Identifying fakes in social media text messages using Stylometry

Cristian G. Butzke[3], Marcelo L. Brocardo[2], Wesley R. Bezerra[1,3], Carlos B. Westphall[1]

[1]LRG – Universidade Federal do Santa Catarina (UFSC)
Campus Universitário – Florianópolis – SC – Brazil

[2]ESAG – Universidade do Estado de Santa Catarina (UDESC)
Itacorubi – Florianópolis – SC – Brazil

[3]Campus Rio do Sul – Instituto Federal Catarinense (IFC)
Rio do Sul – SC – Brazil

{cristianbtzk,wesleybez, carlosbwestphall}@gmail.com,
marcelo.brocardo@udesc.br

Atestar autenticidade de textos na *internet*, principalmente em mídias sociais tem se tornado um desafio nos últimos tempos com a Inteligencia Artificial Generativa [Korobenko et al. 2024]. Isso tem implicado em problema de autenticação em sistema eletrônicos [Capasso et al. 2024], *deepfakes* [Yu et al. 2023, Hutiri et al. 2024] e demais vulnerabilidades decorrentes a *impersonation* de indivíduos através da AI [Ristovska 2023, Bianco et al. 2023].

Nesse aspecto, um dos principais desafios em segurança de aplicações é identificar se um sujeito que realiza determinada ação no sistema é quem afirma ser. A autenticação em meios digitais pode ser um desafio. Sendo assim, o reconhecimento de padrões em autenticação relacionados à biometria comportamental, em contrapartida, utiliza ações realizadas por um usuário, como análise do estilo de escrita para continuamente verificar a autenticidade do usuário [Brocardo 2015]. Métodos baseados nesses padrões têm alcançado altos índices de precisão e tornam-se de mais complexa transgressão em função de seu princípio se dar pelo uso da aplicação.

Dessa forma, este trabalho tem como intuito abordar as etapas desde a extração, transformação e carregamento de dados até a criação do modelo de inferência para a autoria de um texto obtido a partir de redes sociais, em específico a rede X/Twitter[1]. Este estudo apresenta a elaboração de um modelo de autenticação contínua baseada em publicações de usuários da rede social X como prevenção de *impersonation*. Trazendo as seguintes contribuições: **(1)** propor uma solução para identificação de autoria do texto utilizando estilometria; e **(2)** avaliar a solução proposta frente as demais soluções existentes.

Como **discussão**, os melhores resultados para cada kernel utilizado (**i**) são demonstrados na Tabela 1. O melhor resultado foi obtido pela configuração que utilizou o kernel linear, onde a acurácia foi de aproximadamente 90%. Os kernels polinomial, RBF e sigmoide atingiram 87%, 83% e 69%, respectivamente.

Adicionalmente, observou-se que os modelos apresentaram desempenhos distintos na classificação das publicações, variando de acordo com o autor (**ii**). A maior acurácia foi registrada para as publicações de Sebastian Ruder, alcançando 94%, enquanto as de

[1]https://x.com/

Kernel	**Valor de C**	**Valor Gamma**	**Acurácia**	**Precisão**
Linear	1	-	0.90	0.90
Polynomial	0.1	-	0.87	0.89
RBF	1000	0.1	0.83	0.84
Sigmoid	0.1	-	0.69	0.69

Table 1. Melhores resultados durante a etapa de teste

Katy Perry resultaram na menor acurácia, com 86%.

É válido ressaltar, portanto, que os modelos de classificação demonstram acurácias distintas para cada usuário, refletindo particularidades na quantificação de elementos textuais e padrões de escrita.

Quanto a **conclusão**, como pôde-se notar através da Tabela 1 os resultados encontrados para os testes são relevantes e adequados para verificação das identidades dos autores. Ainda que haja uma variação de acordo com as características de construção textual de cada autor, é possível afirmar que a estilometria é adequada ao propósito da autenticação contínua para mídias sociais e na identificação da *fakes* publicados. Isto é reforçado através da validação através do simulador que avalia as publicações no caso de um dos usuários treinados.

Como **Acknowledgments**, o autores agradecem a UFSC e ao IFC. Adicionalmente, informamos que este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC), Edital 20/2024.

References

Bianco, T., Castellano, G., Scaringi, R., Vessio, G., et al. (2023). Identifying ai-generated art with deep learning. In *CREAI@ AI* IA*, pages 16–25.

Brocardo, M. L. (2015). *Continuous Authentication using Stylometry*. PhD thesis.

Capasso, P., Cattaneo, G., and De Marsico, M. (2024). A comprehensive survey on methods for image integrity. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11):1–34.

Hutiri, W., Papakyriakopoulos, O., and Xiang, A. (2024). Not my voice! a taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 359–376.

Korobenko, D., Nikiforova, A., and Sharma, R. (2024). Towards a privacy and security-aware framework for ethical ai: Guiding the development and assessment of ai systems. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, pages 740–753.

Ristovska, S. (2023). Ways of seeing: The power and limitation of video evidence across law and policy. *First Monday*.

Yu, Z., Zhai, S., and Zhang, N. (2023). Antifake: Using adversarial audio to prevent unauthorized speech synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 460–474.