Demografische Einflüsse auf Authorenschaftsidentifikation

Jasmin Wyss Rebekah Overdorf

30. Mai 2025

Zusammenfassung

Auf maschinellem Lernen basierte Authorenschaftsidentifikation, wird seit Längerem in der Computersicherheitsliteratur als Mittel zur Deanonymisierung studiert [1],[4],[2]. Moderne Methoden machen wenig Fehler, auch bei vielen potenziellen Autoren [3]. Doch wie viel von dieser hohen Genauigkeit, ist das Ergebnis der sprachlichen Eigenheiten der Autoren, und wie viel Einfluss haben demografische, kulturelle oder kontextuelle Faktoren?

Diese Frage ist von entscheidender Bedeutung für das Verständnis der Auswirkungen der Authorenschaftsidentifikation auf die Privatsphäre. Sie gibt Aufschluss darüber, ob diese Modelle den wahren Autor identifizieren oder ob sie lediglich die Eigenschaften des Autors widerspiegeln.

In unserer Studie analysieren wir den Einfluss von Muttersprache, Geschlecht und Alter auf die Authorenschaftsidentifikation. Wir analysieren dazu die Resultate von Authorenschaftidentifikationsmodellen, die mit von uns gesammelten Reddit-Datensätzen trainiert wurden.

Unsere Resultate zeigen, dass die Untersuchung von Klassifizierungsfehlern, wie sie in der algorithmischen Fairness-Literatur üblich ist, für die Authorenschaftsidentifikation nicht geeignet ist. Denn wir stellen keinen statistisch signifikanten Einfluss von den gewählten demografischen Faktoren auf das Resultat der Authorenschaftsidentifikation fest, solange der wahre Autor des Textes auch in den Trainingsdaten des Modells ist. Falls jedoch der wahre Autor des Textes nicht in den Trainingsdaten des Modells ist, wird der Text tendenziell einem Autor zugeordnet, der dieselben demografischen Merkmale wie der wahre Autor hat.

Unsere Ergebnisse stellen den Einsatz von auf maschinellem Lernen basierten Authorenschaftsidentifikation infrage. Insbesondere in kritischen Situationen wie dem Strafverfahren. Falls die falschen Autoren als potenzielle Täter identifiziert werden, betrifft eine Fehlidentifikation eher Personen, die demografische Merkmale mit dem wahren Autor eines Textes teilen. Dies bedeutet, dass marginalisierte Gemeinschaften, wie z. B. Einwanderergemeinschaften, die oft eine andere Muttersprache sprechen, unverhältnismäßig stark von Fehlern betroffen sein könnten.

Literatur

[1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29, 2008.

[2] Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. A girl has no name: Automated authorship obfuscation using mutant-x. *Proc. Priv. Enhancing Technol.*, 2019(4):54–71, 2019.

[3] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314. IEEE, 2012.

[4] Rebekah Overdorf and Rachel Greenstadt. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proc. Priv. Enhancing Technol.*, 2016(3):155–171, 2016.