Fraunhofer SIT

ATHENE

WORK IN

CYBERSECURITY

Bachelorarbeit: Angriffe auf Autorschaftsverifikationsmethoden

Was Du bei uns tust

Die Autorschaftverifikation (AV) stellt eine zentrale Herausforderung im Bereich der Digitalen Textforensik dar. Sie wird in zahlreichen Anwendungsfeldern genutzt, wie beispielsweise in der Strafverfolgung zur Bekämpfung von Hassrede im Netz oder im akademischen Bereich zur Überprüfung von Plagiaten. Der Forschungszweig beschäftigt sich mit der zentralen Frage ob zwei oder mehr Texte von derselben Person verfasst wurden. Das Pendant zur AV stellt Authorship Obfuscation (AO) dar, welches eine Teildisziplin der Digitalen Textforensik darstellt. AO-Methoden zielen darauf ab stilistische Merkmale eines Textes so zu verändern, dass es schwierig oder (im Idealfall) unmöglich wird, den ursprünglichen Autor zu identifizieren. Diese Methoden werden insbesondere genutzt, um die Privatsphäre eines Autors zu schützen oder um eine Identifikation durch AV-Methoden zu erschweren.
Ziel dieser Arbeit ist es, die Robustheit existierender AV-Methoden hinsichtlich Angriffsmöglichkeiten seitens von AO-Methoden zu untersuchen. Letztere basieren sowohl auf (1) stilometrische als auch (2) Deep-Learning-basierte Ansätze. Zu (1) zählen etwa Textmodifikationen wie Ersetzungen von Synonymen/Phrasen, Umstellungen von Aufzählungen oder auch das Umschreiben von Kontraktionen. Zu (2) zählen hingegen Textmodifikationen wie Backtranslation (Vor- und Rückwärtsübersetzung mit mehreren Zwischensprachen) mittels Übersetzungsdienste oder auch Paraphrasierung von Sätzen mittels Large Language Models (LLMs). Im ersten Schritt sollen zunächst mehrere aktuelle AV- und AO-Methoden recherchiert und ggf. nachimplementiert werden. Anschließend gilt es aus gegebenen Datensätzen neue, modifizierte Datensätze zu erstellen. Dazu sollen die Original-Texte anhand der recherchierten AO-Methoden entsprechend modifiziert werden. Aufbauend auf den konstruierten Datensätzen sollen anschließend die recherchierten AV-Methoden evaluiert werden, wobei es zu analysieren gilt, wie stark die Erkennungsleistung in Bezug zu den Original-Datensätzen sinkt und womit das zusammenhängt. Aufbauend auf den gewonnenen Erkenntnissen soll eine der recherchierten AV-Methoden ausgewählt und auf diesen Anwendungsfall erweitert bzw. optimiert werden, um die AO-Angriffe entgegenzuwirken. Dies kann z. B. mithilfe von stilistischen Merkmalen bewerkstelligt werden, welche von den betrachteten AO-Methoden nicht berücksichtigt werden. Die Erkenntnisse dieser Arbeit sollen einen Beitrag zur Weiterentwicklung von AO-resistenten AV-Ansätzen beitragen.

Was Du mitbringst

- Studium der Informatik, Mathematik oder eines verwandten Fachgebiets mit Fokus auf Maschinelles Lernen und idealerweise Natural Language Processing (NLP)
- Fundierte Kenntnisse in Machine/Deep Learning
  - Vertraut mit verschiedenen Architekturen von Neuronalen Netze (u.a. CNNs, Transformer, GNNs, xLSTM)
  - Vertraut mit grundlegenden Begriffen und Konzepte wie: Klassifikation, Hyperparameter-Optimierung, Fine-Tuning, Evaluierung von Modellen
- Fundierte Kenntnisse in Python sind zwingend erforderlich
- Von Vorteil: Fähigkeit, Methoden und Verfahren aus wissenschaftlichen Veröffentlichungen eigenständig umzusetzen
- Von Vorteil: Wissen und Erfahrung im Bereich Cybersicherheit
- Bereitschaft, sich neuen Herausforderungen zu stellen
- Ausgeprägtes analytisches Denken

Was Du erwarten kannst

- Selbstständige Arbeitszeiteinteilung
- Einblicke in das Schnittfeld von akademischer Forschung und industrieller Anwendung

Wir wertschätzen und fördern die Vielfalt der Kompetenzen unserer Mitarbeitenden und begrüßen daher alle Bewerbungen – unabhängig von Alter, Geschlecht, Nationalität, ethnischer und sozialer Herkunft, Religion, Weltanschauung, Behinderung sowie sexueller Orientierung und Identität. Schwerbehinderte Menschen werden bei gleicher Eignung bevorzugt eingestellt.
Haben wir Dein Interesse geweckt? Dann bewirb Dich jetzt online mit Deinen aussagekräftigen Bewerbungsunterlagen. Wir freuen uns darauf, Dich kennenzulernen!
Kennziffer: 78185