limitando o resultado aos *tweets* e usuários que espalharam a mesma notícia. Para cada *tweet* retornado são colhidos o usuário que publicou o *tweet* e os usuários que curtiram e/ou compartilharam a *fake news.* O compartilhamento e a curtida são registrados como eventos na base de dados.

No caso de *scrapping* (atividade de coleta automática de informação a partir de algoritmos) de usuários que compartilham notícias legítimas, foram usados tweets de jornais respeitados e reconhecidos (Folha, Estadão, O Globo) e foi aplicado o mesmo procedimento ilustrado anteriormente na Figura 1.

4 - Rotulação de usuários

No caso da rotulação de usuários, foi usada a categorização de usuários semelhante à feita por SHU *et al.* (2018), sendo atribuído a cada usuário um nível de confiança, onde quanto menor esse nível mais eventos de *fake news* foram relacionados ao usuário. Após a rotulação por reincidência de fakes news espalhadas por usuário, sendo uma contagem simples de quantos eventos com *fake news* foram de autoria de cada usuário, os usuários foram rotulados de acordo com as cinco personalidades. A cada usuário foi atribuído um vetor binário com as cinco dimensões e um indicativo dizendo se compartilha *fake news* ou não, sendo que essa classificação de cada personalidade do usuário é feita utilizando o classificador SVM (a partir da utilização do dataset B5-POST). O vetor binário de personalidades foi usado como entrada para o algoritmo de classificação de usuários que compartilham *fake news* de acordo com a personalidade. Neste trabalho, foram utilizadas as bibliotecas Scikit-Learn**,** para os algoritmos de machine learning, e Spacy para pré-processamento do texto e processamento de linguagem natural.

RESULTADOS

A Tabela 1 apresenta os números conseguidos com o *scrapping* de usuários propagadores de *fake news* na rede social *Twitter*. A Figura 2 ilustra a contagem dos eventos correspondentes à interação de usuários com tweets espalhadores de *fake news*.

Tabela 1: Quantidade de objetos na base de dados.

Tipo	Quantidade
Eventos notícias legítimas	8850
Eventos *fake news*	11399
Tweets *fake news*	52
Tweets notícias legítimas	80
Usuários espalhadores de *fake news*	10115
Usuários espalhadores de notícias legítimas	7392

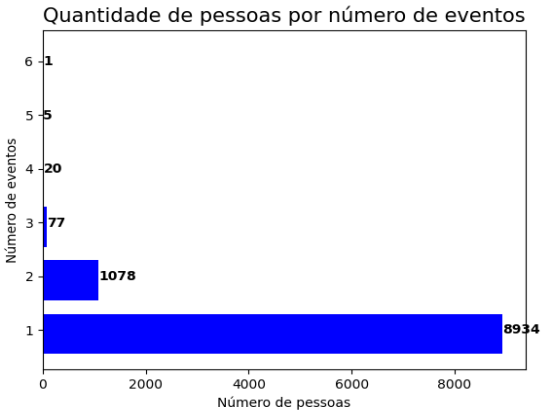

Figura 2: Reincidência em eventos fake.

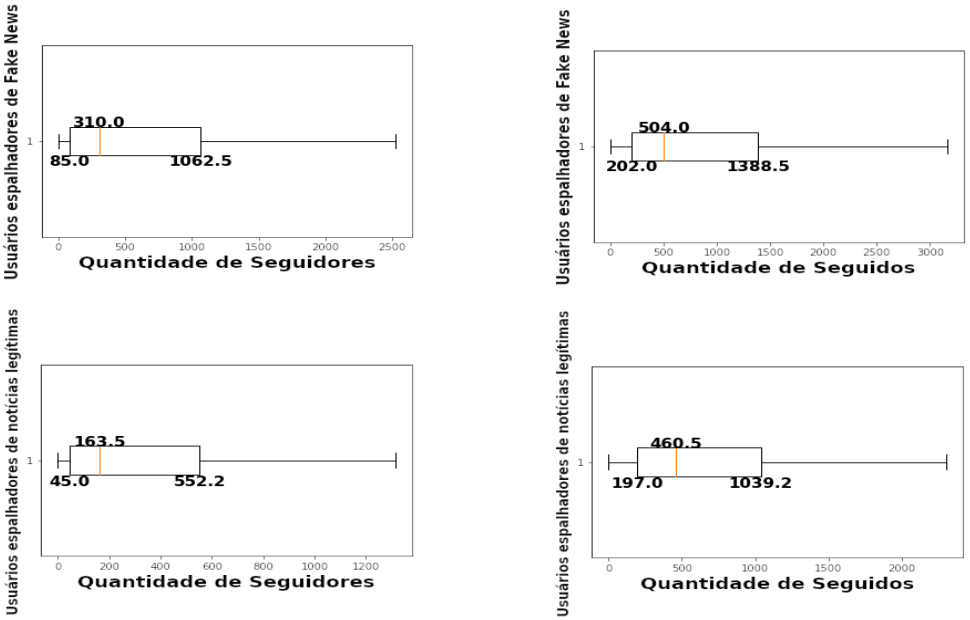

Figura 3: Boxplot com quantidade de seguidores e seguidos de espalhadores de *fake news* e notícias legítimas

A Figura 3 mostra que a grande maioria dos usuários que postaram ou tiveram alguma interação com *fake news* o fizeram apenas uma vez, sem reincidência, resultado semelhante ao que Kai Shu *et al.* (2018) observou em

seus estudos. O maior número de reincidência observado para esta base de dados é de 6 eventos, ou seja, um mesmo usuário interagiu 6 vezes com notícias falsas. Os gráficos de *boxplot* mostram as estatísticas para quantidade de seguidores e seguidos para os usuários propagadores de *fake news*, excluindo-se valores abaixo e acima de 50% do valor do primeiro e terceiro quartil. Kai Shu *et al.* (2018) comentam uma relação entre perfis de usuários propagadores de *fake news* com pouco engajamento à rede social, sendo baixa quantidade de seguidores e seguidos um sinal de falta de engajamento. Porém os gráficos mostram que os usuários desta base de dados possuem em sua maioria considerável engajamento quanto a esses parâmetros. Outro ponto a ser comentado é uma tendência de usuários espalhadores de *fake news* terem mais seguidores e seguidos que usuários espalhadores de notícias legítimas.

A Tabela 2 mostra o desempenho dos classificadores SVM utilizados para a construção dos vetores de personalidade de cada usuário na base de dados utilizando o dataset B5 post. Os resultados encontrados para esta classificação estão de acordo com o esperado pela literatura (SHU et al., 2018; DA SILVA et al., 2018; RAMOS, 2018; HARTMAN et al., 2017).

Tabela 2: Desempenho dos classificadores SVM para rotulação de personalidades de acordo com texto

Dimensão	**Precisão**	**Recall**	**F1-score**	**Acurácia**
Abertura à Experiência	não - 52% sim - 58%	não - 43% sim - 66%	não - 47% sim - 62%	56%
Conscienciosidade	não - 57% sim - 57%	não - 59% sim - 55%	não - 58% sim - 56%	57%
Extroversão	não - 53% sim - 59%	não - 42% sim - 69%	não - 47% sim - 63%	57%
Amabilidade	não - 54% sim - 56%	não - 48% sim - 63%	não - 51% sim - 59%	55%
Neuroticismo	não - 51% sim - 59%	não - 37% sim - 72%	não - 43% sim - 65%	57%

A Tabela 3 demonstra o desempenho do classificador de usuários propagadores de *fake news*, onde os vetores de personalidade são entradas, e como saída, espera-se a catalogação do usuário como provável espalhador de *fake news* ou não. Para estudo comparativo, também foram utilizados os classificadores de árvore binária e *Random Forest.* Na Tabela 3 nota-se que os classificadores obtiveram aproximadamente o mesmo desempenho quando comparados entre si para a tarefa de classificação de prováveis espalhadores de *fake news*.

Tabela 3: Desempenho de classificadores para rotulação de usuários de acordo com texto

Algoritmo de seleção	**F1-Score**	**Acurácia**
Árvore binária	47%	47%
SVM	47%	48%
Random Forest	47%	47%

DISCUSSÃO E CONCLUSÕES

Levando em conta apenas os resultados da Tabela 3, para o classificador de possíveis espalhadores de *fake news*, notamos pela baixa acurácia e F1 Score que a característica de perfil de personalidade do usuário não demonstra uma relação forte com a ação de espalhar notícias falsas.

Entretanto, vale ressaltar que este estudo se deparou com algumas limitações que podem ter influenciado o resultado final. Uma delas é a característica subjetiva de classificação de personalidade, o que torna a tarefa de classificação e construção de um dataset complexa. No caso do B5-POST, o dataset foi construído a partir de autoclassificação de personalidade pelos próprios autores das frases por meio de um questionário (RAMOS, 2018), o que pode introduzir um viés no dataset. Como observa-se da Tabela 2, o classificador de personalidades não desempenha bem para este tipo de classificação como já é dado na literatura para classificador em língua portuguesa, portanto é possível que haja propagação de erro da classificação para o vetor de personalidades.

Dessa forma, também se pode concluir que a ausência de modelos com maior desempenho para classificação de personalidade abre espaço para futuros investimentos em diferentes métodos de classificação e geração de novos datasets ou melhoramento dos já existentes, sendo necessário mais estudos especialmente em língua portuguesa.

BIBLIOGRAFIA

[1] CAMARA DOS DEPUTADOS; SENADO FEDERAL. Redes Sociais, Notícias Falsas e Privacidade de Dados na Internet**. Pesquisa DataSenado.** Nov.de 2019 Disponível em: https://www12.senado.leg.br/institucional/datasenado/arquivos/mais-de-80-dos-brasileiros-acreditam-que-redes-sociais-influenciam-muito-a-opiniao-das-pessoas. Acessado em 28-fev-2021.

[2] FOLHA DE SÃO PAULO. **Agência Lupa.** Disponível em: https://piaui.folha.uol.com.br/lupa/. Acessado em 28-fev-2021.

[3] GRUPO GLOBO. **Fato ou Fake**. Disponível em: https://g1.globo.com/fato-ou-fake/. Acessado em 28-fev-2021.

[4] SHU, K., WANG, S., LEE, D., LIU, H. **Disinformation, Misinformation, and Fake News in Social Media.** 2020, Springer.

[5] MONTEIRO, Rafael A., [et al.]. **Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results.** Computational Processing of Portuguese Language, Springer, Set2018.

[6] SHU, K., et al. **Beyond News Contents: The Role of Social Context for Fake News Detection.** WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Jan2019, p. 312-320.

[7] SHU, K., et al. **Defend: explainable Fake News detection.** KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, jul2019, p. 395−405.

[8] SHU, K., et al. **Understanding User Profiles on Social Media for Fake News Detection.** Proceedings - IEEE 1st Conference on Multimedia Information Processing and Retrieval, Abr2018, p. 430-435.

[9] DA SILVA, B., Paraboni, I. **Personality Recognition from Facebook Text.** Computational Processing of Portuguese Language, Springer, Set2018, p107-113.

[10] RAMOS, R.M.S. **Building a Corpus for Personality-dependent Natural Language Understanding and Generation.** Em: 11th International Conference on Language Resources and Evaluation (LREC-2018), ELRA, Miyazaki, Japan, pp. 1138−1145 (2018).

[11] HARTMAN, N., et al. **Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks.** Em*:* Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), 1. , 2017, Minas Gerais.

[12]GRUPO BOATOS. **Boatos.org**. Disponível em: https://www.boatos.org/. Acessado em 28-fev-2021.

[13] BOISBERRANGER, J., *et al*. **Scikit-Learn**. Disponível em: https://scikit-learn.org/stable/index.html.Acessado em 24-jul-2022.